Thomas S. Shores

Applied Linear Algebra and Matrix Analysis

Second Edition

Springer

Thomas S. Shores
Department of Mathematics
University of Nebraska
Lincoln, NE
USA

ISSN 0172-6056 ISSN 2197-5604 (electronic)
Undergraduate Texts in Mathematics
ISBN 978-3-319-74747-7 ISBN 978-3-319-74748-4
https://doi.org/10.1007/978-3-319-74748-4

Library of Congress Control Number: 2018930352

Preface

Preface to Revised Edition

Times change. So do learning needs, learning styles, students, teachers, authors, and textbooks. The need for a solid understanding of linear algebra and matrix analysis is changing as well. Arguably, as we move deeper into an age of intellectual technology, this need is actually greater. Witness, for example, Google's PageRank technology, an application that has a place in nearly every chapter of this text. In the first edition of this text (henceforth referenced as ALAMA), I suggested that for many students "linear algebra will be as fundamental in their professional work as the tools of calculus." I believe now that this applies to most students of technology. Hence, this revision.

So what has changed in this revision? The objectives of this text, as stated in the preface to ALAMA, have not:

- To provide a balanced blend of applications, theory, and computation that emphasizes their interdependence.
- To assist those who wish to incorporate mathematical experimentation through computer technology into the class. Each chapter has computer exercises sprinkled throughout and an optional section on applications and computational notes. Students should use locally available tools to carry out experiments suggested in projects and use the word processing capabilities of their computer system to create reports of their results.
- To help students to express their thoughts clearly. Requiring written reports is one vehicle for teaching good expression of mathematical ideas.
- To encourage cooperative learning. Mathematics educators have become increasingly appreciative of this powerful mode of learning. Team projects and reports are excellent vehicles for cooperative learning.
- To promote individual learning by providing a complete and readable text. I hope that readers will find this text worthy of being a permanent part of their reference library, particularly for the basic linear algebra needed in the applied mathematical sciences.

What has changed in this revision is that I have incorporated improvements in readability, relevance, and motivation suggested to me by many readers. Readers have also provided many corrections and comments which have been added to the revision. In addition, each chapter of this revised text concludes with introductions to some of the more significant applications of linear algebra in contemporary technology. These include graph theory and network modeling such as Google's PageRank; also included are modeling examples of diffusive processes, linear programming, image processing, digital signal processing, Fourier analysis, and more.

The first edition made specific references to various computer algebra system (CAS) and matrix algebra system (MAS) computer systems. The proliferation of matrix-computing–capable devices (desktop computers, laptops, PDAs, tablets, smartphones, smartwatches, calculators, etc.) and attendant software makes these acronyms too narrow. And besides, who knows what's next ... bionic chip implants? Instructors have a large variety of systems and devices to make available to their students. Therefore, in this revision, I will refer to any such device or software platform as a "technology tool." I will confine occasional specific references to a few freely available tools such as Octave, the R programming language, and the ALAMA Calculator which was written by me specifically for this textbook.

Although calculus is usually a prerequisite for a college-level linear algebra course, this revision could very well be used in a non-calculus–based course without loss of matrix and linear algebra content by skipping any calculus-based text examples or exercises. Indeed, for many students the tools of matrix and linear algebra will be as fundamental in their professional work as the tools of calculus if not more so; thus, it is important to ensure that students appreciate the utility and beauty of these subjects as well as the mechanics. To this end, applied mathematics and mathematical modeling have an important role in an introductory treatment of linear algebra. In this way, students see that concepts of matrix and linear algebra make otherwise intractable concrete problems workable.

The text has a strong orientation toward numerical computation and applied mathematics, which means that matrix analysis plays a central role. All three of the basic components of linear algebra — theory, computation, and applications — receive their due. The proper balance of these components gives students the tools they need as well as the motivation to acquire these tools. Another feature of this text is an emphasis on linear algebra as an experimental science; this emphasis is found in certain examples, computer exercises, and projects. Contemporary mathematical technology tools make ideal "laboratories" for mathematical experimentation. Nonetheless, this text is independent of specific hardware and software platforms. Applications and ideas should take center stage, not hardware or software.

An outline of the book is as follows: Chapter 1 contains a thorough development of Gaussian elimination. Along the way, complex numbers and the basic language of sets are reviewed early on; experience has shown that

this material is frequently long forgotten by many students, so such a review is warranted. Basic properties of matrix arithmetic and determinant algebra are developed in Chapter 2. Special types of matrices, such as elementary and symmetric, are also introduced. Chapter 3 begins with the "standard" Euclidean vector spaces, both real and complex. These provide motivation for the more sophisticated ideas of abstract vector space, subspace, and basis, which are introduced subsequently largely in the context of the standard spaces. Chapter 4 introduces geometrical aspects of standard vector spaces such as norm, dot product, and angle. Chapter 5 introduces eigenvalues and eigenvectors. General norm and inner product concepts for abstract vector spaces are examined in Chapter 6. Each section concludes with a set of exercises and problems.

Each chapter contains a few more optional topics, which are independent of the non-optional sections. Of course, one instructor's optional is another's mandatory. Optional sections cover tensor products, change of basis and linear operators, linear programming, the Schur triangularization theorem, the singular value decomposition, and operator norms. In addition, each chapter has an optional section of applications and computational notes which has been considerably expanded from the first edition along with a concluding section of projects and reports. I employ the convention of marking sections and subsections that I consider optional with an asterisk.

There is more than enough material in this book for a one-semester course. Tastes vary, so there is ample material in the text to accommodate different interests. One could increase emphasis on any one of the theoretical, applied, or computational aspects of linear algebra by the appropriate selection of syllabus topics. The text is well suited to a course with a three-hour lecture and laboratory component, but computer-related material is not mandatory. Every instructor has his/her own idea about how much time to spend on proofs, how much on examples, which sections to skip, etc.; so the amount of material covered will vary considerably. Instructors may mix and match any of the optional sections according to their own interests and needs of their students, since these sections are largely independent of each other. While it would be very time-consuming to cover them all, every instructor ought to use some part of this material. The unstarred sections form the core of the book; most of this material should be covered. There are 27 unstarred sections and 17 optional sections. I hope the optional sections come in enough flavors to please any pure, applied, or computational palate.

Of course, no one size fits all, so I will suggest two examples of how one might use this text for a three-hour one-semester course. Such a course will typically meet three times a week for fifteen weeks, for a total of 45 classes. The material of most of the unstarred sections can be covered at an average rate of about one and one-half class periods per section. Thus, the core material could be covered in about 40 or fewer class periods. This leaves time for extra sections and in-class examinations. In a two-semester course or a course of more than three hours, one could expect to cover most, if not all, of the text.

If the instructor prefers a course that emphasizes the standard Euclidean spaces, and moves at a more leisurely pace, then the core material of the first five chapters of the text is sufficient. This approach reduces the number of unstarred sections to be covered from 27 to 23.

About numbering: Exercises and problems are numbered consecutively in each section. All other numbered items (sections, theorems, definitions, examples, etc.) are numbered consecutively in each chapter and are prefixed by the chapter number in which the item occurs. About examples: In this text, these are illustrative problems, so each is followed by a solution.

I employ the following taxonomy for the reader tasks presented in this text. *Exercises* constitute the usual learning activities for basic skills; these come in pairs, and solutions to the odd-numbered exercises are given in an appendix. More advanced conceptual or computational exercises that ask for explanations or examples are termed *problems*, and solutions for problems are not given, but hints are supplied for those problems marked with an asterisk. Some of these exercises and problems are computer-related. As with pencil-and-paper exercises, these are learning activities for basic skills. The difference is that some computing equipment is required to complete such exercises and problems. At the next level are *projects.* These assignments involve ideas that extend the standard text material, possibly some numerical experimentation and some written exposition in the form of brief project papers. These are analogous to laboratory projects in the physical sciences. Finally, at the top level are *reports.* These require a more detailed exposition of ideas, considerable experimentation — possibly open ended in scope — and a carefully written report document. Reports are comparable to "scientific term papers." They approximate the kind of activity that many students will be involved in throughout their professional lives and are well suited for team efforts. The projects and reports in this text also provide templates for instructors who wish to build their own project/report materials. Students are open to all sorts of technology in mathematics. This openness, together with the availability of inexpensive high-technology tools, has changed how and what we teach in linear algebra.

I would like to thank my colleagues whose encouragement, ideas, and suggestions helped me complete this project, particularly Kristin Pfabe and David Logan. Also, thanks to all those who sent me helpful comments and corrections, particularly David Taylor, David Cox, and Mats Desaix. Finally, I would like to thank the outstanding staff at Springer for their patience and support in bringing this project to completion.

A linear algebra page with some useful materials for instructors and students using this text can be reached at

http://www.math.unl.edu/~tshores1/mylinalg.html

Suggestions, corrections, or comments are welcome. These may be sent to me at tshores1@math.unl.edu.

Contents

1

LINEAR SYSTEMS OF EQUATIONS

Welcome to the world of linear algebra. The two central problems about which much of the theory of linear algebra revolves are the problem of finding all solutions to a linear system and that of finding an eigensystem for a square matrix. The latter problem will not be encountered until Chapter 5; it requires some background development and the motivation for this problem is fairly sophisticated. By contrast, the former problem is easy to understand and motivate. As a matter of fact, simple cases of this problem are a part of most high-school algebra backgrounds. We will address the problem of existence of solutions for a linear system and how to solve such a system for all of its solutions. Examples of linear systems appear in nearly every scientific discipline; we touch on a few in this chapter.

1.1 Some Examples

Here are a few very elementary examples of linear systems:

Example 1.1. For what values of the unknowns x and y are the following equations satisfied?

$$\begin{aligned} x + 2y &= 5 \\ 4x + y &= 6. \end{aligned}$$

Solution. One way that we were taught to solve this problem was the geometrical approach: every equation of the form $ax + by + c = 0$ represents the graph of a straight line. Thus, each equation above represents a line. We need only graph each of the lines, then look for the point where these lines intersect, to find the unique solution to the graph (see Figure 1.1). Of course, the two equations may represent the same line, in which case there are infinitely many solutions, or distinct parallel lines, in which case there are no solutions. These could be viewed as exceptional or "degenerate" cases. Normally, we expect the solution to be unique, which it is in this example.

We also learned how to solve such an equation algebraically: in the present case we may use either equation to solve for one variable, say x, and substitute

the result into the other equation to obtain an equation that is easily solved for y. For example, the first equation above yields $x = 5 - 2y$ and substitution into the second yields $4(5 - 2y) + y = 6$, i.e., $-7y = -14$, so that $y = 2$. Now substitute 2 for y in the first equation and obtain that $x = 5 - 2(2) = 1$. □

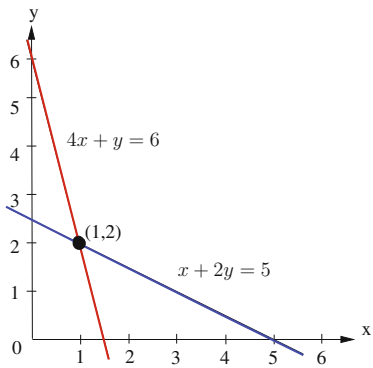


Fig. 1.1: Graphical solution to Example 1.1.

Example 1.2. For what values of the unknowns x, y, and z are the following equations satisfied?

$$\begin{aligned} 2x + 2y + 5z &= 11 \\ 4x + 6y + 8z &= 24 \\ x + y + z &= 4. \end{aligned}$$

Solution. The geometrical approach becomes impractical as a means of obtaining an explicit solution to our problem: graphing in three dimensions on a flat sheet of paper doesn't lead to very accurate answers! Nonetheless, the geometrical approach gives us a qualitative idea of what to expect without actually solving the system of equations.

With reference to our system of three equations in three unknowns, the first fact to take note of is that each of the three equations is an instance of the general equation $ax + by + cz + d = 0$. Now we know from analytical geometry that the graph of this equation is a plane in three dimensions. In general, two planes will intersect in a line, though there are exceptional cases of the two planes represented being identical or distinct and parallel. Similarly, three planes will intersect in a plane, line, point, or nothing. Hence, we know that the above system of three equations has a solution set that is either a plane, line, point, or the empty set.

Which outcome occurs with our system of equations? Figure 1.2 suggests a single point, but graphical methods are not very practical for problems with more than two variables. We need the algebraic point of view to help us calculate the solution. The matter of dealing with three equations and three

unknowns is a bit trickier than the problem of two equations and unknowns. Just as with two equations and unknowns, the key idea is still to use one equation to solve for one unknown. In this problem, subtract 2 times the third equation from the first and 4 times the third equation from the second to obtain the system

$$3z = 3$$
$$2y + 4z = 8,$$

which is easily solved to obtain $z = 1$ and $y = 2$. Now substitute back into the third equation $x + y + z = 4$ and obtain $x = 1$. □

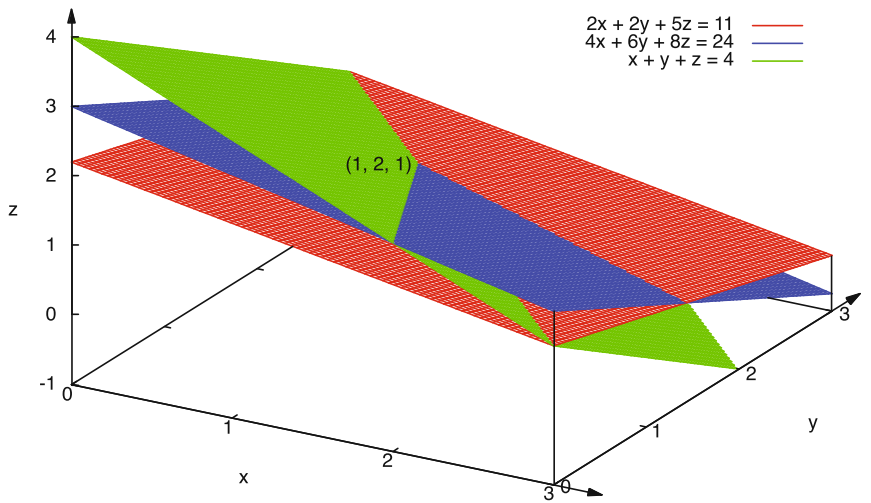


Fig. 1.2: Graphical solution to Example 1.2.

Some Key Notation

Here is a formal statement of the kind of equation that we want to study in this chapter. This formulation gives us the notation for dealing with the general problem later on.

Definition 1.1. Linear Equation A *linear equation* in the variables $x_1, x_2, \ldots, x_n$ is an equation of the form

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b$$

where the coefficients $a_1, a_2, \ldots, a_n$ and term b of the right-hand side are constants.

Of course, there are many interesting and useful nonlinear equations, such as $ax^2 + bx + c = 0$, or $x^2 + y^2 = 1$. But our focus is on systems that consist solely of linear equations. Our next definition describes a general linear system.

Definition 1.2. Linear System A *linear system* of m equations in the n unknowns $x_1, x_2, \ldots, x_n$ is a list of m equations of the form

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1j}x_j + \cdots + a_{1n}x_n &=& b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2j}x_j + \cdots + a_{2n}x_n &=& b_2 \\ \vdots & \vdots & \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ij}x_j + \cdots + a_{in}x_n &=& b_i \\ \vdots & \vdots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mj}x_j + \cdots + a_{mn}x_n &=& b_m. \end{array} \tag{1.1}$$

Row and Column Index

Notice how the coefficients are indexed: in the ith row the coefficient of the jth variable, x_j, is the number a_{ij}, and the right-hand side of the ith equation is b_i. This systematic way of describing the system will come in handy later, when we introduce the matrix concept. About indices: it would be safer — but less convenient — to write $a_{i,j}$ instead of a_{ij}, since ij could be construed to be a single symbol. In those rare situations where confusion is possible, e.g., numeric indices greater than 9, we will separate row and column number with a comma. We call the layout of of this definition the *standard form* of a linear system.

* Examples of Modeling Problems

It is easy to get the impression that linear algebra is only about the simple kinds of problems such as the preceding examples. So why develop a whole subject? We shall consider a few examples whose solutions are not so apparent as those of the previous two examples. The point of this chapter, as well as that of Chapters 2 and 3, is to develop algebraic and geometrical methodologies that are powerful enough to handle problems like these.

Diffusion Processes

Diffusion processes are studied in biology, chemistry, physics, sociology and other areas of science. We shall examine a very simple diffusion problem, that of the flow of heat through a homogeneous material. A basic physical observation is that change in heat is directly proportional to change in temperature. In a wide range of problems this hypothesis is true, and we shall assume that we are modeling such a problem. Thus, we can measure the amount of heat at a point by measuring temperature. To fix ideas, suppose we have a rod of material of unit length, say, situated on the x-axis, on $0 \leq x \leq 1$. Suppose further

that the rod is laterally insulated, but has a known internal heat source that doesn't change with time. When sufficient time passes, the temperature of the rod at each point will settle down to "steady-state" values, dependent only on position x. Say the heat source is described by a function $f(x)$, $0 \le x \le 1$ in heat generated per unit length at the point x. Also suppose that the left and right ends of the rod are held at fixed temperatures y_{left} and y_{right}, respectively.

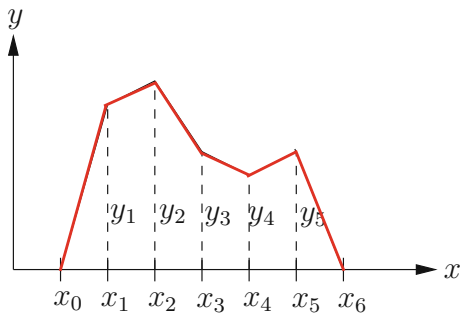


Fig. 1.3: Discrete approximation to temperature function ($n = 5$).

To model a steady state imagine that the rod is divided up into a finite number of segments between equally spaced points, called nodes, namely $x_0 = 0, x_1, x_2, \ldots, x_{n+1} = 1$, and that the heat on the ith segment is well approximated by the temperature at its left node. Assume that the nodes are a distance h apart. Since spacing is equal, the relation between h and n is $h = 1/(n+1)$. Let the temperature function be $y(x)$ and let $y_i = y(x_i)$. Approximate $y(x)$ in between nodes by connecting adjacent points (x_i, y_i) with a line segment. (See Figure 1.3 for a graph of the resulting approximation to $y(x)$.) We know that at the end nodes the temperature is specified: $y(x_0) = y_{\text{left}}$ and $y(x_{n+1}) = y_{\text{right}}$. By examining the process at each interior node, we can obtain the following linear equation for each interior node index $i = 1, 2, \ldots, n$ involving a constant K called the thermal conductivity of the material. (A detailed derivation is given in Section 1.5.) This equation can be understood as balancing the flow of heat from a node to its neighbors:

$$-y_{i-1} + 2y_i - y_{i+1} = \frac{h^2}{K} f(x_i). \tag{1.2}$$

Example 1.3. Suppose we have a rod of material of conductivity $K = 1$ and situated on the x-axis, for $0 \le x \le 1$. Suppose further that the rod is laterally insulated, but has a known internal heat source $f(x)$. The left and right ends of the rod are held at $0\,^\circ C$ (degrees Celsius). With $n = 5$ what are the discretized steady-state equations for this problem?

Solution. Follow the notation of the discussion preceding this example. Notice that in this case $x_i = ih$. Remember that y_0 is given to be 0, so the term y_0 disappears. Also, the value of $y_{n+1} = y_6$ is zero, so it too disappears. Thus

we have from equation (1.2) five equations in the unknowns y_i, $i = 1, 2, \ldots, 5$. The system of five equations in five unknowns becomes

$$\begin{array}{rrrrrl} 2y_1 & -y_2 & & & & = f(1/6)/36 \\ -y_1 & +2y_2 & -y_3 & & & = f(2/6)/36 \\ & -y_2 & +2y_3 & -y_4 & & = f(3/6)/36 \\ & & -y_3 & +2y_4 & -y_5 & = f(4/6)/36 \\ & & & -y_4 & +2y_5 & = f(5/6)/36. \end{array}$$

□

It is reasonable to expect that the smaller h is, the more accurately y_i will approximate $y(x_i)$. This is indeed the case. But consider what we are confronted with when we take $n = 5$, so that $h = 1/(5+1) = 1/6$. This is hardly a small value of h, yet the problem is already about as large as we might want to work by hand, if not larger. The basic ideas of solving systems like this are the same as in Examples 1.1 and 1.2. For very small h, say $h = .01$ and hence $n = 99$, we clearly would need some help from a technology tool.

Leontief Input–Output Models

Here is a simple model of an open economy consisting of three sectors that supply each other and consumers. Suppose the three sectors are (M)aterials, (P)roduction and (S)ervices and that the demands of one sector from all sectors are proportional to its output. This is reasonable; if, for example, the materials sector doubled its output, one would expect its needs for materials, production and services to likewise double. A table of these demand constants of proportionality for production of a unit of sector output is called a *consumption matrix*. Equilibrium of the economy is reached when total production matches consumption. If at some level of output the economy exactly meets some positive demand, we say the system is in equilibrium and call the consumption matrix *productive*. On the other hand, if at some level of output the demands of all sectors exactly equal output, we say the economy is *closed*. Of course we would like to know if the economy is productive or closed.

Consumption Matrix
Productive Matrix
Closed Economy

Example 1.4. Given the following consumption matrix, and that consumer demands for the output of sectors M, P, S are the constant $20, 10, 30$ units, respectively, express the equilibrium of the economy as a system of equations.

		Consumed by		
		M	P	S
Produced by	M	0.2	0.3	0.1
	P	0.1	0.3	0.2
	S	0.4	0.2	0.1

Solution. Let x, y, z be the total outputs of the sectors M, P, and S respectively. Consider how we balance the total supply and demand for materials. The total output of materials is x units. The demands on sector M from the three sectors M, P and S are, according to the table data, $0.2x$, $0.3y$, and $0.1z$, respectively. Further, consumers demand 20 units of energy. In equation form,

$$x = 0.2x + 0.3y + 0.1z + 20.$$

Likewise we can balance the input/output of the sectors P and S to arrive at a system of three equations in three unknowns:

$$\begin{aligned} x &= 0.2x + 0.3y + 0.1z + 20 \\ y &= 0.1x + 0.3y + 0.2z + 10 \\ z &= 0.4x + 0.2y + 0.1z + 30. \end{aligned}$$

The questions that interest economists are whether this system has solutions and if so, how to interpret them. □

Next, consider the situation of a closed economic system, that is, one in which everything produced by the sectors of the system is consumed by those sectors.

Example 1.5. An administrative unit has four divisions serving the internal needs of the unit, labeled (A)ccounting, (M)aintenance, (S)upplies, and (T)raining. Each unit produces the "commodity" its name suggests, and charges the other divisions for its services. The input–output table of demand rates are specified by the following table. Express the equilibrium of this system as a system of equations.

		Consumed by			
		A	M	S	T
Produced by	A	0.2	0.3	0.3	0.2
	M	0.1	0.2	0.2	0.1
	S	0.4	0.2	0.2	0.2
	T	0.4	0.1	0.3	0.2

Solution. Let x, y, z, w be the total outputs of the sectors A, M, S, and T, respectively. The analysis proceeds along the lines of the previous example and results in the system

$$\begin{aligned} x &= 0.2x + 0.3y + 0.3z + 0.2w \\ y &= 0.1x + 0.2y + 0.2z + 0.1w \\ z &= 0.4x + 0.2y + 0.2z + 0.2w \\ w &= 0.4x + 0.1y + 0.3z + 0.2w. \end{aligned}$$

There is an obvious, but useless, solution to this system (all variables equal to zero). One hopes for nontrivial solutions that are meaningful in the sense that each variable takes on a nonnegative value. □

The PageRank Tool

Consider the Google problem of displaying the results of a search on a certain phrase. There could be many thousands of matching web pages. So which ones should be displayed in the user's window? Enter PageRank technology (famously referenced by Kurt Bryan and Tanya Leise in [5] as a "billion dollar eigenvalue") which ranks the pages in terms of an "importance" score. This remarkable technology has found significant application in areas such as chemistry, biology, bioinformatics, neuroscience, complex systems engineering and even sports rankings (a comprehensive summary can be found in [13]).

Let's start small: suppose we have a web of four pages represented as in Figure 1.4 with pages as vertices and links from one page to another as arrows.

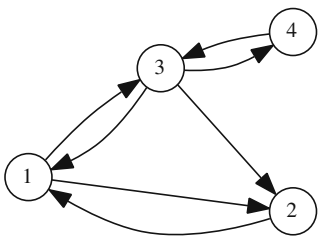


Fig. 1.4: A web with four pages as vertices and links as arrows.

Here is a first pass at page ranking (but not the last, we will return to this significant example with refinements several times more in this text). We could simply count *backlinks* (incoming links) of each page and rank pages according to that score, larger being more important than smaller. One problem with this solution is that it would give equal weight to a link from any page, whether the linking page were of low or high rank. Another problem is that the rank of two pages could be artificially inflated by increasing the number of backlinks and outgoing links between them. So here is a second pass to correct some of these deficiencies: let a page score be the sum of all scores of pages linking to it. For page i let x_i be its score and L_i be the set of all indices of pages linking to it. Then the score for vertex i is given by

$$x_i = \sum_{x_j \in L_i} x_j. \tag{1.3}$$

But that ranking could still give excess influence to a page simply by its linking from many other pages. To correct this deficiency we make a third pass: for page j let n_j be its total number of outgoing links on that page. Then the score for vertex i is given by

$$x_i = \sum_{x_j \in L_i} \frac{x_j}{n_j}. \tag{1.4}$$

The result is that each page divides its one unit of influence among all pages to which it links, so that no page has more influence to distribute than any other.

This is a good start on PageRank. However there are additional problems with these formulations of the ranking problem which we shall resolve with yet another pass at it in Section 2.5 of Chapter 2.

Example 1.6. Exhibit the systems of equations resulting from applying the ranking systems of the preceding discussion to the web of Figure 1.4.

Solution. If we simply count backlinks, then there is nothing to solve since counting links gives $x_1 = 2$, $x_2 = 2$, $x_3 = 2$ and $x_4 = 1$ so that vertices 1, 2 and 3 are tied for most important with two backlinks, while vertex 4 is the least important with only one backlink. If we use the second approach, then we can see from inspection of the graph and equation (1.3) that the resulting linear system is

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

Finally, if we use equation (1.4) for the third approach, the resulting system is

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

□

Note 1.1. In some of the exercises and projects in this text you will find references to "technology tools." This may be a scientific calculator that is required for the course, a math computer program or a computer system for which you are given an account. This includes both hardware and software, which many authors commonly term a "computer algebra system" or "CAS". This textbook does not depend on any particular system, but certain exercises require a suitable computational device. It will occasionally give a few details about using ALAMA Calculator, a software program which was designed with this text in mind.

1.1 Exercises and Problems

Exercise 1. Solve the following systems algebraically.

(a) $\begin{aligned} x + 2y &= 1 \\ 3x - y &= -4 \end{aligned}$ (b) $\begin{aligned} x - y + 2z &= 6 \\ 2x - z &= 3 \\ y + 2z &= 0 \end{aligned}$ (c) $\begin{aligned} x - y &= 1 \\ 2x - y &= 3 \\ x + y &= 3 \end{aligned}$

Exercise 2. Solve the following systems algebraically.

(a) $\begin{aligned} x - y &= -3 \\ x + y &= 1 \end{aligned}$ (b) $\begin{aligned} x - y + 2z &= 0 \\ x - z &= -2 \\ z &= 0 \end{aligned}$ (c) $\begin{aligned} x + 2y &= 1 \\ 2x - y &= 2 \\ x + y &= 2 \end{aligned}$

Exercise 3. Determine whether each of the following systems of equations is linear. If so, put it in standard form.

(a) $\begin{aligned} x + 2 &= y + z \\ 3x - y &= 4 \end{aligned}$ (b) $\begin{aligned} xy + 2 &= 1 \\ 2x - 6 &= y \end{aligned}$ (c) $\begin{aligned} x + 2y &= -2y \\ 2x &= y \\ 2 &= x + y \end{aligned}$

Exercise 4. Determine whether each of the following systems of equations is linear. If so, put it in standard format.

(a) $\begin{aligned} x + 2 &= 1 \\ x + 3 &= y^2 \end{aligned}$ (b) $\begin{aligned} x + 2z &= y \\ 3x - y &= y \end{aligned}$ (c) $\begin{aligned} x + y &= -3y \\ 2x &= xy \end{aligned}$

Exercise 5. Express the following systems of equations in the notation of the definition of linear systems by specifying the numbers m, n, a_{ij}, and b_i.

(a) $\begin{aligned} x_1 - 2x_2 + x_3 &= 2 \\ x_2 &= 1 \\ -x_1 + x_3 &= 1 \end{aligned}$ (b) $\begin{aligned} x_1 - 3x_2 &= 1 \\ x_2 &= 5 \end{aligned}$

Exercise 6. Express the following systems of equations in the notation of the definition of linear systems by specifying the numbers m, n, a_{ij}, and b_i.

(a) $\begin{aligned} x_1 - x_2 &= 1 \\ 2x_1 - x_2 &= 3 \\ x_2 + x_1 &= 3 \end{aligned}$ (b) $\begin{aligned} -2x_1 + x_3 &= 1 \\ x_2 - x_3 &= 5 \end{aligned}$

Exercise 7. Write out the linear system that results from Example 1.3 if we take $n = 4$, $y_5 = 50$ and $f(x) = 3y(x)$.

Exercise 8. Write out the linear systems that result from Example 1.6 if we remove vertex 4 and its connecting edges from Figure 1.4.

Exercise 9. Suppose that in the input–output model of Example 1.4 each sector charges a unit price for its commodity, say p_1, p_2, p_3, and that the MPS columns of the consumption matrix represent the fraction of each producer commodity needed by the consumer to produce one unit of its own commodity. Derive equations for prices that achieve equilibrium, that is, equations that say that the price received for a unit item equals the cost of producing it.

Exercise 10. Suppose that in the input–output model of Example 1.5 each producer charges a unit price for its commodity, say p_1, p_2, p_3, p_4 and that the columns of the table represent fraction of each producer commodity needed by the consumer to produce one unit of its own commodity. Derive equilibrium equations for these prices.

Exercise 11. Solve the system that results from the second pass of Example 1.6 for page ranking.

Exercise 12. Solve the system that results from the third pass of Example 1.6 for page ranking given that x_4 is assigned a value of 1.

Exercise 13. Construct a linear system that has $x_1 = 1$, $x_2 = -1$ as a solution and right-hand side terms $b_1 = 1$, $b_2 = -2$, $b_3 = 3$.

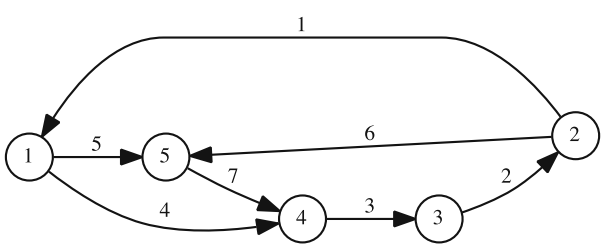
Exercise 14. Construct a linear system that has both $x_1 = 1$, $x_2 = -1$ and $x_1 = 2$, $x_2 = 2$ as solutions and right-hand side terms $b_1 = 3$, $b_2 = 1$, $b_3 = 4$.

Problem 15. Suppose that we construct a web of pages by removing vertex 4 and its connecting edges from Figure 1.4. Write out the system of equations that results from the second and third passes of Example 1.6 for page ranking and solve these systems.

Problem 16. Use ALAMA Calculator or other technology tool to solve the systems of Examples 1.4 and 1.5. Comment on your solutions. Are they sensible?

Problem 17. A polynomial $y = a_0 + a_1x + a_2x^2$ is required to interpolate a function $f(x)$ at $x = 1, 2, 3$, where $f(1) = 1$, $f(2) = 1$, and $f(3) = 2$. Express these three conditions as a linear system of three equations in the unknowns a_0, a_1, a_2. What kind of general system would result from interpolating $f(x)$ with a polynomial at points $x = 1, 2, \ldots, n$ where $f(x)$ is known?

*Problem 18. The topology of a certain network is indicated by the digraph (directed graph) pictured below, where five vertices represent locations of hardware units that receive and transmit data along connection edges to other units in the direction of the arrows. Suppose the system is in a steady state and that the data flow along edge j is the nonnegative quantity x_j. The single law that these flows must obey is this: net flow in equals net flow out at each of the five vertices (like Kirchhoff's first law in electrical circuits). Write out a system of linear equations satisfied by variables $x_1, x_2, x_3, x_4, x_5, x_6, x_7$.



Problem 19. Use ALAMA Calculator or other technology tool to solve the system of Example 1.3 with conductivity $K = 1$ and internal heat source $f(x) = x$ and graph the approximate solution by connecting the points (x_j, y_j) as in Figure 1.3.

1.2 Notation and a Review of Numbers

The Language of Sets

The language of sets pervades all of mathematics. It provides a convenient shorthand for expressing mathematical statements. Loosely speaking, a set can be defined as a collection of objects, called the *members* of the set. This definition will suffice for us. We use some shorthand to indicate certain relationships between sets and elements. Usually, sets will be designated by uppercase letters such as A, B, etc., and elements will be designated by lowercase letters such as a, b, etc. As usual, set A is a *subset* of set B if every element of A is an element of B, and a *proper* subset if it is a subset but not equal to B. Two sets A and B are said to be *equal* if they have exactly the same elements.

Set Symbols Some shorthand:

$\emptyset$ denotes the empty set, i.e., the set with no members.
$a \in A$ means "a is a member of the set A."
$A = B$ means "the set A is equal to the set B."
$A \subseteq B$ means "A is a subset of B."
$A \subset B$ means "A is a proper subset of B."

There are two ways in which we may define a set: we may *list* its elements, such as in the definition $A = \{0, 1, 2, 3\}$, or specify them by *rule* such as in the definition $A = \{x \mid x \text{ is an integer and } 0 \leq x \leq 3\}$. (Read this as "$A$ is the set of x such that x is an integer and $0 \leq x \leq 3$.") With this notation we can give formal definitions of set intersections and unions:

Definition 1.3. **Set Union, Intersection, Difference** Let A and B be sets. Then the *intersection* of A and B is defined to be the set $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$. The *union* of A and B is the set $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$ (inclusive or, which means that $x \in A$ or $x \in B$ or both). The *difference* of A and B is the set $A - B = \{x \mid x \in A \text{ and } x \notin B\}$.

Example 1.7. Let $A = \{0, 1, 3\}$ and $B = \{0, 1, 2, 4\}$. Then

$$
\begin{aligned}
A \cup \emptyset &= A, \\
A \cap \emptyset &= \emptyset, \\
A \cup B &= \{0, 1, 2, 3, 4\}, \\
A \cap B &= \{0, 1\}, \\
A - B &= \{3\}.
\end{aligned}
$$

□

About Numbers

One could spend a whole course fully developing the properties of number systems. We won't do that, but we will review some of the basic sets of numbers, and assume that the reader is familiar with properties of numbers we have not mentioned here. At the start of it all is the kind of numbers that everyone knows something about: the *natural* or *counting* numbers. This is the set

Natural Numbers

$$\mathbb{N} = \{1, 2, \ldots\}.$$

One could view most subsequent expansions of the concept of number as a matter of rising to the challenge of solving new equations. For example, we cannot solve the equation

$$x + m = n, \quad m, n \in \mathbb{N},$$

for the unknown x without introducing subtraction and extending the notion of natural number that of *integer*. The set of integers is denoted by

Integers

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}.$$

Next, we cannot solve the equation

$$ax = b, \ 0 \neq a, b \in \mathbb{Z},$$

for the unknown x without introducing division and extending the notion of integer to that of *rational number*. The set of rationals is denoted by

Rational Numbers

$$\mathbb{Q} = \{a/b \mid a, b \in \mathbb{Z} \text{ and } b \neq 0\}.$$

Rational-number arithmetic has some characteristics that distinguish it from integer arithmetic. The main difference is that nonzero rational numbers have multiplicative inverses: the multiplicative inverse of a/b is b/a. Such a number system is called a *field* of numbers. In a nut-

Field of Numbers

shell, a *field of numbers* is a system of objects, called numbers, together with operations of addition, subtraction, multiplication, and division that satisfy the usual arithmetic laws; in particular, it must be possible to subtract any number from any other and divide any number by a nonzero number to obtain another such number. The associative, commutative, identity, and inverse

laws must hold for each of addition and multiplication; and the distributive law must hold for multiplication over addition. The rationals form a field of numbers; the integers don't since division by nonzero integers does not always yield an integer.

The jump from rational to real numbers cannot be entirely explained by algebra, although algebra offers some insight as to why the number system still needs to be extended. There is no rational number whose square is 2. Thus the equation

$$x^2 = 2$$

cannot be solved using rational numbers alone. (Story has it that this is lethal knowledge, in that followers of a Pythagorean cult claim that the gods threw overboard from a ship one of their followers, Hippasus of Metapontum, who was unfortunate enough to discover that fact.) There is also the problem of numbers like π and the mathematical constant e which do not satisfy any polynomial equation. The heart of the problem is that if we consider only rationals on a number line, there are many "holes" that are filled by numbers like π and $\sqrt{2}$. Filling in these holes leads us to the set $\mathbb{R}$ of real numbers, which are in one-to-one correspondence with the points on a number line. We won't give an exact definition of the set of real numbers. Recall that every real number admits a (possibly infinite) decimal representation, such as $1/3 = 0.333\ldots$ or $\pi = 3.14159\ldots$. This provides us with a loose definition:

Real Numbers Real numbers are numbers that can be expressed by a decimal representation, i.e., limits of finite decimal expansions. Equivalently, real numbers can be thought of as points on the real number line. As usual, the set of all real numbers is denoted by $\mathbb{R}$. In addition, we employ the usual interval notations for real numbers a, b such that $a \leq b$:

$$\begin{aligned}
[a, b] &= \{x \in \mathbb{R} \,|\, a \leq x \leq b\}\,, \\
[a, b) &= \{x \in \mathbb{R} \,|\, a \leq x < b\}\,, \\
(a, b) &= \{x \in \mathbb{R} \,|\, a < x < b\}\,.
\end{aligned}$$

There is one more problem to overcome. How do we solve a system like

$$x^2 + 1 = 0$$

over the reals? The answer is we can't: if x is real, then $x^2 \geq 0$, so $x^2 + 1 > 0$.

Complex Numbers We need to extend our number system one more time, and this leads to the set $\mathbb{C}$ of *complex* numbers. We define i to be a quantity such that $\mathrm{i}^2 = -1$ and

$$\mathbb{C} = \{a + b\mathrm{i} \mid a, b \in \mathbb{R}\}\,.$$

Standard Form We say that the form $z = a + b\mathrm{i}$ is the *standard form* of z. In this case the real part of z is $\Re(z) = a$ and the imaginary part is defined

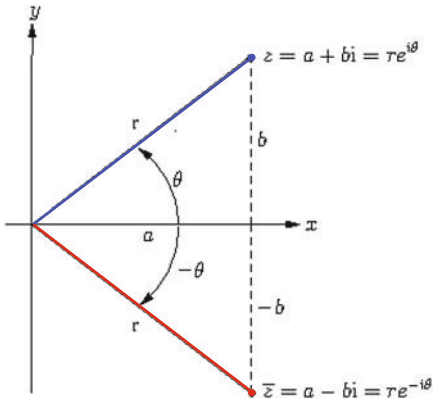


Fig. 1.5: Standard and polar coordinates in the complex plane.

as $\Im(z) = b$. (Notice that the imaginary part of z is a real number: it is the real coefficient of i.) Two complex numbers are *equal* precisely when they have the same real part and the same imaginary part. All of this could be put on a more formal basis by initially defining complex numbers to be ordered pairs of real numbers. We will not do so, but the fact that complex numbers behave like ordered pairs of real numbers leads to an important geometrical insight: complex numbers can be identified with points in the plane.

Real and Imaginary Parts

Instead of an x- and y-axis, one lays out a *real* and an *imaginary* axis (which are still usually labeled with x and y) and plots complex numbers $a + b\mathrm{i}$ as in Figure 1.5. This results in the *complex plane.* Arithmetic in $\mathbb{C}$ is carried out using the usual laws of arithmetic for $\mathbb{R}$ and the algebraic identity $\mathrm{i}^2 = -1$ to reduce the result to standard form. In addition, there are several more useful ideas about complex numbers that we will need.

Absolute Value

The *length,* or *absolute value,* of a complex number in standard form, $z = a + b\mathrm{i}$, is defined as the nonnegative real number $|z| = \sqrt{a^2 + b^2}$, which is the distance from the origin to z. The *complex conjugate* of z is defined as $\overline{z} = a - b\mathrm{i}$ (see Figure 1.5). Thus we have:

Laws of Complex Arithmetic

$$\begin{aligned}(a+b\mathrm{i})+(c+d\mathrm{i}) &= (a+c)+(b+d)\mathrm{i}\\ (a+b\mathrm{i})\cdot(c+d\mathrm{i}) &= (ac-bd)+(ad+bc)\mathrm{i}\\ \overline{a+b\mathrm{i}} &= a-b\mathrm{i}\\ |a+b\mathrm{i}| &= \sqrt{a^2+b^2}\end{aligned}$$

If meaning is clear, the product $z_1 \cdot z_2$ is often abbreviated to $z_1 z_2$.

Example 1.8. Let $z_1 = 2+4\mathrm{i}$ and $z_2 = 1-3\mathrm{i}$. Compute $z_1 - 3z_2$ and $z_1 z_2$.

Solution. We have that

$$z_1 - 3z_2 = (2+4\mathrm{i}) - 3(1-3\mathrm{i}) = 2+4\mathrm{i}-3+9\mathrm{i} = -1+13\mathrm{i}$$

and

$$z_1 z_2 = (2+4\mathrm{i})(1-3\mathrm{i}) = 2+4\mathrm{i}-2\cdot 3\mathrm{i}-4\cdot 3\mathrm{i}^2 = (2+12)+(4-6)\mathrm{i} = 14-2\mathrm{i}. \quad \square$$

Here are some easily checked and very useful facts about absolute value and complex conjugation:

Laws of Conjugation and Absolute Value

$$\begin{aligned}|z_1 z_2| &= |z_1|\,|z_2|\\ |z_1+z_2| &\le |z_1|+|z_2|\\ |z|^2 &= z\overline{z}\\ \overline{z_1+z_2} &= \overline{z_1}+\overline{z_2}\\ \overline{z_1 z_2} &= \overline{z_1}\,\overline{z_2}\\ z_1/z_2 &= z_1\overline{z_2}/|z_2|^2\end{aligned}$$

Example 1.9. Let $z_1 = 2+4\mathrm{i}$ and $z_2 = 1-3\mathrm{i}$. Verify that $|z_1 z_2| = |z_1|\,|z_2|$.

Solution. From Example 1.8, $z_1 z_2 = 14-2\mathrm{i}$, so that $|z_1 z_2| = \sqrt{14^2+(-2)^2} = \sqrt{200}$, while $|z_1| = \sqrt{2^2+4^2} = \sqrt{20}$ and $|z_2| = \sqrt{1^2+(-3)^2} = \sqrt{10}$. Hence $|z_1 z_2| = \sqrt{10}\sqrt{20} = |z_1|\,|z_2|$. $\square$

Example 1.10. Verify that the conjugate of the product is the product of conjugates.

Solution. This is just the fifth fact in the preceding list. Let $z_1 = x_1 + \mathrm{i}y_1$ and $z_2 = x_2 + \mathrm{i}y_2$ be in standard form, so that $\overline{z}_1 = x_1 - \mathrm{i}y_1$ and $\overline{z}_2 = x_2 - \mathrm{i}y_2$. We calculate

$$z_1 z_2 = (x_1 x_2 - y_1 y_2) + \mathrm{i}(x_1 y_2 + x_2 y_1),$$

so that

$$\overline{z_1 z_2} = (x_1 x_2 - y_1 y_2) - \mathrm{i}(x_1 y_2 + x_2 y_1).$$

Also,

$$\overline{z}_1\,\overline{z}_2 = (x_1 - \mathrm{i}y_1)(x_2 - \mathrm{i}y_2) = (x_1 x_2 - y_1 y_2) - \mathrm{i}(x_1 y_2 + x_2 y_1) = \overline{z_1 z_2}. \quad \square$$

The complex number $z = \mathrm{i}$ solves the equation $z^2 + 1 = 0$ (no surprise here: it was invented expressly for that purpose). The big surprise is that once we have the complex numbers in hand, we have a number system so complete that we can solve *any* polynomial equation in it. We won't offer a proof of this fact; it's very nontrivial. Suffice it to say that nineteenth-century mathematicians considered this fact so fundamental that they dubbed it the "Fundamental Theorem of Algebra," a terminology we adopt.

Theorem 1.1. Fundamental Theorem of Algebra Let $p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ be a nonconstant polynomial in the variable z with complex coefficients $a_0, \ldots, a_n$. Then the polynomial equation $p(z) = 0$ has a solution in the field $\mathbb{C}$ of complex numbers.

Note that the fundamental theorem doesn't tell us how to find a root of a polynomial, only that it exists. There are numerical techniques for approximating such roots. But for polynomials of degree greater than four, there are no general algebraic expressions in terms of radicals (like the quadratic formula) for their roots.

In vector space theory the numbers in use are called *scalars*, and we will use this term. Unless otherwise stated or suggested by the presence of i, the field of scalars in which we do arithmetic is assumed to be the field of real numbers. However, we shall see later, when we study eigensystems, that even if we are interested only in real scalars, complex numbers have a way of turning up quite naturally.

Scalars

The following example shows how to "rationalize" a complex denominator.

Example 1.11. Solve the linear equation $(1 - 2\mathrm{i})\, z = (2 + 4\mathrm{i})$ for the complex variable z. Also compute the complex conjugate and absolute value of the solution.

Solution. The solution requires that we put the complex number $z = (2+4\mathrm{i})/(1-2\mathrm{i})$ in standard form. Proceed as follows: multiply both numerator and denominator by $\overline{(1 - 2\mathrm{i})} = 1 + 2\mathrm{i}$ to obtain that

$$z = \frac{2 + 4\mathrm{i}}{1 - 2\mathrm{i}} = \frac{(2 + 4\mathrm{i})(1 + 2\mathrm{i})}{(1 - 2\mathrm{i})(1 + 2\mathrm{i})} = \frac{2 - 8 + (4 + 4)\mathrm{i}}{1 + 4} = \frac{-6}{5} + \frac{8}{5}\mathrm{i}.$$

Next we see that

$$\overline{z} = \overline{\frac{-6}{5} + \frac{8}{5}\mathrm{i}} = -\frac{6}{5} - \frac{8}{5}\mathrm{i}$$

and

$$|z| = \left|\frac{1}{5}(-6 + 8\mathrm{i})\right| = \frac{1}{5}\,|(-6 + 8\mathrm{i})| = \frac{1}{5}\sqrt{(-6)^2 + 8^2} = \frac{10}{5} = 2. \quad \square$$

Practical Complex Arithmetic

We conclude this section with a discussion of the more advanced aspects of complex arithmetic. This material will not be needed until Chapter 4. Recall from basic algebra the *roots theorem*: the linear polynomial $z - a$ is a factor of a polynomial $f(z) = a_0 + a_1 z + \cdots + a_n z^n$ if and only if a is a *root* of the polynomial, i.e., $f(a) = 0$. If we team this fact up with the Fundamental Theorem of Algebra, we see an interesting fact about factoring polynomials over $\mathbb{C}$: every polynomial can be completely factored into a constant times a product of linear polynomials of the form $z - a$. The numbers a that occur are exactly the roots of $f(z)$. Of course, these roots could be repeated roots, as in the case of $f(z) = 3z^2 - 6z + 3 = 3(z-1)^2$. But how can we use the Fundamental Theorem of Algebra in a practical way to find the roots of a polynomial? Unfortunately, the usual proofs of the Fundamental Theorem of Algebra don't offer a clue, because they are *nonconstructive*, i.e., they prove that solutions must exist, but do not show how to explicitly construct such a solution. Usually, we have to resort to numerical methods to get approximate solutions, such as the Newton's method used in calculus. For now, we will settle on a few ad hoc methods for solving some important special cases.

First-degree equations offer little difficulty: the solution to $az = b$ is $z = b/a$, as usual. There is one detail to attend to: what complex number is represented by the expression b/a? We saw how to handle this by the trick of "rationalizing" the denominator in Example 1.11.

Quadratic equations are also simple enough: use the quadratic formula, which says that the solutions to $az^2 + bz + c = 0$, where $a \neq 0$, are given by

Quadratic Formula

$$z = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

One little catch: what does the square root of a complex number mean? For nonnegative real numbers r the expression $\sqrt{r}$ is called the principal square root of r and its meaning is unambiguous. For complex numbers it is not. What we are really asking is this: How do we solve the equation $z^2 = d$ for z, where d is a complex number? Let's try for a little more: How do we solve $z^n = d$ for all possible solutions z, where d is a nonzero complex number? In a few cases, such an equation is quite easy to solve. We know, for example, that $z = \pm\mathrm{i}$ are solutions to $z^2 = -1$, so these are all the solutions. Similarly, one can check by hand that $\pm 1, \pm\mathrm{i}$ are all solutions to $z^4 = 1$. Consequently, $z^4 - 1 = (z-1)(z+1)(z-\mathrm{i})(z+\mathrm{i})$. Roots of the equation $z^n = 1$ are sometimes called the nth roots of unity. Thus the 4th roots of unity are ± 1 and $\pm\mathrm{i}$. But what about something like $z^3 = 1 + \mathrm{i}$?

The key to answering this question is another form of a nonzero complex number $z = a + b\mathrm{i}$. In reference to Figure 1.5 we can write $z = r(\cos\theta + \mathrm{i}\sin\theta) = re^{\mathrm{i}\theta}$, where θ is a real number, r is a positive real, and $e^{\mathrm{i}\theta}$ is *defined* by the following expression, which is called *Euler's formula*:

Definition 1.4. Polar form $e^{i\theta} = \cos\theta + i\sin\theta$.

Notice that $|e^{i\theta}| = \sqrt{\cos^2\theta + \sin^2\theta} = 1$, so that $|re^{i\theta}| = |r||e^{i\theta}| = r$, provided r is nonnegative. The expression $re^{i\theta}$ with $r = |z|$ and the angle θ measured counterclockwise in radians is called the *polar form* of z. The number θ is sometimes called an *argument* of z. It is important to notice that θ is not unique. If the angle θ_0 works for the nonzero complex number z, then so does $\theta = \theta_0 + 2\pi k$, for any integer k, since $\sin\theta$ and $\cos\theta$ are periodic of period 2π. It follows that a complex number may have more than one polar form. For example, $i = e^{i\pi/2} = e^{i5\pi/2}$ (here $r = 1$). In fact, the most general polar expression for i is $i = e^{i(\pi/2+2k\pi)}$, where k is an arbitrary integer.

Example 1.12. Find the possible polar forms of $1 + i$.

Solution. Sketch a picture of $1 + i$ in the complex plane and we see that the angle $\theta_0 = \pi/4$ works fine as a measure of the angle from the positive x-axis to the radial line from the origin to z. Moreover, the absolute value of z is $\sqrt{1+1} = \sqrt{2}$. However, we can adjust the angle θ_0 by any multiple of 2π, a full rotation and get a polar form for z. So the most general form for z is $z = \sqrt{2}e^{i(\pi/4+2k\pi)}$, where k is any integer. □

As the notation suggests, polar forms obey the laws of exponents. A simple application of the laws for the sine and cosine of a sum of angles shows that for angles θ and ψ we have the identity

$$e^{i(\theta+\psi)} = e^{i\theta}e^{i\psi}.$$

By using this formula n times, we obtain that $e^{in\theta} = \left(e^{i\theta}\right)^n$, which can also be expressed as *de Moivre's Formula:*

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

Now for solving $z^n = d$: First, find the general polar form of d, say $d = ae^{i(\theta_0+2k\pi)}$, where θ_0 is the *principal angle* for d, i.e., $0 \le \theta_0 < 2\pi$, and $a = |d|$. Next, write $z = re^{i\theta}$, so that the equation to be solved becomes

$$r^n e^{in\theta} = ae^{i(\theta_0+2k\pi)}.$$

Taking absolute values of both sides yields that $r^n = a$, whence we obtain the unique value of $r = \sqrt[n]{a} = \sqrt[n]{|d|}$. What about θ? The most general form for $n\theta$ is

$$n\theta = \theta_0 + 2k\pi.$$

Hence, we obtain that

$$\theta = \frac{\theta_0}{n} + \frac{2k\pi}{n}.$$

Notice that the values of $e^{i2k\pi/n}$ start repeating themselves as k passes a multiple of n, since $e^{i2\pi} = e^0 = 1$. Therefore, one gets exactly n distinct values for $e^{i\theta}$, namely

$$\theta = \frac{\theta_0}{n} + \frac{2k\pi}{n}, \quad k = 0, \ldots, n-1.$$

These points are equally spaced around the unit circle in the complex plane, starting with the point $e^{i\theta_0}$. Thus we have obtained n distinct solutions to the equation $z^n = d$, where $d \neq 0$, namely

General solution to $z^n = d$

$$z = a^{1/n} e^{i(\theta_0/n + 2k\pi/n)}, \quad k = 0, \ldots, n-1, \text{ where } 0 \neq d = ae^{i\theta_0}$$

Example 1.13. Solve the equation $z^3 = 1 + i$ for the unknown z.

Solution. The solution goes as follows: We have seen that $1+i$ has a polar form

$$1 + i = \sqrt{2}e^{i\pi/4}.$$

Then according to the previous formula, the three solutions to our cubic are

$$z = (\sqrt{2})^{1/3} e^{i(\pi/4 + 2k\pi)/3} = 2^{1/6} e^{i(1+8k)\pi/12}, \quad k = 0, 1, 2.$$

See Figure 1.6 for a graph of these complex roots. □

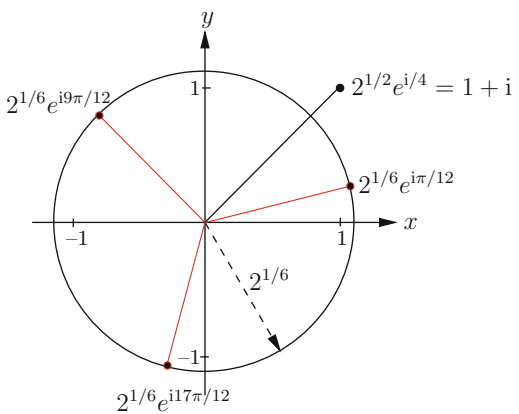


Fig. 1.6: Roots of $z^3 = 1 + i$.

We conclude with a little practice with square roots and the quadratic formula. In regard to square roots, as we have noted, the expression $w = \sqrt{d}$ is ambiguous when dealing with complex numbers. In view of this difference, we will generally avoid using the radical sign with complex numbers (exceptions: the quadratic formula and the case of $\sqrt{-d}$ with positive d, in which case the interpretation is $\sqrt{-d} = i\sqrt{d}$).

Example 1.14. Compute a square root of the numbers -4 and i.

Solution. Observe that $-4 = 4 \cdot (-1)$. It is reasonable to expect the laws of exponents to continue to hold, so we should have $(-4)^{1/2} = 4^{1/2} \cdot (-1)^{1/2}$.

Now we know that $\mathrm{i}^2 = -1$, so we can take $\mathrm{i} = (-1)^{1/2}$ and obtain that $\sqrt{-4} = (-4)^{1/2} = 2\mathrm{i}$. Let's check it: $(2\mathrm{i})^2 = 4\mathrm{i}^2 = -4$.

We have to be a bit more careful with i. We'll just borrow the idea of the formula for solving $z^n = d$. First, put i in polar form as $\mathrm{i} = 1 \cdot \mathrm{e}^{\mathrm{i}\pi/2}$. Now raise each side to the $1/2$ power to obtain

$$\begin{aligned} \mathrm{i}^{1/2} &= 1^{1/2} \cdot (\mathrm{e}^{\mathrm{i}\pi/2})^{1/2} \\ &= 1 \cdot \mathrm{e}^{\mathrm{i}\pi/4} = \cos(\pi/4) + \mathrm{i}\sin(\pi/4) \\ &= \frac{1}{\sqrt{2}}(1 + \mathrm{i}). \end{aligned}$$

A quick check confirms that $((1+\mathrm{i})/\sqrt{2})^2 = 2\mathrm{i}/2 = \mathrm{i}$. □

Example 1.15. Solve the equation $z^2 + z + 1 = 0$.

Solution. According to the quadratic formula, the answer is

$$z = \frac{-1 \pm \sqrt{1^2 - 4}}{2} = -\frac{1}{2} \pm \mathrm{i}\frac{\sqrt{3}}{2}.$$ □

Example 1.16. Solve $z^2 + z + 1 + \mathrm{i} = 0$ and factor the polynomial $z^2 + z + 1 + \mathrm{i}$.

Solution. This time we obtain from the quadratic formula that

$$z = \frac{-1 \pm \sqrt{1 - 4(1+\mathrm{i})}}{2} = \frac{-1 \pm \sqrt{-(3+4\mathrm{i})}}{2}.$$

What is interesting about this problem is that we don't know the polar angle θ for $z = -(3 + 4i)$. Fortunately, we don't have to. We know that $\sin\theta = -4/5$ and $\cos\theta = -3/5$. We also have the standard half angle formulas from trigonometry to help us:

$$\cos^2\theta/2 = \frac{1+\cos\theta}{2} = \frac{1}{5} \text{ and } \sin^2\theta/2 = \frac{1-\cos\theta}{2} = \frac{4}{5}.$$

Since θ is in the third quadrant of the complex plane, $\theta/2$ is in the second, so

$$\cos\theta/2 = \frac{-1}{\sqrt{5}} \text{ and } \sin\theta/2 = \frac{2}{\sqrt{5}}.$$

Notice that $|-(3+4\mathrm{i})| = 5$. Hence, a square root of $-(3+4\mathrm{i})$ is given by

$$w = \sqrt{5}\left(\frac{-1}{\sqrt{5}} + \frac{2}{\sqrt{5}}\mathrm{i}\right) = -1 + 2\mathrm{i}.$$

Check that $w^2 = -(3+4\mathrm{i})$, so the two roots to our quadratic equation are given by

$$z = \frac{-1 \pm (-1 + 2\mathrm{i})}{2} = -1 + \mathrm{i},\ -\mathrm{i}.$$

In particular, we see that $z^2 + z + 1 + \mathrm{i} = (z + 1 - \mathrm{i})(z + \mathrm{i})$. □

1.2 Exercises and Problems

In the following exercises, z is a complex number, and answers should be expressed in standard form if possible.

Exercise 1. Determine the following sets, given that $A = \{x \mid x \in \mathbb{R} \text{ and } x^2 < 3\}$ and $B = \{x \mid x \in \mathbb{Z} \text{ and } x > -1\}$:
(a) $A \cap B$ (b) $B - A$ (c) $\mathbb{Z} - B$ (d) $\mathbb{N} \cup B$ (e) $\mathbb{R} \cap A$

Exercise 2. Let $C = \{x \mid x \in \mathbb{Z} \text{ and } x^2 > 4\}$ and and determine the following sets:
(a) $C \cup D$ (b) $D - C$ (c) $D \cap \emptyset$ (d) $\mathbb{R} \cup D$

Exercise 3. Put the following complex numbers into polar form and sketch them in the complex plane:
(a) $-\mathrm{i}$ (b) $1 + \mathrm{i}$ (c) $-1 + \mathrm{i}\sqrt{3}$ (d) 1 (e) $2 - 2\mathrm{i}$ (f) $2\mathrm{i}$ (g) π

Exercise 4. Put the following complex numbers into polar form and sketch them in the complex plane:
(a) $3 + \mathrm{i}$ (b) i (c) $1 + \mathrm{i}\sqrt{3}$ (d) -1 (e) $3 - \mathrm{i}$ (f) $-\pi$ (g) e^{π}

Exercise 5. Calculate the following:
(a) $(4 + 2\mathrm{i}) - (3 - 6\mathrm{i})$ (b) $(2 + 4\mathrm{i})\,(3 - \mathrm{i})$ (c) $\dfrac{2 + \mathrm{i}}{2 - \mathrm{i}}$ (d) $\dfrac{1 - 2\mathrm{i}}{1 + 2\mathrm{i}}$ (e) $\overline{7\,(6 - \mathrm{i})}$

Exercise 6. Calculate the following:
(a) $|2 + 4\mathrm{i}|$ (b) $-7\mathrm{i}^2 + 6\mathrm{i}^3$ (c) $(3 + 4\mathrm{i})\,(7 - 6\mathrm{i})$ (d) $\overline{\mathrm{i}\,(1 - \mathrm{i})}$

Exercise 7. Solve the following systems for the unknown z:
(a) $(2 + \mathrm{i})z = 4 - 2\mathrm{i}$ (b) $z^4 = -16$ (c) $\dfrac{z + 1}{z} = 2$ (d) $(z + 1)(z^2 + 1) = 0$

Exercise 8. Solve the equations for the unknown z:
(a) $(2 + \mathrm{i})z = 1$ (b) $-\mathrm{i}z = 2z + 5$ (c) $\Im(z) = 2\Re(z) + 1$ (d) $\overline{z} = z$

Exercise 9. Find the polar and standard form of the complex numbers:
(a) $\dfrac{1}{1 - \mathrm{i}}$ (b) $-2\mathrm{e}^{\mathrm{i}\pi/3}$ (c) $\mathrm{i}\left(\mathrm{i} + \sqrt{3}\right)$ (d) $-\mathrm{i}/2$ (e) $\mathrm{i}\mathrm{e}^{\pi/4}$

Exercise 10. Find the polar and standard form of the complex numbers:
(a) $(2 + 4\mathrm{i})\,(3 - \mathrm{i})$ (b) $(2 + 4\mathrm{i}) - (3 + 3\mathrm{i})$ (c) $1/\mathrm{i}$ (d) $-1 + \mathrm{i}$ (e) $\mathrm{i}\mathrm{e}^{\mathrm{i}\pi/4}$

Exercise 11. Find all solutions to the following equations:
(a) $z^2 + z + 3 = 0$ (b) $z^2 - 1 = \mathrm{i}z$ (c) $z^2 - 2z + \mathrm{i} = 0$ (d) $z^2 + 4 = 0$

Exercise 12. Find the solutions to the following equations:
(a) $z^3 = 1$ (b) $z^3 = -8$ (c) $(z-1)^3 = -1$ (d) $z^4 + z^2 + 1 = 0$

Exercise 13. Describe and sketch the set of complex numbers z such that
(a) $|z| = 2$ (b) $|z+1| = |z-1|$ (c) $|z-2| < 1$
Hint: It's easier to work with absolute value squared.

Exercise 14. What is the set of complex numbers z such that
(a) $|z+1| = 2$ (b) $|z+3| = |z-1|$ (c) $|z-2| > 2$
Sketch these sets in the complex plane.

Exercise 15. Let $z_1 = 2 + 4\mathrm{i}$ and $z_2 = 1 - 3\mathrm{i}$. Verify for this z_1 and z_2 that $\overline{z_1} + \overline{z_2} = \overline{z_1 + z_2}$.

Exercise 16. Let $z_1 = 2 + 3\mathrm{i}$ and $z_2 = 2 - 3\mathrm{i}$. Verify for this z_1 and z_2 that $\overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$.

Exercise 17. Find the roots of the polynomial $p(z) = z^2 - 2z + 2$ and use this to factor the polynomial. Verify the factorization by expanding it.

Exercise 18. Show that $1 + \mathrm{i}, 1 - \mathrm{i}$, and 2 are roots of the polynomial $p(z) = z^3 - 4z^2 + 6z - 4$ and use this to factor the polynomial.

Exercise 19. Express the function $f(z)$ of the complex variable $z = x + \mathrm{i}y$ in standard form $g(x,y) + \mathrm{i}h(x,y)$, where $g(x,y)$ and $h(x,y)$ are real-valued functions and
(a) $f(z) = (z-2)^2$ (b) $f(z) = z^3 - 2z + 1$

Exercise 20. Express the function $f(z) = e^{z^2}$ of the complex variable z in standard form $g(x,y) + \mathrm{i}h(x,y)$.

Problem 21. Write out the values of i^k in standard form for integers $k = -1, 0, 1, 2, 3, 4$ and deduce a formula for i^k consistent with these values.

Problem 22. Verify that for any two complex numbers, the sum of the conjugates is the conjugate of the sum.

*Problem 23. Use the notation of Example 1.10 to show that $|z_1 z_2| = |z_1|\,|z_2|$.

Problem 24. Use the definitions of exponentials along with the sum of angles formulas for $\sin(\theta+\psi)$ and $\cos(\theta+\psi)$ to verify the law of addition of exponents: $\mathrm{e}^{\mathrm{i}(\theta+\psi)} = \mathrm{e}^{\mathrm{i}\theta}\mathrm{e}^{\mathrm{i}\psi}$.

Problem 25. Use a technology tool to find all roots to the polynomial equation $z^5 + z + 1 = 0$. How many roots (counting multiplicities) should this equation have? How many of these roots can you find with your system?

*Problem 26. Show that if w is a root of the polynomial $p(z)$, that is, $p(w) = 0$, where $p(z)$ has real coefficients, then $\overline{w}$ is also a root of $p(z)$.

1.3 Gaussian Elimination: Basic Ideas

We return now to the main theme of this chapter, which is the systematic solution of linear systems, as defined in equation (1.1) of Section 1.1. The principal methodology is the method of *Gaussian elimination* and its variants, which we introduce by way of a few simple examples. The idea of this process is to reduce a system of equations by certain legitimate and reversible algebraic operations (called "elementary operations") to a form in which we can easily see what the solutions to the system are, if there are any. Specifically, we want to get the system in a form where the first equation involves all the variables, the second equation involve all but the first, and so forth. Then it will be simple to solve for each variable one at a time, starting with the last equation, which will involve only the last variable. In a nutshell, this is Gaussian elimination.

One more matter that will have an effect on our description of solutions to a linear system is that of the number system in use. As we noted earlier, it is customary in linear algebra to refer to numbers as "scalars." The two basic choices of scalar fields are the real number system and the complex number system. Unless complex numbers occur explicitly in a linear system, we will assume that the scalars to be used in finding a solution come from the field of real numbers. Such will be the case for most of the problems in this chapter.

An Example and Some Shorthand

Example 1.17. Solve the simple system

$$\begin{aligned} 2x - y &= 1 \\ 4x + 4y &= 20. \end{aligned} \tag{1.5}$$

Solution. First, let's switch the equations to obtain

$$\begin{aligned} 4x + 4y &= 20 \\ 2x - y &= 1. \end{aligned} \tag{1.6}$$

Next, multiply the first equation by $1/4$ to obtain

$$\begin{aligned} x + y &= 5 \\ 2x - y &= 1. \end{aligned} \tag{1.7}$$

Now, multiply a copy of the first equation by -2 and add it to the second. We can do this easily if we take care to combine like terms as we go. In particular, the resulting x term in the new second equation will be $-2x + 2x = 0$, the y term will be $-2y - y = -3y$, and the constant term on the right-hand side will be $-2 \cdot 5 + 1 = -9$. Thus, we obtain

$$\begin{aligned} x + y &= 5 \\ 0x - 3y &= -9. \end{aligned} \tag{1.8}$$

This completes the first phase of Gaussian elimination, which is called "forward solving." Note that we have put the system in a form in which only the first equation involves the first variable and only the first and second involve the second variable. The second phase of Gaussian elimination is called "back solving," and it works like it sounds. Use the last equation to solve for the last variable, then work backward, solving for the remaining variables in reverse order. In our case, the second equation is used to solve for y simply by dividing by -3 to obtain that

$$y = \frac{-9}{-3} = 3.$$

Finally, use our knowledge of y and the first equation to solve for x, to obtain

$$x = 5 - y = 5 - 3 = 2. \qquad \square$$

The preceding example may seem like too much work for such a simple system. We could easily scratch out the solution in much less space. But what if the system is larger, say 4 equations in 4 unknowns, or more? How do we proceed then? It pays to have a systematic strategy and notation. We also had an ulterior motive in the way we solved this system. All of the operations we will ever need to solve a linear system were illustrated in the preceding example: switching equations, multiplying equations by nonzero scalars, and adding a multiple of one equation to another.

Before proceeding to another example, let's work on the notation a bit. Take a closer look at the system of equations (1.5). As long as we write numbers down systematically, there is no need to write out all the equal signs or plus signs. Isn't every bit of information that we require contained in the following table of numbers?

$$\begin{bmatrix} 2 & -1 & 1 \\ 4 & 4 & 20 \end{bmatrix}.$$

Of course, we have to remember that each row of numbers represents an equation, the first two columns of numbers are coefficients of x and y, respectively, and the third column consists of terms on right-hand side. So we could embellish the table with a few reminders in an extra top row:

$$\begin{matrix} x & y & = & \text{r.h.s.} \end{matrix}$$
$$\begin{bmatrix} 2 & -1 & & 1 \\ 4 & 4 & & 20 \end{bmatrix}$$

With a little practice, we will find that the reminders are usually unnecessary, so we dispense with them. Rectangular tables of numbers are very useful in representing a system of equations. Such a table is one of the basic objects studied in this text. As such, it warrants a formal definition.

Definition 1.5. Matrices and Vectors A *matrix* is a rectangular array of numbers. If a matrix has m rows and n columns, then the *size* of the matrix is said to be $m \times n$. If the matrix is $1 \times n$ or $m \times 1$, it is called a *vector*. If $m = n$, then it is called a *square matrix of order* n. Finally, the number that occurs in the ith row and jth column is called the (i, j)th *entry* of the matrix.

The objects we have just defined are basic "quantities" of linear algebra and matrix analysis, along with scalar quantities. Although every vector is itself a matrix, we want to single vectors out when they are identified as such. Therefore, we will follow a standard typographical convention: Matrices are usually designated by capital letters, while vectors are usually designated by boldface lowercase letters. In a few cases these conventions are not followed, but the meaning of the symbols should be clear from context.

We shall need to refer to parts of a matrix. As indicated above, the location of each entry of a matrix is determined by the index of the row and column it occupies.

The statement "$A = [a_{ij}]$" means that A is a matrix whose (i, j)th entry is denoted by a_{ij} (or $a_{i,j}$ to separate indices). Generally, the size of A will be clear from context. If we want to indicate that A is an $m \times n$ matrix, we write

$$A = [a_{ij}]_{m,n} .$$

Similarly, the statement "$\mathbf{b} = [b_i]$" means that b is a n-vector whose ith entry is denoted by b_i. In case the type of the vector (row or column) is not clear from context, the default is a column vector. Many of the matrices we encounter will be *square*, that is, $n \times n$. In this case we say that n is the *order* of the matrix. Another term that we will use frequently is the following:

Order of Square Matrix

Definition 1.6. Leading Entry The *leading entry* of a row vector is the first nonzero element of that vector, counting from left to right. If all entries are zero, the vector has no leading entry.

The equations of (1.5) have several matrices associated with them. First is the full matrix that describes the system, which we call the *augmented matrix* of the system. In our previous example, this is the 2×3 matrix

$$\begin{bmatrix} 2 & -1 & 1 \\ 4 & 4 & 20 \end{bmatrix} .$$

Note, for example, that we would say that the $(1, 1)$th entry of this matrix is 2, which is also the leading entry of the first row, and the $(2, 3)$th entry is 20. Next, there is the submatrix consisting of coefficients of the variables. This is called the *coefficient matrix* of the system, in our case the 2×2 matrix

$$\begin{bmatrix} 2 & -1 \\ 4 & 4 \end{bmatrix}.$$

Finally, there is the single column matrix of right-hand-side constants, which we call the right-hand-side vector. In our example, it is the 2×1 vector

$$\begin{bmatrix} 1 \\ 20 \end{bmatrix}.$$

How can we describe the matrices of the general linear system of equations specified by (1.1)?

Coefficient Matrix

First, there is the $m \times n$ *coefficient matrix*

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}.$$

Notice that the way we subscripted entries of this matrix is really very descriptive: the first index indicates the row position of the entry, and the second, the column position of the entry.

Right-Hand-Side Vector

Next, there is the $m \times 1$ *right-hand-side vector* of constants

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_i \\ \vdots \\ b_m \end{bmatrix}.$$

Augmented Matrix

Finally, stack this matrix and vector along side each other (we use a vertical bar below to separate the two symbols) to obtain the $m \times (n+1)$ *augmented matrix*

$$\widetilde{A} = [A \mid \mathbf{b}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} & b_i \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} & b_m \end{bmatrix}.$$

Example 1.18. Describe the associated matrices for a linear system that solves the problem of finding a polynomial that interpolates a specified set of points.

Solution. Suppose that the points in question are (x_i, y_i), $i = 0, 1, \ldots, n$, with all abscissas x_i distinct. Just as it takes two such points to uniquely determine a linear function, three to determine a quadratic function, and so forth, it is reasonable to expect that $n+1$ points will determine an nth degree polynomial of the form

$$p(x) = c_0 + c_1 x + \cdots + c_n x^n.$$

The conditions of interpolation are simply that $p(x_i) = y_i$, $i = 0, 1, \cdots, n$. These conditions lead to the linear system

$$p(x_i) = c_0 + c_1 x_i + \cdots + c_n x_i^n, \ i = 0, 1, \cdots, n$$

in the $n+1$ unknowns $c_0, c_1, \ldots, c_n$. The coefficient matrix for this system is the $(n+1) \times (n+1)$ matrix

$$A = \begin{bmatrix} 1 & x_0 & \cdots & x_0^j & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^j & \cdots & x_1^n \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^j & \cdots & x_n^0 \end{bmatrix}$$

and the augmented matrix for this system is

$$\widetilde{A} = [A \mid \mathbf{b}] = \begin{bmatrix} 1 & x_0 & \cdots & x_0^j & \cdots & x_0^n & y_0 \\ 1 & x_1 & \cdots & x_1^j & \cdots & x_1^n & y_1 \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 1 & x_n & \cdots & x_n^j & \cdots & x_n^0 & y_m \end{bmatrix}.$$

The system coefficient matrix A is called a *Vandermonde matrix.* □

The Elementary Row Operations

Here is more notation that we will find extremely handy in the sequel. This notation is related to the operations that we performed on the preceding example. Now that we have the matrix notation, we could just as well perform these operations on each row of the augmented matrix, since a row corresponds to an equation in the original system. Three types of operations were used. We shall catalog these and give them names, so that we can document our work in solving a system of equations in a concise way. Here are the three elementary operations we shall use, described in terms of their action on rows of a matrix; an entirely equivalent description applies to the equations of the linear system whose augmented matrix is the matrix below.

Notation for Elementary Operations

- E_{ij}: This is shorthand for the elementary operation of *switching the ith and jth rows* of the matrix. For instance, in Example 1.17 we moved from equation (1.5) to equation (1.6) by using the elementary operation E_{12}.
- $E_i(c)$: This is shorthand for the elementary operation of *multiplying the ith row by the nonzero constant c*. For instance, we moved from equation (1.6) to equation (1.7) by using the elementary operation $E_1(1/4)$.
- $E_{ij}(d)$: This is shorthand for the elementary operation of *adding d times the jth row to the ith row.* (Read the symbols from right to left to get the correct order.) For instance, we moved from equation (1.7) to equation (1.8) by using the elementary operation $E_{21}(-2)$.

Now let's put it all together. The whole forward-solving phase of Example 1.17 could be described concisely with the notation we have developed:

$$\begin{bmatrix} 2 & -1 & 1 \\ 4 & 4 & 20 \end{bmatrix} \xrightarrow{E_{12}} \begin{bmatrix} 4 & 4 & 20 \\ 2 & -1 & 1 \end{bmatrix} \xrightarrow{E_1(1/4)} \begin{bmatrix} 1 & 1 & 5 \\ 2 & -1 & 1 \end{bmatrix} \xrightarrow{E_{21}(-2)} \begin{bmatrix} 1 & 1 & 5 \\ 0 & -3 & -9 \end{bmatrix}.$$

This is a big improvement over our first description of the solution. There is still the job of back solving, which is the second phase of Gaussian elimination. When doing hand calculations, we're right back to writing out a bunch of extra symbols again, which is exactly what we set out to avoid by using matrix notation.

Gauss–Jordan Elimination

Here's a better way to do the second phase by hand: Stick with the augmented matrix. Starting with the last nonzero row, convert the leading entry (this means the first nonzero entry in the row) to a 1 by an elementary operation, and then use elementary operations to convert all entries above this 1 entry to 0's. Now work backward, row by row, up to the first row. At this point we can read off the solution to the system. Let's see how it works with Example 1.17. Here are the details using our shorthand for elementary operations:

$$\begin{bmatrix} 1 & 1 & 5 \\ 0 & -3 & -9 \end{bmatrix} \xrightarrow{E_2(-1/3)} \begin{bmatrix} 1 & 1 & 5 \\ 0 & 1 & 3 \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix}.$$

All we have to do is remember the function of each column in order to read off the answer from this last matrix. The underlying system that is represented is

$$\begin{aligned} 1 \cdot x + 0 \cdot y &= 2 \\ 0 \cdot x + 1 \cdot y &= 3. \end{aligned}$$

This is, of course, the answer we found earlier: $x = 2$, $y = 3$.

The method of combining forward and back solving into elementary operations on the augmented matrix has a name: It is called *Gauss–Jordan elimination*, and it is the method of choice for solving many linear systems. Let's see how it works on an example.

Example 1.19. Solve the following system by Gauss–Jordan elimination:

$$\begin{aligned} x + y + z &= 4 \\ 2x + 2y + 4z &= 11 \\ 4x + 6y + 8z &= 24 \end{aligned}$$

Solution. First form the augmented matrix of the system, the 3×4 matrix

$$\begin{bmatrix} 1 & 1 & 1 & 4 \\ 2 & 2 & 4 & 11 \\ 4 & 6 & 8 & 24 \end{bmatrix}.$$

Now forward solve:

$$\begin{bmatrix} 1 & 1 & 1 & 4 \\ 2 & 2 & 4 & 11 \\ 4 & 6 & 8 & 24 \end{bmatrix} \xrightarrow{E_{21}(-2)} \begin{bmatrix} 1 & 1 & 1 & 4 \\ 0 & 0 & 2 & 3 \\ 4 & 6 & 8 & 24 \end{bmatrix} \xrightarrow{E_{31}(-4)} \begin{bmatrix} 1 & 1 & 1 & 4 \\ 0 & 0 & 2 & 3 \\ 0 & 2 & 4 & 8 \end{bmatrix} \xrightarrow{E_{23}} \begin{bmatrix} 1 & 1 & 1 & 4 \\ 0 & 2 & 4 & 8 \\ 0 & 0 & 2 & 3 \end{bmatrix}.$$

Notice, by the way, that the row switch of the third step is essential. Otherwise, we cannot use the second equation to solve for the second variable, y. Next back solve:

$$\begin{bmatrix} 1 & 1 & 1 & 4 \\ 0 & 2 & 4 & 8 \\ 0 & 0 & 2 & 3 \end{bmatrix} \xrightarrow{E_3(1/2)} \begin{bmatrix} 1 & 1 & 1 & 4 \\ 0 & 2 & 4 & 8 \\ 0 & 0 & 1 & \frac{3}{2} \end{bmatrix} \xrightarrow{E_{23}(-4)} \begin{bmatrix} 1 & 1 & 1 & 4 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 1 & \frac{3}{2} \end{bmatrix}$$

$$\xrightarrow{E_{13}(-1)} \begin{bmatrix} 1 & 1 & 0 & \frac{5}{2} \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 1 & \frac{3}{2} \end{bmatrix} \xrightarrow{E_2(1/2)} \begin{bmatrix} 1 & 1 & 0 & \frac{5}{2} \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & \frac{3}{2} \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} 1 & 0 & 0 & \frac{3}{2} \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & \frac{3}{2} \end{bmatrix}.$$

At this point we read off the solution to the system: $x = 3/2$, $y = 1$, $z = 3/2$. □

Systems with Non-unique Solutions

Next, we consider an example that will pose a new kind of difficulty, namely, that of infinitely many solutions. Here is some handy terminology.

Pivots An entry of a matrix used to zero out entries above or below it by means of elementary row operations is called a *pivot*.

The entries that we use in Gaussian or Gauss–Jordan elimination for pivots are always leading entries in the row that they occupy. For the sake of emphasis, in the next few examples we will put a circle around the pivot entries as they occur.

Example 1.20. Solve for the variables x, y, and z in the system

$$\begin{aligned} z &= 2 \\ x + y + z &= 2 \\ 2x + 2y + 4z &= 8 \end{aligned}$$

.

Solution. Here the augmented matrix of the system is

$$\begin{bmatrix} 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 2 \\ 2 & 2 & 4 & 8 \end{bmatrix}.$$

Now proceed to use Gaussian elimination on the matrix:

$$\begin{bmatrix} 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 2 \\ 2 & 2 & 4 & 8 \end{bmatrix} \xrightarrow{E_{12}} \begin{bmatrix} \textcircled{1} & 1 & 1 & 2 \\ 0 & 0 & 1 & 2 \\ 2 & 2 & 4 & 8 \end{bmatrix} \xrightarrow{E_{31}(-2)} \begin{bmatrix} \textcircled{1} & 1 & 1 & 2 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

What do we do next? Neither the second nor the third row corresponds to equations that involve the variable y. Switching the second and third equations won't help, either. So here is the point of view that we adopt in applying Gaussian elimination to this system: The first equation has already been "used up" and is reserved for eventually solving for x. We now restrict our attention to the "unused" second and third equations. Perform the following operations to do Gauss–Jordan elimination on the system:

$$\begin{bmatrix} \textcircled{1} & 1 & 1 & 2 \\ 0 & 0 & \textcircled{2} & 4 \\ 0 & 0 & 1 & 2 \end{bmatrix} \xrightarrow{E_2(1/2)} \begin{bmatrix} \textcircled{1} & 1 & 1 & 2 \\ 0 & 0 & \textcircled{1} & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$$\xrightarrow{E_{32}(-1)} \begin{bmatrix} \textcircled{1} & 1 & 1 & 2 \\ 0 & 0 & \textcircled{1} & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} \textcircled{1} & 1 & 0 & 0 \\ 0 & 0 & \textcircled{1} & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

How do we interpret this result? We take the point of view that the first row represents an equation to be used in solving for x since the leading entry of the row is in the column of coefficients of x. Similarly, the second row represents an equation to be used in solving for z, since the leading entry of that row is in the column of coefficients of z. What about y? Notice that the third equation represented by this matrix is simply $0 = 0$, which carries no information. The point is that there is not enough information in the system to solve for the variable y, even though we started with three distinct equations. Somehow, they contained redundant information. Therefore, we take the point of view that y is not to be solved for; it is a *free* variable in the sense that we can assign it any value whatsoever and obtain a legitimate solution to the system. On the other hand, the variables x and z are *bound* in the sense that they will be solved for in terms of constants and free variables. The equations represented by the last matrix above are

Free and Bound Variables

$$\begin{aligned} x + y &= 0 \\ z &= 2 \\ 0 &= 0. \end{aligned}$$

Use the first equation to solve for x and the second to solve for z to obtain the *general form* of a solution to the system:

$$\begin{aligned} x &= -y \\ z &= 2 \\ y &\text{ is free.} \end{aligned}$$ □

In the preceding example y can take on any scalar value. For example, $x = 0$, $z = 2$, $y = 0$ is a solution to the original system (check this). Likewise, $x = -5$, $z = 2$, $y = 5$ is a solution to the system. Clearly, we have an infinite number of solutions to the system, thanks to the appearance of free variables. Up to this point, the linear systems we have considered had unique solutions, so every variable was solved for, and hence bound. Another point to note, incidentally, is that the scalar field we choose to work with has an effect on our answer. The default is that y is allowed to take on any *real* value from $\mathbb{R}$. But if, for some reason, we choose to work with the complex numbers as our scalars, then y would be allowed to take on any *complex* value from $\mathbb{C}$. In this case, another solution to the system would be given by $x = -3 - \mathrm{i}$, $z = 2$, $y = 3 + \mathrm{i}$, for example.

To summarize, once we have completed Gauss–Jordan elimination on an augmented matrix, we can immediately spot the free and bound variables of the system: The column of a bound variable will have a pivot in it, while the column of a free variable will not. Another example will illustrate the point.

Example 1.21. Suppose the augmented matrix of a linear system of three equations involving variables x, y, z, w becomes, after applying suitable elementary row operations,

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

Describe the general solution to the system.

Solution. We solve this problem by observing that the first and third columns have pivots in them, which the second and fourth do not. The fifth column represents the right-hand side. Put our little reminder labels in the matrix, and we obtain

$$\begin{bmatrix} x & y & z & w & \text{rhs} \\ \textcircled{1} & 2 & 0 & -1 & 2 \\ 0 & 0 & \textcircled{1} & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Hence, x and z are bound variables, while y and w are free. The two nontrivial equations that are represented by this matrix are

$$\begin{aligned} x + 2y - w &= 2 \\ z + 3w &= 0. \end{aligned}$$

Use the first to solve for x and the second to solve for z to obtain the general solution

$$\begin{aligned} x &= 2 - 2y + w \\ z &= -3w \\ y, w &\text{ are free.} \end{aligned}$$

□

We have seen so far that a linear system may have exactly one solution or infinitely many. Actually, there is only one more possibility, which is illustrated by the following example.

Example 1.22. Solve the linear system

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

Solution. We extract the augmented matrix and proceed with Gauss–Jordan elimination. This time we'll save a little space by writing more than one elementary operation between matrices. It is understood that they are done in order, starting with the top one. This is a very efficient way of doing hand calculations and minimizing the amount of rewriting of matrices as we go:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 3 & 2 & 5 \end{bmatrix} \xrightarrow[E_{31}(-3)]{E_{21}(-2)} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & 0 \\ 0 & -1 & 2 \end{bmatrix} \xrightarrow{E_{32}(-1)} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Stop everything! We aren't done with Gauss–Jordan elimination yet, since we've only done the forward-solving portion. But something strange is going on here. Notice that the third row of the last matrix above stands for the equation $0x+0y = 2$, i.e., $0 = 2$. This is impossible. What this matrix is telling us is that the original system has no solution, i.e., it is *inconsistent*. A system can be identified as inconsistent as soon as one encounters a leading entry in the column of constant terms. For this always means that an equation of the form $0 =$ nonzero constant has been formed from the system by legitimate algebraic operations. Thus, one need proceed no further. The system has no solutions. □

Definition 1.7. **Consistent System** A system of equations is *consistent* if it has at least one solution. Otherwise it is called *inconsistent*.

Our last example is one involving complex numbers explicitly.

Example 1.23. Solve the following system of equations:

$$\begin{aligned} x + y &= 4 \\ (-1 + \mathrm{i})x + y &= -1. \end{aligned}$$

Solution. The procedure is the same, no matter what the field of scalars is. Of course, the arithmetic is a bit harder. Gauss–Jordan elimination yields

$$\begin{bmatrix} 1 & 1 & 4 \\ -1+\mathrm{i} & 1 & -1 \end{bmatrix} \xrightarrow{E_{21}(1-\mathrm{i})} \begin{bmatrix} 1 & 1 & 4 \\ 0 & 2-\mathrm{i} & 3-4\mathrm{i} \end{bmatrix}$$

$$\xrightarrow{E_2(1/(2-\mathrm{i}))} \begin{bmatrix} 1 & 1 & 4 \\ 0 & 1 & 2-\mathrm{i} \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} 1 & 0 & 2+\mathrm{i} \\ 0 & 1 & 2-\mathrm{i} \end{bmatrix}.$$

Here we used the fact that

$$\frac{3-4\mathrm{i}}{2-\mathrm{i}} = \frac{(3-4\mathrm{i})(2+\mathrm{i})}{(2-\mathrm{i})(2+\mathrm{i})} = \frac{10-5\mathrm{i}}{5} = 2-\mathrm{i}.$$

Thus, we see that the system has the unique solution

$$\begin{aligned} x &= 2+\mathrm{i} \\ y &= 2-\mathrm{i}. \end{aligned}$$

□

1.3 Exercises and Problems

Exercise 1. For each of the following matrices identify the size and the (i,j)th entry for all relevant indices i and j:

(a) $\begin{bmatrix} 1 & -1 & 2 & 1 \\ -2 & 2 & 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 1 \\ 2 & -1 \\ 0 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} -2 \\ 3 \end{bmatrix}$ (d) $[1+\mathrm{i}]$

Exercise 2. For each of the following matrices identify the size and the (i,j)th entry for all relevant indices i and j:

(a) $\begin{bmatrix} 1 & -1 & 0 \\ 0 & 2 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 1 & 3 \end{bmatrix}$ (d) $\begin{bmatrix} 3 \\ \mathrm{i} \end{bmatrix}$

Exercise 3. Exhibit the augmented matrix of each system and give its size. Then use Gaussian elimination and back solving to find the general solution to the systems.

(a) $\begin{aligned} 2x+3y &= 7 \\ x+2y &= -2 \end{aligned}$ (b) $\begin{aligned} 3x_1+6x_2-x_3 &= -4 \\ -2x_1-4x_2+x_3 &= 3 \\ x_3 &= 1 \end{aligned}$ (c) $\begin{aligned} x_1+x_2 &= -2 \\ 5x_1+2x_2 &= 5 \\ x_1+2x_2 &= -7 \end{aligned}$

Exercise 4. Use Gaussian elimination and back solving to find the general solution to the systems.

(a) $\begin{aligned} x+3y &= 7 \\ x+2y &= 1 \end{aligned}$ (b) $\begin{aligned} 2x_1+6x_2 &= 2 \\ -2x_1+x_2 &= 1 \end{aligned}$ (c) $\begin{aligned} x_1+x_2 &= 1 \\ 5x_1+2x_2 &= 5 \\ x_1+2x_2 &= -7 \end{aligned}$

Exercise 5. Use Gauss–Jordan elimination to find the general solution to the systems. Show the elementary operations you use.

$$\text{(a)}\quad \begin{aligned} x_1 + x_2 &= 1 \\ 2x_1 + 2x_2 + x_3 &= 1 \\ 2x_1 + 2x_2 &= 2 \end{aligned} \qquad \text{(b)}\quad \begin{aligned} x_3 + x_4 &= 1 \\ -2x_1 - 4x_2 + x_3 &= 0 \\ 3x_1 + 6x_2 + x_4 &= 0 \end{aligned} \qquad \text{(c)}\quad \begin{aligned} x_1 + x_2 + 3x_3 &= 2 \\ 2x_1 + 5x_2 + 9x_3 &= 1 \\ x_1 + 2x_2 + 4x_3 &= 1 \end{aligned}$$

$$\text{(d)}\quad \begin{aligned} x_1 - x_2 &= \mathrm{i} \\ 2x_1 + x_2 &= 3 + \mathrm{i} \end{aligned} \qquad \text{(e)}\quad \begin{aligned} x_1 + x_2 + x_3 - x_4 &= 0 \\ -2x_1 - 4x_2 + x_3 &= 0 \\ x_1 + 6x_2 - x_3 + x_4 &= 0 \end{aligned}$$

Exercise 6. Use Gauss–Jordan elimination to find the general solution to the systems.

$$\text{(a)}\quad \begin{aligned} x_1 + x_2 + x_4 &= 1 \\ 2x_1 + 2x_2 + x_3 + x_4 &= 1 \\ 2x_1 + 2x_2 + 2x_4 &= 2 \end{aligned} \qquad \text{(b)}\quad \begin{aligned} x_3 + x_4 &= 0 \\ -2x_1 - 4x_2 + x_3 &= 0 \\ -x_3 + x_4 &= 0 \end{aligned} \qquad \text{(c)}\quad \begin{aligned} x_1 + x_2 + 3x_3 &= 2 \\ 2x_1 + 5x_2 + 9x_3 &= 1 \\ x_1 + 2x_2 + 4x_3 &= 1 \end{aligned}$$

$$\text{(d)}\quad \begin{aligned} 2x_1 + x_2 + 7x_3 &= -1 \\ 3x_1 + 2x_2 - 2x_4 &= 1 \\ 2x_1 + 2x_2 + 2x_3 - 2x_4 &= 4 \end{aligned} \qquad \text{(e)}\quad \begin{aligned} x_1 + x_2 + x_3 &= 2 \\ 2x_1 + x_2 &= \mathrm{i} \\ 2x_1 + 2x_2 + \mathrm{i}x_3 &= 4 \end{aligned}$$

Exercise 7. Each of the following matrices results from applying Gauss–Jordan elimination to the augmented matrix of a linear system. In each case, write out the general solution to the system or indicate that it is inconsistent.

$$\text{(a)}\ \begin{bmatrix} 1 & 0 & 0 & 4 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \text{(b)}\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix} \qquad \text{(c)}\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \text{(d)}\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Exercise 8. Write out the general solution to the system with the following augmented matrix or indicate that it is inconsistent.

$$\text{(a)}\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \text{(b)}\ \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix} \qquad \text{(c)}\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -2 \\ 1 & 0 & 0 & 1 \end{bmatrix} \qquad \text{(d)}\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Exercise 9. Use Gauss–Jordan elimination to solve the system resulting from the second approach to page ranking described in Example 1.6. Discuss the result.

Exercise 10. Use Gauss–Jordan elimination to solve the system resulting from the third approach to page ranking described in Example 1.6. Discuss the result.

Exercise 11. Use any method to find the solution to each of the following systems. Here, b_1, b_2 are constants and x_1, x_2 are the unknowns.

$$\text{(a)}\quad \begin{aligned} x_1 - x_2 &= b_1 \\ x_1 + 2x_2 &= b_2 \end{aligned} \qquad \text{(b)}\quad \begin{aligned} x_1 - x_2 &= b_1 \\ 2x_1 - 2x_2 &= b_2 \end{aligned} \qquad \text{(c)}\quad \begin{aligned} \mathrm{i}x_1 - x_2 &= b_1 \\ 2x_1 + 2x_2 &= b_2 \end{aligned}$$

Exercise 12. Apply the operations used in Exercise 5 (a), (c) in the same order to the right-hand-side vector $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$. What does this tell you about each system's consistency?

Exercise 13. Solve the three systems

$$\text{(a)}\ \begin{aligned} x_1 - x_2 &= 1 \\ x_1 - 2x_2 &= 0 \end{aligned} \qquad \text{(b)}\ \begin{aligned} x_1 - x_2 &= 0 \\ x_1 - 2x_2 &= 1 \end{aligned} \qquad \text{(c)}\ \begin{aligned} x_1 - x_2 &= 2 \\ x_1 - 2x_2 &= 3 \end{aligned}$$

by using a single augmented matrix that has all three right-hand sides in it.

Exercise 14. Set up a single augmented matrix for the three systems

$$\text{(a)}\ \begin{aligned} x_1 + x_2 &= 1 \\ x_2 + 2x_3 &= 0 \\ 2x_2 + x_3 &= 0 \end{aligned} \qquad \text{(b)}\ \begin{aligned} x_1 + x_2 &= 0 \\ x_2 + 2x_3 &= 0 \\ 2x_2 + x_3 &= 0 \end{aligned} \qquad \text{(c)}\ \begin{aligned} x_1 + x_2 &= 2 \\ x_2 + 2x_3 &= 3 \\ 2x_2 + x_3 &= 3 \end{aligned}$$

and use it to solve the three systems simultaneously.

Exercise 15. Find the general solution to the linear system of Exercise 9 of Section 1.1. Are there any meaningful solutions?

Exercise 16. Find the general solution to the linear system of Example 1.4 of Section 1.1. Are there any meaningful solutions?

Exercise 17. Show that the following nonlinear systems become linear if we view the unknowns as $1/x$, $1/y$, and $1/z$ rather than x, y, and z. Use this to find the solution sets of the nonlinear systems. (You must also account for the possibilities that one of x, y, z is zero.)

$$\text{(a)}\ \begin{aligned} 2x - y + 3xy &= 0 \\ 4x + 2y - xy &= 0 \end{aligned} \qquad \text{(b)}\ \begin{aligned} yz + 3xz - xy &= 0 \\ yz + 2xy &= 0 \end{aligned}$$

Exercise 18. Show that the following nonlinear systems become linear if we make the right choice of unknowns from x, y, z, $1/x$, $1/y$, and $1/z$ rather than x, y, and z. Use this to find the solution sets of these nonlinear systems.

$$\text{(a)}\ \begin{aligned} 3x - xy &= 1 \\ 4x - xy &= 2 \end{aligned} \qquad \text{(b)}\ \begin{aligned} 2xy &= 1 \\ y + z - 3yz &= 0 \\ xz - 2z &= -1 \end{aligned}$$

Exercise 19. Exhibit the coefficient and augmented matrix for the system that finds a quadratic polynomial interpolating the points $(0, 2)$, $(1, 2)$ and $(2, 4)$ as in Example 1.18. Solve this system to determine the polynomial.

Exercise 20. There is no quadratic polynomial interpolating the points $(0, 2)$, $(2, 2)$, $(-1, 4)$ and $(2, 3)$. Exhibit the coefficient and augmented matrix for the interpolating system and show how this conclusion can be drawn from inspection of the augmented matrix of this system.

*Problem 21. Suppose that the input–output table of Example 1.5 of Section 1.1 is modified so that all entries are nonnegative, but the sum of the entries in each row is smaller than 1. Show that the only solution to the corresponding system with nonnegative values is the solution with all variables equal to zero.

Problem 22. Use a technology tool to solve the system of Example 1.3 with $f(x) = \sin(\pi x)$. Graph this approximation along with the true solution, which is $y(x) = \sin(\pi x)/\pi^2$.

*Problem 23. Suppose the function $f(x)$ is to be interpolated at three interpolating points x_0, x_1, x_2 by a quadratic polynomial $p(x) = a + bx + cx^2$, that is, $f(x_i) = p(x_i), i = 0, 1, 2$. As in Exercise 17 of Section 1.1, this leads to a system of three linear equations in the three unknowns a, b, c.
(a) Solve these equations in the case that $f(x) = e^x$, $0 \le x \le 1$, and $x_j = 0, \frac{1}{2}, 1$.
(b) Plot the error function $f(x) - p(x)$ and estimate the largest value of the error function (in absolute value).
(c) Use trial and error to find three points x_1, x_2, x_3 on the interval $0 \le x \le 1$ for which the resulting interpolating quadratic gives an error function with a largest absolute error that is less than half of that found in part (b).

Problem 24. Solve the network system of Problem 18 of Section 1 and exhibit all physically meaningful solutions.

Problem 25. Suppose one wants to solve the integral equation $\int_0^1 e^{st} x(s) ds = 1 + t^2$ for the unknown function $x(t)$. If we only want to approximate the values of $x(t)$ at $x = 0, \frac{1}{2}, 1$, derive and solve a system of equations for these three values by evaluating the integral equation at $t = 0, \frac{1}{2}, 1$, and using the trapezoidal method to approximate the integrals with the values of $x(s)$, $s = 0, \frac{1}{2}, 1$.

*Problem 26. Treat the network system of Problem 18 of Section 1.1 as a web network of pages and apply the third approach to page ranking described on page 9 of Section 1.1 to rank the importance of the hardware unit locations.

1.4 Gaussian Elimination: General Procedure

The preceding section introduced Gaussian elimination and Gauss–Jordan elimination at a practical level. In this section we will see why these methods work and what they really mean in matrix terms. Then we will find conditions of a very general nature under which a linear system has either no, one, or infinitely many solutions. A key idea that comes out of this section is the notion of the *rank* of a matrix.

Equivalent Systems

The first question to be considered is this: How is it that Gaussian elimination or Gauss–Jordan elimination gives us *every* solution of the system we begin with and *only* solutions to that system? To see that linear systems are special, consider the following nonlinear system of equations.

Example 1.24. Solve for the real roots of the system

$$\begin{aligned} x + y &= 2 \\ \sqrt{x} &= y. \end{aligned}$$

Solution. Let's follow the Gauss–Jordan elimination philosophy of using one equation to solve for one unknown. The first equation enables us to solve for y to get $y = 2 - x$. Substitute this into the second equation to obtain $\sqrt{x} = 2 - x$. Then square both sides to obtain $x = (2 - x)^2$, or

$$0 = x^2 - 5x + 4 = (x - 1)(x - 4).$$

Now $x = 1$ leads to $y = 1$, a solution to the system. But $x = 4$ gives $y = -2$, which is not a solution to the system since $\sqrt{x}$ cannot be negative. □

What went wrong in this example is that the squaring step, which does not correspond to any elementary operation, introduced extraneous solutions to the system. Is Gaussian or Gauss–Jordan elimination safe from this kind of difficulty? The answer lies in examining the kinds of operations we perform with these methods. First, we need some terminology. Up to this point we have always described a solution to a linear system in terms of a list of equations. For general problems this is a bit of a nuisance. Since we are using the matrix/vector notation, we may as well go all the way and use it to concisely describe solutions as well. We will use column vectors to define solutions as follows.

Definition 1.8. **Solution Vector** A *solution vector* for the general linear system of equation (1.1) is a vector

$$\mathbf{x} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix}$$

such that the resulting equations are satisfied for these choices of the variables. The set of all such solutions is called the *solution set* of the linear system, and two linear systems are said to be *equivalent* if they have the same solution set.

We will want to make frequent reference to vectors without having to display them in the text. Of course, for $1 \times n$ row vectors this is no problem.

To save space in referring to column vectors, we shall adopt the convention that a column vector will also be denoted by a tuple with the same entries. The n-tuple $(x_1, x_2, \ldots, x_n)$ is a shorthand for the $n \times 1$ column vector $\mathbf{x}$ with entries $x_1, x_2, \ldots, x_n$. For example, we can write $(1, 3, 2)$ in place of $\begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$.

Tuple Convention

Example 1.25. Describe the solution sets of all the examples worked out in the previous section.

Solution. Here is the solution set to Example 1.17. It is the singleton set

$$S = \left\{ \begin{bmatrix} 2 \\ 3 \end{bmatrix} \right\} = \{(2, 3)\}.$$

The solution set for Example 1.19 is $S = \left\{ \left(\frac{3}{2}, 1, \frac{3}{2} \right) \right\}$; remember that we can designate column vectors by tuples if we wish.

For Example 1.20 the solution set requires some fancier set notation, since it is an infinite set. Here it is:

$$S = \left\{ \begin{bmatrix} -y \\ y \\ 2 \end{bmatrix} \mid y \in \mathbb{R} \right\} = \{(-y, y, 2) \mid y \in \mathbb{R}\}.$$

Example 1.22 is an inconsistent system, so has no solutions. Hence, its solution set is $S = \emptyset$. Finally, the solution set for Example 1.23 is the singleton set $S = \{(2 + \mathrm{i}, 2 - \mathrm{i})\}$. □

A key question about Gaussian elimination and equivalent systems: What happens to a system if we change it by performing one elementary row operation? After all, Gaussian and Gauss–Jordan elimination amount to a sequence of elementary row operations applied to the augmented matrix of a linear system. Answer: Nothing happens to the solution set!

Theorem 1.2. **Equivalent Systems** Suppose a linear system has augmented matrix $\widetilde{A}$ upon which an elementary row operation is applied to yield a new augmented matrix $\widetilde{B}$ corresponding to a new linear system. Then these two linear systems are equivalent, i.e., have the same solution set.

Proof. If we replace the variables in the system corresponding to $\widetilde{A}$ by the values of a solution, the resulting equations will be satisfied. Now perform the elementary operation in question on this system of equations to obtain that the equations for the system corresponding to the augmented matrix $\widetilde{B}$ are also satisfied. Thus, every solution to the old system is also a solution to the new system resulting from performing an elementary operation. For the converse, it is sufficient for us to show that the old system can be obtained

from the new one by another elementary operation. In other words, we need to show that the effect of any elementary operation can be undone by another elementary operation. This will show that every solution to the new system is also a solution to the old system. If E represents an elementary operation, then the operation that undoes it could reasonably be designated as E^{-1}, since the effect of the inverse operation is rather like canceling a number by multiplying by its inverse. Let us examine each elementary operation in turn:

Inverse Elementary Operations

- E_{ij}: The elementary operation of switching the ith and jth rows of the matrix. Notice that the effect of this operation is undone by performing the same operation, E_{ij}, again. This switches the rows back. Symbolically we write $E_{ij}^{-1} = E_{ij}$.
- $E_i(c)$: The elementary operation of multiplying the ith row by the nonzero constant c. This elementary operation is undone by performing the elementary operation $E_i(1/c)$; in other words, by multiplying the ith row by the nonzero constant $1/c$. We write $E_i(c)^{-1} = E_i(1/c)$.
- $E_{ij}(d)$: The elementary operation of adding d times the jth row to the ith row. This operation is undone by adding $-d$ times the jth row to the ith row. We write $E_{ij}(d)^{-1} = E_{ij}(-d)$.

Thus, in all cases the effects of an elementary operation can be undone by applying another elementary operation of the same type, which is what we wanted to show. □

The inverse notation we used here doesn't do much for us yet. In Chapter 2 this notation will take on an entirely new and richer meaning.

The Reduced Row Echelon Form

Theorem 1.2 tells us that the methods of Gaussian and Gauss–Jordan elimination do not alter the solution set we are interested in finding. Our next objective is to describe the end result of these methods in a precise way. That is, we want to give a careful definition of the form of the matrix that these methods lead us to, starting with the augmented matrix of the original system. Recall that the *leading entry* of a row is the first nonzero entry of that row. (So a row of zeros has no leading entry.)

Definition 1.9. Reduced Row (Echelon) Form A matrix R is said to be in *reduced row form* if:

(1) The nonzero rows of R precede the zero rows.
(2) The column numbers of the leading entries of the nonzero rows, say rows $1, 2, \ldots, r$, form an increasing sequence of numbers $c_1 < c_2 < \cdots < c_r$.

The matrix R is said to be in *reduced row echelon form* if in addition to the above:

(3) Each leading entry is a 1.
(4) Each leading entry has only zeros above it.

Example 1.26. Consider the following matrices (whose leading entries are enclosed in a circle). Which are in reduced row form? Reduced row echelon form?

(a) $\begin{bmatrix} \textcircled{1} & 2 \\ 0 & \textcircled{3} \end{bmatrix}$ (b) $\begin{bmatrix} \textcircled{1} & 2 & 0 \\ 0 & 0 & \textcircled{3} \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & 0 \\ \textcircled{1} & 0 & 0 \end{bmatrix}$

(d) $\begin{bmatrix} \textcircled{1} & 2 & 0 \\ 0 & 0 & \textcircled{1} \\ 0 & 0 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} \textcircled{1} & 0 & 0 \\ 0 & 0 & \textcircled{1} \\ 0 & \textcircled{1} & 0 \end{bmatrix}$

Solution. Checking through (1)–(2), we see that (a), (b), and (d) fulfill the conditions for reduced row matrices. But (c) fails, since a zero row precedes the nonzero ones; matrix (e) fails to be in reduced row form because the column numbers of the leading entries do not form an increasing sequence. Matrices (a) and (b) don't satisfy (3), so matrix (d) is the only one that satisfies (3)–(4). Hence, it is the only matrix in the list in reduced row echelon form. □

We can now describe the goal of Gaussian elimination as follows: Use elementary row operations to reduce the augmented matrix of a linear system to reduced row form; then back solve the resulting system. On the other hand, the goal of Gauss–Jordan elimination is to use elementary operations to reduce the augmented matrix of a linear system to reduced row echelon form. From this form one can read off the solution(s) to the system.

Is it always possible to reduce a matrix to a reduced row form or row echelon form? If so, to how many such forms? These are important questions. If we take the matrix in question to be the augmented matrix of a linear system, what we are really asking becomes, how reliable is Gaussian elimination? Does it always lead us to answers that have the same form? Certainly, matrices can be transformed by elementary row operations to different reduced row forms, as the following simple example shows:

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 2 & -1 \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} 1 & 0 & 5 \\ 0 & 2 & -1 \end{bmatrix} \xrightarrow{E_2(1/2)} \begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & -\frac{1}{2} \end{bmatrix}.$$

Every matrix of this example is already in reduced row form. The last matrix is also in reduced row echelon form. Yet all three of these matrices can be obtained from each other by elementary row operations. It is significant that only one of the three matrices is in reduced row echelon form. As a matter of fact, any matrix can be reduced by elementary row operations to *one and only one* reduced row echelon form, which we can call *the* reduced row echelon form of the matrix. The fact that at least one such form is always possible is justified by an algorithm which starts with a matrix and terminates in a finite number of steps with a reduced row echelon form for the matrix. Here is an informal description of one such algorithm which could easily be programmed:

Algorithm RREF

Input: $m \times n$ matrix $A = [a_{ij}]$.
Output: reduced row echelon form matrix $R = [r_{ij}]$.
Procedure:
Set $p = 1$, $q = 1$, $R = A$.
While $p \leq m$ and $q \leq n$:
  Search for index $i \geq p$ such that $r_{iq} \neq 0$.
  If none found
    set $q = q + 1$
  else
    interchange rows i and p with E_{ip}
    convert (p, q)th entry to 1 with $E_p\left(\frac{1}{r_{pq}}\right)$
    zero out entries above and below (p, q)th entry with suitable $E_{kp}\,(-r_{kq})$
    set $p = p + 1$, $q = j + 1$.
end while

This algorithm must terminate in finitely many steps and replaces the matrix A with a reduced row echelon form E. So A has at least one such form. In fact it is the only one:

Theorem 1.3. **Uniqueness of Reduced Row Echelon Form** Every matrix can be reduced by a sequence of elementary row operations to one and only one reduced row echelon form.

Proof. Algorithm RREF yields one such form. Suppose that some matrix could be reduced to two distinct reduced row echelon forms. Then there is such an $m \times n$ matrix $\widetilde{A}$ with the fewest possible columns n; that is, the theorem is true for every matrix with fewer columns. A single column matrix can be reduced to only one reduced row echelon form, namely the 0 column if it is a 0 column, or a column with first entry 1 and the other entries 0 otherwise. Hence $n > 1$. The matrix $\widetilde{A}$ can be reduced to two different reduced row echelon forms, say R_1 and R_2, with $R_1 \neq R_2$. Write $\widetilde{A} = [A \mid \mathbf{b}]$, so that we can think of $\widetilde{A}$ as the augmented matrix of a linear system (1.1). Now for $i = 1, 2$ write each R_i as $R_i = [L_i \mid \mathbf{b}_i]$, where $\mathbf{b}_i$ is the last column of the

$m \times n$ matrix R_i, and L_i is the $m \times (n-1)$ matrix formed from the first $n-1$ columns of R_i. Each L_i satisfies the definition of reduced row echelon form, since each R_i is in reduced row echelon form. Also, each L_i results from performing elementary row operations on the matrix A, which has only $n-1$ columns. By the minimum columns hypothesis, we have that $L_1 = L_2$. There are two possibilities to consider.

Case 1: The last column $\mathbf{b}_i$ of either R_i has a leading entry in it. Then the system of equations represented by $\widetilde{A}$ is inconsistent. It follows that both columns $\mathbf{b}_i$ have a leading entry in them, which must be a 1 in the first row whose portion in L_i consists of zeros, and the entries above and below this leading entry must be 0. Since $L_1 = L_2$, it follows that $\mathbf{b}_1 = \mathbf{b}_2$, and thus $R_1 = R_2$, a contradiction. So this case can't occur.

Case 2: Each $\mathbf{b}_i$ has no leading entry in it. Then the system of equations represented by $\widetilde{A}$ is consistent. Both augmented matrices have the same basic and free variables since $L_1 = L_2$. Hence, we obtain the same solution with either augmented matrix by setting the free variables of the system equal to 0. When we do so, the bound variables are uniquely determined: The first equation says that the first bound variable equals the first entry in the right-hand-side vector since all other variables will either be zero or have zero coefficient in the first equation of the system. Similarly, the second says that the second bound variable equals the second entry in the right-hand-side vector, and so forth. Therefore, $\mathbf{b}_1 = \mathbf{b}_2$ and thus $R_1 = R_2$, a contradiction again. Hence, there can be no counterexample to the theorem, which completes the proof. □

The following consequence of the preceding theorem is a fact that we will find very useful in Chapter 2.

Corollary 1.1. Let the matrix B be obtained from the matrix A by performing a sequence of elementary row operations on A. Then B and A have the same reduced row echelon form.

Proof. To see this, perform the elementary operations on B that undo the ones originally performed on A to get B. The matrix A results from these operations. Now perform whatever elementary row operations are needed to reduce A to its reduced row echelon form. Since B can be reduced to one and only one reduced row echelon form, the reduced row echelon forms of A and B coincide. □

Rank and Nullity of a Matrix

Now that we have Theorem 1.3 in hand, we can introduce the notion of *rank* of a matrix, for it uses the fact that A has exactly one reduced row echelon form.

Definition 1.10. Matrix Rank The *rank* of a matrix A is the number of nonzero rows of the reduced row echelon form of A. This number is written as rank A.

There are other ways to describe the rank of a matrix. For example, rank can also be defined as the number of nonzero rows in any reduced row form of a matrix. One has to check that any two reduced row forms have the same number of nonzero rows (they do). Rank can also be defined as the number of columns of the reduced row echelon form with leading entries in them, since each entry of a reduced row echelon form occupies a unique column. The number of remaining columns also has a name:

Definition 1.11. Matrix Nullity The *nullity* of a matrix A is the number of columns of the reduced row echelon form of A that do not contain a leading entry. This number is written as null A.

In the case that A is the coefficient matrix of a linear system, we can interpret the rank of A as the number of bound variables of the system and the nullity of A as the number of free variables of the system.

Example 1.27. Find the rank and nullity of the matrix $A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 2 & 5 \\ 3 & 3 & 2 \end{bmatrix}$.

Solution. Elementary row operations give

$$\begin{bmatrix} 1 & 1 & 2 \\ 2 & 2 & 5 \\ 3 & 3 & 2 \end{bmatrix} \xrightarrow{E_{21}(-2)} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 3 & 3 & 2 \end{bmatrix} \xrightarrow{E_{31}(-3)} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & -4 \end{bmatrix} \xrightarrow[E_{12}(-2)]{E_{32}(4)} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

From the reduced row echelon form of A at the far right we see that the rank of A is 2, that is, rank $A = 2$. Since only one column does not contain a pivot, we see that the nullity of A is 1, that is, null $A = 1$. □

One point that the previous example makes is that one cannot determine the rank of a matrix by counting nonzero rows of the original matrix.

Caution: Remember that the rank of A is the number of nonzero rows in one of its reduced row forms, and *not* the number of nonzero rows of A itself.

The rank of a matrix is a nonnegative number, but it *could* be 0! This happens if the matrix has only zero entries, so that it has no nonzero rows. In this case, the nullity of the matrix is as large as possible, namely the number of columns of the matrix. There are some simple limits on the size of rank A and null A. First, we need a notation that occurs frequently throughout the text.

Definition 1.12. Min Max Given a list of real numbers $a_1, a_2, \ldots, a_m$, the smallest number in the list is $\min\{a_1, a_2, \ldots, a_m\}$, and $\max\{a_1, a_2, \ldots, a_m\}$ is the largest number in the list.

Theorem 1.4. Let A be an $m \times n$ matrix. Then

(1) $0 \leq \text{rank } A \leq \min\{m, n\}$.
(2) $\text{rank } A + \text{null } A = n$.

Proof. By definition, rank A is the number of nonzero rows of the reduced row echelon form of A, which is itself an $m \times n$ matrix. There can be no more leading entries than rows; hence rank $A \leq m$. Also, each leading entry of a matrix in reduced row echelon form is the unique nonzero entry in its column. So there can be no more leading entries than columns n. Since rank A is less than or equal to both m and n, it is less than or equal to their minimum, which is the first inequality. The number of pivot columns is rank A and the number of non-pivot columns is null A. The sum of these numbers is n. □

In words, item (1) of Theorem 1.4 says that the rank of a matrix cannot exceed the number of rows or columns of the matrix. If the rank of a matrix equals its column number we say that the matrix has *full column rank*. Similarly, a matrix has *full row rank* if its rank equals the row number of the matrix. For example, matrix A of Example 1.27 is 3×3 of rank 2. Since this rank is smaller than 3, A does not have full column or row rank. Here is an application of the rank concept to systems.

Theorem 1.5. Consistency in Terms of Rank The general linear system (1.1) with $m \times n$ coefficient matrix A, right-hand-side vector $\mathbf{b}$, and augmented matrix $\widetilde{A} = [A \mid \mathbf{b}]$ is consistent if and only if rank A = rank $\widetilde{A}$, in which case either

(1) rank $A = n$, in which case the system has a unique solution, or
(2) rank $A < n$, in which case the system has infinitely many solutions.

Proof. We can reduce $\widetilde{A}$ to reduced row echelon form by first doing the elementary operations that reduce the A part of the matrix to reduced row echelon form, then attending to the last column. Hence, it is always the case that rank $A \leq$ rank $\widetilde{A}$. The only way to get strict inequality is to have a leading entry in the last column, which means that some equation in the equivalent system corresponding to the reduced augmented matrix is $0 = 1$, which implies that the system is inconsistent. On the other hand, we have already seen (in the proof of Theorem 1.3, for example) that if the last column does not contain a leading entry, then the system is consistent. This establishes the first statement of the theorem.

Now suppose that rank A = rank $\widetilde{A}$, so that the system is consistent. By Theorem 1.4, rank $A \leq n$, so that either rank $A < n$ or rank $A = n$. The

number of variables of the system is n. Also, the number of leading entries (equivalently, pivots) of the reduced row form of $\widetilde{A}$, which is rank A, is equal to the number of bound variables; the remaining $n - \text{rank } A$ variables are the free variables of the system. Thus, to say that rank $A = n$ is to say that no variables are free; that is, solving the system leads to a unique solution. And to say that rank $A < n$ is to say that there is at least one free variable, in which case the system has infinitely many solutions. □

Here is an example of how this theorem can be put to work. It confirms our intuition that if a system does not have "enough" equations, then it can't have a unique solution:

Corollary 1.2. If a consistent linear system of equations has more unknowns than equations, then the system has infinitely many solutions.

Proof. In the notation of the previous theorem, the hypothesis simply means that $m < n$. But we know from Theorem 1.4 that rank $A \leq \min\{m, n\}$. Thus, rank $A < n$ and the part (2) of Theorem 1.5 yields the desired result. □

Of course, there is still the question of when a system is consistent. In general, there isn't an easy way to see when this is so outside of explicitly solving the system. However, in some cases there are easy answers. One such important special case is given by the following definition.

Definition 1.13. **Homogeneous System** The general linear system (1.1) with $m \times n$ coefficient matrix A and right-hand-side vector $\mathbf{b}$ is said to be *homogeneous* if the entries of $\mathbf{b}$ are all zero. Otherwise, the system is said to be *inhomogeneous*.

Trivial Solution

The nice feature of homogeneous systems is that they are always consistent! In fact, it is easy to exhibit a specific solution to the system, namely, take the value of all the variables to be zero. For obvious reasons this solution is called the *trivial* solution to the system. Thus, the previous corollary implies that a homogeneous linear system with fewer equations than unknowns must have infinitely many solutions. Of course, if we want to find all the solutions, we will have to do the work of Gauss–Jordan elimination. However, we acquire a small notational convenience in dealing with homogeneous systems. Notice that the right-hand side of zeros is never changed by an elementary row operation. So why bother writing out the augmented matrix of such a system? It suffices to perform elementary operations on the coefficient matrix alone. In the end, the right-hand side is still a column of zeros.

Example 1.28. Solve and describe the solution set of the homogeneous system

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

Solution. Since row operations will not change the right-hand side, we need only perform them on the coefficient matrix to obtain

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \xrightarrow[E_{31}(-1)]{E_{21}(-1)} \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix} \xrightarrow[E_3(-1)]{E_2(1/2)} \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix} \xrightarrow[E_{13}(-1)]{E_{23}(1/2)} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

One has to be a little careful here: The leading entry in the fourth column does not indicate that the system is inconsistent, since we deleted the right-hand-side column of the system. Had we carried it along in the calculations above, we would have obtained

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

which is the matrix of a consistent system. We see from the reduced row echelon form of the coefficient matrix that x_2 is free and the other variables are bound. The general solution is

$$\begin{aligned} x_1 &= -x_2 \\ x_3 &= 0 \\ x_4 &= 0 \\ x_2 &\text{ is free.} \end{aligned}$$

Finally, the solution set S of this system can be described as

$$S = \{(-x_2, x_2, 0, 0) \mid x_2 \in \mathbb{R}\}. \qquad \square$$

For many practical problems solution sets as described in the preceding example require a more sophisticated solution description. For example, consider this variation on the Leontief input-output model of Example 1.4 (highly simplified, since economists might be interested in a much more complex model involving hundreds of variables):

Example 1.29. We are given the following consumption matrix for the commodity output of three sectors M, P, S in a closed economy. Suppose that each producer charges a unit price for its commodity, say p_1, p_2, p_3, and that the columns of the table represent fraction of each producer commodity needed by the consumer to produce one unit of its own commodity. Find all possible equilibrium prices for these products, i.e., prices such that the cost of production of an item is equal to its price.

		Consumed by M	P	S
Produced by	M	0.0	0.5	0.5
	P	0.4	0.2	0.4
	S	0.6	0.2	0.2

Solution. Sector M will charge a price of p_1 for for a unit of output and this must be balanced by its payments, $0.4p_2$ and $0.6p_3$. The same applies to the other sectors, which yields the following system of equations:

$$\begin{aligned} p_1 &= 0.4p_2 + 0.6p_3 \\ p_2 &= 0.5p_1 + 0.2p_2 + 0.2p_3 \\ p_3 &= 0.5p_1 + 0.4p_2 + 0.2p_3. \end{aligned}$$

Subtract terms on the right to end up with a homogeneous system. Row operations do not change the right-hand side, so perform row operations on the coefficient matrix (whose entries we convert to fractions for convenience) to obtain

$$\begin{bmatrix} 1 & -\frac{2}{5} & -\frac{3}{5} \\ -\frac{1}{2} & \frac{4}{5} & -\frac{1}{5} \\ -\frac{1}{2} & -\frac{2}{5} & \frac{4}{5} \end{bmatrix} \xrightarrow[E_{31}(\frac{1}{2})]{E_{21}(\frac{1}{2})} \begin{bmatrix} 1 & -\frac{2}{5} & -\frac{3}{5} \\ 0 & \frac{3}{5} & -\frac{1}{2} \\ 0 & -\frac{3}{5} & \frac{1}{2} \end{bmatrix} \xrightarrow[E_{12}(\frac{2}{3})]{E_{32}(1)} \begin{bmatrix} 1 & 0 & -\frac{14}{15} \\ 0 & \frac{3}{5} & -\frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{E_{23}(\frac{5}{3})} \begin{bmatrix} 1 & 0 & -\frac{14}{15} \\ 0 & 1 & -\frac{5}{6} \\ 0 & 0 & 0 \end{bmatrix}.$$

Thus, p_3 is free and hence the solution set of interest to economists is $p_1 = \frac{14}{15}p_3$, $p_2 = \frac{5}{6}p_3$ and $p_3 > 0$. □

1.4 Exercises and Problems

Exercise 1. Circle leading entries and determine which of the following matrices are in reduced row form or reduced row echelon form.

(a) $\begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (g) $\begin{bmatrix} 1 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$ (h) $[1\ 3]$

Exercise 2. Circle leading entries and determine which of the following matrices can be put into reduced row echelon form with at most one elementary operation.

(a) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 2 \end{bmatrix}$

(d) $\begin{bmatrix} 2 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$

Exercise 3. The rank of the following matrices can be determined by inspection. Inspect these matrices and specify their rank.

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

Exercise 4. Inspect these matrices and specify their rank without pencil and paper calculation.

(a) $\begin{bmatrix} 1 & 3 & 3 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 2 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

Exercise 5. Show that the elementary operations you use to find the reduced row echelon form of the following matrices. Give the rank and nullity of each matrix.

(a) $\begin{bmatrix} 1 & -1 & 2 \\ 1 & 3 & 4 \\ 2 & 2 & 6 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 1 & 9 & 2 \\ -3 & 0 & 6 & -5 \\ 0 & 0 & 1 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 2 \end{bmatrix}$

(d) $\begin{bmatrix} 2 & 4 & 2 \\ 4 & 9 & 3 \\ 2 & 3 & 3 \end{bmatrix}$ (e) $\begin{bmatrix} 2 & 2 & 5 & 6 \\ 1 & 1 & -2 & 2 \end{bmatrix}$ (f) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$

Exercise 6. Compute a reduced row form that can be reached in a minimum number of steps and the reduced row echelon forms of the following matrices. Given that the matrices are augmented matrices for a linear system, write out the general solution to the system.

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

(d) $\begin{bmatrix} 2 & 4 & 2 \\ 2 & 1 & 1 \\ 1 & 1 & 3 \end{bmatrix}$ (e) $\begin{bmatrix} 2 & 2 \\ 3 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 2 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}$

Exercise 7. Find the rank of the augmented and coefficient matrix of the following linear systems and the solution sets to the following systems. Are these systems equivalent?

(a)
$$\begin{aligned} x_1 + x_2 + x_3 - x_4 &= 2 \\ 2x_1 + x_2 - 2x_4 &= 1 \\ 2x_1 + 2x_2 + 2x_3 - 2x_4 &= 4 \end{aligned}$$

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

Exercise 8. Show that the following systems are equivalent and find a sequence of elementary operations that transforms the augmented matrix of (a) into that of (b).

(a)
$$\begin{aligned} x_1 + x_2 + x_3 - x_4 &= 2 \\ 2x_1 + x_2 - 2x_4 &= 1 \\ 2x_1 + 2x_2 + 2x_3 - 2x_4 &= 4 \end{aligned}$$

(b)
$$\begin{aligned} x_1 + x_2 + x_3 - x_4 &= 2 \\ 4x_1 + 3x_2 + 2x_3 - 4x_4 &= 5 \\ 7x_1 + 6x_2 + 5x_3 - 7x_4 &= 11 \end{aligned}$$

Exercise 9. Find upper and lower bounds on the rank of the 4×3 matrix A, given that some system with coefficient matrix A has infinitely many solutions.

Exercise 10. Find upper and lower bounds on the rank of matrix A, where A has four rows and some system of equations with coefficient matrix A has a unique solution.

Exercise 11. For what values of c are the following systems inconsistent, with unique solution or with infinitely many solutions?

(a) $\begin{aligned} x_2 + cx_3 &= 0 \\ x_1 - cx_2 &= 1 \end{aligned}$ (b) $\begin{aligned} x_1 + 2x_2 - x_3 &= c \\ x_1 + 3x_2 + x_3 &= 1 \\ 3x_1 + 7x_2 - x_3 &= 4 \end{aligned}$ (c) $\begin{aligned} cx_1 + x_2 + x_3 &= 2 \\ x_1 + cx_2 + x_3 &= 2 \\ x_1 + x_2 + cx_3 &= 2 \end{aligned}$

Exercise 12. Consider the system

$$\begin{aligned} ax + by &= c \\ bx + cy &= d \end{aligned}$$

in the unknowns x, y, where $a \neq 0$. Use the reduced row echelon form to determine conditions on the other constants such that the system has no, one, or infinitely many solutions.

Exercise 13. Consider the system

$$\begin{aligned} x_1 + 2x_2 &= a \\ x_1 + x_2 + x_3 - x_4 &= b \\ 2x_3 + 2x_4 &= c \end{aligned}$$

in the unknowns x_1, x_2, x_3, x_4. Solve this system by reducing the augmented matrix to reduced row echelon form. This system will have solutions for any right-hand side. Justify this fact in terms of rank.

Exercise 14. Give a rank condition for a linear homogeneous system that is equivalent to the system having a unique solution. Justify your answer.

Exercise 15. Fill in the blanks:

(a) If A is a 3×7 matrix then the rank of A is at most _____.
(b) Equivalent systems have the same _____.
(c) The inverse of the elementary operation $E_{23}(5)$ is _____.
(d) The rank of a nonzero 3×3 matrix with all entries equal is _____.

Exercise 16. Fill in the blanks:

(a) If A is a 4×8 matrix, then the nullity of A is larger than _____.
(b) The rank of a nonzero 4×3 matrix with constant entries in each column is _____.
(c) An example of a matrix with nullity 1 and rank 2 is _____.

(d) The size of the matrix $\begin{bmatrix} 0 & -1 & 2 \\ 0 & 3 & 4 \end{bmatrix}$ is _____.

Exercise 17. Consider the system

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 0 \\ 2x_1 + x_2 - x_3 &= a \\ x_2 + x_3 &= b \end{aligned}$$

in the unknowns x_1, x_2, x_3. Find a reduced row form for the augmented matrix of this system with the fewest elementary row operations possible. What can you deduce from this about the solutions to the system?

Exercise 18. Consider the system

$$\begin{aligned} 2x_1 + x_2 - x_3 &= 0 \\ x_1 + x_2 + x_3 &= a \\ x_2 + cx_3 &= b \end{aligned}$$

in the unknowns x_1, x_2, x_3. Find a reduced row form for the augmented matrix of this system with the fewest elementary row operations possible. What can you deduce from this about the solutions to the system?

*Problem 19. Answer True/False and explain your answers:

(a) If a linear system is inconsistent, then the rank of the augmented matrix exceeds the number of unknowns.
(b) Any homogeneous linear system is consistent.
(c) A system of 3 linear equations in 4 unknowns has infinitely many solutions.
(d) Every matrix can be reduced to only one matrix in reduced row form.
(e) Any homogeneous linear system with more equations than unknowns has a nontrivial solution.

Problem 20. Show that a system of linear equations has a unique solution if and only if every column, except the last one, of the reduced row echelon form of the augmented matrix has a pivot entry in it.

Problem 21. Prove or disprove by example: If two linear systems are equivalent, then they must have the same size augmented matrix.

*Problem 22. Use Theorem 1.3 to show that any two reduced row forms for a matrix A must have the same number of nonzero rows.

Problem 23. Suppose that the matrix C can be written in the augmented form $C = [A \mid B]$, where the matrix B may have more than one column. Prove that rank $C \leq$ rank $A +$ rank B.

Problem 24. Suppose that in Example 1.29 sector P no longer needs to consume any of the commodity of sector S to produce its output. Reformulate the consumption matrix for this example and determine equilibrium prices, if any.

1.5 *Applications and Computational Notes

Roundoff Errors

In many practical problems, calculations are not exact. There are several reasons for this unfortunate fact. For one, the input data used to construct a system matrix or coefficient may be inexact (the GIGO principle, garbage in, garbage out, is at least marginally true in such cases). For another, scientific calculators are by their very nature only finite-precision machines. That is, only a fixed number of significant digits of the numbers we are calculating may be used in any calculation. For instance, verify this simple arithmetic fact on a (floating point, not symbolic) technology tool:

$$\left(\left(\frac{2}{3}+100\right)-100\right)-\frac{2}{3}=0.$$

In many cases this calculation will not yield 0. The problem is that if, for example, a calculator uses 6-digit accuracy, then $\frac{2}{3}$ is calculated as 0.666667, which is really incorrect. Even if arithmetic calculations were exact, the data that form the basis of our calculations are often derived from scientific measurements that themselves will almost certainly be in error. Starting with erroneous data and doing an exact calculation can be as bad as starting with exact data and doing an inexact calculation. In fact, in a certain sense they are equivalent to each other. Error resulting from truncating data for storage or finite-precision arithmetic calculations is called *roundoff error*. **Roundoff Error**

We will not give an elaborate treatment of roundoff error. A thorough analysis can be found in the Golub and Van Loan text [14] of the bibliography. The subject of this book, numerical linear algebra, is a part of an entire field of applied mathematics known as numerical analysis. The texts [25] and [9] provide excellent treatments of this subject. Consider this question: Could roundoff error be a significant problem in Gaussian elimination? It isn't at all clear that there is a problem. After all, even in the above example, the final error is relatively small. Is it possible that with all the arithmetic performed in Gaussian elimination the errors pile up and become large? The answer is yes. With the advent of computers came a heightened interest in these questions. In the early 1950s numerical analysts intensified efforts to determine whether Gaussian elimination can reliably solve larger linear systems. In fact, we don't really have to look at complicated examples to realize that there are potential difficulties. Consider the following example.

Example 1.30. Let ϵ be a number so small that our calculator yields $1+\epsilon=1$. This equation appears a bit odd, but from the calculator's point of view it may be perfectly correct; if, for example, our calculator performs 6-digit decimal arithmetic, then $\epsilon=10^{-6}$ will do nicely. Notice that with such a calculator, $1+1/\epsilon=(\epsilon+1)/\epsilon=1/\epsilon$. Now solve the linear system

$$\begin{aligned}\epsilon x_1+x_2&=1\\ x_1-x_2&=0.\end{aligned}\tag{1.9}$$

Solution. Let's solve this system by Gauss–Jordan elimination with our calculator to obtain

$$\begin{bmatrix} \epsilon & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix} \xrightarrow{E_{21}\left(-\frac{1}{\epsilon}\right)} \begin{bmatrix} \epsilon & 1 & 1 \\ 0 & \frac{1}{\epsilon} & -\frac{1}{\epsilon} \end{bmatrix} \xrightarrow{E_2(\epsilon)} \begin{bmatrix} \epsilon & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} \epsilon & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \xrightarrow{E_1\left(\frac{1}{\epsilon}\right)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

Thus, we obtain the calculated solution $x_1 = 0$, $x_2 = 1$. This answer is spectacularly bad! If $\epsilon = 10^{-6}$ as above, then the correct answer is

$$x_1 = x_2 = \frac{1}{1+\epsilon} = 0.99999909999990\ldots.$$

Our calculated answer is not even good to one digit. So we see that there can be serious problems with Gaussian or Gauss–Jordan elimination on finite-precision machines. □

Partial Pivoting

It turns out that information that would be significant for x_1 in the first equation is lost in the truncated arithmetic that says that $1 + 1/\epsilon = 1/\epsilon$. There is a fix for problems such as this, namely a technique called *partial pivoting*. The idea is fairly simple: Do not choose the next available column entry for a pivot. Rather, search down the column in question for the largest entry (in absolute value). Then switch rows, if necessary, and use this entry as a pivot. For instance, in the preceding example, we would not pivot off the ϵ entry of the first column. Since the entry of the second row, first column, is larger in absolute value, we would switch rows and then do the usual Gaussian elimination step. Here is what we would get (remember that with our calculator $1+\epsilon = 1$):

$$\begin{bmatrix} \epsilon & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix} \xrightarrow{E_{21}} \begin{bmatrix} 1 & -1 & 0 \\ \epsilon & 1 & 1 \end{bmatrix} \xrightarrow{E_{21}(-\epsilon)} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \xrightarrow{E_{12}(1)} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Now we get the quite acceptable answer $x_1 = x_2 = 1$.

But partial pivoting is not a panacea for numerical problems. In fact, it can be easily defeated. Multiply the second equation of equation (1.9) by ϵ^2, and we get a system for which partial pivoting still picks the wrong pivot. Here the problem is a matter of scale. It can be cured by dividing each row by the largest entry of the row before beginning the Gaussian elimination process. This procedure is known as *row scaling*. The combination of row scaling and partial pivoting overcomes many of the numerical problems of Gaussian and Gauss–Jordan elimination (but not all!). There is a more drastic procedure, known as *complete pivoting*. In this procedure one searches all the unused rows (excluding the right-hand sides) for the largest entry, then uses it as a pivot for Gaussian elimination. The columns used in this procedure do not move in that left-to-right fashion we are used to seeing in system solving. It can be shown rigorously that the error of roundoff propagates in a predictable and controlled fashion with complete pivoting; in contrast, we do not really have a satisfactory explanation as to why row scaling and partial pivoting work well. Yet

Complete Pivoting

in most cases they do reasonably well. Since this combination involves much less calculation than complete pivoting, it is the method of choice for many problems.

There are deeper reasons for numerical problems in solving some systems than the one the preceding example illustrates. One difficulty has to do with the "sensitivity" of the coefficient matrix to small changes. That is, in some systems, small changes in the coefficient matrix lead to dramatic changes in the *exact* answer. The practical effect of roundoff error can be shown to be equivalent to introducing small changes in the coefficient matrix and obtaining an exact answer to the perturbed (changed) system. There is no cure for these difficulties, short of computing in higher precision. A classical example of this type of problem, the Hilbert matrix, is discussed in one of the projects below. We will attempt to quantify this "sensitivity" in Chapter 5.

Computational Efficiency of Gaussian Elimination

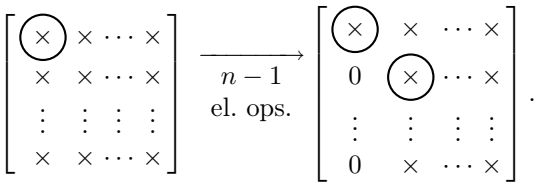
How much work is it to solve a linear system and how does the amount of work grow with the dimensions of the system? The first thing we need is a unit of work. In computer science one of the principal units of work of numerical computation is a *flop* (floating point operation), namely a single $+, -, \times$, or $\div$. For example, we say that the amount of work in computing $e + \pi$ or $e \times \pi$ is one flop, while the work in calculating $e + 3 \times \pi$ is two flops. The cost of performing an operation is called its *flop count*. The following example is extremely useful.

Flop Count

Example 1.31. How many flops does it cost to add a multiple of one row to another, as in Gaussian elimination, if the rows have n elements?

Solution. Say that row $\mathbf{a}$ is to be multiplied by the scalar α, and added to the row $\mathbf{b}$. Designate the row $\mathbf{a} = [a_i]$ and the row $\mathbf{b} = [b_i]$. We have n entries to worry about. Consider a typical one, say the ith one. The ith entry of $\mathbf{b}$, namely b_i, will be replaced by the quantity $b_i + \alpha a_i$. The amount of work in this calculation is two flops. Since there are n entries to compute, the total work is $2n$ flops. □

Our goal is to determine the expense of solving a system by Gauss–Jordan elimination. For the sake of simplicity, let's assume that the system under consideration has n equations in n unknowns and the coefficient matrix has rank n. This ensures that we will have a pivot in every row of the matrix. We won't count row exchanges either, since they don't involve any flops. (This may not be realistic on a fast computer, since memory fetches and stores may not take significantly less time than a floating-point operation.) Now consider the expense of clearing out the entries under the first pivot. A picture of the augmented matrix looks something like this, where an $\times$ is an entry that may not be 0 and an $\textcircled{\times}$ is a nonzero pivot entry:



Each elementary operation will involve adding a multiple of the first row, starting with the *second* entry, since we don't need to do arithmetic in the first column — we know what goes there — to the $n-1$ subsequent rows. By the preceding example, each of these elementary operations will cost $2(n-1)$ flops. Add 1 flop for the cost of determining the multiplier to obtain $2n-1$. So the total cost of zeroing out the first column is $(n-1)(2n-1)$ flops. Now examine the lower unfinished block in the above figure. Notice that it's as though we were starting over with the row and column dimensions reduced by 1. Therefore, the total cost of the next phase is $(n-2)(2(n-1)+1)$ flops. Continue in this fashion, and we obtain a count of

$$0+\sum_{j=2}^{n}(j-1)(2j-1)=\sum_{j=1}^{n}(j-1)(2j-1)=\sum_{j=1}^{n}2j^2-3j+1$$

flops. Recall the identities for sums of consecutive integers and their squares:

$$\sum_{j=1}^{n}j=\frac{n(n+1)}{2}\quad\text{and}\quad\sum_{j=1}^{n}j^2=\frac{n(n+1)(2n+1)}{6}.$$

Thus, we have a total flop count of

$$\sum_{j=1}^{n}2j^2-3j+1=2\frac{n(n+1)(2n+1)}{6}-3\frac{n(n+1)}{2}+n=\frac{2n^3}{3}-\frac{n^2}{2}-\frac{n}{6}.$$

This is the cost of forward solving. Now let's simplify our answer a bit more. For large n we have that n^3 is much larger than n or n^2 (e.g., for $n=10$ compare 1000 to 10 or 100). Hence, we ignore the lower-degree terms and arrive at a simple approximation to the number of flops required to forward solve a linear system of n equations in n unknowns using Gauss–Jordan elimination. There remains the matter of back solving. We leave as an exercise to show that the total work of back solving is quadratic in n. Therefore, the "leading-order" approximation that we found for forward solving remains unchanged. Hence, we have the following estimate of the *complexity* of Gaussian elimination.

Theorem 1.6. Computational Efficiency The number of flops required to solve a linear system of n equations in n unknowns using Gaussian or Gauss–Jordan elimination without row exchanges is *approximately* $2n^3/3$.

Thus, for example, the work of forward solving a system of 21 equations in 21 unknowns is approximately $2\cdot 21^3/3=6174$ flops. Exact answer: 5950.

Derivation of a Time Dependent form of the Diffusion Equation

The term "diffusion" refers to the spatial flow of a material in time from higher to lower concentrations. The "material" could be a chemical, fluid, electrical charge, heat or a population of organisms, to name but a few of the materials to which the notion of diffusion could be applied. The study of diffusion is a vast enterprise. We shall confine ourselves to a simplified discrete one-dimensional version of the so-called reaction-diffusion equation derived from Fick's first law, which can be stated simply as

Fick's First Law: Diffusive flux of a material is directly proportional to concentration gradient of the material.

Here we understand that *flux* means the amount of material that flows across a surface per unit surface area per unit time, so that it has the units of material/(area·time) and gradient means the change in concentration (density) per unit length, so that it has units of material/(volume·length). Hence, the coefficient of proportion (diffusion coefficient) has units of area/time. Fick's law has many variants in science such as Fourier's law in heat conduction, Ohm's law in electrical current and Darcy's law in groundwater flow, among others.

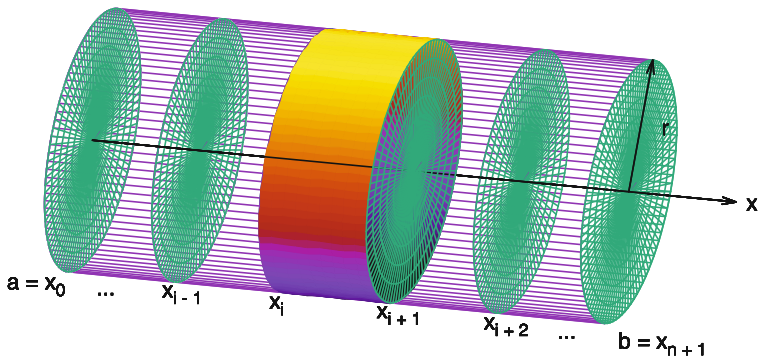


Fig. 1.7: Rod centered along the x-axis on the interval $[a, b]$ with circular cross-section of radius r and area $A = \pi r^2$.

The material movement we consider is one dimensional, say along the x-axis on the interval $[a, b]$. For example, think of the material as moving along a rod, say a wire or tube, of constant cross section with area A placed along the interval $[a, b]$. We shall assume that the medium in which the material is diffused is uniform in its properties. In particular, we assume a constant diffusion coefficient across the rod. Let the density function be $y(x, t)$, a function of position x and time t. Assume that y is specified at the endpoints: $y(x_0, t) = y_{\text{left}}(t)$ and $y(x_{n+1}, t) = y_{\text{right}}(t)$. Also assume that y is specified at initial time $t = 0$: $y(x, 0) = y_0(x)$. Finally, we suppose that there is source

of material (a reaction term) along the rod that could vary with position, time and material density, namely a function $f(x,t,y)$ measured in units of material/volume.

As a practical matter we cannot measure the density continuously in either space or time. Therefore, we "discretize" the problem as follows: Divide the interval into a finite number of equal sized subintervals delimited by points called *nodes*, namely $a = x_0, x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{n+1} = b$. Assume that the nodes are a distance h apart, so that for indices $i = 0, 1, ..., n$, $x_{i+1} = x_i + h$. Since spacing is equal, the relation between h and n is $h = (b-a)/(n+1)$. Further suppose that we are interested in determining the material density at each spatial node in discrete points in time which are uniformly spaced by time difference k, yielding times $t_0 = 0, t_1, t_2, \ldots, t_j, \ldots$. Here $t_{j+1} = t_j + k$ for all indices j.

Now we are in a position to develop approximate solutions to our diffusion problem. We assume that h and k are sufficiently small that the material density is well approximated over the space interval $[x_i, x_{i+1})$ at a specified time. Denote our approximations to the material density $y(x_i, t_j)$ by $y_{i,j}$ for all indices i and j. Suppose that we have calculated values $y_{i,j}$ for a time index j and all space indices $i = 0, 1, ..., n+1$. How can we obtain values of $y_{i,j+1}$, $i = 0, 1, ..., n+1$, from this information?

To answer this question, we examine a reference volume of the rod occupying the space between x_i and x_{i+1} (see Figure 1.7). The volume of this region (really a cylinder) is Ah. During the time interval $[t_j, t_{j+1}]$ material flows by diffusion across the left and right faces of the volume at x_i and x_{i+1}, respectively. In addition, material is added by the source term $f(x_i, t_j, y_{i,j}) = f_{i,j}$ in units of material per unit volume. In these terms, the density gradient over the distance h is $(y_m - y_{m-1})/h$ for any m such that $1 \le m \le n$. Thus, Fick's law says that the flow of material/(area·time) across the mth face is given by $-D(y_m - y_{m-1})/h$, where D is a positive constant called the *diffusion coefficient*. The reason for the minus sign is that we regard flow to the right as positive; a positive gradient would mandate a negative flow to the left (higher to lower concentration), so the minus sign corrects for this. Hence, the total influx of material into the reference volume in this time interval is

$$\begin{aligned}\text{left inflow} + \text{right inflow} + \text{source} &= \left(-D\frac{y_{i,j} - y_{i-1,j}}{h} + D\frac{y_{i+1,j} - y_{i,j}}{h}\right) Ak \\ &\quad + f_{i,j}Ahk \\ &= Ahk\left(D\frac{y_{i-1,j} - 2y_{i,j} + y_{i+1,j}}{h^2} + f_{i,j}\right)\end{aligned}$$

However the influx of material over the time interval $[t_j, t_{j+1}]$ is simply the change in material over that interval, $y_{i,j+1}Ah - y_{i,j}Ah$. Equating these two terms and cancelling the common factor of volume Ah yields the equation

$$y_{i,j+1} - y_{i,j} = k\left(D\frac{y_{i-1,j} - 2y_{i,j} + y_{i+1,j}}{h^2} + f_{i,j}\right),$$

that is,

Time Dependent Reaction-Diffusion

$$y_{i,j+1} = y_{i,j} + \frac{kD}{h^2}\left(y_{i-1,j} - 2y_{i,j} + y_{i+1,j}\right) + kf_{i,j}. \tag{1.10}$$

This equation allows us to solve for values of $y_{i+1,j}$ given that we know all values of $y_{i,j}$ at the previous time step. Such a method is commonly called a *marching method*. It is termed *explicit* because it yields results at time level $k+1$ explicitly in terms of time level k. This particular method is sometimes called the (explicit) Euler method.

One can reasonably expect that as we make h and k smaller, our approximations should improve in accuracy. But as we let $h, k \to 0$, how should they relate to each other? It is beyond the scope of this discussion, but in the discipline of numerical PDEs the number $\sigma = kD/h^2$ is called the Courant number of the problem, and the requirement for a stable convergence to the solution to the problem is that $\sigma \leq 1/2$, or equivalently,

$$k \leq \frac{h^2}{2D}, \text{ as } h, k \to 0.$$

A simple plausibility argument follows from rewriting equation (1.10) in the form

$$y_{i,j+1} = \sigma y_{i-1,j} + (1 - 2\sigma)\, y_{i,j} + \sigma y_{i+1,j} + kf_{i,j}. \tag{1.11}$$

One can see from this that if $\sigma > 1/2$ then, independently of any source or decay term, the net contribution of material to $y_{i,j+1}$ from $y_{i,j}$ would be negative. This is counter-intuitive since no matter how much of the material in the ith reference volume at time step j were distributed to its neighbors, one would expect a nonnegative amount to remain in it at the $(j+1)$th time step. The approximate equality $k \approx \frac{h^2}{2D}$ is sometimes used as a rough estimate of how much time k it takes for a material to diffuse over a distance h.

Derivation of the Diffusion Equation for Steady-State Heat Flow

The idea behind steady state diffusion is that over time, the diffusive process has settled down to a point where material and source term densities do not change with time. In terms of equation (1.10), this means that $y_{i,j+1} \approx y_{i,j}$ for all indices i and j. So subtract these terms from both sides of equation (1.10) to obtain

$$0 = \frac{kD}{h^2}\left(y_{i-1,j} - 2y_{i,j} + y_{i+1,j}\right) + kf_{i,j}.$$

Since $y_{i,j}$ and $\mathrm{f}(x_i, t_j, y_{i,j})$ do not vary with j, we may as well set $y_{i,j} = y_i$ and $f(x_i, t_j, y_{i,j}) = f_i$. Cancel the factor k from both sides (time is irrelevant here) and rearrange a bit to obtain the steady state diffusion equation

Steady State Reaction-Diffusion

$$\frac{D}{h^2}\left(-y_{i-1}+2y_i-y_{i+1}\right)=f_i. \tag{1.12}$$

With reference to Example 1.3, the relevant variant of Fick's first law is *Fourier's heat law*. For a one-dimensional flow, it says that the flow of internal heat per unit length from one point to another is proportional to the negative rate of change in temperature with respect to directed distance from the one point to the other. The positive constant of proportionality α is known as the *thermal conductivity* of the material. This law uses temperature as a proxy for heat. So if one wants a time dependent equation for temperature alone, one can use the fact that a change in heat is proportional to a change in temperature with a constant of proportionality $c_p\rho$ (called the *volumetric heat capacity*), where c_p is the specific heat and ρ is the mass density of the conducting medium. If the heat density source is $q(x,t)$, then the corresponding temperature source is $c_p\rho q(x,t)$. In this case the time dependent heat equation in terms of *temperature* $y(x_i,t_j)\approx y_{i,j}$ becomes

$$c_p\rho y_{i,j+1}=c_p\rho y_{i,j}+\frac{k\alpha}{h^2}\left(y_{i-1,j}-2y_{i,j}+y_{i+1,j}\right)+kq\left(x_i,t_j\right)$$

or

$$y_{i,j+1}=y_{i,j}+\frac{kK}{h^2}\left(y_{i-1,j}-2y_{i,j}+y_{i+1,j}\right)+kf\left(x_i,t_j\right), \tag{1.13}$$

where the so-called *thermal diffusivity* is $K=\alpha/\left(c_p\rho\right)$ and the temperature flux term is $f(x,t)=q(x,t)/\left(c_p\rho\right)$. If we are in a steady state of temperature, then $y_{i,j+1}=y_{i,j}$, so just as in the case of general steady state diffusion we can cancel these terms, set $y_{i,j+1}=y_{i,j}$, $f(x_i,t_j)=f(x_i)$ and rearrange a bit to obtain the system of equations for temperature used in Example 1.3:

Steady-State Heat Flow

$$-y_{i-1}+2y_i-y_{i+1}=\frac{h^2}{K}f\left(x_i\right),\ i=1,2,...,n. \tag{1.14}$$

Equivalently, one can use the heat source function in this equation since

$$\frac{h^2}{K}f\left(x_i\right)=\frac{h^2}{\alpha/\left(c_p\rho\right)}\frac{q\left(x_i\right)}{c_p\rho}=\frac{h^2}{\alpha}q\left(x_i\right). \tag{1.15}$$

Although one does not need to have studied partial differential equations to understand equations (1.10–1.14), what we have really developed here amounts to numerical methods for solving several partial differential equations (PDEs). There are many fine texts on the theory, applications and numerics of partial differential equations. Readers with a calculus background who might be interested in pursuing these important topics further can consult, e.g., [20] for an introduction to applied PDEs, [23] for an advanced study of applied PDEs and their theory, or [24] for an introduction to numerical PDEs.

1.5 Exercises and Problems

Exercise 1. What is the flop count for calculating the value of the expression $x\,(x\,(ax+b)+c)+d$ as written?

Exercise 2. What is the flop count for calculating the value of the expression ax^3+bx^2+cx+d as written and how does the value of the expression compare to that of Exercise 1?

Exercise 3. Carry out the calculation $((\frac{2}{3}+100)-100)-\frac{2}{3}$ on ALAMA Calculator or other technology tools that use floating point. Do you get the correct answer?

Exercise 4. Use Gaussian elimination with partial pivoting and technology tool to solve the system (1.9) with $\epsilon = 10^{-14}$. How many digits of accuracy does your answer contain?

Problem 5. Let c be a positive constant with $0 < c < 1$. Compute the RREF of the following matrix by hand:

$$A = \begin{bmatrix} 1 & c & & 1 \\ 0 & c & 1-c \\ c & c & & 1 \end{bmatrix}.$$

Now use a floating point technology tool with a built-in RREF command (such as ALAMA calculator) to find the RREF of A with $c = 10^{-8}$ and $c = 10^{-15}$. Do the results confirm your calculation?

*Problem 6. Show that the flop count for back solving an $n \times n$ system is quadratic in n.

Problem 7. Compare the strategy of Gauss–Jordan elimination by using each pivot to zero out all entries above and below before proceeding to the next pivot to the forward solve/back solve strategy. Which is computationally more expensive? Illustrate both strategies with the matrix

$$A = \begin{bmatrix} 3 & 1 & 9 & 2 \\ -3 & 0 & 6 & -5 \\ 6 & 1 & 3 & 0 \end{bmatrix}.$$

*Problem 8. Modify equation (1.10) by evaluating the gradient terms (coefficients of kD/h^2) and source terms at time level $k+1$ instead of level k. Such an equation is called an *implicit* marching method (in this case the *implicit Euler method*). Write out the resulting equation as a linear equation with unknowns on the left and knowns on the right.

Problem 9. Obtain another implicit marching method by averaging the implicit and explicit Euler methods for general diffusion. Write out the resulting linear system as a linear system of equations with unknowns on the left and knowns on the right.

1.6 *Projects and Reports

In this section we give a few samples of project material (reports too – these are just promoted projects). These projects provide an opportunity to explore a subject in a greater depth than exercises permit. Your instructor will define her/his own expectations for projects. Also, the technology tools used for the projects will vary.

About writing project/reports: Here are a few suggestions.

- *Know your audience.* Usually, you may assume that your report will be read by your supervisors, who are technical people such as yourself. Therefore, you should write a brief statement of the problem and discussion of methodology. In practice, reports assume physical laws and assumptions without further justification, but in real life you might be expected to offer some explanation of physical principles used in your model.
- *Structure your paper.* Stream of consciousness doesn't work here. Have in mind a target length for your paper. Don't clutter your work with long lists of numbers and try to keep the length at a minimum rather than maximum. Generally, a discourse should have three parts: beginning, middle, and end. Roughly, a beginning should consist of introductory material. In the middle you develop the ideas described or theses proposed in the introduction, and in the end you summarize your work and tie up loose ends.
- *Pay attention to appearance and neatness*, but don't be overly concerned about your writing style. Remember that "simpler is better." Prefer short and straightforward sentences to convoluted ones. Use a vocabulary with which you are comfortable. Use a spell-checker if one is available.
- *Pay attention to format.* A project/report assignment may be supplied with a report template by your instructor or carry explicit instructions about format, intended audience, etc. Read and follow these instructions carefully.
- *Acknowledge your sources.* Use every available resource, of course. In particular, we all know that the internet is a gold mine of information (and disinformation!). Utilize it and other resources fully, but give appropriate references and credits, just as you would with a textbook source.

Of course, rules about paper writing are not set in concrete. Also, a part can be quite short; for example, an introduction might only be a paragraph or two. Here is a sample skeleton for a report (perhaps rather more elaborate than you need): 1. Introduction (title page, summary, and conclusions); 2. Main sections (problem statement, assumptions, methodology, results, conclusions); 3. Appendices (such as mathematical analysis, graphs, possible extensions, etc.), and References.

Project: The Accuracy of Gaussian Elimination

Problem Description: This project is concerned with determining the accuracy of Gaussian elimination as applied to two linear systems, one of which is known

to be difficult to solve numerically. Both of these systems will be square (equal number of unknowns and equations) and have a unique solution. Also, both of these systems are to be solved for various sizes, namely $n = 4, 8, 12, 16, 24$. In order to get a handle on the error, our main interest, we shall start with a known answer. The answer shall consist in setting all variables equal to 1. So it is the solution vector $(1, 1, \ldots, 1)$. The coefficient matrix shall be one of two types:

(1) A Hilbert matrix, i.e., an $n \times n$ matrix given by the formula $H_n = \left[\frac{1}{i+j-1}\right]$.
(2) An $n \times n$ matrix with random entries from a uniform distribution on $[0, 1]$.

The right-hand-side vector $\mathbf{b}$ is uniquely determined by the coefficient matrix and solution. In fact, the entries of $\mathbf{b}$ are easy to obtain: Simply add up all the entries in the ith row of the coefficient matrix to obtain the ith entry of $\mathbf{b}$.

The problem is to measure the error of Gaussian elimination. This is done by finding the largest (in absolute value) difference between the computed value of each variable and actual value, which in all cases is 1. Discuss your results and draw conclusions from your experiments. Be sure that the technology tools that you use employ floating point calculations and Gaussian elimination as in RREF or LU factorization. (ALAMA calculator has an explicit RREF command).

Implementation Notes: Consult your technology tool user guide for instructions on generating Hilbert matrices and generating random matrices. Finding error vectors involves coordinate-wise subtraction of two vectors. Though not necessary for this project, Chapter 2 examines the arithmetic of matrices and vectors is the subject matter of Chapter 2.

Project: An Inverse Problem

Problem Description: You are given a long tube of still dry air in which there are 7 sampling/insertion points equally spaced 1/6 meters apart from each other. The position of each point is measured by setting the leftmost point at 0.0 meters and rightmost at 1.0 meters. Initially, a small amount of a certain gas is inserted in the central insertion point. Subsequently, measurements of the concentration of the gas at each sampling/insertion point are taken at later times in seconds. The results of these measurements, which you may assume are accurate to about 2-3 digits, are specified in Table 1.1. Based on this information, your task is to determine the best estimate you can find for the true value of the diffusion coefficient D of this gas in a motionless air medium. Use this estimate and a marching method to calculate values of the material density function on the interval $[0, 1]$ at times $t = 210$ and $t = 300$ and at the given spatial nodes.

Procedure: You should use equation (1.10) or some variant to move backward and forward in time. These will result in linear systems, which ALAMA calculator or another technology tool can solve. One way to proceed is simply to use trial and error until you think you've hit on a reasonable value of D, that is, the one that gives the best approximation to $t = 180$ from the $t = 360$

Sec \ Meter	0	1/6	1/3	1/2	2/3	5/6	1
$t = 240$	0.0	0.032	1.23	3.69	1.23	0.032	0.0
$t = 270$	0.0	0.051	1.21	3.48	1.21	0.051	0.0

Table 1.1: Concentration data measurements of a gaseous material.

values. Do *not* expect perfect matches – the data is relatively sparse. Then march backwards in time once more to get the initial values at $t = 0$. Finally, march forward in time to compute and plot the resulting approximate density function.

Output: Discus your results and provide a graph of profiles of the material density function at times in the data table along with your computed profiles.

Comments: This project introduces you to a very interesting area of mathematics called "inverse theory." The idea is, rather than proceeding from problem (the governing equations for concentration values) to solution (concentration profiles), you are given the "solution," namely the measured solution values at various points, and are to determine from this information the "problem," i.e., the diffusion coefficient needed to define the governing equations.

Report: Heat Flow

Problem Description: You are working for the firm Universal Dynamics on a project that has a number of components. You have been assigned the analysis of a component that is similar to a laterally insulated rod. The problem: Part of the specs for the rod dictate that *no point of the rod should stay at temperatures above* 10 *degrees Celsius for a long period of time.* You must decide whether any of the materials listed below are acceptable for making the rod and write a report on your findings. You may assume that the rod is one meter in length. Suppose further that internal heat sources come from a position-dependent function $q(x) = 6600 \sin(\pi x^2)$, $0 \le x \le 1$. Also suppose that the left and right ends of the rod are held at 0 and 10 degrees Celsius, respectively. (You may assume that these numbers are appropriate for the unspecified SI units.) When sufficient time passes, the temperature of the rod at each point will settle down to "steady-state" values, dependent only on position x. These are the temperatures you are interested in. Refer to the discussion in Section 1.5 for the details of the descriptive equations that result from discretizing the problem into finitely many nodes.

The heat diffusivity constants for the materials under consideration for the rod are contained in Table 1.2. Based on this data, which of these materials (if any) are acceptable?

Procedure: For the solution of the problem, formulate a discrete approximation to the problem using equations (1.14) and (1.15). Choose an integer n and divide the interval $[0, 1]$ into $n + 1$ equal subintervals with endpoints $0 = x_0, x_1, \ldots, x_{n+1} = 1$. Then the width of each subinterval is $h = 1/(n+1)$. Next let y_i be your approximation to $y(x_i)$ and proceed as in Example 1.3. There results a linear system of n equations in the n unknowns $y_1, y_2, \ldots, y_n$.

Metal	Thermal Conductivity α
Aluminum	204
Copper	386
Iron	72.7
Silver	419

Table 1.2: Conductivity coefficients for selected metals.

For this problem take $n = 4$. Use the largest y_i as an estimate of the highest temperature at any point in the rod. Now double the number of subintervals and see whether your values for y change appreciably at a value of x. If they do, you may want to repeat this procedure until you obtain numbers that you judge to be satisfactory.

Output: Return your methods, results and conclusions in the form of a written report which should be intelligible to your fellow students. Steady state temperature profiles for each of the metals are very informative. Data presented in tabular or graphical (or both) formats would be a nice plus. ALAMA calculator is capable of producing the desired data output, but your instructors may have own requirements for technology tool usage.

2

MATRIX ALGEBRA

In Chapter 1 we used matrices and vectors as simple storage devices. In this chapter matrices and vectors take on a life of their own. We develop the arithmetic of matrices and vectors. Much of what we do is motivated by a desire to extend the ideas of ordinary arithmetic to matrices. Our notational style of writing a matrix in the form $A = [a_{ij}]$ hints that a matrix could be treated like a single number. What if we could manipulate equations with matrix and vector quantities in the same way that we do equations with scalars? We shall see that this powerful idea gives us now methods for formulating and solving practical problems. In this chapter we use it to find effective methods for solving linear and nonlinear systems, solve problems of graph theory and analyze an important modeling tool of applied mathematics called a Markov chain.

2.1 Matrix Addition and Scalar Multiplication

To begin our discussion of arithmetic we consider the matter of equality of matrices. Suppose that A and B represent two matrices. When do we declare them to be equal? The answer is, of course, if they represent the same matrix! Thus, we expect that all the usual laws of equalities will hold (e.g., equals may be substituted for equals) and in fact, they do. There are times, however, when we need to prove that two symbolic matrices are equal. For this purpose, we need something a little more precise. So we have the following definition, which includes vectors as a special case of matrices.

Definition 2.1. **Matrix Equality** Two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are said to be *equal* if these matrices have the same size, and for each index pair (i, j), $a_{ij} = b_{ij}$, that is, corresponding entries of A and B are equal.

Example 2.1. Which of the following matrices are equal, if any?

(a) $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 \\ 0 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1 \\ 1-1 & 1+1 \end{bmatrix}$

Solution. The answer is that only (c) and (d) have any chance of being equal, since they are the only matrices in the list with the same size (2×2). As a matter of fact, an entry-by-entry check verifies that they really are equal. □

Matrix Addition and Subtraction

How should we define addition or subtraction of matrices? We take a clue from elementary two- and three-dimensional vectors, such as the type we would encounter in geometry or calculus. There, in order to add two vectors, one condition has to hold: the vectors have to be the same size. If they are the same size, we simply add the vectors coordinate by coordinate to obtain a new vector of the same size, which is what the following definition does.

Definition 2.2. **Matrix Addition and Subtraction** Let $A = [a_{ij}]$ and $B = [b_{ij}]$ be $m \times n$ matrices. Then the *sum* of the matrices, denoted by $A + B$, is the $m \times n$ matrix defined by the formula

$$A + B = [a_{ij} + b_{ij}] .$$

The *negative* of the matrix A, denoted by $-A$, is defined by the formula

$$-A = [-a_{ij}] .$$

Finally, the *difference* of A and B, denoted by $A - B$, is defined by the formula

$$A - B = [a_{ij} - b_{ij}] .$$

Notice that matrices must be the same size before we attempt to add them. We say that two such matrices or vectors are *conformable for addition*.

Example 2.2. Let

$$A = \begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -3 & 2 & 1 \\ 1 & 4 & 0 \end{bmatrix} .$$

Find $A + B$, $A - B$, and $-A$.

Solution. Here we see that

$$A + B = \begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} + \begin{bmatrix} -3 & 2 & 1 \\ 1 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 3-3 & 1+2 & 0+1 \\ -2+1 & 0+4 & 1+0 \end{bmatrix} = \begin{bmatrix} 0 & 3 & 1 \\ -1 & 4 & 1 \end{bmatrix} .$$

Likewise,

$$A - B = \begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} - \begin{bmatrix} -3 & 2 & 1 \\ 1 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 3 - -3 & 1 - 2 & 0 - 1 \\ -2 - 1 & 0 - 4 & 1 - 0 \end{bmatrix} = \begin{bmatrix} 6 & -1 & -1 \\ -3 & -4 & 1 \end{bmatrix}.$$

The negative of A is even simpler:

$$-A = \begin{bmatrix} -3 & -1 & -0 \\ - & -2 & -0 & -1 \end{bmatrix} = \begin{bmatrix} -3 & -1 & 0 \\ 2 & 0 & -1 \end{bmatrix}. \qquad \square$$

Scalar Multiplication

The next arithmetic concept we want to explore is that of scalar multiplication. Once again, we take a clue from the elementary vectors, where the idea behind scalar multiplication is simply to "scale" a vector a certain amount by multiplying each of its coordinates by that amount, which is what the following definition says.

Definition 2.3. Scalar Multiplication Let $A = [a_{ij}]$ be an $m \times n$ matrix and c a scalar. The *product* of the scalar c with the matrix A, denoted by cA, is defined by the formula

$$cA = [ca_{ij}].$$

Recall that the default scalars are real numbers, but they could also be complex numbers.

Example 2.3. Let

$$A = \begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \quad \text{and} \quad c = 3.$$

Find cA, $0A$, and $-1A$.

Solution. Here we see that

$$cA = 3\begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 \cdot 3 & 3 \cdot 1 & 3 \cdot 0 \\ 3 \cdot -2 & 3 \cdot 0 & 3 \cdot 1 \end{bmatrix} = \begin{bmatrix} 9 & 3 & 0 \\ -6 & 0 & 3 \end{bmatrix},$$

while

$$0A = 0\begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$(-1)\,A = (-1)\begin{bmatrix} 3 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -1 & 0 \\ 2 & 0 & -1 \end{bmatrix} = -A. \qquad \square$$

Linear Combinations

Now that we have a notion of scalar multiplication and addition, we can blend these two ideas to yield a very fundamental notion in linear algebra, that of a *linear combination*.

Definition 2.4. Linear Combination A *linear combination* of the matrices $A_1, A_2, \ldots, A_n$ is an expression of the form

$$c_1A_1 + c_2A_2 + \cdots + c_nA_n$$

where $c_1, c_2, \ldots, c_n$ are scalars and $A_1, A_2, \ldots, A_n$ are all of the same size.

Example 2.4. Given that

$$A_1 = \begin{bmatrix} 2 \\ 6 \\ 4 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix}, \quad \text{and} \quad A_3 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

compute the linear combination $-2A_1 + 3A_2 - 2A_3$.

Solution. The solution is that

$$\begin{aligned} -2A_1 + 3A_2 - 2A_3 &= -2 \begin{bmatrix} 2 \\ 6 \\ 4 \end{bmatrix} + 3 \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} - 2 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} -2 \cdot 2 + 3 \cdot 2 - 2 \cdot 1 \\ -2 \cdot 6 + 3 \cdot 4 - 2 \cdot 0 \\ -2 \cdot 4 + 3 \cdot 2 - 2 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \end{aligned}$$

□

It seems like too much work to write out objects such as the vector $(0, 0, 0)$ that occurred in the last equation; after all, we know that all the entries are all 0. So we make the following notational convention. A *zero matrix* is a matrix whose every entry is 0. We shall denote such matrices by the symbol 0.

Zero Matrix

Caution: This convention makes the symbol 0 ambiguous, but the meaning of the symbol will be clear from context, and the convenience gained is worth the potential ambiguity. For example, the equation of the preceding example is stated very simply as $-2A_1 + 3A_2 - 2A_3 = 0$, where we understand from context that 0 has to mean the 3×1 column vector of zeros. If we use boldface for vectors, we will also then use boldface for the vector zero, so some distinction is regained.

Example 2.5. Use the identity $-2A_1 + 3A_2 - 2A_3 = 0$ of the preceding example to express A_1 in terms of A_2 and A_3.

Solution. To solve this problem, just forget that the quantities A_1, A_2, A_3 are anything special and use ordinary algebra. First, add $-3A_2 + 2A_3$ to both sides to obtain

$$-2A_1 + 3A_2 - 2A_3 - 3A_2 + 2A_3 = -3A_2 + 2A_3,$$

so that

$$-2A_1 = -3A_2 + 2A_3,$$

and multiply both sides by the scalar $-\frac{1}{2}$ to obtain the identity

$$A_1 = \frac{-1}{2}(-2A_1) = \frac{-1}{2}(-3A_2 + 2A_3) = \frac{3}{2}A_2 - A_3. \qquad \square$$

The linear combination idea has a really useful application to linear systems, namely, it gives us another way to express the solution set of a linear system that clearly identifies the role of free variables. The following example illustrates this point.

Example 2.6. Suppose that a linear system in the unknowns x_1, x_2, x_3, x_4 has general solution $(x_2 + 3x_4, x_2, 2x_2 - x_4, x_4)$, where the variables x_2, x_4 are free. Describe the solution set of this linear system in terms of linear combinations with free variables as coefficients.

Solution. The trick here is to use only the parts of the general solution involving x_2 for one vector and the parts involving x_4 as the other vectors in such a way that these vectors add up to the general solution. In our case

$$\begin{bmatrix} x_2 + 3x_4 \\ x_2 \\ 2x_2 - x_4 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_2 \\ 2x_2 \\ 0 \end{bmatrix} + \begin{bmatrix} 3x_4 \\ 0 \\ -x_4 \\ x_4 \end{bmatrix} = x_2 \begin{bmatrix} 1 \\ 1 \\ 2 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 3 \\ 0 \\ -1 \\ 1 \end{bmatrix}.$$

Now simply define vectors $A_1 = (1, 1, 2, 0)$, $A_2 = (3, 0, -1, 1)$, and we see that since x_2 and x_4 are arbitrary, the solution set is

$$S = \{x_2A_1 + x_4A_2 \mid x_2, x_4 \in \mathbb{R}\}.$$

In other words, the solution set to the system is the set of all possible linear combinations of the vectors A_1 and A_2. $\square$

The idea of solution sets as linear combinations is an important one that we will return to in later chapters. You might notice that once we have the general form of a solution vector we can see that there is an easier way to determine the constant vectors A_1 and A_2. Simply set $x_2 = 1$ and the other free variable(s) equal to zero—in this case just x_4—to get the solution vector A_1, and set $x_4 = 1$ and $x_2 = 0$ to get the solution vector A_2.

Laws of Arithmetic

The last example brings up an important point: to what extent can we rely on the ordinary laws of arithmetic and algebra in our calculations with matrices and vectors? For matrix *multiplication* there are some surprises. On the other hand, the laws for addition and scalar multiplication are pretty much what

we would expect them to be. Here are the laws with their customary names. These same names can apply to more than one operation. For instance, there is a closure law for addition and one for scalar multiplication as well.

Laws of Matrix Addition and Scalar Multiplication Let A, B, C be matrices of the same size $m \times n$, 0 the $m \times n$ zero matrix, and c and d scalars.

(1) (Closure Law) $A + B$ is an $m \times n$ matrix.
(2) (Associative Law) $(A + B) + C = A + (B + C)$
(3) (Commutative Law) $A + B = B + A$
(4) (Identity Law) $A + 0 = A$
(5) (Inverse Law) $A + (-A) = 0$
(6) (Closure Law) cA is an $m \times n$ matrix.
(7) (Associative Law) $c(dA) = (cd)A$
(8) (Distributive Law) $(c + d)A = cA + dA$
(9) (Distributive Law) $c(A + B) = cA + cB$
(10) (Monoidal Law) $1A = A$

It is fairly straightforward to prove from definitions that these laws are valid. The verifications all follow a similar pattern, which we illustrate by verifying the commutative law for addition: let $A = [a_{ij}]$ and $B = [b_{ij}]$ be $m \times n$ matrices. Then we have that

$$\begin{aligned} A + B &= [a_{ij} + b_{ij}] \\ &= [b_{ij} + a_{ij}] \\ &= B + A. \end{aligned}$$

where the first and third equalities come from the definition of matrix addition, and the second equality follows from the fact that for all indices i and j, $a_{ij} + b_{ij} = b_{ij} + a_{ij}$ by the commutative law for addition of scalars.

2.1 Exercises and Problems

Exercise 1. Calculate the following where possible.

(a) $\begin{bmatrix} 1 & 2 & -1 \\ 0 & 2 & 2 \end{bmatrix} - \begin{bmatrix} 3 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ (b) $2\begin{bmatrix} 1 \\ 3 \end{bmatrix} - 5\begin{bmatrix} 2 \\ 2 \end{bmatrix} + 3\begin{bmatrix} 4 \\ 1 \end{bmatrix}$ (c) $2\begin{bmatrix} 1 & 4 \\ 0 & 0 \end{bmatrix} + 3\begin{bmatrix} 0 & 0 \\ 2 & 1 \end{bmatrix}$

(d) $a\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + b\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 2 & -1 \\ 0 & 0 & 2 \\ 0 & 2 & -2 \end{bmatrix} + 2\begin{bmatrix} 3 & 1 & 0 \\ 5 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ (f) $x\begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} + y\begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix}$

Exercise 2. Calculate the following where possible.

(a) $8\begin{bmatrix} 1 & 2 & -1 \\ 1 & 0 & 0 \\ 2 & -1 & 3 \end{bmatrix}$ (b) $-\begin{bmatrix} 2 \\ 3 \end{bmatrix} + 3\begin{bmatrix} 2 \\ -1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 4 & 2 \\ 1 & 0 & 3 \end{bmatrix} + (-4)\begin{bmatrix} 0 & 0 & 1 \\ 2 & 1 & -2 \end{bmatrix}$

(d) $4\begin{bmatrix} 0 & 1 & -1 \\ 2 & 0 & 2 \\ 0 & 2 & 0 \end{bmatrix} - 2\begin{bmatrix} 0 & 2 & 0 \\ -3 & 0 & 1 \\ 1 & -2 & 0 \end{bmatrix}$ (e) $2\begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} + u\begin{bmatrix} -2 \\ 2 \\ 3 \end{bmatrix} + v\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$

Exercise 3. Let $A = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 2 \\ 1 & -2 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}$, and compute the following, where possible.

(a) $A + 3B$ (b) $2A - 3C$ (c) $A - C$ (d) $6B + C$ (e) $2C - 3(A - 2C)$

Exercise 4. With A, B, C as in Exercise 3, solve for the unknown matrix X in the equations

(a) $X + 3A = C$ (b) $A - 3X = 3C$ (c) $2X + \begin{bmatrix} 2 & 2 \\ 1 & -2 \end{bmatrix} = B$.

Exercise 5. Write the following vectors as a linear combination of constant vectors with scalar coefficients x, y, or z.

(a) $\begin{bmatrix} x + 2y \\ 2x - z \end{bmatrix}$ (b) $\begin{bmatrix} x - y \\ 2x + 3y \end{bmatrix}$ (c) $\begin{bmatrix} 3x + 2y \\ -z \\ x + y + 5z \end{bmatrix}$ (d) $\begin{bmatrix} x - 3y \\ 4x + z \\ 2y - z \end{bmatrix}$

Exercise 6. Write the following vectors as a linear combination of constant vectors with scalar coefficients x, y, z, or w.

(a) $\begin{bmatrix} 3x + y \\ x + y + z \end{bmatrix}$ (b) $\begin{bmatrix} 3x + 2y - w \\ w - z \\ x + y - 2w \end{bmatrix}$ (c) $\begin{bmatrix} x + 3y \\ 2y - x \end{bmatrix}$ (d) $\begin{bmatrix} x - 2y \\ 4x + z \\ 3w - z \end{bmatrix}$

Exercise 7. Find scalars a, b, c such that

$$\begin{bmatrix} c & b \\ 0 & c \end{bmatrix} = \begin{bmatrix} a - b & c + 2 \\ a + b & a - b \end{bmatrix}.$$

Exercise 8. Find scalars a, b, c, d such that

$$\begin{bmatrix} d & 2a \\ 2d & a \end{bmatrix} = \begin{bmatrix} a - b & b + c \\ a + b & c - b + 1 \end{bmatrix}.$$

Exercise 9. Express the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ as a linear combination of the four matrices $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$, and $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$.

Exercise 10. Express the matrix $D = \begin{bmatrix} 3 & 3 \\ 1 & -3 \end{bmatrix}$ as a linear combination of the matrices $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$, and $C = \begin{bmatrix} 0 & 2 \\ 0 & -1 \end{bmatrix}$.

Exercise 11. Verify that the associative law and commutative laws for addition hold for

$$A = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & -1 \\ 4 & 1 & 3 \end{bmatrix}, \quad C = \begin{bmatrix} -1 & 0 & -1 \\ 1 & -1 & 0 \end{bmatrix}.$$

Exercise 12. Verify that both distributive laws for addition hold for $c = 2$, $d = -3$, and A, B, and C as in Exercise 11.

Exercise 13. Given that a linear system in the unknowns x_1, x_2, x_3, x_4 has general solution $(x_2 + 3x_4 + 4, x_2, 2 - x_4, x_4)$ for free variables x_2, x_4, find a minimal reduced row echelon for this system.

Exercise 14. Given that a linear system in the unknowns x_1, x_2, x_3, x_4 has general solution $(x_2 + 3x_4 + 4, x_2, x_4 - 2x_2, x_4)$ for arbitrary x_2, x_4, find a minimal reduced row echelon for this system.

Problem 15. Show by examples that it is false that for arbitrary matrices A and B, and constant c,
(a) $\operatorname{rank}(cA) = \operatorname{rank} A$ (b) $\operatorname{rank}(A + B) \geq \operatorname{rank} A + \operatorname{rank} B$.

Problem 16. Prove that the associative law for addition of matrices holds.

Problem 17. Prove that both distributive laws hold.

*Problem 18. Prove that if A and B are matrices such that $2A - 4B = 0$ and $A + 2B = I$, then $A = \frac{1}{2}I$.

Problem 19. Prove the following assertions for $m \times n$ matrices A and B by using the laws of matrix addition and scalar multiplication. Clearly specify each law that you use.
(a) If $A = -A$, then $A = 0$.
(b) If $cA = 0$ for some scalar c, then either $c = 0$ or $A = 0$.
(c) If $B = cB$ for some scalar $c \neq 1$, then $B = 0$.

2.2 Matrix Multiplication

Matrix multiplication is somewhat more subtle than matrix addition and scalar multiplication. Of course, we could define matrix multiplication to be a coordinatewise operation, just as addition is (there is such a thing, called Hadamard multiplication). But our motivation is not merely to make definitions, but rather to make *useful* definitions for basic problems.

Definition of Multiplication

To motivate the definition, let us consider a single linear equation

$$2x - 3y + 4z = 5.$$

We will find it handy to think of the left-hand side of the equation as a "product" of the coefficient matrix $[2, -3, 4]$ and the column matrix of unknowns $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$. Thus, we have that the product of this row and column is

$$[2, -3, 4] \begin{bmatrix} x \\ y \\ z \end{bmatrix} = [2x - 3y + 4z].$$

Notice that we have made the result of the product into a 1×1 matrix. This introduces us to a permanent abuse of notation that is almost always used in linear algebra: we don't distinguish between the scalar a and the 1×1 matrix $[a]$, though technically perhaps we should. In the same spirit, we make the following definition.

Definition 2.5. Row Column Product The *product* of the $1 \times n$ row $[a_1, a_2, \ldots, a_n]$ with the $n \times 1$ column $\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$ is defined to be the 1×1 matrix $[a_1 b_1 + a_2 b_2 + \cdots + a_n b_n]$.

It is this row-column product strategy that guides us to the general definition. Notice how the column number of the first matrix had to match the row number of the second, and that this number disappears in the size of the resulting product. This is exactly what happens in general.

Definition 2.6. Matrix Product Let $A = [a_{ij}]$ be an $m \times p$ matrix and $B = [b_{ij}]$ a $p \times n$ matrix. Then the *product* of the matrices A and B, denoted by AB, is the $m \times n$ matrix whose (i, j)th entry, for $1 \leq i \leq m$ and $1 \leq j \leq n$, is the entry of the product of the ith row of A and the jth column of B; more specifically, the (i, j)th entry of AB is

$$a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{ip} b_{pj}.$$

Notice that, in contrast to the case of addition, two matrices may be of different sizes when we can multiply them together. If A is $m \times p$ and B is $p \times n$, we say that A and B are *conformable* for multiplication. It is also worth noticing that if A and B are square *and of the same size,* then the products AB and BA are always defined.

Some Illustrative Examples

Let's check our understanding with a few examples.

Example 2.7. Compute, if possible, the products AB of the following pairs of matrices A, B.

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

(d) $\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 2 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & -1 \end{bmatrix}$ (f) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$

Solution. First check conformability for multiplication. In part (a) A is 2×3 and B is 3×2. Stack these dimensions alongside each other and see that the 3's match; now "cancel" the matching middle 3's to obtain that the dimension of the product is $2 \times \not{3} \ \not{3} \times 2 = 2 \times 2$. For example, multiply the first row of A by the second column of B to obtain the $(1,2)$th entry of the product matrix:

$$[1,2,1] \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} = [1 \cdot (-2) + 2 \cdot 1 + 1 \cdot 1] = [1].$$

Similarly, the full product calculation looks like this:

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} 4 & -2 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 4 + 2 \cdot 0 + 1 \cdot 2 & 1 \cdot (-2) + 2 \cdot 1 + 1 \cdot 1 \\ 2 \cdot 4 + 3 \cdot 0 + (-1) \cdot 2 & 2 \cdot (-2) + 3 \cdot 1 + (-1) \cdot 1 \end{bmatrix}$$
$$= \begin{bmatrix} 6 & 1 \\ 6 & -2 \end{bmatrix}.$$

A size check of part (b) reveals a mismatch between the column number of the first matrix (3) and the row number (2) of the second matrix. Thus, these matrices are *not conformable* for multiplication in the specified order. Hence, the product

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

is undefined.

In part (c) a size check shows that the product has size $2 \times \not{1} \ \not{1} \times 2 = 2 \times 2$. The calculation gives

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 \cdot 1 & 0 \cdot 2 \\ 0 \cdot 1 & 0 \cdot 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

For part (d) the size check shows gives $1 \times \not{2} \ \not{2} \times 1 = 1 \times 1$. Hence, the product exists and is 1×1. The calculation gives

$$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = [1 \cdot 0 + 2 \cdot 0] = [0] .$$

Matrix Multiplication Not Commutative or Cancellative

Something very interesting comes out of parts (c) and (d). Notice that AB and BA are *not* the same matrices—never mind that their entries are all 0's—the important point is that these matrices are not even the same size! Thus, a very familiar law of arithmetic, the commutativity of multiplication, has just fallen by the wayside.

Things work well in (e), where the size check gives $2 \times \not{2} \quad \not{2} \times 3 = 2 \times 3$ as the size of the product. As a matter of fact, this is a rather interesting calculation:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 1 + 0 \cdot 2 & 1 \cdot 2 + 0 \cdot 3 & 1 \cdot 1 + 0 \cdot (-1) \\ 0 \cdot 1 + 1 \cdot 2 & 0 \cdot 2 + 1 \cdot 3 & 0 \cdot 1 + 1 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & -1 \end{bmatrix} .$$

Notice that we end up with the second matrix in the product. This is similar to the arithmetic fact that $1 \cdot x = x$ for a real number x. So the matrix on the left acted like a multiplicative identity. We'll see that this is no accident.

Finally, for the calculation in (f), notice that

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 1 + 1 \cdot -1 & 1 \cdot 1 + 1 \cdot -1 \\ 1 \cdot 1 + 1 \cdot -1 & 1 \cdot 1 + 1 \cdot -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} .$$

There's something very curious here, too. Notice that two nonzero matrices of the same size multiplied together to give a zero matrix. This kind of thing never happens in ordinary arithmetic, where the cancellation law assures that if $a \cdot b = 0$ then $a = 0$ or $b = 0$. □

The calculation in (e) inspires some more notation. The left-hand matrix of this product has a very important property. It acts like a "1" for matrix multiplication. So it deserves its own name. A matrix of the form

Identity Matrix

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots 0 \\ \vdots & & \ddots & \\ 0 & \dots & & 1 \ 0 \\ 0 & 0 & \dots & 0 \ 1 \end{bmatrix} = [\delta_{ij}]$$

is called an $n \times n$ *identity matrix*.

The (i, j)th entry of I_n is designated by the Kronecker symbol δ_{ij}, which is 1 if $i = j$ and 0 otherwise. If n is clear from context, we simply write I in place of I_n.

Kronecker Symbol

So we see in the previous example that the left-hand matrix of part (e) is

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2 .$$

Linear Systems as a Matrix Product

Let's have another look at a system we examined in Chapter 1. We'll change the names of the variables from x, y, z to x_1, x_2, x_3 in anticipation of a notation that will work with any number of variables.

Example 2.8. Express the following linear system as a matrix product:

$$\begin{aligned} x_1 + x_2 + x_3 &= 4 \\ 2x_1 + 2x_2 + 5x_3 &= 11 \\ 4x_1 + 6x_2 + 8x_3 &= 24 \end{aligned}$$

Solution. Recall how we defined multiplication of a row vector and column vector at the beginning of this section. We use that as our inspiration. Define

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 4 \\ 11 \\ 24 \end{bmatrix}, \quad \text{and } A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 5 \\ 4 & 6 & 8 \end{bmatrix}.$$

Of course, A is just the coefficient matrix of the system and b is the right-hand-side vector, which we have seen several times before. But now these take on a new significance. Notice that if we take the first row of A and multiply it by $\mathbf{x}$ we get the left-hand side of the first equation of our system. Likewise for the second and third rows. Therefore, we may write in the language of matrices that

$$A\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 5 \\ 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 11 \\ 24 \end{bmatrix} = \mathbf{b}.$$

Thus, the system is represented very succinctly as $A\mathbf{x} = \mathbf{b}$. □

Once we understand this example, it is easy to see that the general abstract system that we examined in Section 1.1 can just as easily be abbreviated. Now we have a new way of looking at a system of equations: it is just like a simple first-degree equation in one variable. Of course, the catch is that the symbols $A, \mathbf{x}, \mathbf{b}$ now represent an $m \times n$ matrix, and $n \times 1$ and $m \times 1$ vectors, respectively. In spite of this, the matrix multiplication idea is very appealing. For instance, it might inspire us to ask whether we could somehow solve the system $A\mathbf{x} = \mathbf{b}$ by multiplying both sides of the equation by some kind of matrix "$1/A$" so as to cancel the A and get

$$(1/A)A\mathbf{x} = I\mathbf{x} = \mathbf{x} = (1/A)\mathbf{b}.$$

We'll follow up on this idea in Section 2.5.

Here is another perspective on matrix–vector multiplication that gives a powerful way of thinking about such multiplications.

Example 2.9. Interpret the matrix product of Example 2.8 as a linear combination of column vectors.

Solution. Examine the system of this example and we see that the column $(1, 2, 4)$ appears to be multiplied by x_1. Similarly, the column $(1, 2, 6)$ is multiplied by x_2 and the column $(1, 5, 8)$ by x_3. Hence, if we use the same right-hand-side column $(4, 11, 24)$ as before, we obtain that this column can be expressed as a linear combination of column vectors, namely

$$x_1 \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 2 \\ 6 \end{bmatrix} + x_3 \begin{bmatrix} 1 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 4 \\ 11 \\ 24 \end{bmatrix}. \qquad \square$$

We could write the equation of the previous example very succinctly as follows: let A have columns $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$, so that $A = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$, and let $\mathbf{x} = (x_1, x_2, x_3)$. Then

Matrix-Vector Multiplication

$$A\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3.$$

This formula extends to general matrix–vector multiplication. It is extremely useful in interpreting such products, so we will elevate its status to that of a theorem worth remembering.

Theorem 2.1. Let $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ be an $m \times n$ matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ and let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Then

$$A\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n.$$

Example 2.10. Apply Theorem 2.1 to the following chemistry problem: Benzoic acid (chemical formula $C_7H_6O_2$) oxidizes to carbon dioxide and water. The appropriate chemical equation is

$$C_7H_6O_2 + O_2 \rightarrow CO_2 + H_2O.$$

Balance this equation.

Solution. Here the term "equation" refers to a description of the reactants on the left and products on the right. To balance this equation we must describe how many of each molecule is required on each side in order to make the number of atoms of each element match on both sides. To this end, we describe each molecule in the equation by the a vector in $\mathbb{R}^3$ of the form (c, o, h), where c is the number of carbon atoms, o the number of oxygen atoms and h the number of hydrogen atoms in the molecule. Next let x_1, x_2, x_3, and x_4 represent the number of molecules of benzoic acid, oxygen, carbon dioxide and water, respectively, needed to balance the equation of this reaction. Then the correct balance equation can be described as

$$x_1 \begin{bmatrix} 7 \\ 2 \\ 6 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} = x_3 \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}.$$

If we transport the right-hand side terms to the left and use Theorem 2.1 with $\mathbf{x} = (x_1, x_2, x_3, x_4)$, the resulting system becomes

$$A\mathbf{x} = \begin{bmatrix} 7 & 0 & -1 & 0 \\ 2 & 2 & -2 & -1 \\ 6 & 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Row reduction of the coefficient matrix A yields

$$\begin{bmatrix} 7 & 0 & -1 & 0 \\ 2 & 2 & -2 & -1 \\ 6 & 0 & 0 & -2 \end{bmatrix} \xrightarrow[E_{31}(-\frac{6}{7})]{E_{21}(-\frac{2}{7})} \begin{bmatrix} 7 & 0 & -1 & 0 \\ 0 & 2 & -\frac{12}{7} & -1 \\ 0 & 0 & \frac{6}{7} & -2 \end{bmatrix} \xrightarrow[E_3(\frac{7}{6})]{E_1(\frac{1}{7})\ E_2(\frac{1}{2})} \begin{bmatrix} 1 & 0 & -\frac{1}{7} & 0 \\ 0 & 1 & -\frac{6}{7} & -\frac{1}{2} \\ 0 & 0 & 1 & -\frac{7}{3} \end{bmatrix}$$

$$\xrightarrow[E_{13}(\frac{1}{7})]{E_{23}(\frac{6}{7})} \begin{bmatrix} 1 & 0 & 0 & -\frac{1}{3} \\ 0 & 1 & 0 & -\frac{5}{2} \\ 0 & 0 & 1 & -\frac{7}{3} \end{bmatrix}$$

Thus, x_4 is free, while $x_3 = \frac{7}{3}x_4$, $x_2 = \frac{5}{2}x_4$ and $x_1 = \frac{1}{3}x_4$. However the solution should consist of positive integers only. Therefore, the smallest choice for x_4 is $x_4 = 6$ in which case $x_3 = 14$, $x_2 = 15$ and $x_1 = 2$. This results in the balanced chemical equation

$$2C_7H_6O_2 + 15O_2 \rightarrow 14CO_2 + 2H_2O.$$

□

Laws of Arithmetic

We have already seen that the laws of matrix arithmetic may not be quite the same as the ordinary arithmetic laws that we are used to. Nonetheless, as long as we don't assume a cancellation law or a commutative law for multiplication, things are pretty much what one might expect.

Laws of Matrix Multiplication

Let A, B, C be matrices of the appropriate sizes so that the following multiplications make sense, I a suitably sized identity matrix, and c and d scalars.

(1) (Closure Law) The product AB is a matrix.
(2) (Associative Law) $(AB)C = A(BC)$
(3) (Identity Law) $AI = A$ and $IB = B$
(4) (Associative Law for Scalars) $c(AB) = (cA)B = A(cB)$
(5) (Distributive Law) $(A + B)C = AC + BC$
(6) (Distributive Law) $A(B + C) = AB + AC$

One can formally verify these laws by working through the definitions. For example, to verify the first half of the identity law, let $A = [a_{ij}]$ be an $m \times n$

matrix, so that $I = [\delta_{ij}]$ has to be I_n in order for the product AI to make sense. Now we see from the formal definition of matrix multiplication that

$$AI = \left[\sum_{k=1}^{n} a_{ik}\delta_{kj}\right] = [a_{ij} \cdot 1] = A.$$

The middle equality follows from the fact that δ_{kj} is 0 unless $k = j$. Thus, the sum collapses to a single term. A similar calculation verifies the other laws.

We end our discussion of matrix multiplication with a familiar-looking notation that will prove to be extremely handy in the sequel. This notation applies only to *square* matrices. Let A be a square $n \times n$ matrix and k a nonnegative integer. Then we define the *kth power* of A to be

Exponent Notation

$$A^k = \begin{cases} I_n & \text{if } k = 0, \\ \underbrace{A \cdot A \cdots A}_{k \quad \text{times}} & \text{if } k > 0. \end{cases}$$

As a simple consequence of this definition we have standard exponent laws.

Laws of Exponents
For nonnegative integers i, j and square matrix A:
(1) $A^{i+j} = A^i \cdot A^j$
(2) $A^{ij} = (A^i)^j$

Notice that the law $(AB)^i = A^iB^i$ is missing. It won't work with matrices. (Why not?) The following example illustrates a useful application of exponent notation.

Example 2.11. Let $f(x) = 1 - 2x + 3x^2$ be a polynomial function. Use the definition of matrix powers to derive a sensible interpretation of $f(A)$, where A is a square matrix. Evaluate $f\left(\left[\begin{smallmatrix} 2 & -1 \\ 0 & 1 \end{smallmatrix}\right]\right)$ explicitly with this interpretation.

Solution. Let's take a closer look at the polynomial expression

$$f(x) = 1 - 2x + 3x^2 = 1x^0 - 2x^1 + 3x^2.$$

Since $A^0 = I$ and A is square, the interpretation is easy:

$$f(A) = A^0 - 2A^1 + 3A^2 = I - 2A + 3A^2.$$

In particular, for a 2×2 matrix we take $I = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and obtain

$$\begin{aligned} f\left(\begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix}\right) &= I - 2\begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix} + 3\begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix}^2 \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2\begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix} + 3\begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 4 & -2 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 12 & -9 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 9 & -7 \\ 0 & 2 \end{bmatrix}. \end{aligned}$$

□

2.2 Exercises and Problems

Exercise 1. Carry out these calculations or indicate they are impossible, given that $\mathbf{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 3 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 2 & 1+\mathrm{i} \\ 0 & -1 \end{bmatrix}$.
(a) $\mathbf{b}C\mathbf{a}$ (b) $\mathbf{ab}$ (c) $C\mathbf{b}$ (d) $(\mathbf{a}C)\,\mathbf{b}$ (e) $C\mathbf{a}$ (f) $C\,(\mathbf{ab})$ (g) $\mathbf{ba}$ (h) $C\,(\mathbf{a}+\mathbf{b})$

Exercise 2. For each pair of matrices A, B, calculate the product AB or indicate that the product is undefined.

(a) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 3 & -2 & 0 \\ -2 & 5 & 8 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 & 0 \\ 0 & 8 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 3 & 1 & 2 \\ 1 & 0 & 0 \\ 4 & 3 & 2 \end{bmatrix}, \begin{bmatrix} -5 & 4 & -2 \\ -2 & 3 & 1 \\ 1 & 0 & 4 \end{bmatrix}$

(d) $\begin{bmatrix} 3 & 1 \\ 1 & 0 \\ 4 & 3 \end{bmatrix}, \begin{bmatrix} -5 & 4 & -2 \\ -2 & 3 & 1 \end{bmatrix}$ (e) $\begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}, \begin{bmatrix} -5 & 4 \\ -2 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 2 & 0 \\ 2 & 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix}$

Exercise 3. Express these systems of equations in the notation of matrix multiplication and as a linear combination of vectors as in Example 2.8.

(a) $\begin{aligned} x_1 - 2x_2 + 4x_3 &= 3 \\ x_2 - x_3 &= 2 \\ -x_1 + 4x_4 &= 1 \end{aligned}$ (b) $\begin{aligned} x - y - 3z &= 3 \\ 2x + 2y + 4z &= 10 \\ -x + z &= 3 \end{aligned}$ (c) $\begin{aligned} x - 3y + 1 &= 0 \\ 2y &= 0 \\ -x + 3y &= 0 \end{aligned}$

Exercise 4. Express these systems of equations in the notation of matrix multiplication and as a linear combination of vectors as in Example 2.8.

(a) $\begin{aligned} x_1 + x_3 &= -1 \\ x_2 + x_3 &= 0 \\ x_1 + x_3 &= 1 \end{aligned}$ (b) $\begin{aligned} x - y - 3z &= 1 \\ z &= 0 \\ -x + y &= 3 \end{aligned}$ (c) $\begin{aligned} x - 4y &= 0 \\ 2y &= 0 \\ -x + 3y &= 0 \end{aligned}$

Exercise 5. Let $A = \begin{bmatrix} 2 & -1 & 1 \\ 2 & 3 & -2 \\ 4 & 2 & -2 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$, and $X = \begin{bmatrix} x & 0 & 0 \\ 0 & y & 0 \\ 0 & 0 & z \end{bmatrix}$. Find the coefficient matrix of the linear system $XA\mathbf{b} + A\mathbf{x} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$ in the variables x, y, z.

Exercise 6. Let $A = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}$ and $X = \begin{bmatrix} x & y \\ z & w \end{bmatrix}$. Find the coefficient matrix of the linear system $AX - XA = I_2$ in the variables x, y, z, w.

Exercise 7. Let $\mathbf{u} = (1, 1, 0)$, $\mathbf{v} = (0, 1, 1)$, and $\mathbf{w} = (1, 3, 1)$. Write each of the following expressions as single matrix product.
(a) $2\mathbf{u} - 4\mathbf{v} - 3\mathbf{w}$ (b) $\mathbf{w} - \mathbf{v} + 2\mathrm{i}\mathbf{u}$ (c) $x_1\mathbf{u} - 3x_2\mathbf{v} + x_3\mathbf{w}$

Exercise 8. Express the following matrix products as linear combinations of vectors.

(a) $\begin{bmatrix} 2 & 1 \\ 0 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ -5 \\ 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 \\ 1 & 1+\mathrm{i} \end{bmatrix} \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix}$

Exercise 9. Let $A = \begin{bmatrix} 0 & 2 \\ 1 & 1 \end{bmatrix}$, $f(x) = 1+x+x^2$, $g(x) = 1-x$, and $h(x) = 1-x^3$. Verify that $f(A)\,g(A) = h(A)$.

Exercise 10. Let $A = \begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{5}{2} & -\frac{3}{2} & 0 \end{bmatrix}$. Compute $f(A)$ and $f(B)$, where $f(x) = 2x^3 + 3x - 5$.

Exercise 11. Find all possible products of two matrices from among the following:

$A = \begin{bmatrix} 1 & -2 \\ 1 & 3 \end{bmatrix}$ $B = \begin{bmatrix} 2 & 4 \end{bmatrix}$ $C = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$ $D = \begin{bmatrix} 1 & 3 & 0 \\ -1 & 2 & 1 \end{bmatrix}$

Exercise 12. Find all possible products of three matrices from among the following:

$A = \begin{bmatrix} -1 & 2 \\ 0 & 2 \end{bmatrix}$ $B = \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ 2 & 3 \end{bmatrix}$ $C = \begin{bmatrix} -3 \\ 2 \end{bmatrix}$ $D = \begin{bmatrix} 2 & 3 & -1 \\ 1 & 2 & 1 \end{bmatrix}$ $E = \begin{bmatrix} -2 & 4 \end{bmatrix}$

Exercise 13. A square matrix A is said to be *nilpotent* if there is a positive integer k such that $A^k = 0$. Determine which of the following matrices are nilpotent. (You may assume that if A is $n \times n$ nilpotent, then $A^n = 0$.)

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

Exercise 14. A square matrix A is *idempotent* if $A^2 = A$. Determine which of the following matrices are idempotent.

(a) $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 0 & 2 \\ 1 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}$

Exercise 15. Show by example that a sum of nilpotent matrices need not be nilpotent.

Exercise 16. Show by example that a product of idempotent matrices need not be idempotent.

Exercise 17. Verify that the product $\mathbf{uv}$, where $\mathbf{u} = (1, 0, 2)$ and $\mathbf{v} = \begin{bmatrix} -1 & 1 & 1 \end{bmatrix}$, is a rank-one matrix.

Exercise 18. Verify that the product $\mathbf{uv} + \mathbf{wu}^T$, where $\mathbf{u} = (1, 0, 2)$, $\mathbf{v} = \begin{bmatrix} -1 & 1 & 1 \end{bmatrix}$, and $\mathbf{w} = (1, 0, 1)$, is a matrix of rank at most two.

Exercise 19. Given that $A = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix}$ and $AB = \begin{bmatrix} 5 & -2 & 3 \\ -1 & 1 & -6 \end{bmatrix}$ for a suitable matrix B, find the third column of B.

Exercise 20. Given that $B = \begin{bmatrix} 4 & -4 \\ 2 & 1 \end{bmatrix}$ and $AB = \begin{bmatrix} 10 & -7 \\ 4 & -4 \end{bmatrix}$ for a suitable matrix A, find the first row of A.

Exercise 21. Verify that both associative laws of multiplication hold for

$$c = 4, \qquad A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 2 \\ 0 & 3 \end{bmatrix}, \qquad C = \begin{bmatrix} 1+\mathrm{i} & 1 \\ 1 & 2 \end{bmatrix}.$$

Exercise 22. Verify that both distributive laws of multiplication hold for

$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 2 \\ 0 & 3 \end{bmatrix}, \qquad C = \begin{bmatrix} 1+\mathrm{i} & 1 \\ 1 & 2 \end{bmatrix}.$$

Problem 23. Use the technique of Example 2.10 to balance the following chemical equation:

$$C_8H_{18} + O_2 \rightarrow CO_2 + H_2O.$$

Problem 24. Find examples of 2×2 matrices A and B that fulfill each of the following conditions.

(a) $(AB)^2 \neq A^2B^2$ (b) $AB \neq BA$

Problem 25. Find examples of nonzero 2×2 matrices A, B, and C that fulfill each of the following conditions.

(a) $A^2 = 0,\ B^2 = 0$ (b) $(AB)^2 \neq 0$

*Problem 26. Show that if A is a 2×2 matrix such that $AB = BA$ for every 2×2 matrix B, then A is a multiple of I_2.

Problem 27. Prove that the associative law for scalars is valid.

Problem 28. Prove that both distributive laws for matrix multiplication are valid.

Problem 29. Show that if A is a square matrix such that $A^{k+1} = 0$, then

$$(I - A)\left(I + A + A^2 + \cdots + A^k\right) = I.$$

*Problem 30. Show that if two matrices A and B of the same size have the property that $A\mathbf{b} = B\mathbf{b}$ for every column vector $\mathbf{b}$ of the correct size for multiplication, then $A = B$.

Problem 31. Determine the flop count for multiplication of $m \times p$ matrix A by $p \times n$ matrix B. (See page 54.)

2.3 Applications of Matrix Arithmetic

We next examine a few more applications of the matrix multiplication idea that should reinforce the importance of this idea and provide us with some interpretations of matrix multiplication.

Matrix Multiplication as Function

The function idea is basic to mathematics. Recall that a *function* f is a rule of correspondence that assigns to each argument x in a set called its *domain*, a unique value $y = f(x)$ from a set called its *target*. Each branch of mathematics has its own special functions; for example, in calculus differentiable functions $f(x)$ are fundamental.

Linear algebra also has its special functions. Suppose that $T(\mathbf{u})$ represents a function whose arguments $\mathbf{u}$ and values $\mathbf{v} = T(\mathbf{u})$ are vectors. We say that the function T is *linear* if T preserves linear combinations, that is, for all vectors $\mathbf{u}, \mathbf{v}$ in the domain of T, and scalars c, d, we have that $c\mathbf{u} + d\mathbf{v}$ is in the domain of T and

Linear Functions

$$T(c\mathbf{u} + d\mathbf{v}) = cT(\mathbf{u}) + dT(\mathbf{v}).$$

Example 2.12. Show that the function T, whose domain is the set of 2×1 vectors and definition is

$$T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = x$$

is a linear function.

Solution. Let (x, y) and (z, w) be two elements in the domain of T and c, d any two scalars. Now compute

$$\begin{aligned} T\left(c\begin{bmatrix} x \\ y \end{bmatrix} + d\begin{bmatrix} z \\ w \end{bmatrix}\right) &= T\left(\begin{bmatrix} cx \\ cy \end{bmatrix} + \begin{bmatrix} dz \\ dw \end{bmatrix}\right) = T\left(\begin{bmatrix} cx + dz \\ cy + dw \end{bmatrix}\right) \\ &= cx + dz = cT\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) + dT\left(\begin{bmatrix} z \\ w \end{bmatrix}\right). \end{aligned}$$

Thus, T satisfies the definition of linear function. □

One can check that the function T just defined can be expressed as a matrix multiplication, namely,

$$T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}.$$

This example gives yet another reason for defining matrix multiplication in the way that we do. Here is a general definition for these kinds of functions (also known as linear transformations or linear operators).

Definition 2.7. Matrix Operator Let A be an $m \times n$ matrix. The function T_A that maps $n \times 1$ vectors to $m \times 1$ vectors according to the formula

$$T_A(\mathbf{u}) = A\mathbf{u}$$

is called the *linear function (operator or transformation)* associated with the matrix A or simply a *matrix operator*.

Let's verify that this function T actually is linear. Use the definition of T_A along with the distributive law of multiplication and associative law for scalars to obtain that

$$\begin{aligned} T_A(c\mathbf{u} + d\mathbf{v}) &= A(c\mathbf{u} + d\mathbf{v}) = A(c\mathbf{u}) + A(d\mathbf{v}) \\ &= c(A\mathbf{u}) + d(A\mathbf{v}) = cT_A(\mathbf{u}) + dT_A(\mathbf{v}). \end{aligned}$$

Thus, multiplication of vectors by a fixed matrix A is a linear function. Notice that this result contains Example 2.12 as a special case.

Function Composition Notation Recall that the composition of functions f and g is the function $f \circ g$ whose definition is $(f \circ g)(x) = f(g(x))$ for all x in the domain of g.

Example 2.13. Use the associative law of matrix multiplication to show that the composition of matrix multiplication functions corresponds to the matrix product.

Solution. For all vectors $\mathbf{u}$ and for suitably sized matrices A, B, we have by the associative law that $A(B\mathbf{u}) = (AB)\mathbf{u}$. In function terms, this means that $T_A(T_B(\mathbf{u})) = T_{AB}(\mathbf{u})$. Since this is true for all arguments $\mathbf{u}$, it follows that $T_A \circ T_B = T_{AB}$, which is what we were to show. □

We will have more to say about linear functions in Chapters 3 and 6, where they will go by the name of linear operators. Here is an example that gives another slant on why the "linear" in "linear function."

Example 2.14. Describe the action of the matrix operator T_A on the x-axis and y-axis, where $A = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}$.

Solution. A typical element of the x-axis has the form $\mathbf{v} = (x, 0)$. Thus, we have that $T(\mathbf{v}) = T((x, 0))$. Now calculate

$$T(\mathbf{v}) = T_A((x, 0)) = A\mathbf{v} = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} 2x \\ 4x \end{bmatrix} = x \begin{bmatrix} 2 \\ 4 \end{bmatrix}.$$

Thus, the x-axis is mapped to all multiples of the vector $(2, 4)$. Set $t = 2x$, and we see that $x(2, 4) = (t, 2t)$. Hence, these are simply points on the line $x = t$, $y = 2t$. Equivalently, this is the line $y = 2x$. Similarly, one checks that the y-axis is mapped to the line $y = 2x$ as well. □

Example 2.15. Let L be set of points (x, y) defined by the equation $y = x+1$ and let $T_A(L) = \{T(((x, y)) \mid (x, y) \in L\}$, where $A = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}$. Describe and sketch these sets in the plane.

Solution. Of course, the set L is just the straight line defined by the linear equation $y = x + 1$. To see what $T_A(L)$ looks like, write a typical element of L in the form $(x, x + 1)$. Now calculate

$$T_A((x, x+1)) = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} x \\ x+1 \end{bmatrix} = \begin{bmatrix} 3x+1 \\ 6x+2 \end{bmatrix}.$$

Next make the substitution $t = 3x + 1$, and we see that a typical element of $T_A(L)$ has the form $(t, 2t)$, where t is any real number. We recognize these points as exactly the points on the line $y = 2x$. Thus, the function T_A maps the line $y = x + 1$ to the line $y = 2x$. Figure 2.1 illustrates this mapping as well as the fact that T_A maps the line segment from $\left(\frac{-1}{3}, \frac{2}{3}\right)$ to $\left(\frac{1}{6}, \frac{7}{6}\right)$ on L to the line segment from $(0, 0)$ to $\left(\frac{3}{2}, 3\right)$ on $T_A(L)$. □

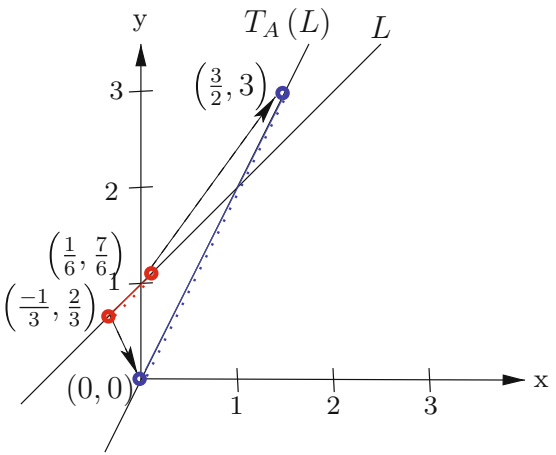


Fig. 2.1: Action of T_A on line L given by $y = x + 1$, points on L, and the segment between them

Graphics specialists and game programmers have a special interest in *real-time rendering*, the discipline concerned with algorithms that create synthetic images fast enough that the viewer can interact with a virtual environment. For a comprehensive treatment of this subject, consult the text [2]. A number of fundamental matrix-defined operators are used in real-time rendering, where they are called *transforms*. Here are a few examples of such operators. A *scaling operator* is effected by multiplying each coordinate of a point by a

Real-Time Rendering

fixed (positive) scale factor. A *shearing operator* is effected by adding a constant shear factor times one coordinate to another coordinate of the point. A *rotation operator* is effected by rotating each point a fixed angle θ in the counterclockwise direction about the origin.

Example 2.16. Let the scaling operator S on points in two dimensions have scale factors of $\frac{3}{2}$ in the x-direction and $\frac{1}{2}$ in the y-direction. Let the shearing operator H on these points have a shear factor of $\frac{1}{2}$ by the y-coordinate on the x-coordinate. Express these operators as matrix operators and graph their action on four unit squares situated diagonally from the origin.

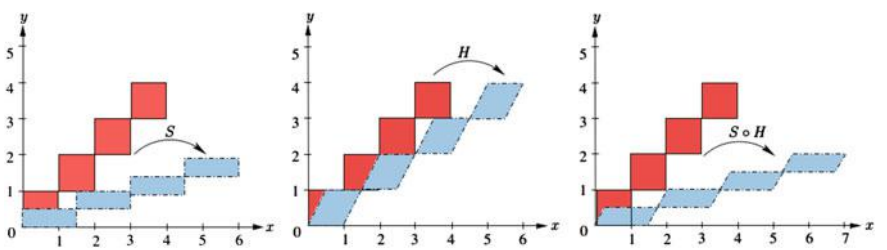
Solution. First consider the scaling operator. The point (x, y) will be transformed into the point $\left(\frac{3}{2}x, \frac{1}{2}y\right)$. Observe that

$$S\left((x,y)\right) = \begin{bmatrix} \frac{3}{2}x \\ \frac{1}{2}y \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = T_A\left((x,y)\right),$$

where $A = \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$. Similarly, the shearing operator transforms the point (x, y) into the point $\left(x + \frac{1}{2}y, y\right)$. Thus, we have

$$H\left((x,y)\right) = \begin{bmatrix} x + \frac{1}{2}y \\ y \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = T_B\left((x,y)\right),$$

where $B = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$. The action of these operators on four unit squares is illustrated in Figure 2.2. □



(a) Scaling in x-direction by $\frac{3}{2}$, y-direction by $\frac{1}{2}$.

(b) Shearing in x-direction by y with shear factor $\frac{1}{2}$.

(c) Concatenation of operators S and H.

Fig. 2.2: Action of scaling operator, shearing operator, and concatenation.

Example 2.17. Express the concatenation $S \circ H$ of the scaling operator S and shearing operator H of Example 2.16 as a matrix operator and graph its action on four unit squares situated diagonally from the origin.

Solution. From Example 2.16 we have that $S = T_A$, where $A = \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$, and $H = T_B$, where $B = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$. From Example 2.13 we know that function composition corresponds to matrix multiplication, that is,

$$S \circ H\left((x,y)\right) = T_A \circ T_B\left((x,y)\right) = T_{AB}\left((x,y)\right)$$
$$= \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & \frac{3}{4} \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = T_C\left((x,y)\right),$$

where

$$C = AB = \begin{bmatrix} \frac{3}{2} & \frac{3}{4} \\ 0 & \frac{1}{2} \end{bmatrix}.$$

The action of $S \circ H$ is illustrated in Figure 2.2. □

Example 2.18. Describe the plane rotation (about the origin) operator.

Solution. Consult Figure 2.3. Observe that if the point (x,y) is given by $(r\cos\phi, r\sin\phi)$ in polar coordinates, then the rotated point (x',y') has coordinates $(r\cos(\theta+\phi), r\sin(\theta+\phi))$. Now use the double-angle formula for angles and obtain that

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} r\cos(\theta+\phi) \\ r\sin(\theta+\phi) \end{bmatrix} = \begin{bmatrix} r\cos\theta\cos\phi - r\sin\theta\sin\phi \\ r\sin\theta\cos\phi + r\cos\theta\sin\phi \end{bmatrix}$$
$$= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} r\cos\phi \\ r\sin\phi \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Now define the *rotation matrix* $R(\theta)$ by

Rotation Matrix

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

It follows that $(x',y') = T_{R(\theta)}\left((x,y)\right)$. □

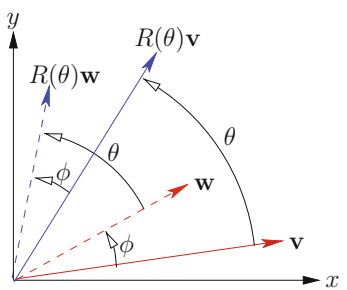


Fig. 2.3: Action of rotation matrix $R(\theta)$ on vectors $\mathbf{v}$ and $\mathbf{w}$

Discrete Dynamical Systems

Discrete linear dynamical systems are an extremely useful modeling tool in a wide variety of disciplines. Here is the definition of such a system.

Definition 2.8. Discrete Dynamical System A *discrete linear dynamical system* is a sequence of vectors $\mathbf{x}^{(k)}$, $k = 0, 1, \ldots$, called *states*, which is defined by an initial vector $\mathbf{x}^{(0)}$ and by the rule

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} + \mathbf{b}_k, \qquad k = 0, 1, \ldots,$$

where A is a fixed square matrix, called the *transition matrix* of the system and the vectors $\mathbf{b}_k$, $k = 0, 1, \ldots$ are called the *input vectors* of the system.

If input vectors are not explicitly specified, we shall assume that $\mathbf{b}_k = \mathbf{0}$ for all k. In this case we call the system a *homogeneous* dynamical system.

Homogeneous Dynamical System

A particularly important question about these systems is whether or not they are stable in the sense that the states $\mathbf{x}^{(k)}$ tend towards a constant state $\mathbf{x}$. In the case of a *homogeneous* dynamical system such a state should have the property that if it is the initial state, then it equals all subsequent states. This observation motivates the following definition:

Definition 2.9. Stationary Vector A vector $\mathbf{x}$ satisfying $\mathbf{x} = A\mathbf{x}$, for a square matrix A, is called a *stationary vector* for A.

In the case that A is the transition matrix for a homogeneous discrete dynamical system, we also call such a vector a *stationary state*. For Markov chains (defined below) we add the condition that the vector be a distribution vector.

Example 2.19. Suppose two toothpaste companies compete for customers in a fixed market in which each customer uses either Brand A or Brand B. Suppose also that a market analysis shows that the buying habits of the customers fit the following pattern in the quarters that were analyzed: each quarter (three-month period), 30% of A users will switch to B, while the rest stay with A. Moreover, 40% of B users will switch to A in a given quarter, while the remaining B users will stay with B. If we *assume* that this pattern does not vary from quarter to quarter, we have an example of what is called a *Markov chain model*. Express the data of this model in matrix–vector language.

Solution. Notice that if a_0 and b_0 are the fractions of the customers using A and B, respectively, in a given quarter, a_1 and b_1 the fractions of customers using A and B in the next quarter, then our hypotheses say that

$$\begin{aligned} a_1 &= 0.7a_0 + 0.4b_0 \\ b_1 &= 0.3a_0 + 0.6b_0. \end{aligned}$$

We could figure out what happens in the quarter after this by replacing the indices 1 and 0 by 2 and 1, respectively, in the preceding formula. In general, we replace the indices $1, 0$ by $k + 1, k$ to obtain

$$a_{k+1} = 0.7a_k + 0.4b_k$$
$$b_{k+1} = 0.3a_k + 0.6b_k.$$

We express this system in matrix form as follows: let

$$\mathbf{x}^{(k)} = \begin{bmatrix} a_k \\ b_k \end{bmatrix} \text{ and } A = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix}.$$

Then the system may be expressed in the matrix form

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}. \qquad \square$$

In light of our interpretation of a linear system as a matrix product, we see that the two equations of Example 2.19 can be written simply as $\mathbf{x}^{(1)} = Ax^{(0)}$. A little more calculation shows that

$$\mathbf{x}^{(2)} = A\mathbf{x}^{(1)} = A \cdot (A\mathbf{x}^{(0)}) = A^2\mathbf{x}^{(0)}$$

and in general,

$$\mathbf{x}^{(k)} = A\mathbf{x}^{(k-1)} = A^2\mathbf{x}^{(k-2)} = \cdots = A^k\mathbf{x}^{(0)}.$$

This is true of any discrete dynamical system and we record this as a *key fact:*

Computing DDS States
For any positive integer k and discrete dynamical system with transition matrix A and initial state $\mathbf{x}^{(0)}$, the k-th state is given by

$$\mathbf{x}^{(k)} = A^k\mathbf{x}^{(0)}.$$

Distribution Vector and Stochastic Matrix

The state vectors $\mathbf{x}^{(k)}$ of the preceding example have the following property: they are column vectors with nonnegative coordinates that sum to 1. Such a vector is called a *distribution (stochastic or probability distribution) vector*. Also, the matrix A has the property that each of its columns is a distribution vector. Such a square matrix is called a *stochastic* matrix. In these terms we now give a precise definition of a Markov chain. (This is really only a special case of what statisticians term a Markov chain, namely a discrete-time finite-state Markov chain.)

Definition 2.10. **Markov Chain** A *Markov chain* is a discrete dynamical system whose initial state $\mathbf{x}^{(0)}$ is a distribution vector and whose transition matrix A is stochastic, i.e., each column of A is a distribution vector.

In the preceding example, we can think of using either Brand A or Brand B as the *events* of the system. A key assumption of Markov chains is that its events are mutually exclusive.

Note that in general the entries of a stochastic matrix $P = [p_{ij}]$ have a simple interpretation when viewed as transition matrix of a Markov chain: Let $\mathbf{e}_j$ be the distribution vector with 1 in the jth entry and 0 in all others. Then as a state vector the meaning of $\mathbf{e}_j$ is that the system has selected jth event exclusively. The next state is $\mathbf{p}_j = P\mathbf{e}_j$, that is, the jth column of P. This implies that the entry p_{ij} is the probability that the ith event will occur, given that the jth event has just occurred. Since events are mutually exclusive and some subsequent event must occur, the sum of these probabilities is 1.

How do subsequent states beyond the first relate to probability distributions in a Markov chain? This is an important question whose answer can be given in a larger context. Suppose we measure the size of a vector by the sum of the absolute values of its coordinates (in Chapter 6 this notion is studied as the 1-*norm* of the vector and denoted by $\|\mathbf{x}\|_1$). For example, $\|(1,-2,3)\|_1 = |1| + |-2| + |3| = 6$. Then we have the following key fact:

Matrix 1-Norm

Stochastic Matrix Inequality For any stochastic matrix P and compatible vector $\mathbf{x}$, $\|P\mathbf{x}\|_1 \leq \|\mathbf{x}\|_1$, with equality if the coordinates of $\mathbf{x}$ are all nonnegative.

This is easily checked. Let $P = [p_{ij}]_{n,n}$ be stochastic, $\mathbf{x} = [x_j]_n$ and use the facts that the entries of P are nonnegative and its columns sum to one to calculate that

$$\sum_{i=1}^{n} \left| \sum_{j=1}^{n} p_{ij} x_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} |x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} p_{ij} = \sum_{j=1}^{n} |x_j| \cdot 1 = \sum_{j=1}^{n} |x_j| .$$

In particular, if $\mathbf{x}$ has all nonnegative coordinates, then we can drop the absolute value signs and the inequality becomes an equality. (Note that if some entry of $\mathbf{x}$ were negative, it is possible for the inequality to be strict.) This shows that all subsequent states in a Markov chain are themselves distribution vectors. Now we really have a very good handle on the Markov chain problem. Consider the following instance of our example.

Markov Chain State

Example 2.20. In the notation of Example 2.19 suppose that initially Brand A has all the customers (i.e., Brand B is just entering the market). What are the market shares 2 quarters later? 20 quarters? Answer the same questions if initially Brand B has all the customers.

Solution. To say that initially Brand A has all the customers is to say that the initial state vector is $\mathbf{x}^{(0)} = (1, 0)$. Now do the arithmetic to find $\mathbf{x}^{(2)}$:

$$\begin{bmatrix} a_2 \\ b_2 \end{bmatrix} = \mathbf{x}^{(2)} = A^2\mathbf{x}^{(0)} = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \left(\begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right)$$
$$= \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix} = \begin{bmatrix} .61 \\ .39 \end{bmatrix}.$$

Thus, Brand A will have 61% of the market and Brand B will have 39% of the market in the second quarter. We did not try to do the next calculation by hand, but rather used a technology tool to get the approximate answer:

$$\mathbf{x}^{(20)} = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix}^{20} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} .57143 \\ .42857 \end{bmatrix}.$$

Thus, after 20 quarters, Brand A's share will have fallen to about 57% of the market and Brand B's share will have risen to about 43%. Now consider what happens if the initial scenario is completely different, i.e., $\mathbf{x}^{(0)} = (0, 1)$. We compute by hand to find that

$$\mathbf{x}^{(2)} = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \left(\begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \begin{bmatrix} .52 \\ .48 \end{bmatrix}.$$

Then we use a technology tool to find that

$$\mathbf{x}^{(20)} = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix}^{20} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} .57143 \\ .42857 \end{bmatrix}.$$

Surprise! For $k = 20$ we get the same answer as we did with a completely different initial condition. There appears to be a unique stationary state which is a steady state for this system. Coincidence? We will return to this example again in Chapters 3 and 5, where concepts introduced therein will cast new light on this model (no, it isn't a coincidence). □

Another important type of model is a so-called *structured population model*. In such a model a population of organisms is divided into a finite number of disjoint states, such as age by year or weight by pound, so that the entire population is described by a state vector that represents the population at discrete times that occur at a constant period such as every day or year. A comprehensive development of this concept can be found in Hal Caswell's text [6]. Here is an example.

Structured Population Model

Example 2.21. A certain insect has three life stages: egg, juvenile, and adult. A population is observed in a certain environment to have the following properties in a two-day time steps: 20% of the eggs will not survive, and 60% will move to the juvenile stage. In the same time-span 10% of the juveniles will not survive, and 60% will move to the adult stage, while 80% of the adults will survive. Also, in the same time-span adults will product about 0.25 eggs per adult. Assume that initially, there are 10, 8, and 6 eggs, juveniles, and adults (measured in thousands), respectively. Model this population as a discrete dynamical system and compute populations total in 2, 10, and 100 days.

Solution. We start time at day 0 and the kth stage is day $2k$. Here the time period is two days and a state vector has the form $\mathbf{x}^{(k)} = (a_k, b_k, c_k)$, where a_k is the number of eggs, b_k the number of juveniles, and c_k the number of adults (all in thousands) on day $2k$. We are given that $\mathbf{x}^{(0)} = (10, 8, 6)$ on day 0. Furthermore, the transition matrix has the form

$$A = \begin{bmatrix} 0.2 & 0 & 0.25 \\ 0.6 & 0.3 & 0 \\ 0 & 0.6 & 0.8 \end{bmatrix}.$$

The first column says that 20% of the eggs will remain eggs over one time period, 60% will progress to juveniles, and the rest do not survive. The second column says that juveniles produce no offspring, 30% will remain juveniles, 60% will become adults, and the rest do not survive. The third column says that .25 eggs results from one adult, no adult becomes a juvenile, and 80% survive. Now do the arithmetic to find the state $\mathbf{x}^{(1)}$ on day 1:

$$\mathbf{x}^{(1)} = \begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix} = A^1\mathbf{x}^{(0)} = \begin{bmatrix} 0.2 & 0 & 0.25 \\ 0.6 & 0.3 & 0 \\ 0 & 0.6 & 0.8 \end{bmatrix} \begin{bmatrix} 10 \\ 8 \\ 6 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 8.4 \\ 9.6 \end{bmatrix}.$$

For the remaining calculations we use a computer (you should check these results with your own calculator or computer) to obtain approximate answers (we use $\approx$ for approximate equality)

$$\mathbf{x}^{(10)} = \begin{bmatrix} a_{10} \\ b_{10} \\ c_{10} \end{bmatrix} = A^{10}\mathbf{x}^{(0)} \approx \begin{bmatrix} 3.33 \\ 2.97 \\ 10.3 \end{bmatrix},$$

$$\mathbf{x}^{(100)} = \begin{bmatrix} a_{100} \\ b_{100} \\ c_{100} \end{bmatrix} = A^{100}\mathbf{x}^{(0)} \approx \begin{bmatrix} 0.284 \\ 0.253 \\ 0.877 \end{bmatrix}.$$

It appears that the population is declining with time. □

The next example constitutes a nonhomogeneous dynamical equation. This important example of difference equations has already made its appearance in Section 1.5 of Chapter 1.

Example 2.22. Consider the equations

$$y_{i,j+1} = \sigma y_{i-1,j} + (1 - 2\sigma)\, y_{i,j} + \sigma y_{i+1,j} + k f_{i,j},\ i = 1, 2, 3, 4 \text{ and } j = 0, 1, 2, \ldots.$$

Here the variables $y_{0,j}$ and $y_{5,j}$ are known for all j, as are the constants k and $f_{i,j}$. Express this system of equations as a dynamical system in matrix-vector form.

Solution. What we are given is a system of four equations in the unknowns $y_{i,j}$, $i = 1, 2, 3, 4$, namely

$$\begin{aligned} y_{1,,j+1} &= (1-2\sigma)\, y_{1,j} + \sigma y_{2,j} + k f_{1,j} + \sigma y_{0,j} \\ y_{2,,j+1} &= \sigma y_{1,j} + (1-2\sigma)\, y_{2,j} + \sigma y_{3,j} + k f_{2,j} \\ y_{3,,j+1} &= \sigma y_{2,j} + (1-2\sigma)\, y_{3,j} + \sigma y_{4,j} + k f_{3,j} \\ y_{4,,j+1} &= \sigma y_{3,j} + (1-2\sigma)\, y_{4,j} + k f_{4,j} + \sigma y_{5,j} \end{aligned}$$

The variables $y_{i,j}$, $i = 1, 2, 3, 4$, play the part of the state vector, while the index j plays the part of a state index. Therefore, we define the state and input vectors to be

$$\mathbf{y}^{(j)} = \begin{bmatrix} y_{1,j} \\ y_{2,j} \\ y_{3,j} \\ y_{4,j} \end{bmatrix} \text{ and } \mathbf{b}^{(j)} = \begin{bmatrix} k f_{1,j} + \sigma y_{0,j} \\ k f_{2,j} \\ k f_{3,j} \\ k f_{4,j} + \sigma y_{5,j} \end{bmatrix}.$$

Next we define the transition matrix for this system to be

$$A = \begin{bmatrix} (1-2\sigma) & \sigma & 0 & 0 \\ \sigma & (1-2\sigma) & \sigma & 0 \\ 0 & \sigma & (1-2\sigma) & \sigma \\ 0 & 0 & \sigma & (1-2\sigma) \end{bmatrix}.$$

Thus, the dynamical system becomes

$$\mathbf{y}^{(j+1)} = A\mathbf{y}^{(j)} + \mathbf{b}^{(j)},\ j = 0, 1, 2, \ldots.$$

As in the preceding examples, in order to actually solve this system we would also need to know the initial state vector $\mathbf{y}^{(0)}$. □

Graphs and Digraphs

We are going to introduce some concepts from graph theory that are useful modeling tools for many practical problems and will accordingly will make their appearance in numerous applications presented in this text.

Definition 2.11. **Graph and Digraph** A *graph* is a set V, whose elements are called *vertices (*or *nodes),* together with a set or list (to allow for repeated edges) E of *unordered* pairs with coordinates in V, called *edges.* If the edges are considered to be ordered pairs, the pair V, E is called a *directed graph* (digraph for short) whose edges are called *directed edges.*

Another useful idea is the following: a *walk* in the graph G is a sequence of graph edges $\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{m-1}, v_m\}$ that goes from vertex v_0 to vertex v_m. The *length* of the walk is m. In the case of a

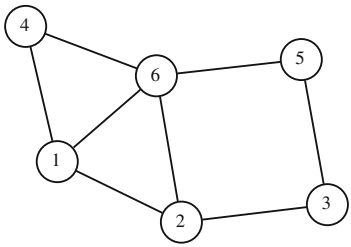


Fig. 2.4: A communications network graph

(Directed) Walk digraph a *(directed) walk* is a sequence of directed edges $(v_0, v_1), (v_1, v_2), \ldots, (v_{m-1}, v_m)$.

Note 2.1. About notation: In subsequent discussion the term "graph" may refer to both directed and undirected graphs. Should we wish to refer specifically to directed graphs, we employ the term "digraph". Otherwise the meaning will be clear from context. For example, specification of the edge set will indicate whether we are referencing a directed or undirected graph.

Example 2.23. Figure 2.4 is a visual representation of a graph in which vertices are exhibited as circled numbers and edges as line segments connecting these vertices. Describe the vertex set V, edge set E and a walk of length five in this graph.

Solution. The figure suggests a graph with vertex set V and edge set E, where

$$V = \{1, 2, 3, 4, 5, 6\}$$
$$E = \{\{1, 2\}, \{2, 3\}, \{3, 5\}, \{5, 6\}, \{6, 4\}, \{4, 1\}, \{1, 6\}, \{6, 2\}\}.$$

The sequence of edges $\{1, 2\}, \{2, 6\}, \{6, 5\}, \{5, 3\}, \{3, 2\}$ or, in terms of the edge indices, $1, 8, 4, 3, 2$, represents one walk of length five. N.B.: The *set* $\{6, 5\}$ is the same as the set $\{5, 6\}$. Were these *directed* edges, the ordered pair $(6, 5)$ would not be the same as the ordered pair $(5, 6)$. □

Example 2.24. You have incomplete data about six teams who have played each other in matches. Each match produces a winner and a loser, with no score attached. Identify the teams by labels $1, 2, 3,\ 4,\ 5,\ 6$. We could describe a match by a pair of numbers (i, j), where team i played and defeated team j (no ties allowed). Here are the data:

$$\{(1, 2), (2, 3), (3, 4), (4, 2), (1, 4), (3, 1), (3, 6), (4, 5), (5, 6)\}.$$

Give a reasonable graphical representation of these data.

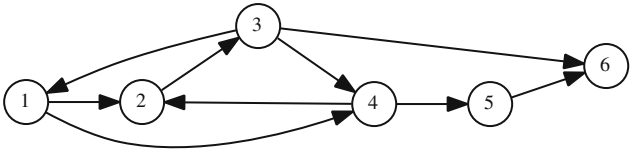


Fig. 2.5: Graph of data from Example 2.24

Solution. The data suggests a digraph with vertices $\{1, 2, 3, 4, 5, 6\}$ and the data as edges, where direction of the arrow along an edge points from winner to loser in a match. We can visually represent all the data by drawing a diagram of this digraph as in Figure 2.5. □

Next, consider the following question relating to Example 2.24. Given this incomplete data, how could we determine a ranking of each team in some sensible way? In order to answer this question, we introduce a notion of "power" that has proved to be useful in many situations. The *power* of a vertex in a digraph is the number of directed walks of length 1 or 2 originating at the vertex. In Figure 2.5, the power of vertex 1 is 5. Why only walks of length 1 or 2? One good reason is that walks of length 3 introduce the possibility of *loops*, i.e., walks that "loop around" to the same point. It isn't very definitive to find out that team 1 beat team 2 beat team 3 beat team 1.

Vertex Power

The digraph of Example 2.24 has no edges from a vertex to itself (called *self-loops*), and for a pair of distinct vertices, at most one edge connecting the two vertices. In other words, a team doesn't play itself and plays another team at most once. Such a digraph is called a *dominance-directed graph*. Although the notion of power of a point is defined for any digraph, it makes the most sense for dominance-directed graphs, like that of Figure 2.5.

Dominance Directed Graph

Example 2.25. Find the power of each vertex in the digraph of Example 2.24 and use this information to rank the teams.

Solution. In this example we could find the power of all points by inspection of Figure 2.5. Let's do it: careful counting gives that the power of vertex 1 is 5, vertex 2 is 4, vertex 3 is 7, vertex 4 is 4, vertex 5 is 1, and the power of vertex 6 is 0. Consequently, team 3 is ranked first, team 1 is second, teams 2 and 4 are tied for third, team 5 is fourth and team 6 is last. □

One can imagine situations (like describing the structure of the communications network pictured in Figure 2.4) in which the edges shouldn't really have a direction, since connections are bidirectional. For such situations the graph concept is a more natural tool.

A practical question: how could we write a computer program to compute powers? More generally, how can we compute the total number of walks of a

certain length? Here is a key tool for the answer: all the information about our graph (or digraph) can be stored in its *adjacency matrix*:

Definition 2.12. **Adjacency Matrix** The *adjacency matrix* of a graph is defined to be a square matrix whose rows and columns are indexed by the vertices of the graph and whose (i, j)th entry is the number of edges going from vertex i to vertex j (this entry is 0 if there are none).

Note that in this definition it is understand that a directed edge (v_i, v_j) of a digraph must start at v_i and end at v_j, while no such restriction applies to the edges of a graph. Thus, an edge $\{v_i, v_j\}$ in a graph is counted twice, namely as going both from vertex v_i to v_j and from vertex v_j to v_i. However, in a digraph an edge (v_i, v_j) is only counted once in the direction from vertex v_i to v_j.

If we designate the adjacency matrix of the digraph of Figure 2.5 by A and the adjacency matrix of the graph of Figure 2.4 by B, then

$$A = \begin{bmatrix} 0&1&0&1&0&0 \\ 0&0&1&0&0&0 \\ 1&0&0&1&0&1 \\ 0&1&0&0&1&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0&1&0&1&0&1 \\ 1&0&1&0&0&1 \\ 0&1&0&0&1&0 \\ 1&0&0&0&0&1 \\ 0&0&1&0&0&1 \\ 1&1&0&1&1&0 \end{bmatrix}.$$

Notice that we could reconstruct the entire digraph or graph from these matrices alone. Thus, all the information about the graph or digraph is, in principle, encapsulated in these matrices.

As an example, consider the problem of counting the number of paths of a given length in a graph. If the graph is very large, it might be quite difficult to track all possible such paths from a picture of the graph. However, it would be much easier to construct an adjacency matrix from such a picture. For a general graph with n vertices and adjacency matrix $A = [a_{ij}]$, we can use this matrix to compute powers of vertices. To count up the walks of length 1 emanating from vertex i, simply add up the elements of the ith row of A. Now what about the paths of length 2? Observe that there is an edge from i to k and then from k to j precisely when the product $a_{ik}a_{kj}$ is equal to 1. Otherwise, one of the factors will be 0 and therefore the product is 0. So the number of paths of length 2 from vertex i to vertex j is the familiar sum

$$a_{i1}a_{1j} + a_{i2}a_{2j} + \cdots + a_{in}a_{nj}.$$

This is just the (i, j)th entry of the matrix A^2. A similar argument shows the following fact:

Theorem 2.2. If A is the adjacency matrix of the graph G, then the (i, j)th entry of A^r gives the number of (directed) walks of length r starting at vertex i and ending at vertex j.

Since the power of vertex i is the number of all paths of length 1 or 2 emanating from vertex i, we have the following key fact:

Corollary 2.1. Vertex Power If A is the adjacency matrix of the digraph G, then the power of the ith vertex is the sum of all entries in the ith row of the matrix $A + A^2$.

Example 2.26. Use the preceding facts to calculate the powers of all the vertices in the digraph of Example 2.24.

Solution. Using the matrix A above we calculate that

$$A + A^2 = \begin{bmatrix} 0&1&0&1&0&0 \\ 0&0&1&0&0&0 \\ 1&0&0&1&0&1 \\ 0&1&0&0&1&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&0 \end{bmatrix} + \begin{bmatrix} 0&1&0&1&0&0 \\ 0&0&1&0&0&0 \\ 1&0&0&1&0&1 \\ 0&1&0&0&1&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&0 \end{bmatrix} \begin{bmatrix} 0&1&0&1&0&0 \\ 0&0&1&0&0&0 \\ 1&0&0&1&0&1 \\ 0&1&0&0&1&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&0 \end{bmatrix} = \begin{bmatrix} 0&2&1&1&1&0 \\ 1&0&1&1&0&1 \\ 1&2&0&2&1&1 \\ 0&1&1&0&1&1 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&0 \end{bmatrix}.$$

An easy way to sum each row with matrix arithmetic is to multiply $A+A^2$ on the right by a column of 1's, but in this case we see immediately that the power of the vertices are consistent with what we observed earlier by inspection of the graph (row 1 sums to 5, 2 to 4, 3 to 7, 4 to 4, 5 to 1 and 6 to 0). □

Difference Equations

The idea of a difference equation has numerous applications in mathematics and computer science. In the latter field, these equations often go by the name of "recurrence relations." They can be used for a variety of applications ranging from population modeling to analysis of complexity of algorithms. We will introduce them by way of a simple financial model.

Example 2.27. Suppose that you invest in a contractual fund where you must invest in the funds for three years before you can receive any return on your investment (with a positive first-year investment). Thereafter, you are vested in the fund and may remove your money at any time. While you are vested in the fund, annual returns are calculated as follows: money that was in the fund one year ago earns nothing, while money that was in the fund two years ago earns 6% of its value and money that was in the fund three years ago earns 12% of its value. Find an equation that describes your investment's growth.

Solution. Let y_k be the amount of your investment in the kth year. The numbers y_0, y_1, y_2 represent your investments for the first three years (we're counting from 0). Consider the third year amount y_3. According to your contract, your total funds in the third year will be

$$y_3 = y_2 + 0.06y_1 + 0.12y_0.$$

Similarly, a general formula for y_{k+3} in terms of the preceding three terms is

$$y_{k+3} = y_{k+2} + 0.06y_{k+1} + 0.12y_k, \qquad k = 0, 1, 2, \ldots. \tag{2.1}$$

□

Definition 2.13. Linear Difference Equations A *linear difference equation* (or *recurrence relation*) of order m in the variables $y_0, y_1, \ldots$ is an equation of the form

$$a_m y_{k+m} + a_{m-1} y_{k+m-1} + \cdots + a_1 y_{k+1} + a_0 y_k = b_k, \quad k = 0, 1, 2, \ldots, \tag{2.2}$$

where $a_0, a_1, \ldots, a_m$ and b_k, $k = 0, 1, 2, \ldots$, are coefficients with $a_0 \neq 0$, $a_m \neq 0$ and $y_0, y_1, \ldots, y_{m-1}$ are initial values. If the coefficients are independent of k, the difference equation is said to have *constant coefficients*. If $b_k = 0$, $k = 0, 1, 2 \ldots$, the difference equation is said to be *homogeneous*, otherwise *nonhomogeneous*.

Notice that such an equation cannot determine the numbers $y_0, y_1, \ldots, y_{m-1}$. These values have to be initially specified, just as in our fund example. Notice that in our fund example, we have to bring all terms of equation (2.1) to the left-hand side to obtain the (constant coefficient homogeneous) third order difference equation form

$$y_{k+3} - y_{k+2} - 0.06y_{k+1} - 0.12y_k = 0.$$

Now we see that $a_3 = 1$, $a_2 = -1$, $a_1 = -0.06$, and $a_0 = -0.12$.

There are many ways to solve difference equations. We are not going to give a complete solution to this problem at this point; we postpone this issue to Chapters 3 and 5, where we introduce vector spaces, eigenvalues and eigenvectors. However, we can now show how to turn a difference equation as given above into a matrix equation. Consider our fund example. The secret is to identify the right vector variables. To this end, define an indexed vector $\mathbf{x}_k$ by the formula

$$\mathbf{x}_k = \begin{bmatrix} y_k \\ y_{k+1} \\ y_{k+2} \end{bmatrix}, \qquad k = 0, 1, 2, \ldots$$

from which it is easy to check that since $a_{k+3} = a_{k+2} + 0.06a_{k+1} + 0.12a_k$, we have

$$\mathbf{x}_{k+1} = \begin{bmatrix} y_{k+1} \\ y_{k+2} \\ y_{k+3} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.12 & 0.06 & 1 \end{bmatrix} \mathbf{x}_k = A\mathbf{x}_k.$$

This is the matrix form we seek. It appears to have a lot in common with the Markov chains examined earlier in this section, in that we pass from one "state vector" to another by multiplication by a fixed "transition matrix" A.

There is a fairly straightforward way to generate nonzero solutions to some homogeneous difference equations: Given equation (2.2) of order m with $b_k = 0$ and all coefficients constant, substitute the values $y_j = r^j$ with $r \neq 0$ to obtain the polynomial equation

$$a_m r^{k+m} + a_{m-1} r^{k+m-1} + \cdots + a_1 r^{k+1} + a_0 r^k = 0.$$

Cancel the common factor of r^k from both sides to obtain the *characteristic polynomial* $p(r)$ and *characteristic polynomial* $p(r) = 0$ of the difference equation:

Characteristic Polynomial

$$p(r) = a_m r^m + a_{m-1} r^{m-1} + \cdots + a_1 r + a_0 = 0.$$

It follows that any root of this polynomial i.e., solution of the polynomial equation, gives us a solution to the difference equation, namely $y_k = r^k$, $k = 0, 1, \ldots$. Note however that this also assigns specific values to $y_0, y_1, \ldots y_{m-1}$; in fact, this kind of particular solution is simply a geometric sequence. We leave it as an exercise to show that if y_k, $k = 0, 1, 2, \ldots$, is a solution to a homogeneous difference equation, any constant multiple ay_k, $k = 0, 1, 2, \ldots$, is also solution to the difference equation.

Example 2.28. Express the linear difference equation $2y_{k+2} - 3y_{k+1} - 2y_k = 0$ in matrix form and find particular solutions to this to this equation.

Solution. For the matrix form, solve for y_{k+2} to obtain $y_{k+2} = \frac{3}{2}y_{k+1} + y_k$, so set $\mathbf{x}_k = (y_k, y_{k+1})$ to obtain

$$\mathbf{x}_{k+1} = \begin{bmatrix} y_{k+1} \\ y_{k+2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} y_k \\ y_{k+1} \end{bmatrix} = A\mathbf{x}_k$$

The polynomial equation that results from this difference equation is

$$0 = 2x^2 - 3x - 2 = 2\left(x^2 - \frac{3}{2}x - 1\right) = 2(x-2)\left(x + \frac{1}{2}\right).$$

Thus, we obtain two particular solutions to this difference equation, namely $y_k = 2^k$ and $y_k = \left(-\frac{1}{2}\right)^k$, $k = 0, 1, 2, \ldots$. □

2.3 Exercises and Problems

Exercise 1. Determine the effect of the matrix operator T_A on the x-axis, y-axis, and the points $(\pm 1, \pm 1)$, where A is one of the following.

(a) $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ (b) $\frac{1}{5}\begin{bmatrix} -3 & -4 \\ -4 & 3 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$

Exercise 2. Determine the effect of the matrix operator T_A on the x-axis, y-axis, and the points $(\pm 1, \pm 1)$, where A is one of the following. Plot the images of the squares with corners $(\pm 1, \pm 1)$.

(a) $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}$

Exercise 3. Express the following functions, if linear, as matrix operators.

(a) $T((x_1, x_2)) = (x_1 + x_2, 2x_1, 4x_2 - x_1)$ (b) $T((x_1, x_2)) = (x_1 + x_2, 2x_1x_2)$
(c) $T((x_1, x_2, x_3)) = (2x_3, -x_1)$ (d) $T((x_1, x_2, x_3)) = (x_2 - x_1, x_3, x_2 + x_3)$

Exercise 4. Express the following functions, if linear, as matrix operators.

(a) $T((x_1, x_2, x_3)) = x_1 - x_3 + 2x_2$ (b) $T((x_1, x_2)) = (|x_1|, 2x_2, x_1 + 3x_2)$
(c) $T((x_1, x_2)) = (x_1, 2x_1, -x_1)$ (d) $T((x_1, x_2, x_3)) = (-x_3, x_1, 4x_2)$

Exercise 5. A linear operator on $\mathbb{R}^2$ is defined by first applying a scaling operator with scale factors of 2 in the x-direction and 4 in the y-direction, followed by a counterclockwise rotation about the origin of $\pi/6$ radians. Express this operator and the operator that results from reversing the order of the scaling and rotation as matrix operators.

Exercise 6. A linear operator on $\mathbb{R}^2$ is defined by first applying a shear in the x-direction with a shear factor of 3 followed by a clockwise rotation about the origin of $\pi/4$ radians. Express this operator and the operator that results from reversing the order of the shear and rotation as matrix operators.

Exercise 7. Find a scaling operator S and shearing operator H such that the concatenation $S \circ H$ maps the points $(1, 0)$ to $(2, 0)$ and $(0, 1)$ to $(4, 3)$.

Exercise 8. Find a scaling operator S and shearing operator H such that the concatenation $S \circ H$ maps the points $(1, 0)$ to $(2, 8)$, $(0, 1)$ to $(0, 4)$ and $(-1, 2)$ to $(-2, 0)$.

Exercise 9. A *fixed-point* of a linear operator T_A is a vector $\mathbf{x}$ such that $T_A(\mathbf{x}) = \mathbf{x}$. Find all fixed points, if any, of the linear operators in Exercise 3.

Exercise 10. Find all fixed points, if any, of the linear operators in Exercise 4.

Exercise 11. Given transition matrices for discrete dynamical systems

(a) $\begin{bmatrix} .1 & .3 & 0 \\ 0 & .4 & 1 \\ .9 & .3 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} .5 & .3 & 0 \\ 0 & .4 & 0 \\ .5 & .3 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 0 & 0.9 \\ 0.5 & 0 & 0 \\ 0 & 0.5 & 0.1 \end{bmatrix}$

and initial state vector $\mathbf{x}^{(0)} = \frac{1}{2}(1, 1, 0)$, calculate the first and second state vector for each system and determine whether it is a Markov chain.

Exercise 12. For each of the dynamical systems of Exercise 11, determine by calculation whether the system tends to a limiting steady-state vector. If so, what is it?

Exercise 13. A population is modeled with two states, immature and mature, and the resulting structured population model transition matrix is $\begin{bmatrix} \frac{1}{2} & 1 \\ \frac{1}{2} & 0 \end{bmatrix}$.
(a) Explain what this matrix says about the two states.
(b) Starting with a population of $(30, 100)$, does the population stabilize, increase or decrease over time? If it stabilizes, to what distribution?

Exercise 14. A population is modeled with three larva, pupa and adult, and the resulting structured population model transition matrix is $\begin{bmatrix} 0 & 0 & 0.6 \\ 0.5 & 0 & 0 \\ 0 & 0.9 & 0.8 \end{bmatrix}$.
(a) Explain what this matrix says about the three states.
(b) Starting with a population of $(0, 30, 100)$, does the population stabilize, increase or decrease over time? If it stabilizes, to what distribution?

Exercise 15. A digraph G has vertex set $V = \{1, 2, 3, 4, 5\}$ and edge set $E = \{(2, 1), (1, 5), (2, 5), (5, 4), (4, 2), (4, 3), (3, 2)\}$. Sketch a picture of the graph G and find its adjacency matrix. Use this to find the power of each vertex of the graph and determine whether this graph is dominance-directed.

Exercise 16. A digraph has adjacency matrix

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix}.$$

Sketch a picture of this digraph and find the total number of directed walks of length at most 3.

Exercise 17. Convert these difference equations into matrix–vector form.
(a) $2y_{k+3} + 3y_{k+2} - 4y_{k+1} + 5y_k = 0$ (b) $y_{k+2} - y_{k+1} + 2y_k = 1$

Exercise 18. Convert these difference equations into matrix–vector form.
(a) $2y_{k+3} + 2y_{k+1} - 3y_k = 0$ (b) $y_{k+2} + y_{k+1} - 2y_k = 3$

Exercise 19. Consider the linear difference $y_{k+2} - y_{k+1} - y_k = 0$.
(a) Express this difference in matrix form.
(b) Find the first ten terms of the solution to this difference given the initial conditions $y_0 = 0$, $y_1 = 1$. (This is the well-known Fibonacci sequence.)

Exercise 20. Find two geometric solutions $y_k = r^k$ to the difference $y_{k+2} - y_{k+1} - y_k = 0$ and show that the difference of these two solutions is also a solution to the difference.

*Problem 21. Show that if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is a real 2×2 matrix, then the matrix multiplication function maps a line through the origin onto a line through the origin or a point.

Problem 22. Show how the transition matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ for a Markov chain can be described using only two variables.

*Problem 23. Use the definition of matrix multiplication function to show that if $T_A = T_B$, then $A = B$.

Problem 24. Suppose that in Example 2.27 you invest $1,000 initially (the zeroth year) and no further amounts. Make a table of the value of your investment for years 0 to 12. Also include a column that calculates the annual interest rate that your investment is earning each year, based on the current and previous year's values. What conclusions do you draw? You will need a technology tool for this exercise.

Problem 25. Show that if the state vector $\mathbf{x}^{(k)} = (a_k, b_k, c_k)$ in a Markov chain is a probability distribution vector, then so is $\mathbf{x}^{(k+1)}$.

*Problem 26. Show if A and B are $n \times n$ stochastic matrices and $0 \leq \alpha \leq 1$, then the matrix $\alpha A + (1 - \alpha)$B is also stochastic.

Problem 27. Suppose that the difference (2.2) is nonhomogeneous, the values y_k, $k = 0, 1, \ldots$, tend to a constant value y_* and the values b_k tend to the constant value $b_* \neq 0$. Find a formula for y_* in terms of the coefficients of (2.2).

Problem 28. Suppose that you invest in the contractual fund of Example (2.27) as follows: In the first year you invest $2,000 and in the next two years you invest $1,000 each. How many years after the third will you have to wait before you break even on your investment?

Problem 29. Show that if $\{y_j\}_{j=0}^{\infty}$ is a solution to a homogeneous difference equation, then so is $\{ay_j\}_{j=0}^{\infty}$ for any constant a.

2.4 Special Matrices and Transposes

There are certain types of matrices that are so important that they have acquired names of their own. We introduce some of these in this section, as well as one more matrix operation that has proved to be a very practical tool in matrix analysis, namely the operation of transposing a matrix.

Elementary Matrices and Gaussian Elimination

We are going to show a new way to execute the elementary row operations used in Gaussian elimination. Recall the shorthand we used:

- E_{ij}: The elementary operation of *switching the ith and jth rows* of the matrix.
- $E_i(c)$: The elementary operation of *multiplying the ith row by the nonzero constant c.*
- $E_{ij}(d)$: The elementary operation of *adding d times the jth row to the ith row.*

From now on we will use the very same symbols to represent matrices. The size of the matrix will depend on the context of our discussion, so the notation is ambiguous, but it is still very useful.

Elementary Matrix

An *elementary matrix* of size n is obtained by performing the corresponding elementary row operation on the identity matrix I_n. We denote the resulting matrix by the same symbol as the corresponding row operation.

Example 2.29. Describe the following elementary matrices of size $n = 3$:
(a) $E_{13}(-4)$ (b) $E_{21}(3)$ (c) E_{23} (d) $E_1(\frac{1}{2})$

Solution. We start with $I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

For part (a) we add -4 times the 3rd row of I_3 to its first row to obtain

$$E_{13}(-4) = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For part (b) add 3 times the first row of I_3 to its second row to obtain

$$E_{21}(3) = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For part (c) interchange the second and third rows of I_3 to obtain that

$$E_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Finally, for part (d) we multiply the first row of I_3 by $\frac{1}{2}$ to obtain

$$E_1\left(\frac{1}{2}\right) = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \qquad \square$$

What good are these matrices? One sees that that following fact is true:

Theorem 2.3. Let $C = BA$ be a product of two matrices and perform an elementary row operation on C. Then the same result is obtained if one performs the same elementary operation on the matrix B and multiplies the result by A on the right.

We won't give a formal proof of this statement, but it isn't hard to see why it is true. For example, suppose one interchanges two rows, say the ith and jth, of $C = BA$ to obtain a new matrix D. How do we get the ith or jth row of C? Answer: multiply the corresponding row of B by the matrix A. Therefore, we would obtain D by interchanging the ith and jth rows of B and multiplying the result by the matrix A, which is exactly what the theorem says. Similar arguments apply to the other elementary operations.

Now take $B = I$, and we see from the definition of elementary matrix and Theorem 2.3 that the following is true.

Corollary 2.2. If an elementary row operation is performed on a matrix A to obtain a matrix A', then $A' = EA$, where E is the elementary matrix corresponding to the elementary row operation performed.

Elementary Operations as Matrix Multiplication

The meaning of this corollary is that we accomplish an elementary row operation by multiplying by the corresponding elementary matrix on the left. Of course, we don't need elementary matrices to accomplish row operations; but they give us another perspective on row operations.

Example 2.30. Express these calculations of Example 1.17 in matrix product form:

$$\begin{bmatrix} 2 & -1 & 1 \\ 4 & 4 & 20 \end{bmatrix} \xrightarrow{E_{12}} \begin{bmatrix} 4 & 4 & 20 \\ 2 & -1 & 1 \end{bmatrix} \xrightarrow{E_1\,(1/4)} \begin{bmatrix} 1 & 1 & 5 \\ 2 & -1 & 1 \end{bmatrix}$$

$$\xrightarrow{E_{21}\,(-2)} \begin{bmatrix} 1 & 1 & 5 \\ 0 & -3 & -9 \end{bmatrix} \xrightarrow{E_2\,(-1/3)} \begin{bmatrix} 1 & 1 & 5 \\ 0 & 1 & 3 \end{bmatrix} \xrightarrow{E_{12}\,(-1)} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix}.$$

Solution. One point to observe: the order of elementary operations. We compose the elementary matrices on the left in the same order that the operations are done. Thus, we may state the above calculations in the concise form

$$\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix} = E_{12}\,(-1)\,E_2\,(-1/3)\,E_{21}\,(-2)\,E_1\,(1/4)\,E_{12} \begin{bmatrix} 2 & -1 & 1 \\ 4 & 4 & 20 \end{bmatrix}. \qquad \square$$

It is important to read the preceding line carefully and understand how it follows from the long form above. This conversion of row operations to matrix multiplication will prove to be very useful in the next section.

Some Matrices with Simple Structure

Certain types of matrices have already been named in our discussions. For example, the identity and zero matrices are particularly useful. Another example is the reduced row echelon form. What's next? Let us classify some simple matrices and attach names to them. For square matrices, we have the following definitions, in ascending order of complexity.

Definition 2.14. Simple Structure Matrices Let $A = [a_{ij}]$ be a square $n \times n$ matrix. Then A is

- *Scalar* if $a_{ij} = 0$ and $a_{ii} = a_{jj}$ for all $i \neq j$. (Equivalently: $A = cI_n$ for some scalar c, which explains the term "scalar.")
- *Diagonal* if $a_{ij} = 0$ for all $i \neq j$. (Equivalently: Off-diagonal entries of A are 0.)
- *(Upper) triangular* if $a_{ij} = 0$ for all $i > j$. (Equivalently: Subdiagonal entries of A are 0.)
- *(Lower) triangular* if $a_{ij} = 0$ for all $i < j$. (Equivalently: Superdiagonal entries of A are 0.)
- *Triangular* if the matrix is upper or lower triangular.
- *Strictly triangular* if it is triangular and the diagonal entries are also zero.
- *Tridiagonal* if $a_{ij} = 0$ when $j > i + 1$ or $j < i - 1$. (Equivalently: Entries off the main diagonal, first subdiagonal, and first superdiagonal are zero.)

The index conditions that we use above have simple interpretations. For example, the entry a_{ij} with $i > j$ is located further down than over, since the row number is larger than the column number. Hence, it resides in the "lower triangle" of the matrix. Similarly, the entry a_{ij} with $i < j$ resides in the "upper triangle." Entries a_{ij} with $i = j$ reside along the main diagonal of the matrix. See Figure 2.6 for a picture of these triangular regions of the matrix.

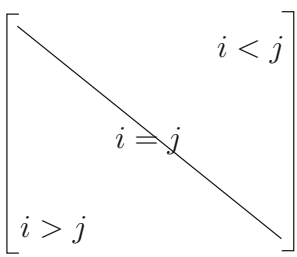


Fig. 2.6: Matrix regions

Example 2.31. Classify the following matrices (elementary matrices are understood to be 3×3) in the terminology of Definition 2.14.

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

(e) $\begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 4 \\ 0 & 0 & 0 \end{bmatrix}$ (f) $E_{21}(3)$ (g) $E_2(-3)$ (h) $\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$

Solution. Notice that (a) is not scalar, since diagonal entries differ from each other, but it is a diagonal matrix, since the off-diagonal entries are all 0. On the other hand, the matrix of (b) is really just $2I_3$, so this matrix is a scalar matrix. Matrix (c) has all terms below the main diagonal equal to 0, so this matrix is triangular and, specifically, upper triangular. Similarly, matrix (d) is lower triangular. Matrix (e) is clearly upper triangular, but it is also strictly upper triangular since the diagonal terms themselves are 0. Next, we have

$$E_{21}(3) = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad E_2(-3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

so that $E_{21}(3)$ is (lower) triangular and $E_2(-3)$ is a diagonal matrix. Matrix (h) comes from Example 1.3, where we saw that an approximation to a certain diffusion problem led to matrices of that form. This matrix is clearly tridiagonal. In fact, note that the matrices of (a), (b), (f), and (g) also can also be classified as tridiagonal. □

Block Matrices

Another type of matrix that occurs frequently enough to be discussed is a *block matrix*. Actually, we already used the idea of blocks when we described the augmented matrix of the system $A\mathbf{x} = \mathbf{b}$ as the matrix $\tilde{A} = [A \,|\, \mathbf{b}]$. The *block*s of $\tilde{A}$ in *partitioned* form $[A, \mathbf{b}]$ are A and $\mathbf{b}$. There is no reason we couldn't partition by inserting more vertical lines or horizontal lines as well, and this partitioning leads to the blocks. The main point to bear in mind when using the block notation is that the blocks must be correctly sized so that the resulting matrix makes sense. One virtue of the block form that results from partitioning is that for purposes of matrix addition or multiplication, we can treat the blocks rather like scalars, provided the addition or multiplication that results makes sense. We will use this idea from time to time without fanfare. One could go through a formal description of partitioning and proofs; we won't. Rather, we'll show how this idea can be used by example.

Block Notation

Example 2.32. Use block multiplication to simplify this multiplication:

$$\begin{bmatrix} 1&2&0&0\\ 3&4&0&0\\ 0&0&1&0 \end{bmatrix} \begin{bmatrix} 0&0&2&1\\ 0&0&1&1\\ 0&0&1&0\\ 0&0&0&1 \end{bmatrix}.$$

Solution. The blocking we want to use makes the column numbers of the blocks on the left match the row numbers of the blocks on the right and looks like this:

$$\left[\begin{array}{cc|c} 1&2&0\ 0\\ 3&4&0\ 0\\ \hline 0&0&1\ 0 \end{array}\right] \left[\begin{array}{cc|c} 0&0&2\ \ 1\\ 0&0&-1\ 1\\ \hline 0&0&1\ 0\\ 0&0&0\ 1 \end{array}\right].$$

We see that these submatrices are built from zero matrices and these blocks:

$$A = \begin{bmatrix} 1&2\\ 3&4 \end{bmatrix}, \qquad B = \begin{bmatrix} 1&0 \end{bmatrix}, \qquad C = \begin{bmatrix} 2&1\\ -1&1 \end{bmatrix}, \qquad I_2 = \begin{bmatrix} 1&0\\ 0&1 \end{bmatrix}.$$

Now we can work this product out by interpreting it as

$$\begin{bmatrix} A&0\\ 0&B \end{bmatrix} \begin{bmatrix} 0&C\\ 0&I_2 \end{bmatrix} = \begin{bmatrix} A\cdot 0+0\cdot 0 & A\cdot C+0\cdot I_2\\ 0\cdot 0+B\cdot 0 & 0\cdot C+B\cdot I_2 \end{bmatrix} = \begin{bmatrix} 0&0&0&3\\ 0&0&2&7\\ 0&0&1&0 \end{bmatrix}. \qquad \square$$

For another (important!) example of block arithmetic, examine Example 2.9 and the discussion following it. There we view a matrix as blocked into its respective columns, and a column vector as blocked into its rows, to obtain

$$A\mathbf{x} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] \begin{bmatrix} x_1\\ x_2\\ x_3 \end{bmatrix} = \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3.$$

Transpose of a Matrix

Sometimes we prefer to work with a different form of a matrix that contains the same information as the matrix. Transposes are operations that allow us to do that. The idea is simple: Interchange rows and columns. It turns out that for complex matrices, there is an analogue that is not quite the same thing as transposing, though it yields the same result when applied to real matrices. This analogue is called the conjugate (or Hermitian) transpose. Here are the appropriate definitions.

Definition 2.15. **Transpose and Conjugate Matrices** Let $A = [a_{ij}]$ be an $m \times n$ matrix with (possibly) complex entries. Then the *transpose* of A is the $n \times m$ matrix A^T obtained by interchanging the rows and columns of A, so that the (i, j)th entry of A^T is a_{ji}. The *conjugate* of A is the matrix $\overline{A} = [\overline{a_{ij}}]$. Finally, the *conjugate (Hermitian) transpose* of A is the matrix $A^* = \overline{A}^T$.

Notice that in the case of a real matrix (that is, a matrix with real entries) A there is no difference between transpose and conjugate transpose, since in this case $A = \overline{A}$. Consider these examples.

Example 2.33. Compute the transpose and conjugate transpose of the following matrices:

$$\text{(a)} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \end{bmatrix}, \quad \text{(b)} \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}, \quad \text{(c)} \begin{bmatrix} 1 & 1+\mathrm{i} \\ 0 & 2\mathrm{i} \end{bmatrix}.$$

Solution. For matrix (a) we have

$$\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \end{bmatrix}^* = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 1 \end{bmatrix}.$$

Notice how the dimensions of a transpose get switched from the original.

For matrix (b) we have

$$\begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}^* = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}^T = \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix},$$

and for matrix (c) we have

$$\begin{bmatrix} 1 & 1+\mathrm{i} \\ 0 & 2\mathrm{i} \end{bmatrix}^* = \begin{bmatrix} 1 & 0 \\ 1-\mathrm{i} & -2\mathrm{i} \end{bmatrix}, \quad \begin{bmatrix} 1 & 1+\mathrm{i} \\ 0 & 2\mathrm{i} \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \\ 1+\mathrm{i} & 2\mathrm{i} \end{bmatrix}.$$

In this case, transpose and conjugate transpose are not the same. □

Even when dealing with vectors alone, the transpose notation is handy. For example, there is a bit of terminology that comes from tensor analysis (a branch of higher linear algebra used in many fields including differential geometry, engineering mechanics, and relativity) that can be expressed very concisely with transposes:

Definition 2.16. **Inner and Outer Products** Let $\mathbf{u}$ and $\mathbf{v}$ be column vectors of the same size, say $n \times 1$. Then the *inner product* of $\mathbf{u}$ and $\mathbf{v}$ is the scalar quantity $\mathbf{u}^T\mathbf{v}$, and the *outer product* of $\mathbf{u}$ and $\mathbf{v}$ is the $n \times n$ matrix $\mathbf{u}\mathbf{v}^T$.

Example 2.34. Compute the inner and outer products of the vectors

$$\mathbf{u} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix}.$$

Solution. Here we have the inner product

$$\mathbf{u}^T\mathbf{v} = [2, -1, 1] \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix} = 2 \cdot 3 + (-1)4 + 1 \cdot 1 = 3,$$

while the outer product is

$$\mathbf{u}\mathbf{v}^T = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} [3,4,1] = \begin{bmatrix} 2\cdot 3 & 2\cdot 4 & 2\cdot 1 \\ -1\cdot 3 & -1\cdot 4 & -1\cdot 1 \\ 1\cdot 3 & 1\cdot 4 & 1\cdot 1 \end{bmatrix} = \begin{bmatrix} 6 & 8 & 2 \\ -3 & -4 & -1 \\ 3 & 4 & 1 \end{bmatrix}. \qquad \square$$

Here are a few basic laws relating transposes to other matrix arithmetic that we have learned. These laws remain correct if transpose is replaced by conjugate transpose, with one exception: $(cA)^* = \overline{c}A^*$.

Laws of Matrix Transpose

Let A and B be matrices of the appropriate sizes so that the following operations make sense, and c a scalar.

(1) $(A+B)^T = A^T + B^T$
(2) $(AB)^T = B^T A^T$
(3) $(cA)^T = cA^T$
(4) $(A^T)^T = A$

These laws are easily verified directly from definition. For example, if $A = [a_{ij}]$ and $B = [b_{ij}]$ are $m \times n$ matrices, then we have that $(A+B)^T$ is the $n \times m$ matrix

$$\begin{aligned}(A+B)^T &= [a_{ij} + b_{ij}]^T = [a_{ji} + b_{ji}] \\ &= [a_{ji}] + [b_{ji}] = A^T + B^T.\end{aligned}$$

The other laws are proved similarly.

Transposes of Elementary Matrices

We will require explicit formulas for transposes of the elementary matrices in some later calculations. Notice that the matrix $E_{ij}(c)$ is a matrix with 1's on the diagonal and 0's elsewhere, except that the (i,j)th entry is c. Therefore, transposing switches the entry c to the (j,i)th position and leaves all other entries unchanged. Hence, $E_{ij}(c)^T = E_{ji}(c)$. With similar calculations we have these facts:

- $E_{ij}^T = E_{ij}$
- $E_i(c)^T = E_i(c)$
- $E_{ij}(c)^T = E_{ji}(c)$

Elementary Column Operations

These formulas have an interesting application. Up to this point we have considered only elementary row operations. However, there are situations in which elementary *column* operations on the columns of a matrix are useful. If we want to use such operations, do we have to start over, reinvent elementary column matrices, and so forth? The answer is no and the following example gives an indication of why the transpose idea is useful. This example shows how to do column operations in the language of matrix arithmetic.

Here's the basic idea: Suppose we want to do an elementary column operation on a matrix A corresponding to elementary row operation E to get a new matrix B from A. To do this, turn the columns of A into rows, do the row operation, and then transpose the result back to get the matrix B that we want. In algebraic terms

$$B = \left(EA^T\right)^T = \left(A^T\right)^T E^T = AE^T.$$

So all we have to do to perform an elementary column operation is multiply by the transpose of the corresponding elementary row matrix on the right. Thus, we see that the transposes of elementary row matrices could reasonably be called *elementary column matrices.*

Elementary Column Matrix

Example 2.35. Let A be a matrix. Suppose that we wish to express the result B of swapping the second and third columns of A, followed by adding -2 times the first column to the second, as a product of matrices. How can this be done? Illustrate the procedure with the matrix

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix}.$$

Solution. Apply the preceding remark twice to obtain that

$$B = AE_{23}^T E_{21}\left(-2\right)^T = AE_{23}E_{12}\left(-2\right).$$

Thus, we have

$$B = \begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -3 & 2 \\ 1 & 0 & -1 \end{bmatrix}$$

as a matrix product. □

A very important type of special matrix is one that is invariant under the operation of transposing. It turns out that these matrices are fundamental in certain applications and they have some very remarkable properties that we will study in Chapters 4, 5, and 6.

Definition 2.17. **Symmetric and Hermitian Matrices** The matrix A is said to be *symmetric* if $A^T = A$ and *Hermitian* if $A^* = A$. (Equivalently, $a_{ij} = a_{ji}$ and $a_{ij} = \overline{a_{ji}}$, for all i, j, respectively.)

From the laws of transposing elementary matrices above we see right away that E_{ij} and $E_i(c)$ supply us with examples of symmetric matrices. Also the adjacency matrix of a graph is always symmetric, unlike those of digraphs. Here are a few more examples.

Example 2.36. Are the following matrices symmetric or Hermitian?

$$\text{(a)}\ \begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & 2 \end{bmatrix}, \quad \text{(b)}\ \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}, \quad \text{(c)}\ \begin{bmatrix} 1 & 1+\mathrm{i} \\ 1+\mathrm{i} & 2\mathrm{i} \end{bmatrix}$$

Solution. For matrix (a) we have

$$\begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & 2 \end{bmatrix}^* = \begin{bmatrix} 1 & \overline{1+\mathrm{i}} \\ \overline{1-\mathrm{i}} & 2 \end{bmatrix}^T = \begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & 2 \end{bmatrix}.$$

Hence, this matrix is Hermitian. However, it is *not* symmetric since the $(1,2)$th and $(2,1)$th entries differ. Matrix (b) is easily seen to be symmetric by inspection and Hermitian as well. Matrix (c) is symmetric since the $(1,2)$th and $(2,1)$th entries agree, but it is not Hermitian since

$$\begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & 2\mathrm{i} \end{bmatrix}^* = \begin{bmatrix} 1 & \overline{1+\mathrm{i}} \\ \overline{1-\mathrm{i}} & \overline{2\mathrm{i}} \end{bmatrix}^T = \begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & -2\mathrm{i} \end{bmatrix},$$

and this last matrix is clearly not equal to matrix (c). □

Example 2.37. Consider the quadratic form (this means a homogeneous second-degree polynomial in its variables)

$$Q(x,y,z) = x^2 + 2y^2 + z^2 + 2xy + yz + 3xz.$$

Express this function in terms of matrix products and transposes.

Solution. Write the quadratic form as

$$\begin{aligned} x(x+2y+3z) + y(2y+z) + z^2 &= \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} x+2y+3z \\ 2y+z \\ z \end{bmatrix} \\ &= \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{x}^T A \mathbf{x}, \end{aligned}$$

where

$$\mathbf{x} = (x,y,z) \text{ and } A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

□

Rank of the Matrix Transpose

A basic question is how the rank of a matrix transpose (or Hermitian transpose) is connected to the rank of the matrix. There is a nice answer. We will focus on transposes. First we need the following theorem.

Theorem 2.4. Let A, B be matrices such that the product AB is defined. Then

$$\operatorname{rank} AB \leq \operatorname{rank} A.$$

Proof. Let E be a product of elementary matrices such that $EA = R$, where R is the reduced row echelon form of A. If $\operatorname{rank} A = r$, then the first r rows of R have leading entries of 1, while the remaining rows are zero rows. Also, we saw in Chapter 1 that elementary row operations do not change the rank of a matrix, since according to Corollary 1.1 they do not change the reduced row echelon form of a matrix. Therefore,

$$\operatorname{rank} AB = \operatorname{rank} E(AB) = \operatorname{rank}(EA)B = \operatorname{rank} RB.$$

Now the matrix RB has the same number of rows as R, and the first r of these rows may or may not be nonzero, but the remaining rows must be zero rows, since they result from multiplying columns of B by the zero rows of R. If we perform elementary row operations to reduce RB to its reduced row echelon form we will possibly introduce more zero rows than R has. Consequently, $\operatorname{rank} RB \leq r = \operatorname{rank} A$, which completes the proof. □

Theorem 2.5. Rank Invariant Under Transpose For any matrix A,

$$\operatorname{rank} A = \operatorname{rank} A^T.$$

Proof. As in the previous theorem, let E be a product of elementary matrices such that $EA = R$, where R is the reduced row echelon form of A. If $\operatorname{rank} A = r$, then the first r rows of R have leading entries of 1 whose column numbers form an increasing sequence, while the remaining rows are zero rows. Therefore, R^T is a matrix whose columns have leading entries of 1 and whose row numbers form an increasing sequence. Use elementary row operations to clear out the nonzero entries below each column with a leading 1 to obtain a matrix whose rank is equal to the number of such leading entries, i.e., equal to r. Thus, $\operatorname{rank} R^T = r$.

From Theorem 2.4 we have that $\operatorname{rank} A^T E^T \leq \operatorname{rank} A^T$. It follows that

$$\operatorname{rank} A = \operatorname{rank} R^T = \operatorname{rank} A^T E^T \leq \operatorname{rank} A^T.$$

Substitute the matrix A^T for the matrix A in this inequality, to obtain that

$$\operatorname{rank} A^T \leq \operatorname{rank}(A^T)^T = \operatorname{rank} A.$$

It follows from these two inequalities that $\operatorname{rank} A = \operatorname{rank} A^T$. □

It is instructive to see how a specific example might work out in the preceding proof. For example, R might look like this, where an x designates an arbitrary entry:

$$R = \begin{bmatrix} 1 & 0 & x & 0 & x \\ 0 & 1 & x & 0 & x \\ 0 & 0 & 0 & 1 & x \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

so that R^T is

$$R^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ x & x & 0 & 0 \\ 0 & 0 & 1 & 0 \\ x & x & x & 0 \end{bmatrix}.$$

Thus, if we use elementary row operations to zero out the entries below a column pivot, all entries to the right and below this pivot are unaffected by these operations. Now start with the leftmost column and proceed to the right, zeroing out all entries under each column pivot. The result is a matrix that looks like

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Now swap rows to move the zero rows to the bottom if necessary, and we see that the reduced row echelon form of R^T has exactly as many nonzero rows as did R, that is, r nonzero rows.

A first application of this important fact is to give a fuller picture of the rank of a product of matrices than that given by Theorem 2.4:

Corollary 2.3. Rank of Matrix Product If the product AB is defined, then

$$\operatorname{rank} AB \leq \min\{\operatorname{rank} A, \operatorname{rank} B\}.$$

Proof. We know from Theorem 2.4 that

$$\operatorname{rank} AB \leq \operatorname{rank} A \text{ and } \operatorname{rank} B^T A^T \leq \operatorname{rank} B^T.$$

Since $B^T A^T = (AB)^T$, Theorem 2.5 tells us that

$$\operatorname{rank} B^T A^T = \operatorname{rank} AB \text{ and } \operatorname{rank} B^T = \operatorname{rank} B.$$

Put all this together, and we have

$$\operatorname{rank} AB = \operatorname{rank} B^T A^T \leq \operatorname{rank} B^T = \operatorname{rank} B.$$

It follows that $\operatorname{rank} AB$ is at most the smaller of $\operatorname{rank} A$ and $\operatorname{rank} B$, which is what the corollary asserts. □

Another useful application of this result sheds some light on certain kinds of matrix inverses that are discussed in the next section.

Corollary 2.4. Let A be an $m \times n$ matrix. If there exists a matrix B such that $AB = I_m$, then $m \leq n$ and $\operatorname{rank} A = m$; if there exists a matrix B such that $BA = I_n$, then $n \leq m$ and $\operatorname{rank} A = n$.

Proof. (Note that if A is $m \times n$ and $AB = I$, then a size check shows that B is $n \times m$ and we must have $I = I_m$.) From Corollary 2.3 we obtain that

$$\operatorname{rank} I_m = m = \operatorname{rank} AB \leq \min\{\operatorname{rank} A, \operatorname{rank} B\} \leq \operatorname{rank} A \leq \min\{m, n\} \leq n,$$

from which the first statement follows. For the second, note that if $BA = I$, then $(BA)^T = A^T B^T = I^T = I$, and since $\operatorname{rank} A = \operatorname{rank} A^T$ by Theorem 2.5, the result follows from the first statement by interchanging the roles of m and n. □

2.4 Exercises and Problems

Exercise 1. Convert the following 3×3 elementary operations to matrix form and convert matrices to elementary operation form.

(a) $E_{23}(3)$ (b) E_{13} (c) $E_3(2)$ (d) $E_{23}^T(-1)$

(e) $\begin{bmatrix} 1 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (f) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -a & 0 & 1 \end{bmatrix}$ (g) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (h) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$

Exercise 2. Convert the following 4×4 elementary operations to matrix form and convert matrices to elementary operation form.

(a) E_{24}^T (b) $E_4(-1)$ (c) $E_3^T(2)$ (d) $E_{14}(-1)$

(e) $\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$ (g) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{bmatrix}$

Exercise 3. Describe the effect of multiplying the 3×3 matrix A by each matrix in Exercise 1 on the left.

Exercise 4. Describe the effect of multiplying the 4×4 matrix A by each matrix in Exercise 2 on the right.

Exercise 5. Compute the reduced row echelon form of the following matrices and express each form as a product of elementary matrices and the original matrix.

(a) $\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & -2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1+\mathrm{i} & \mathrm{i} \\ 1 & 0 & -2 \end{bmatrix}$

Exercise 6. Compute the reduced row echelon form of the following matrices and express each form as a product of elementary matrices and the original matrix.

(a) $\begin{bmatrix} 2 & 1 \\ 0 & 1 \\ 0 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 \\ 1 & 1+\mathrm{i} \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 2 & 0 & 2 \\ 1 & 1 & -4 & 3 \end{bmatrix}$

Exercise 7. Identify a complete but minimal list of simple structure descriptions that apply to these matrices (e.g., if "upper triangular," omit "triangular.")

(a) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 & 4 & 2 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (c) I_3 (d) $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ (e) $\begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix}$

Exercise 8. Identify the minimal list of simple structure descriptions that apply to these matrices.

(a) $\begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}$

Exercise 9. Identify the appropriate blocking and calculate the matrix product AB using block multiplication, where

$$A = \begin{bmatrix} 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 4 & 1 & 2 & 1 & 3 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 2 \\ 2 & 2 & -1 & 1 \\ 1 & 1 & 3 & 2 \end{bmatrix},$$

and as many submatrices that form scalar matrices or zero matrices are blocked out as possible.

Exercise 10. Confirm that sizes are correct for block multiplication and calculate the matrix product AB, where $A = \begin{bmatrix} R & 0 \\ S & T \end{bmatrix}$, $B = \begin{bmatrix} U \\ V \end{bmatrix}$, $R = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$, $S = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix}$, $T = \begin{bmatrix} 1 & -1 \\ 2 & 2 \end{bmatrix}$, $U = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}$, and $V = \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}$.

Exercise 11. Express the matrix $\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 2 & 4 & 2 \end{bmatrix}$ as an outer product of two vectors.

Exercise 12. Express the rank-two matrix $\begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix}$ as the sum of two outer products of vectors.

Exercise 13. Compute the transpose and conjugate transpose of the following matrices and determine which are symmetric or Hermitian.

(a) $\begin{bmatrix} 1 & -3 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 \\ 0 & 3 \\ 1 & -4 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & \mathrm{i} \\ -\mathrm{i} & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1 & 3 \\ 1 & 0 & 0 \\ 3 & 0 & 2 \end{bmatrix}$

Exercise 14. Determine which of the following matrices are symmetric or Hermitian.

(a) $\begin{bmatrix} 1 & -3 & 2 \\ -3 & 0 & 0 \\ 2 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} \mathrm{i} & 1 \\ -1 & \mathrm{i} \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

Exercise 15. Answer True/False.
(a) $E_{ij}\,(c)^T = E_{ji}\,(c)$.
(b) Every elementary matrix is symmetric.
(c) The rank of the matrix A may differ from the rank of A^T.
(d) Every real diagonal matrix is Hermitian.
(e) For matrix A and scalar c, $(cA)^* = cA^*$.

Exercise 16. Answer True/False and give reasons.
(a) For matrices A, B, $(AB)^T = B^T A^T$.
(b) Every diagonal matrix is symmetric.
(c) $\operatorname{rank}(AB) = \min\{\operatorname{rank} A, \operatorname{rank} B\}$.
(d) Every diagonal matrix is Hermitian.
(e) Every tridiagonal matrix is symmetric.

Exercise 17. Express the quadratic form $Q(x, y, z) = 2x^2 + y^2 + z^2 + 2xy + 4yz - 6xz$ in the matrix form $\mathbf{x}^T A\mathbf{x}$, where A has as few nonzero entries as possible.

Exercise 18. Express the quadratic form $Q(x, y, z) = x^2 + y^2 - z^2 + 4yz - 6xz$ in the matrix form $\mathbf{x}^T A\mathbf{x}$, where A is a lower triangular matrix.

Exercise 19. Let $A = \begin{bmatrix} -2 & 1 - 2\mathrm{i} \\ 0 & 3 \end{bmatrix}$ and verify that both A^*A and AA^* are Hermitian.

Exercise 20. A square matrix A is called *normal* if $A^*A = AA^*$. Determine which of the following matrices are normal.

(a) $\begin{bmatrix} 2 & \mathrm{i} \\ 1 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & \mathrm{i} \\ 1 & 2+\mathrm{i} \end{bmatrix}$ (d) $\begin{bmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{bmatrix}$

Exercise 21. Let A be an $m \times p$ matrix and $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]$ a $p \times n$ matrix with columns $\mathbf{b}_j$. Justify the block multiplication $AB = [A\mathbf{b}_1, A\mathbf{b}_2, \ldots, A\mathbf{b}_n]$ and illustrate it in the case that $A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 0 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 1 \\ 0 & 3 \\ 1 & -4 \end{bmatrix}$.

Exercise 22. Let $A = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]$ be an $m \times n$ matrix and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ an $n \times 1$ vector. Justify the block multiplication $A\mathbf{b} = Ab_1 + \ldots Ab_n$ and illustrate it in the case that $A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 0 & 4 \end{bmatrix}$ and $\mathbf{b} = (1, -1, 2)$.

Problem 23. Show that a matrix is diagonal if it is both triangular and symmetric.

*Problem 24. Let A and C be square matrices and suppose that the matrix $M = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$ is in block form. Show that for some matrix D, $M^2 = \begin{bmatrix} A^2 & D \\ 0 & C^2 \end{bmatrix}$.

Problem 25. Show that if $C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ in block form, then $\operatorname{rank} C = \operatorname{rank} A + \operatorname{rank} B$.

Problem 26. Prove from the definition that $(A^T)^T = A$.

Problem 27. Let A be an $m \times n$ matrix. Show that both A^*A and AA^* are Hermitian.

Problem 28. Use Corollary 2.3 to prove that the outer product of any two vectors is either a rank-one matrix or zero.

*Problem 29. Show that if P and Q are stochastic matrices of the same size, then PQ is also stochastic.

Problem 30. Let A be a square real matrix. Show the following.
(a) The matrix $B = \frac{1}{2}\left(A + A^T\right)$ is symmetric.
(b) The matrix $C = \frac{1}{2}\left(A - A^T\right)$ is skew-symmetric (a matrix C is *skew-symmetric* if $C^T = -C$.)
(c) The matrix A can be expressed as the sum of a symmetric matrix and a skew-symmetric matrix.
(d) With B and C as in parts (a) and (b), show that for any vector $\mathbf{x}$ of conformable size, $\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T B\mathbf{x}$.
(e) Express $A = \begin{bmatrix} 2 & 2 & -6 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$ as a sum of a symmetric and a skew-symmetric matrix.

Problem 31. Find all 2×2 idempotent upper triangular matrices A (idempotent means $A^2 = A$).

*Problem 32. Let D be a diagonal matrix with distinct entries on the diagonal and B any other matrix of the same size. Show that $DB = BD$ if and only if B is diagonal.

Problem 33. Show that if U is an $n \times n$ strictly upper triangular matrix, then $U^n = 0$, so U is nilpotent. (It might help to see what happens in a 2×2 and a 3×3 case first.)

Problem 34. Use Problem 30 to show that every quadratic form $Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ defined by matrix A can be defined by a symmetric matrix $B = (A + A^T)/2$ as well. Apply this result to the matrix of Example 2.37.

*Problem 35. Suppose that $A = B + C$, where B is a symmetric matrix and C is a skew-symmetric matrix. Show that $B = \frac{1}{2}(A + A^T)$ and $C = \frac{1}{2}(A - A^T)$.

*Problem 36. The digraph H that results from reversing all the arrows in a digraph G is called *reverse digraph* of G. Show that if A is the adjacency matrix for G then A^T is the adjacency matrix for the reverse digraph H.

2.5 Matrix Inverses

Definitions

We have seen that if we could make sense of "$1/A$," then we could write the solution to the linear system $A\mathbf{x} = \mathbf{b}$ as simply $\mathbf{x} = (1/A)\mathbf{b}$. We are going to tackle this problem now. First, we need a definition of the object that we are trying to uncover. Notice that "inverses" could work only on one side. For example,

$$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = [1] = \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

which suggests that both $\begin{bmatrix} 1 & 2 \end{bmatrix}$ and $\begin{bmatrix} 2 & 3 \end{bmatrix}$ should qualify as left inverses of the matrix $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$, since multiplication on the left by them results in a 1×1 identity matrix. As a matter of fact, right and left inverses are studied and do have applications. But they have some unusual properties such as non-uniqueness. We are going to focus on a type of inverse that is closer to the familiar inverses in fields of numbers, namely, *two-sided* inverses. These make sense only for square matrices, so the non-square example above is ruled out.

Definition 2.18. **Invertible Matrix** Let A be a square matrix. Then a *(two-sided) inverse* for A is a square matrix B of the same size as A such that $AB = I = BA$. If such a B exists, then the matrix A is said to be *invertible*.

Of course, any nonsquare matrix is non-invertible. Square matrices are classified as either "*singular*" (non-invertible), or "*nonsingular*" (invertible).

Nonsingular Matrix

Since we will mostly be concerned with two-sided inverses, the unqualified term "inverse" will be understood to mean a "two-sided inverse." Notice that this definition is actually symmetric in A and B. In other words, if B is an inverse for A, then A is an inverse for B.

Examples of Inverses

Example 2.38. Show that $B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ is an inverse for $A = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$.

Solution. All we have to do is check the definition. But remember that there are *two* multiplications to confirm. (We'll show later that this isn't necessary, but right now we are working strictly from the definition.) We have

$$AB = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 - 1 \cdot 1 & 2 \cdot 1 - 1 \cdot 2 \\ -1 \cdot 1 + 1 \cdot 1 & -1 \cdot 1 + 1 \cdot 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

and similarly

$$BA = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 1 \cdot (-1) & 1 \cdot (-1) + 1 \cdot 1 \\ 1 \cdot 2 + 2 \cdot (-1) & 1 \cdot (-1) + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I.$$

Therefore, the definition for inverse is satisfied, so that A and B work as inverses to each other. □

Of course not every square matrix is invertible: Consider, e.g., zero matrices. However it is sometimes not entirely obvious why a matrix should not be invertible.

Example 2.39. Show that the matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ cannot have an inverse.

Solution. How do we get our hands on a "non-inverse"? We try an indirect approach. If A had an inverse B, then we could always find a solution to the linear system $A\mathbf{x} = \mathbf{b}$ by multiplying each side on the left by B to obtain that $BA\mathbf{x} = I\mathbf{x} = \mathbf{x} = B\mathbf{b}$, *no matter what right-hand-side vector* $\mathbf{b}$ we used. Yet it is easy to come up with right-hand-side vectors for which the system has no solution, e.g., try $\mathbf{b} = (1, 2)$. Since the resulting system is clearly inconsistent, there cannot be an inverse matrix B, which is what we wanted to show. □

One moral of this last example is that it is not enough for every entry of a matrix to be nonzero for the matrix itself to be invertible. Our next example produces a gold mine of invertible matrices, namely any elementary matrix we can construct.

Example 2.40. Find formulas for inverses of all the elementary matrices.

Solution. Recall from Corollary 2.2 that left multiplication by an elementary matrix is the same as performing the corresponding elementary

Elementary Matrix Inverses

row operation. Furthermore, from the discussion following Theorem 1.2 we see the following:

- E_{ij}: The elementary operation of switching the ith and jth rows is undone by applying E_{ij}. Hence,

$$E_{ij}E_{ij} = E_{ij}E_{ij}I = I,$$

  so that E_{ij} works as its own inverse. (This is rather like -1, since $(-1)\cdot(-1) = 1$.)
- $E_i(c)$: The elementary operation of multiplying the ith row by the nonzero constant c, is undone by applying $E_i(1/c)$. Hence,

$$E_i(1/c)E_i(c) = E_i(1/c)E_i(c)I = I.$$

- $E_{ij}(d)$: The elementary operation of adding d times the jth row to the ith row is undone by applying $E_{ij}(-d)$. Hence,

$$E_{ij}(-d)E_{ij}(d) = E_{ij}(-d)E_{ij}(d)I = I. \qquad \square$$

Example 2.41. Show that if D is a diagonal matrix with nonzero diagonal entries, then D is invertible.

Solution. Suppose that

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}.$$

Diagonal Matrix Shorthand and Inverse

For a convenient shorthand, we write $D = \operatorname{diag}\{d_1, d_2, \ldots, d_n\}$. It is easily checked that if $E = \operatorname{diag}\{e_1, e_2, \ldots, e_n\}$, then

$$DE = \operatorname{diag}\{d_1e_1, d_2e_2, \ldots, d_ne_n\} = \operatorname{diag}\{e_1d_1, e_2d_2, \ldots, e_nd_n\} = ED.$$

Therefore, if $d_i \neq 0$, for $i = 1, \ldots, n$, then

$$\operatorname{diag}\{d_1, d_2, \ldots, d_n\} \operatorname{diag}\{1/d_1, 1/d_2, \ldots, 1/d_n\} = \operatorname{diag}\{1, 1, \ldots, 1\} = I_n,$$

which shows that $\operatorname{diag}\{1/d_1, 1/d_2, \ldots, 1/d_n\}$ is an inverse of D. $\square$

Laws of Inverses

Here are some of the basic laws of inverse calculations.

Laws of Matrix Inverses

Let A, B, C be matrices of the appropriate sizes so that the following multiplications make sense, I a suitably sized identity matrix, and c a nonzero scalar. Then

(1) (Uniqueness) If the matrix A is invertible, then it has only one inverse, which is denoted by A^{-1}.
(2) (Double Inverse) If A is invertible, then $\left(A^{-1}\right)^{-1} = A$.
(3) (2/3 Rule) If any two of the three matrices A, B, and AB are invertible, then so is the third, and moreover, $(AB)^{-1} = B^{-1}A^{-1}$.
(4) If A is invertible and $c \neq 0$, then $(cA)^{-1} = (1/c)A^{-1}$.
(5) (Inverse/Transpose) If A is invertible, then $(A^T)^{-1} = (A^{-1})^T$ and $(A^*)^{-1} = (A^{-1})^*$.
(6) (Cancellation) Suppose A is invertible. If $AB = AC$ or $BA = CA$, then $B = C$.
(7) (Rank) If A is invertible, then $\operatorname{rank} A = n$ and the reduced row echelon form of A is I_n.

Note 2.2. Observe that the 2/3 Rule reverses order when taking the inverse of a product. This should remind you of the operation of transposing a product. A common mistake is to forget to reverse the order. Secondly, notice that the cancellation law restores something that appeared to be lost when we first discussed matrices. Yes, we can cancel a common factor from both sides of an equation, but (1) the factor must be on the same side and (2) the factor must be an invertible matrix.

Verification of Laws: Suppose that both B and C work as inverses to the matrix A. We will show that these matrices must be identical. The associative and identity laws of matrices yield

$$B = BI = B(AC) = (BA)C = IC = C.$$

Henceforth, we shall write A^{-1} for the unique (two-sided) inverse of the square matrix A, provided of course that there is an inverse at all (remember that existence of inverses is not a sure thing).

Matrix Inverse Notation

The double inverse law is a matter of examining the definition of inverse:

$$AA^{-1} = I = A^{-1}A$$

shows that A is an inverse matrix for A^{-1}. Hence, $(A^{-1})^{-1} = A$.

Now suppose that A and B are both invertible and of the same size. Using the laws of matrix arithmetic, we see that

$$AB(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$$

and that

$$(B^{-1}A^{-1})AB = B^{-1}(A^{-1}A)B = B^{-1}IB = B^{-1}B = I.$$

In other words, the matrix $B^{-1}A^{-1}$ works as an inverse for the matrix AB, which is what we wanted to show. We leave the remaining cases of the 2/3 Rule as an exercise.

Suppose that c is nonzero and perform the calculation

$$(cA)(1/c)A^{-1} = (c/c)AA^{-1} = 1 \cdot I = I.$$

A similar calculation on the other side shows that $(cA)^{-1} = (1/c)A^{-1}$.

Next, apply the transpose operator to the definition of inverse ((2.18)) and use the law of transpose products to obtain that

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

This shows that the definition of inverse is satisfied for $(A^{-1})^T$ relative to A^T, that is, that $(A^T)^{-1} = (A^{-1})^T$, which is the inverse/transpose law. The same argument works with conjugate transpose in place of transpose.

Next, if A is invertible and $AB = AC$, then multiply both sides of this equation on the left by A^{-1} to obtain that

$$A^{-1}(AB) = (A^{-1}A)B = B = A^{-1}(AC) = (A^{-1}A)C = C,$$

which is the cancellation that we want.

Finally, if A is $n \times n$ invertible with inverse B, then $AB = I_n$, so $n = \operatorname{rank} I_n \leq \operatorname{rank} A \leq n$, with the first inequality due to Theorem 2.4 and the second due to the size of A (Theorem 1.4). Hence, $\operatorname{rank} A = n$; but the only $n \times n$ full rank reduced row echelon form is I_n, so that must be the reduced row echelon form of A. □

We can now extend the power notation to negative exponents. Let A be an invertible matrix and k a positive integer. Then we write

Negative Matrix Power

$$A^{-k} = A^{-1}A^{-1}\cdots A^{-1},$$

where the product is taken over k terms.

The laws of exponents that we saw earlier can now be expressed for arbitrary integers, *provided* that A is invertible. Here is an example of how we can use the various laws of arithmetic and inverses to carry out an inverse calculation.

Example 2.42. Let

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Show that $(I - A)^3 = 0$ and use this to find A^{-1}.

Solution. First we calculate that

$$(I-A)=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}-\begin{bmatrix}1&2&0\\0&1&1\\0&0&1\end{bmatrix}=\begin{bmatrix}0&-2&0\\0&0&-1\\0&0&0\end{bmatrix}$$

and check that

$$\begin{aligned}(I-A)^3&=\begin{bmatrix}0&-2&0\\0&0&-1\\0&0&0\end{bmatrix}\begin{bmatrix}0&-2&0\\0&0&-1\\0&0&0\end{bmatrix}\begin{bmatrix}0&-2&0\\0&0&-1\\0&0&0\end{bmatrix}\\&=\begin{bmatrix}0&0&2\\0&0&0\\0&0&0\end{bmatrix}\begin{bmatrix}0&-2&0\\0&0&-1\\0&0&0\end{bmatrix}=\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\end{bmatrix}.\end{aligned}$$

Next we do some symbolic algebra, using the laws of matrix arithmetic:

$$0=(I-A)^3=(I-A)(I^2-2A+A^2)=I-3A+3A^2-A^3.$$

Subtract all terms involving A from both sides to obtain that

$$3A-3A^2+A^3=A\cdot 3I-3A^2+A^3=A(3I-3A+A^2)=I.$$

Since $A(3I-3A+A^2)=(3I-3A+A^2)A$, we see from definition of inverse that

$$A^{-1}=3I-3A+A^2=\begin{bmatrix}1&-2&2\\0&1&-1\\0&0&1\end{bmatrix}.$$ □

Notice that in the preceding example we were careful not to leave a "3" behind when we factored out A from $3A$. The reason is that $3+3A+A^2$ makes no sense as a sum, since one term is a scalar and the other two are matrices.

Rank and Inverse Calculation

Although we can calculate a few examples of inverses such as the last example, we really need a general procedure. So let's get right to the heart of the matter. How can we find the inverse of a matrix, or decide that none exists? Actually, we already have done all the hard work necessary to understand computing inverses. The secret is in the notions of reduced row echelon form and rank. (Remember, we use elementary row operations to reduce a matrix to its reduced row echelon form. Once we have done so, the rank of the matrix is simply the number of nonzero rows in the reduced row echelon form.) Let's recall the results of Example 2.30:

$$\begin{bmatrix}1&0&2\\0&1&3\end{bmatrix}=E_{12}(-1)E_2(-1/3)E_{21}(-2)E_1(1/4)E_{12}\begin{bmatrix}2&-1&1\\4&4&20\end{bmatrix}.$$

Now remove the last column from each of the matrices at the right of each side and we have this result:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = E_{12}(-1)E_2(-1/3)E_{21}(-2)E_1(1/4)E_{12}\begin{bmatrix} 2 & -1 \\ 4 & 4 \end{bmatrix}.$$

This suggests that if $A = \begin{bmatrix} 2 & -1 \\ 4 & 4 \end{bmatrix}$, then

$$A^{-1} = E_{12}(-1)E_2(-1/3)E_{21}(-2)E_1(1/4)E_{12}.$$

To prove this, we argue in the general case as follows: Let A be an $n \times n$ matrix and suppose that by a succession of elementary row operations $E_1, E_2, \ldots, E_k$, we reduce A to its reduced row echelon form R, which happens to be I. In the language of matrix multiplication, what we have obtained is

$$I = E_k E_{k-1} \cdots E_1 A.$$

Now let $B = E_k E_{k-1} \cdots E_1$. By repeated application of the 2/3 rule, we see that a product of any number of invertible matrices is invertible. Since each elementary matrix is invertible, it follows that B is. Multiply both sides of the equation $I = BA$ by B^{-1} to obtain that $B^{-1}I = B^{-1} = B^{-1}BA = A$. Therefore, A is the inverse of the matrix B, hence is itself invertible.

A practical trick for storing this product of elementary matrices on the fly:

Superaugmented Matrix

Form what we term the *superaugmented matrix* $[A \mid I]$. If we perform the elementary operation E on the superaugmented matrix, we have the same result as

$$E[A \mid I] = [EA \mid EI] = [EA \mid E].$$

So the matrix occupied by the I part of the superaugmented matrix is just the product of the elementary matrices that we have used so far. Now continue applying elementary row operations until the part of the matrix originally occupied by A is reduced to the reduced row echelon form of A. We end up with this schematic picture of our calculations:

$$\left[\, A \mid I \,\right] \xrightarrow{E_1, E_2, \ldots, E_k} \left[\, I \mid B \,\right],$$

where $B = E_k E_{k-1} \cdots E_1$ is the product of the various elementary matrices we used, composed in the correct order of usage. We can summarize this discussion with the following algorithm:

Inverse Algorithm

Given an $n \times n$ matrix A, to compute A^{-1}:

(1) Form the superaugmented matrix $\widetilde{A} = [A \mid I_n]$.
(2) Reduce the first n columns of $\widetilde{A}$ to reduced row echelon form by performing elementary operations on the matrix $\widetilde{A}$ resulting in the matrix $[R \mid B]$.
(3) If $R = I_n$ then set $A^{-1} = B$; otherwise, A is singular and A^{-1} does not exist.

Example 2.43. Use the inverse algorithm to compute the inverse of Example 2.8,

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Solution. Notice that this matrix is already upper triangular. Therefore, as in Gaussian elimination, it is a bit more efficient to start with the bottom pivot and clear out entries above in reverse order. So we compute

$$[A \mid I_3] = \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow{E_{23}(-1)} \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow{E_{1,2}(-2)} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & -2 & 2 \\ 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right].$$

We conclude that A is indeed invertible and

$$A^{-1} = \begin{bmatrix} 1 & -2 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}. \qquad \square$$

There is a simple formula for the inverse of a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Set $D = ad - bc$. It is easy to verify that if $D \neq 0$, then

Two by Two Inverse

$$A^{-1} = \frac{1}{D} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Example 2.44. Use the 2×2 inverse formula to find the inverse of the matrix $A = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix}$, and verify that the same answer results if we use the inverse algorithm.

Solution. First we apply the inverse algorithm:

$$\left[\begin{array}{cc|cc} 1 & -1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{array}\right] \xrightarrow{E_{21}(-1)} \left[\begin{array}{cc|cc} 1 & -1 & 1 & 0 \\ 0 & 3 & -1 & 1 \end{array}\right] \xrightarrow{E3(1/3)} \left[\begin{array}{cc|cc} 1 & -1 & 1 & 0 \\ 0 & 1 & -1/3 & 1/3 \end{array}\right]$$
$$\xrightarrow{E_{12}(1)} \left[\begin{array}{cc|cc} 1 & 0 & 2/3 & 1/3 \\ 0 & 1 & -1/3 & 1/3 \end{array}\right].$$

Thus, we have found that

$$\begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix}^{-1} = \tfrac{1}{3} \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix}.$$

To apply the inverse formula, calculate $D = 1 \cdot 2 - 1 \cdot (-1) = 3$. Swap diagonal entries of A, negate the off-diagonal entries, and divide by D to get the same result as in the preceding equation for the inverse. $\square$

The formula of the preceding example is well worth memorizing, since we will frequently need to find the inverse of a 2×2 matrix. Notice that in order

for it to make sense, we have to have D nonzero. The number D is called the *determinant* of the matrix A. We will have more to say about this number in the next section. It is fairly easy to see why A must have $D \neq 0$ in order for its inverse to exist if we look ahead to the next theorem. Notice in the above elementary operation calculations that if $D = 0$ then elementary operations on A lead to a matrix with a row of zeros. Therefore, the rank of A will be smaller than 2. Here is a summary of our current knowledge of the invertibility of a square matrix.

Theorem 2.6. Conditions for Invertibility The following are equivalent conditions on the square $n \times n$ matrix A:

(1) The matrix A is invertible.
(2) There is a square matrix B such that $BA = I$.
(3) The linear system $A\mathbf{x} = \mathbf{b}$ has a unique solution for every right-hand-side vector $\mathbf{b}$.
(4) The linear system $A\mathbf{x} = \mathbf{b}$ has a unique solution for some right-hand-side vector $\mathbf{b}$.
(5) The linear system $A\mathbf{x} = 0$ has only the trivial solution.
(6) rank $A = n$.
(7) The reduced row echelon form of A is I_n.
(8) The matrix A is a product of elementary matrices.
(9) There is a square matrix B such that $AB = I$.

Proof. The method of proof is to show that each of conditions (1)–(8) implies the next, and that condition (9) implies (1). This connects (1)–(9) in a circle, so that any one condition will imply any other and therefore all are equivalent to each other. Here is our chain of reasoning:

(1) implies (2): Assume A is invertible. Then the choice $B = A^{-1}$ satisfies condition (2).

(2) implies (3): Assume (2) is true. We can multiply both sides of the equation $A\mathbf{x} = \mathbf{b}$ on the left by B to get that $B\mathbf{b} = BA\mathbf{x} = I\mathbf{x} = \mathbf{x}$. So there is only one solution, if any. On the other hand, if the system were inconsistent then we would have rank $A < n$. By Corollary 2.3, rank $BA < n$, contradicting the fact that rank $I_n = n$. Hence, there is a solution, which proves (3).

(3) implies (4): This statement is obvious.

(4) implies (5): Assume (4) is true. Say the unique solution to $A\mathbf{x} = \mathbf{b}$ is $\mathbf{x}_0$. If the system $A\mathbf{x} = 0$ had a nontrivial solution, say $\mathbf{z}$, then we could add $\mathbf{z}$ to $\mathbf{x}_0$ to obtain a different solution $\mathbf{x}_0 + \mathbf{z}$ of the system $A\mathbf{x} = \mathbf{b}$ (check: $A(\mathbf{z} + \mathbf{x}_0) = A\mathbf{z} + A\mathbf{x}_0 = 0 + \mathbf{b} = \mathbf{b}$). This is impossible since (4) is true, so (5) follows.

(5) implies (6): Assume (5) is true. We know from Theorem 1.5 that the consistent system $A\mathbf{x} = 0$ has a unique solution precisely when the rank of A is n. Hence, (6) must be true.

(6) implies (7): Assume (6) is true. The reduced row echelon form of A is the same size as A, that is, $n \times n$, and must have a row pivot entry 1 in every row. Also, the pivot entry must be the only nonzero entry in its column. This exactly describes the matrix I_n, so that (7) is true.

(7) implies (8): This is an immediate consequence of the Inverse Algorithm.

(8) implies (9): Assume (8) is true. Repeated application of the 2/3 Rule shows that the product of any number of invertible matrices is itself invertible. Since elementary matrices are invertible, so is A in which case the choice $B = A^{-1}$ yields (9).

(9) implies (1): Assume (9) is true. Then $I^T = I = (AB)^T = B^T A^T$, so that A^T satisfies (2) which implies (8). Therefore, the matrix A^T is a product of elementary matrices, say $A^T = E_1 E_2 \cdots E_m$. It follows from Law (4) and repeated application of Law (2) of Matrix Transpose that

$$A = \left(A^T\right)^T = (E_1 E_2 \cdots E_m)^T = E_m^T E_{m-1}^T \cdots E_1^T.$$

However, we already know that products of invertible matrices are invertible by the 2/3 Rule and that transposes of elementary matrices are themselves elementary, hence invertible. It follows that A is invertible, which is condition (1). □

Notice that Theorem 2.6 relieves us of the responsibility of checking that a square one-sided inverse of a square matrix is a two-sided inverse: This is now automatic in view of conditions (2) and (9). Another interesting consequence of this theorem that has been found to be useful is an either/or statement, so it will always have something to say about any square linear system. This type of statement is sometimes called a *Fredholm alternative*. Many theorems go by this name, and we'll state another one in Chapter 5. Notice that a matrix is not invertible if and only if one of the conditions of the theorem fails. Certainly, it is true that a square matrix is either invertible or not invertible. That's all the Fredholm alternative really says, but it uses the equivalent conditions (3) and (5) of Theorem 2.6 to say it in a different way:

Corollary 2.5. **Fredholm Alternative** The square linear system $A\mathbf{x} = \mathbf{b}$ either has a unique solution for every right-hand-side vector $\mathbf{b}$ or there is a nonzero solution $\mathbf{x} = \mathbf{x}_0$ to the homogeneous system $A\mathbf{x} = \mathbf{0}$.

The PageRank Problem

We now have sufficient machinery to provide a fuller description and analysis of the PageRank technology which was introduced in Section 1.1 of Chapter 1. We shall tackle the ranking problem from a perspective somewhat different from the introduction to the PageRank in Chapter 1, page 8, though curiously, we will end up with essentially the same system of equations. To illustrate, we consider the page ranking problem of six web pages that are shown as vertices of the digraph Figure 2.7 and linked accordingly.

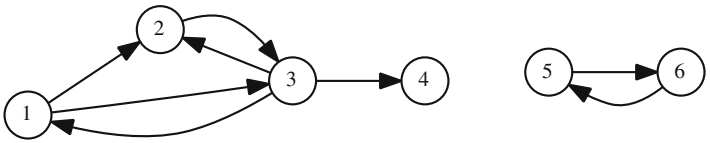


Fig. 2.7: A web with six pages as vertices and links as arrows

Example 2.45. Set up the system of page ranking for the digraph Figure 2.7 as in the third pass of Example 1.6 and express this system as a matrix product.

Solution. Recall that each page obtains its importance from its weighted backlinks. Let x_i be the ranking of page i. Since each page divides its one unit of influence among all pages to which it links we obtain this system:

$$\begin{aligned} x_1 &= \frac{x_3}{3} \\ x_2 &= \frac{x_1}{2} + \frac{x_3}{3} \\ x_3 &= \frac{x_1}{2} + \frac{x_2}{1} \\ x_4 &= \frac{x_3}{3} \\ x_5 &= \frac{x_6}{1} \\ x_6 &= \frac{x_5}{1} \end{aligned}$$

Define the matrix Q and vector $\mathbf{x}$ by

$$Q = \begin{bmatrix} 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$$

so that the system can be expressed as

$$\mathbf{x} = Q\mathbf{x}. \tag{2.3}$$

□

What the preceding example tells us is that the desired ranking vector $\mathbf{x}$ is really a stationary vector for the transition matrix Q. Moreover, since the third pass ranking for any web graph consists of a system of equations like

those of the example, it is clear that such a ranking can always be expressed in the form of equation (2.3) and hence completely defined by its transition matrix Q.

One can also apply the PageRank idea to graphs by thinking of a graph as a digraph in which every link from one node to another is matched a link in the opposite direction. This follows immediately from the fact that the adjacency matrix of a graph is symmetric, so that if we construe it to be the adjacency matrix of a digraph then for every edge from vertex i to j there is a corresponding edge from vertex j to i. The connection between adjacency matrices and transition matrices for a page ranking can be rendered explicit by the following theorem, whose proof is left as an exercise.

Theorem 2.7. Let A be the adjacency matrix of a graph or digraph and let D be a diagonal matrix whose ith entry is either the inverse of the sum of all entries in the ith row of A or zero if if this sum is zero. Then $Q = A^T D$ is the transition matrix for the page ranking of this graph.

Next suppose that we think of web surfing as a random process with each page as a state and the probabilities of moving from one state to another given by a stochastic transition matrix P. Pages are then ranked in importance by the probabilities of visiting them in the long run. Perhaps the simplest way to build such a matrix from a digraph is assume that the probability of transitioning from a vertex along any outgoing edge is equally likely. If the resulting matrix is stochastic, we call it the *surfing matrix* of this digraph. The transition matrix Q of the preceding example is *nearly* a surfing matrix in that its entries are all nonnegative with column sums 0 or 1. (Such a matrix is termed *substochastic*, in the sense that it is a square matrix with nonnegative entries whose columns sum to at most 1.) The 0 sums are the problem, so some repair is required in order to transform Q into a surfing matrix. In the case of the graph in Figure 2.7, the problem is only with vertex (or web page) 4: It is a *dangling node*, that is, a vertex with no transitions to other vertices; hence, the fourth column consists of zeros.

Surfing Matrix

Example 2.46. Repair the dangling node problem for Example 2.45 and exhibit the resulting stochastic surfing matrix.

Solution. This problem can be fixed easily enough: Introduce a *correction probability distribution vector* $\mathbf{u}$ and insert it into the fourth column. A common choice in this situation is $\mathbf{u} = \frac{1}{n}\mathbf{e}$ where n is the number of vertices of the graph and $\mathbf{e}$ is a vector of ones of length n. In terms of the underlying graph this amounts to adding links from the dangling node to every other node, which might be perfectly sensible in the absence of other information about the dangling node. Another common choice is to impose equally likely transitions to some particular related vertices, such as connected ones. Effectively, any of these choices adds links to the original

Correction Vector

graph originating at the dangling node in some controlled fashion. In our example we shall select only the vertices that connect by some directed path to vertex 4, so the resulting correction vector is $u = \frac{1}{3}(1,1,1,0,0,0)$ and the resulting transition matrix from one state to another is

$$P = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & 1 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

which is a (stochastic) surfing matrix suitable for a Markov chain. □

As we have noted, what the solution to the dangling node problem in the preceding example really amounts to is to introduce additional connections from the dangling node to other nodes in the original graph. For example, application of Theorem 2.7 to the adjacency matrix

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

yields the matrices D and P of the preceding example. The graph represented by this adjacency matrix is that of Figure 2.7 with additional outgoing edges from vertex 4 to vertices 1, 2 and 3 itself. If instead we prefer to use the choice $\mathbf{u} = \frac{1}{n}\mathbf{e}$ referenced above, the resulting adjacency matrix and surfing matrix from application of Theorem 2.7 are

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \; D = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } P = A^T D = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{2} & 1 & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{6} & 0 & 1 \\ 0 & 0 & 0 & \frac{1}{6} & 1 & 0 \end{bmatrix}.$$

We now have a stochastic matrix, but there is another problem: There may not be a unique steady state for a Markov chain with this transition matrix, which means no unique ranking. Here is one possible solution: At any vertex also give the surfer equal opportunity to randomly "teleport" from one node to any other according to the probability distribution vector $\mathbf{v}$, called the *teleportation vector*. Next select a *teleportation parameter* α

with $0 < \alpha < 1$ and let $\mathbf{e} = (1, 1, \ldots, 1)$ be a column vector of ones the same size as $\mathbf{v}$, so that the outer product $\mathbf{v}\mathbf{e}^T$ is a stochastic matrix whose columns are all equal to $\mathbf{v}$.

Teleportation Vector and Parameter

Definition 2.19. PageRank Matrix Let P be a stochastic matrix, $\mathbf{v}$ a distribution vector of compatible size and α a teleportation parameter with $0 < \alpha < 1$. Then $\alpha P + (1-\alpha)\mathbf{v}\mathbf{e}^T$ is a PageRank matrix and the corresponding PageRank problem is to find the stationary distribution vector $\mathbf{x}$ that solves the equation

$$\left(\alpha P + (1-\alpha)\mathbf{v}\mathbf{e}^T\right)\mathbf{x} = \mathbf{x}. \tag{2.4}$$

Notice that if $\mathbf{x}$ is a distribution vector, then $\mathbf{e}^T\mathbf{x} = 1$ since $\mathbf{e}^T\mathbf{x}$ is just the sum of the coordinates of $\mathbf{x}$. Thus, equation (2.4) becomes $\alpha P\mathbf{x} + (1-\alpha)\mathbf{v} = \mathbf{x}$ or, upon rearrangement of terms,

$$(I - \alpha P)\,\mathbf{x} = (1-\alpha)\mathbf{v}. \tag{2.5}$$

Note that the right-hand side of the identity

$$\left(I + \alpha P + \cdots + \alpha^k P^k\right)(I - \alpha P) = I - \alpha^{k+1}P^{k+1} \tag{2.6}$$

consists of entries $1 - \alpha^{k+1}q_{ii}$ along the diagonal and $-\alpha^{k+1}q_{ij}$ elsewhere, with all q_{ij} positive and at most 1 since a product of stochastic matrices is stochastic (see Problem 29 of Section 2.4). Since $0 < \alpha < 1$, passing to the limit in each coordinate yields that

$$\left(I + \alpha P + \cdots + \alpha^k P^k + \cdots\right)(I - \alpha P) = I.$$

Hence, $I - \alpha P$ is invertible and equation (2.5) has unique solution

$$\mathbf{x} = (1-\alpha)\left(I + \alpha P + \cdots + \alpha^k P^k + \cdots\right)\mathbf{v}. \tag{2.7}$$

Since all terms on the right-hand side are nonnegative, $\mathbf{x}$ has nonnegative entries.

So let $\mathbf{x}$ be the unique solution to equation (2.5) and multiply both sides of the equation on the left by $\mathbf{e}^T$. Note that $\mathbf{e}^T P = \mathbf{e}^T$ since each column of P is a distribution vector. We obtain

$$\mathbf{e}^T(I - \alpha P)\,\mathbf{x} = \mathbf{e}^T\mathbf{x} - \alpha\mathbf{e}^T\mathbf{x} = (1-\alpha)\,\mathbf{e}^T\mathbf{x} = (1-\alpha)\,\mathbf{e}^T\mathbf{v} = (1-\alpha)\cdot 1.$$

It follows that $\mathbf{e}^T\mathbf{x} = 1$, that is, the coordinates of $\mathbf{x}$ sum to one. Therefore, $\mathbf{x}$ is a distribution vector solving equation (2.5) and equation (2.4) as well, since these equations are equivalent for such $\mathbf{x}$.

We use these facts to obtain the key properties of the PageRank problem:

Theorem 2.8. PageRank Theorem The PageRank matrix is stochastic and the PageRank problem has a unique solution to which all Markov chains with PageRank matrix as transition matrix converge.

Proof. Let $B = \left(\alpha P + (1-\alpha)\mathbf{v}\mathbf{e}^T\right)$ with P, $\mathbf{v}$ and α meeting the conditions of a PageRank problem. The matrix $\mathbf{v}\mathbf{e}^T$ is a square matrix whose columns are all equal to the probability distribution vector $\mathbf{v}$. Therefore, $\mathbf{v}\mathbf{e}^T$ is also stochastic by definition. It follows from Problem 26 of Section 2.3 that B is also stochastic.

From the preceding discussion we know that equation (2.5) has a unique solution $\mathbf{x}$ which is a distribution vector, so it also solves equation (2.4). So consider a Markov chain $\left\{\mathbf{x}^{(k)}\right\}_{k=0}^{\infty}$ with PageRank matrix B as transition matrix. The stationary vector $\mathbf{x}$ satisfies

$$\mathbf{x} = \left(\alpha P + (1-\alpha)\mathbf{v}\mathbf{e}^T\right)\mathbf{x}$$

and the Markov chain $\left\{\mathbf{x}^{(k)}\right\}_{k=0}^{\infty}$ satisfies

$$\mathbf{x}^{(k+1)} = \left(\alpha P + (1-\alpha)\mathbf{v}\mathbf{e}^T\right)\mathbf{x}^{(k)}$$

for all nonnegative integers k. Subtract the second equation from the first and use the fact that $\mathbf{e}^T\mathbf{x} = \mathbf{e}^T\mathbf{x}^{(k)} = 1$ to obtain

$$\mathbf{x} - \mathbf{x}^{(k+1)} = \alpha P\left(\mathbf{x} - \mathbf{x}^{(k)}\right).$$

Repeated application of this equation starting at $k = 0$ yields

$$\mathbf{x} - \mathbf{x}^{(k+1)} = \alpha^k P^k\left(\mathbf{x} - \mathbf{x}^{(0)}\right).$$

The matrix P^k is itself stochastic (Problem 29 of Section 2.4). Hence, the 1-norm of $P^k\left(\mathbf{x} - \mathbf{x}^{(0)}\right)$ is at most that of $\mathbf{x} - \mathbf{x}^{(0)}$ by the stochastic matrix inequality (page 90). Therefore, the 1-norm of $\mathbf{x} - \mathbf{x}^{(k+1)}$ is at most α^k times a constant. Since $\alpha < 1$, the numbers α^k tend to zero, so that the coordinates of $\mathbf{x}^{(k)}$ tend to those of $\mathbf{x}$, which completes the proof. □

Practical values of the parameter α typically lie in the range $0.5 \leq \alpha < 1$ with $\alpha = 0.85$ as a fairly common choice in practice. Intuitively, small values of α would exaggerate the importance of the teleportation vector $\mathbf{v}$. In the case of a substochastic transition matrix, one possible choice for correction vectors $\mathbf{u}$ accounting for dangling nodes is to set $\mathbf{u} = \mathbf{v}$ for all such nodes. This choice is called *strongly preferential* PageRank, while any other choice of corrections is called a *weakly preferential* PageRank. (See the survey [13] by David Gleich for a detailed discussion of these concepts.)

Next suppose that rather than identifying the most important pages as in a web search for pages that are most pointed towards by backlinks, we are interested in finding nodes in a digraph that are most influential as measured

by outgoing links. One application of this notion of counting outgoing links would be rankings of the power of vertices in a graph, analogous to the power ratings discussed in Section 2.3. The PageRank idea is easily adapted to this sort of problem: Reverse the direction of each link in the digraph. In terms of the edge set of a digraph this simply means that the edge (j, k) of the digraph becomes the edge (k, j) of the reverse graph, followed by an application of the PageRank methodology to that reverse graph. In terms of the adjacency matrix A of the digraph, A^T is the adjacency matrix of the reverse digraph.

Reverse Page Rank

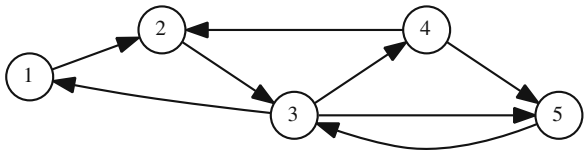


Fig. 2.8: An influence network with five influence nodes

Example 2.47. Sociologists have identified influence tendencies among five groups of individuals according to the graph of Figure 2.8 in which an edge (j, k) means that group j influences group k. Use this graph to rank the influences of each group using reverse PageRank and compare it to the power rating methodology of Section 2.3.

Solution. The first step is construct the adjacency matrix of this graph:

$$A = \begin{bmatrix} 0\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0 \\ 1\,0\,0\,1\,1 \\ 0\,1\,0\,0\,1 \\ 0\,0\,1\,0\,0 \end{bmatrix}.$$

From this we can calculate the power of each vertex from the sums of the row entries of matrix A:

$$\text{Sum}\left(A + A^2\right) = \text{Sum} \begin{bmatrix} 0\,1\,1\,0\,0 \\ 1\,0\,1\,1\,1 \\ 1\,2\,1\,1\,2 \\ 0\,1\,2\,0\,1 \\ 1\,0\,1\,1\,1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 7 \\ 4 \\ 4 \end{bmatrix}.$$

Thus, vertex 3 has the highest power rating of 7, vertices 2, 4 and 5 have a power rating of 4 and vertex 1 has a power rating of 2.

Next we use the reverse PageRank to rank the influence of each group. The adjacency matrix for the reverse graph is simply A^T, which has no dangling nodes. Therefore, the page ranking transition matrix for this graph is

$P = \left(A^T\right)^T D = AD$, where D is the diagonal matrix with diagonal entries $\left(1, \frac{1}{2}, \frac{1}{2}, 1, \frac{1}{2}\right)$, i.e.,

$$P = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix}.$$

Next select a teleportation vector $\mathbf{v} = \left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$ and a teleportation parameter $\alpha = 0.85$. Solving the resulting PageRank system of equation (2.5) with ALAMA calculator or other technology tool yields the (rounded) solution

$$\mathbf{x} = (0.106, 0.180, 0.352, 0.183, 0.180).$$

Thus, the reverse PageRank solution is a bit more refined than the power ranking: Vertex 3 has the highest rank of 0.352, vertex 4 is second with rank 0.183, vertices 2 and 5 are tied for third with rank 0.180 and vertex 1 is last with rank 0.106. Notice that the ordering 3, 4, 2, 5, 1 is consistent with an ordering from the power ranking. □

***Solving Nonlinear Equations**

We conclude this section with an application to the problem of solving systems of nonlinear equations. Although we focus on two equations in two unknowns, the same ideas can be extended to any number of equations in as many unknowns.

Recall from calculus that we could solve the one-variable equation $f(x) = 0$ for a solution point x_1 at which $f(x_1) = 0$ from a "nearby" point x_0 by setting $dx = x_1 - x_0$, and assuming that the change in f is

$$\begin{aligned} \Delta f &= f(x_1) - f(x_0) = 0 - f(x_0) \\ &\approx df = f'(x_0)\, dx = f'(x_0)(x_1 - x_0). \end{aligned}$$

Now solve for x_1 in the equation $-f(x_0) = f'(x_0)(x_1 - x_0)$ and get the equation

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Replace 1 by $n+1$ and 0 by n to obtain the famous Newton formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{2.8}$$

The idea is to start with x_0, use the formula to get x_1 and if $f(x_1)$ is not close enough to 0, then repeat the calculation with x_1 in place of x_0, and so forth until a satisfactory value of $x = x_n$ is reached. How does this relate to a multi-variable problem? We illustrate the basic idea in two variables.

Newton's Method for Systems

Example 2.48. Describe concisely an algorithm analogous to Newton's method in one variable to solve the two-variable problem

$$x^2 + \sin(\pi xy) = 1$$
$$x + y^2 + e^{x+y} = 3.$$

Solution. Our problem can be written as a system of two (nonlinear) equations in two unknowns, namely

$$f(x,y) = x^2 + \sin(\pi xy) - 1 = 0$$
$$g(x,y) = x + y^2 + e^{x+y} - 3 = 0.$$

Now we can pull the same trick with differentials as we did in the one-variable problem by setting $dx = x_1 - x_0, dy = y_1 - y_0$, where $f(x_1, y_1) = 0$, approximating the change in both f and g by total differentials, and recalling the definition of these total differentials in terms of partial derivatives. This leads to the system

$$df = f_x(x_0,y_0)\,dx + f_y(x_0,y_0)\,dy \approx f(x_1,y_1) - f(x_0,y_0) = -f((x_0,y_0))$$
$$dg = g_x(x_0,y_0)\,dx + g_y(x_0,y_0)\,dy \approx f(x_1,y_1) - g(x_0,y_0) = -g(x_0,y_0).$$

Next, write everything in vector style, say

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}, \quad \mathbf{x}^{(0)} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \quad \mathbf{x}^{(1)} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}.$$

Now we can write the *vector* differentials in the forms

$$d\mathbf{F} = \begin{bmatrix} df \\ dg \end{bmatrix} \quad \text{and} \quad d\mathbf{x} = \begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} x_1 - x_0 \\ y_1 - y_0 \end{bmatrix} = \mathbf{x}^{(1)} - \mathbf{x}^{(0)}.$$

The original Newton equations now look like a matrix multiplication involving $d\,\mathbf{x}$, $\mathbf{F}$, and a matrix of derivatives of $\mathbf{F}$, namely the *Jacobian matrix*

$$J_{\mathbf{F}}(x_0,y_0) = \begin{bmatrix} f_x(x_0,y_0) & f_y(x_0,y_0) \\ g_x(x_0,y_0) & g_y(x_0,y_0) \end{bmatrix}.$$

Specifically, we see from the definition of matrix multiplication that the Newton equations are equivalent to the vector equations

$$J_{\mathbf{F}}(\mathbf{x}_0)\left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right) = J_{\mathbf{F}}(\mathbf{x}_0)\,d\,\mathbf{x} = -\mathbf{F}\left(\mathbf{x}^{(0)}\right).$$

Replace $1, 0$ by $n+1, n$, apply the inverse of the Jacobian and add $\mathbf{x}^{(n)}$ to both sides to obtain the famous Newton formula for systems:

Newton's Formula in Vector Form

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} - J_{\mathbf{F}}\left(\mathbf{x}^{(n)}\right)^{-1} \mathbf{F}\left(\mathbf{x}^{(n)}\right).$$

This beautiful analogy to the Newton formula of (2.8) needs the language and algebra of vectors and matrices. One can now calculate the Jacobian for our particular $\mathbf{F}\left(\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]\right)$ and apply this formula, which we leave as an exercise. □

Finally, recall from calculus that we could solve the one-variable problem of finding the extrema (minimum or maximum values in a specified range) of a function $f(x)$ by observing geometrically that at such "hilltops" or "valley bottoms" of the graph of $f(x)$ the slope of the curve is flat, i.e., $f'(x) = 0$. Of course there is a caution here: A *critical point* of $f(x)$, i.e., a point x_0 at which $f'(x_0) = 0$, need not be an extremum: Consider the critical point $x = 0$ of the function $f(x) = x^3$. Now that we have a methodology for solving systems of equations we can apply it to the problem of solving optimization problems of finding the extrema of a function of more than one variable. We illustrate this in the case of two two variables: The extrema (if any) of the function $f(x, y)$ must occur at points (x, y) where the system of equations

$$\begin{aligned} f_x(x, y) &= 0 \\ f_y(x, y) &= 0 \end{aligned}$$

is satisfied.

Example 2.49. Use Newton's method to show that any quadratic function in two variables has at most one isolated extremum.

Solution. The general form of such a function is

$$f(x, y) = ax^2 + bxy + cy^2 + dx + ey + f,$$

where at least one of a, b, c is nonzero. Therefore, the system of equations to be solved for critical points (x, y) is

$$\begin{aligned} f_x(x, y) &= 2ax + by + d = 0 \\ f_y(x, y) &= bx + 2cy + e = 0. \end{aligned}$$

This is a linear system of the form $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 2a & b \\ b & 2c \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} -d \\ -e \end{bmatrix}.$$

We know from Example 2.44 that this matrix is invertible precisely if $D = 4ac - b^2 \neq 0$. In this case there is a unique critical point which may or may not be an extremum. In the case that $D = 0$ the system may be inconsistent or have infinitely many solutions. In the latter case one of the variables x, y is free and may take on any value, so no critical point is isolated. □

2.5 Exercises and Problems

Exercise 1. Find the inverse or show that it does not exist.

(a) $\begin{bmatrix} 1 & -2 & 1 \\ 0 & 2 & 0 \\ -1 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & \mathrm{i} \\ 0 & 4 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & -2 & 1 \\ 0 & 2 & 0 \\ 2 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (e) $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

Exercise 2. Find the inverse or show that it does not exist.

(a) $\begin{bmatrix} 1 & 3 & 0 \\ 0 & 4 & 10 \\ 9 & 3 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$ (e) $\begin{bmatrix} \mathrm{i}+1 & 0 \\ 1 & \mathrm{i} \end{bmatrix}$

Exercise 3. Express the following systems in matrix form and solve by inverting the coefficient matrix of the system.

(a) $\begin{aligned} 2x + 3y &= 7 \\ x + 2y &= -2 \end{aligned}$ (b) $\begin{aligned} 3x_1 + 6x_2 - x_3 &= -4 \\ -2x_1 + x_2 + x_3 &= 3 \\ x_3 &= 1 \end{aligned}$ (c) $\begin{aligned} x_1 + x_2 &= -2 \\ 5x_1 + 2x_2 &= 5 \end{aligned}$

Exercise 4. Solve the following systems by matrix inversion.

(a) $\begin{aligned} 2x_1 + 3x_2 &= 7 \\ x_2 + x_3 &= 1 \\ x_2 - x_3 &= 1 \end{aligned}$ (b) $\begin{aligned} x_1 + 6x_2 - x_3 &= 4 \\ x_1 + x_2 &= 0 \\ x_2 &= 1 \end{aligned}$ (c) $\begin{aligned} x_1 - x_2 &= 2 \\ x_1 + 2x_2 &= 11 \end{aligned}$

Exercise 5. Express inverses of the following matrices as products of elementary matrices using the notation of elementary matrices.

(a) $\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$ (e) $\begin{bmatrix} -1 & 0 \\ \mathrm{i} & 3 \end{bmatrix}$

Exercise 6. Show that the following matrices are invertible by expressing them as products of elementary matrices.

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

Exercise 7. Find $A^{-1}C$ if $A = \begin{bmatrix} 1 & 2 & -3 \\ 0 & -1 & 1 \\ 2 & 5 & -6 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & -1 & 1 & 1 \\ 2 & 0 & -6 & 0 \end{bmatrix}$.

Exercise 8. Solve $AX = B$ for X, where $A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 1 & 0 & -2 \\ 2 & -1 & 1 & 1 \end{bmatrix}$.

Exercise 9. Determine if the following matrices have right inverses and if so, exhibit one.

(a) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & -1 \\ 2 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}$

Exercise 10. Determine if the following matrices have right, left or two-sided inverses and if so, exhibit one.

(a) $\begin{bmatrix} 3 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ 1 & 1 & 3 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$

Exercise 11. Verify the matrix law $\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$ with $A = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix}$.

Exercise 12. Verify the matrix law $\left(A^*\right)^{-1} = \left(A^{-1}\right)^*$ with $A = \begin{bmatrix} 2 & 1-2\mathrm{i} \\ 0 & \mathrm{i} \end{bmatrix}$.

Exercise 13. Verify the matrix law $(AB)^{-1} = B^{-1}A^{-1}$ in the case that $A = \begin{bmatrix} 1 & 2 & -3 \\ 1 & 0 & 1 \\ 2 & 4 & -2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 & 2 \\ 0 & -3 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

Exercise 14. Verify the matrix law $(cA)^{-1} = (1/c)\,A^{-1}$ in the case that $A = \begin{bmatrix} 1 & 2-\mathrm{i} & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ and $c = 2+\mathrm{i}$.

Exercise 15. Determine for what values of k the following matrices are invertible and find the inverse in that case.

(a) $\begin{bmatrix} 1 & k \\ 0 & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ k & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -6 & 0 \\ 0 & 0 & 0 & k \end{bmatrix}$

Exercise 16. Determine the inverses for the following matrices in terms of the parameter c and conditions on c for which the matrix has an inverse.

(a) $\begin{bmatrix} 1 & 2 \\ c & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 & c+1 \\ 0 & 1 & 1 \\ 0 & 0 & c \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & c+\mathrm{i} \\ 0 & -1 & 0 \\ 0 & c & c \end{bmatrix}$

Exercise 17. Give a 2×2 example showing that the sum of invertible matrices need not be invertible.

Exercise 18. Give a 2×2 example showing that the sum of singular matrices need not be singular.

Exercise 19. Problem 29 of Section 2.2 yields a formula for the inverse of the matrix $I-N$, where N is nilpotent, namely, $(I-N)^{-1} = I+N+N^2+\cdots+N^k$. Apply this formula to matrices (a) $\begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ and (b) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$.

Exercise 20. If a matrix can be written as $A = D(I - N)$, where D is diagonal with nonzero entries and N is nilpotent, then $A^{-1} = (I - N)^{-1} D^{-1}$. Use this fact and the formulas of Exercise 19 and Example 2.41 to find inverses of the matrices (a) $\begin{bmatrix} 2 & 2 & 4 \\ 0 & 2 & -2 \\ 0 & 0 & 3 \end{bmatrix}$ and (b) $\begin{bmatrix} 2 & 0 \\ \mathrm{i} & 3 \end{bmatrix}$.

Exercise 21. Solve the PageRank problem with P as in Example 2.46, teleportation vector $\mathbf{v} = \frac{1}{6}\mathbf{e}$ and teleportation parameter $\alpha = 0.8$.

Exercise 22. Modify the surfing matrix P of Example 2.46 by using the correction vector $\frac{1}{5}(1, 1, 1, 0, 1, 1)$ and solve the resulting PageRank problem with teleportation vector $\mathbf{v} = \frac{1}{6}\mathbf{e}$ and teleportation parameter $\alpha = 0.8$.

Exercise 23. Use the adjacency matrix of the digraph G of Exercise 15 of Section 2.3 and Theorem 2.7 to find the surfing matrix of this digraph.

Exercise 24. Convert the digraph G of Exercise 15 of Section 2.3 to a graph by making all the edges unordered. Use the adjacency matrix of the resulting graph G and Theorem 2.7 to find the surfing matrix of this graph.

Exercise 25. Solve the nonlinear system of equations of Example 2.48 by using nine iterations of the vector Newton formula (2.5), starting with the initial guess $\mathbf{x}^{(0)} = (0, 1)$. Evaluate $F\left(\mathbf{x}^{(9)}\right)$.

Exercise 26. Find the minimum value of the function $F(x, y) = \left(x^2 + y + 1\right)^2 + x^4 + y^4$ by using the Newton method to find critical points of the function $F(x, y)$, i.e., points where $f(x, y) = F_x(x, y) = 0$ and $g(x, y) = F_y(x, y) = 0$.

Exercise 27. Use Example 2.49 to exhibit a quadratic function $f(x, y)$ that has no critical point.

Exercise 28. Use Example 2.49 to exhibit a quadratic function $f(x, y)$ that has infinitely many critical points.

Exercise 29. Show that there is more than one stationary state for the Markov chain of Example 2.46.

Exercise 30. Repair the dangling node problem of the graph of Figure 2.7 by using transition to all nodes as equally likely and find all stationary states for the resulting Markov chain.

*Problem 31. Show from the definition that if a square matrix A satisfies the equation $A^3 - 2A + 3I = 0$, then the matrix A must be invertible.

Problem 32. Verify directly from the definition of inverse that the two by two inverse formula gives the inverse of a 2×2 matrix.

Problem 33. Assume that the product of invertible matrices is invertible and deduce that if A and B are matrices of the same size and both B and AB are invertible, then so is A.

*Problem 34. Let A be an invertible matrix. Show that if the product of matrices AB is defined, then $\text{rank}\,(AB) = \text{rank}\,(B)$, and if BA is defined, then $\text{rank}\,(BA) = \text{rank}\,(B)$.

Problem 35. Prove that if $D = ABC$, where A, C, and D are invertible matrices, then B is invertible.

Problem 36. Given that $C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ in block form with A and B square, show that C is invertible if and only if A and B are, in which case $C^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{bmatrix}$.

Problem 37. Let T be an upper triangular matrix, say $T = D + U$, where D is diagonal and U is strictly upper triangular.
(a) Show that if D is invertible, then $T = D(I - N)$, where $N = -D^{-1}U$ is strictly upper triangular.
(b) Assume that D is invertible and use part (a), Problem 29 of Section 2.2 and Problem 33 to obtain a formula for T^{-1} involving D and N.

Problem 38. Show that if the product of matrices BA is defined and A is invertible, then $\text{rank}(BA) = \text{rank}(B)$.

*Problem 39. Given the matrix $M = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$, where the blocks A and C are invertible matrices, find a formula for M^{-1} in terms of A, B, and C.

Problem 40. Let A be the adjacency matrix for a digraph with no dangling nodes. Show that the resulting Markov chain matrix for a surfing model is $P = A^T D^{-1}$ where D is a diagonal matrix whose ith entry is the sum of the entries in the ith row of A.

Problem 41. Apply PageRank directly to the graph of Figure 2.8. How does the resulting ranking compare to power ranking of vertices and reverse PageRank ranking of Example 2.47?

2.6 Determinants

What Are They?

Many students have already had some experience with determinants and may have used them to solve square systems of equations. Why have we waited until now to introduce them? In fact, they are not really the best tool for solving systems: Gaussian elimination is better. Were it not for the *theoretical* usefulness of determinants they might be consigned to a footnote in introductory linear algebra texts as a historical artifact of linear algebra. Perhaps a better question than "What are they?" is "Why are they?".

To motivate determinants, consider Example 2.44. Something remarkable happened in that example. Not only were we able to find a formula for the inverse of a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, but we were able to compute a single number $D = ad - bc$ that told us whether A was invertible. The condition of noninvertibility, namely that $D = 0$, has a very simple interpretation: This happens exactly when one row of A is a multiple of the other, since the example showed that this is when elementary operations use the first row to zero out the second row. Can we extend this idea? Is there a single number that will tell us whether or not the square matrix A is invertible? Yes, this is exactly what determinants were invented for. The concept of determinant is subtle and not intuitive, but a large body of experience led researchers to a "correct" definition of it. There are alternative definitions, but the following, sometimes referred to as "expansion down the first column", will suit our purposes.

Definition 2.20. **Determinant** The *determinant* of a square $n \times n$ matrix $A = [a_{ij}]$ is the scalar quantity $\det A$ defined recursively as follows: If $n = 1$ then $\det A = a_{11}$; otherwise, we suppose that determinants are defined for all square matrices of size less than n and specify that

$$\begin{aligned} \det A &= \sum_{k=1}^{n} a_{k1}(-1)^{k+1} M_{k1}(A) \\ &= a_{11}M_{11}(A) - a_{21}M_{21}(A) + \cdots + (-1)^{n+1} a_{n1} M_{n1}(A), \end{aligned}$$

where $M_{ij}(A)$ is the determinant of the $(n-1) \times (n-1)$ matrix obtained from A by deleting the ith row and jth column of A.

Caution: The determinant of a matrix A is viewed as a scalar number, *not* a matrix.

Example 2.50. Describe the quantities $M_{21}(A)$ and $M_{22}(A)$, where

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 2 \end{bmatrix}.$$

Solution. Erase the second row and first column of A to obtain

$$\begin{bmatrix} 1 & 0 \\ & \\ 1 & 2 \end{bmatrix}.$$

Next collapse the remaining entries together to obtain the matrix

$$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}.$$

Similarly, erase the second row and column of A to obtain the matrix

$$\begin{bmatrix} 2 & 0 \\ & \\ 0 & 2 \end{bmatrix}.$$

Thus, we obtain that

$$M_{21}(A) = \det \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} = 2 \text{ and } M_{22}(A) = \det \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 4. \qquad \square$$

Now how did we calculate these determinants? Part (b) of the next example answers the question.

Example 2.51. Use the definition to compute the determinants of the following matrices.

$$\text{(a) } [-4] \qquad \text{(b) } \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{(c) } \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 2 \end{bmatrix}$$

Solution. (a) From the first part of the definition we have $\det[-4] = -4$.

For (b) we set $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ and use the formula of the definition to obtain that

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = a_{11}M_{11}(A) - a_{21}M_{21}(A) = a \det[d] - c \det[b] = ad - cb.$$

This calculation gives a handy formula for the determinant of a 2×2 matrix. For (c) use the definition to obtain that

$$\begin{aligned} \det \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 2 \end{bmatrix} &= 2 \det \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix} - 1 \det \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} + 0 \det \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \\ &= 2(1 \cdot 2 - 1 \cdot (-1)) - 1(1 \cdot 2 - 1 \cdot 0) + 0(1 \cdot (-1) - 1 \cdot 0) \\ &= 2 \cdot 3 - 1 \cdot 2 + 0 \cdot (-1) \\ &= 4. \end{aligned}$$

A point worth observing here is that we didn't really have to calculate the determinant of any matrix if it is multiplied by a zero. Hence, the more zeros our matrix has, the easier we expect the determinant calculation to be! □

Another common symbol for $\det A$ is $|A|$, which is also written with respect to the elements of A by suppressing matrix brackets:

$$\det A = |A| = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}.$$

This notation invites a certain oddity, if not abuse, of language: We sometimes refer to things like the "second row" or "$(2,3)$th element" or the "size" of the determinant. Yet the determinant is only a number and in that sense doesn't really have rows or entries or a size. Rather, it is the underlying matrix whose determinant is being calculated that has these properties. So be careful of this notation; we plan to use it frequently because it's handy, but you should bear in mind that determinants and matrices are *not* the same thing! Another reason that this notation can be tricky is the case of a one-dimensional matrix, say $A = [a_{11}]$. Here it is definitely *not* a good idea to forget the brackets, since we already understand $|a_{11}|$ to be the absolute value of the scalar a_{11}, a non-negative number. In the 1×1 case use $|[a_{11}]|$ for the determinant, which is just the number a_{11} and may be positive or negative.

The number $M_{ij}(A)$ is called the (i,j)th *minor* of the matrix A. If we collect the sign term in the definition of determinant together with the minor we obtain the (i,j)th *cofactor* $A_{ij} = (-1)^{i+j} M_{ij}(A)$ of the matrix A. In the terminology of cofactors,

Minors and Cofactors

$$\det A = \sum_{k=1}^{n} a_{k1} A_{k1}.$$

Laws of Determinants

Our primary goal here is to show that determinants have the magical property we promised: A matrix is singular exactly when its determinant is 0. Along the way we will examine some useful properties of determinants. There is a lot of clever algebra that can be done here; we will try to keep matters straightforward (if that's possible with determinants). In order to focus on the main ideas, we place most of the verifications of key facts at the end of the section, where we also give a concise summary of the basic determinantal laws. Unless otherwise stated, we assume throughout this section that matrices are square, and that $A = [a_{ij}]$ is an $n \times n$ matrix.

For starters, let's observe that it's very easy to calculate the determinant of upper triangular matrices. Let A be such a matrix. Then $a_{k1} = 0$ if $k > 1$, so

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{vmatrix}$$
$$= \cdots = a_{11} \cdot a_{22} \cdots a_{nn}.$$

Hence, we have established our first determinantal law:

D1: If A is an upper triangular matrix, then the determinant of A is the product of all the diagonal elements of A.

Example 2.52. Compute $D = \begin{vmatrix} 4 & 4 & 1 & 1 \\ 0 & -1 & 2 & 3 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 2 \end{vmatrix}$ and $|I_n| = \det I_n$.

Solution. By D1 we can do this at a glance: $D = 4 \cdot (-1) \cdot 2 \cdot 2 = -16$. Since I_n is diagonal, it is certainly upper triangular. Moreover, the entries down the diagonal of this matrix are 1's, so D1 implies that $|I_n| = 1$. □

Next, suppose that we notice a common factor of the scalar c in a row, say for convenience, the first one. How does this affect the determinantal calculation? In the case of a 1×1 determinant, we could simply factor it out of the original determinant. The general situation is covered by this law:

D2: If B is obtained from A by multiplying one row of A by the scalar c, then $\det B = c \cdot \det A$.

Here is a simple illustration:

Example 2.53. Compute $D = \begin{vmatrix} 5 & 0 & 10 \\ 5 & 5 & 5 \\ 0 & 0 & 2 \end{vmatrix}$.

Solution. Put another way, D2 says that scalars may be factored out of individual rows of a determinant. So use D2 on the first and second rows and then use the definition of determinant to obtain

$$\begin{vmatrix} 5 & 0 & 10 \\ 5 & 5 & 5 \\ 0 & 0 & 2 \end{vmatrix} = 5 \cdot \begin{vmatrix} 1 & 0 & 2 \\ 5 & 5 & 5 \\ 0 & 0 & 2 \end{vmatrix} = 5 \cdot 5 \cdot \begin{vmatrix} 1 & 0 & 2 \\ 1 & 1 & 1 \\ 0 & 0 & 2 \end{vmatrix} = 25 \cdot \left(1 \cdot \begin{vmatrix} 1 & 1 \\ 0 & 2 \end{vmatrix} - 1 \cdot \begin{vmatrix} 0 & 2 \\ 0 & 2 \end{vmatrix} + 0 \cdot \begin{vmatrix} 0 & 2 \\ 1 & 1 \end{vmatrix} \right)$$
$$= 50.$$

One can easily check that this is the same answer we get by working the determinant directly from the definition. □

Next, suppose we interchange two rows of a determinant.

D3: If B is obtained from A by interchanging two rows of A, then $\det B = -\det A$.

Example 2.54. Use D3 to show the following handy fact: If a determinant has a repeated row, then it must be 0.

Solution. Suppose that the ith and jth rows of the matrix A are identical, and B is obtained by switching these two rows of A. Clearly $B = A$. Yet, according to D3, $\det B = -\det A$. It follows that $\det A = -\det A$, i.e., if we add $\det A$ to both sides, $2 \cdot \det A = 0$, so that $\det A = 0$, which is what we wanted to show. □

What happens to a determinant if we add a multiple of one row to another?

D4: If B is obtained from A by adding a multiple of one row of A to another row of A, then $\det B = \det A$.

Example 2.55. Compute $D = \begin{vmatrix} 1 & 4 & 1 & 1 \\ 1 & -1 & 2 & 3 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 1 & 2 \end{vmatrix}$.

Solution. What D4 really says is that any elementary row operation $E_{ij}(c)$ can be applied to the matrix behind a determinant and the determinant will be unchanged. So in this case, add -1 times the first row to the second and $-\frac{1}{2}$ times the third row to the fourth, then apply D1 to obtain

$$\begin{vmatrix} 1 & 4 & 1 & 1 \\ 1 & -1 & 2 & 3 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 1 & 2 \end{vmatrix} = \begin{vmatrix} 1 & 4 & 1 & 1 \\ 0 & -5 & 1 & 2 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 1/2 \end{vmatrix} = 1 \cdot (-5) \cdot 2 \cdot \frac{1}{2} = -5. \qquad \square$$

Example 2.56. Use D3 to show that a matrix with a row of zeros has zero determinant.

Solution. Suppose A has a row of zeros. Add any other row of the matrix A to this zero row to obtain a matrix B with repeated rows. □

We now have enough machinery to establish the most important property of determinants. First of all, we can restate laws D2–D4 in the language of elementary matrices as follows:

Determinants of Elementary Matrices

- D2: $\det(E_i(c)A) = c \cdot \det A$ (remember that for $E_i(c)$ to be an elementary matrix, $c \neq 0$).
- D3: $\det(E_{ij}A) = -\det A$.
- D4: $\det(E_{ij}(s)A) = \det A$.

Apply a sequence of elementary row operations on the $n \times n$ matrix A to reduce it to its reduced row echelon form R, or equivalently, multiply A on the left by elementary matrices $E_1, E_2, \ldots, E_k$ and obtain

$$R = E_1 E_2 \cdots E_k A.$$

Take the determinant of both sides to obtain

$$\det R = \det(E_1 E_2 \cdots E_k A) = \pm(\text{nonzero constant}) \cdot \det A.$$

Therefore, $\det A = 0$ precisely when $\det R = 0$. Now the reduced row echelon form of A is upper triangular. In fact, it is guaranteed to have zeros on the diagonal, and therefore have zero determinant by D1, unless rank $A = n$, in which case $R = I_n$. According to Theorem 2.6 this happens precisely when A is invertible. Thus:

D5: The matrix A is invertible if and only if $\det A \neq 0$.

Example 2.57. Determine whether the following matrices are invertible without actually finding the inverse.

$$\text{(a)} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 2 \end{bmatrix} \qquad \text{(b)} \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Solution. Compute the determinants:

$$\begin{vmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 2 \end{vmatrix} = 2 \begin{vmatrix} 1 & -1 \\ 1 & 2 \end{vmatrix} - 1 \begin{vmatrix} 1 & 0 \\ 1 & 2 \end{vmatrix} = 2 \cdot 3 - 2 = 4,$$

$$\begin{vmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & 2 \end{vmatrix} = 2 \begin{vmatrix} 1 & -1 \\ -1 & 2 \end{vmatrix} - 1 \begin{vmatrix} 1 & 0 \\ -1 & 2 \end{vmatrix} = 2 \cdot 1 - 1 \cdot 2 = 0.$$

Hence, by D5, matrix (a) is invertible and matrix (b) is not invertible. □

There are two more surprising properties of determinants that we now discuss. Their proofs involve using determinantal properties of elementary matrices (see the next section for details).

D6: Given matrices A, B of the same size,

$$\det AB = \det A \det B.$$

Example 2.58. Verify D6 in the case that $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$. How do $\det(A + B)$ and $\det A + \det B$ compare in this case?

Solution. We have easily that $\det A = 1$ and $\det B = 2$. Therefore, $\det A + \det B = 1 + 2 = 3$, while $\det A \cdot \det B = 1 \cdot 2 = 2$. On the other hand,

$$AB = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix},$$

$$A + B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix},$$

so that $\det AB = 2 \cdot 3 - 4 \cdot 1 = 2 = \det A \cdot \det B$, as expected. On the other hand, we have that $\det(A + B) = 3 \cdot 2 - 1 \cdot 1 = 5 \neq \det A + \det B$. □

This example raises a very important point.

Caution: In general, $\det A + \det B \neq \det(A + B)$, though there are occasional exceptions.

In other words, determinants do not distribute over sums. (It is true, however, that the determinant is additive in *one row at a time*. See the proof of D4 for details.)

Finally, we ask how $\det A^T$ compares to $\det A$. Simple cases suggest that there is no difference in determinant. This is exactly what happens in general:

D7: For all square matrices A, $\det A^T = \det A$.

Example 2.59. Compute $D = \begin{vmatrix} 4 & 0 & 0 & 0 \\ 4 & 1 & 0 & 0 \\ 1 & 2 & -2 & 0 \\ 1 & 0 & 1 & 2 \end{vmatrix}$.

Solution. By D7 and D1 we see immediately that $D = 4 \cdot 1 \cdot (-2) \cdot 2 = -16$. □

D7 is a very useful fact. Let's look at it from this point of view: Transposing a matrix interchanges the rows and columns of the matrix. Therefore, everything that we have said about rows of determinants applies equally well to the columns, *including the definition of determinant itself!* Therefore, we could have given the definition of determinant in terms of expanding across the first row instead of down the first column and gotten the same answers. Likewise, we could perform elementary column operations instead of row operations and get the same results as D2–D4. Furthermore, the determinant of a lower triangular matrix is the product of its diagonal elements thanks to D7+D1. By interchanging rows or columns then expanding by first row or column, we see that the same effect is obtained by simply expanding the determinant down any column or across any row. We have to alternate signs starting with the sign $(-1)^{i+j}$ of the first term we use.

Now we can really put it all together and compute determinants to our heart's content with a good deal less effort than the original definition specified. We can use D1–D4 in particular to make a determinant calculation no worse than Gaussian elimination in the amount of work we have to do. We simply reduce a matrix to triangular form by elementary operations, then take the product of the diagonal terms.

Example 2.60. Calculate $D = \begin{vmatrix} 3 & 0 & 6 & 6 \\ 1 & 0 & 2 & 1 \\ 2 & 0 & 0 & 1 \\ -1 & 2 & 0 & 0 \end{vmatrix}$.

Solution. We are going to do this calculation two ways. We may as well use the same elementary operation notation that we have employed in Gaussian

elimination. The only difference is that we have equality instead of arrows, provided that we modify the value of the new determinant in accordance with the laws D1–D3. So here is the straightforward method:

$$D = 3\begin{vmatrix} 1 & 0 & 2 & 2 \\ 1 & 0 & 2 & 1 \\ 2 & 0 & 0 & 1 \\ -1 & 2 & 0 & 0 \end{vmatrix} \underset{\substack{E_{21}(-1)\\ E_{31}(-2)\\ E_{41}(1)}}{=} 3\begin{vmatrix} 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -4 & -3 \\ 0 & 2 & 2 & 2 \end{vmatrix} \underset{E_{24}}{=} -3\begin{vmatrix} 1 & 0 & 2 & 2 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & -4 & -3 \\ 0 & 0 & 0 & -1 \end{vmatrix} = -24.$$

Here is another approach: Let's expand the determinant down the second column, since it is mostly 0's. Remember that the sign in front of the first minor must be $(-1)^{1+2} = -1$. Also, the coefficients of the first three minors are 0, so we need only write down the last one in the second column:

$$D = +2\begin{vmatrix} 3 & 6 & 6 \\ 1 & 2 & 1 \\ 2 & 0 & 1 \end{vmatrix}.$$

Expand down the second column again:

$$D = 2\left(-6\begin{vmatrix} 1 & 1 \\ 2 & 1 \end{vmatrix} + 2\begin{vmatrix} 3 & 6 \\ 2 & 1 \end{vmatrix}\right) = 2(-6 \cdot (-1) + 2 \cdot (-9)) = -24. \qquad \square$$

An Inverse Formula

Let $A = [a_{ij}]$ be an $n \times n$ matrix. We have already seen that we can expand the determinant of A down any column of A (see the discussion following Example 2.59). These expansions lead to cofactor formulas for each column number j:

$$\det A = \sum_{k=1}^{n} a_{kj}A_{kj} = \sum_{k=1}^{n} A_{kj}a_{kj}.$$

This formula resembles a matrix multiplication formula. Consider the slightly altered sum

$$\sum_{k=1}^{n} A_{ki}a_{kj} = A_{1i}a_{1j} + A_{2i}a_{2j} + \cdots + A_{ni}a_{nj}.$$

The key to understanding this expression is to realize that it is exactly what we would get if we replaced the ith column of the matrix A by its jth column and then computed the determinant of the resulting matrix by expansion down the ith column. But such a matrix has two equal columns and therefore has a zero determinant, which we can see by applying Example 2.54 to the transpose of the matrix and using D7. So this sum must be 0 if $i \neq j$. We can combine these two sums by means of the

Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise) in the formula

Kronecker Delta

$$\sum_{k=1}^{n} A_{ki} a_{kj} = \delta_{ij} \det A. \tag{2.9}$$

In order to exploit this formula we make the following definitions:

Definition 2.21. Adjoint, Minor, and Cofactor Matrices The *matrix of minors* of the $n \times n$ matrix $A = [a_{ij}]$ is the matrix $M(A) = [M_{ij}(A)]$ of the same size. The *matrix of cofactors* of A is the matrix $A_{\text{cof}} = [A_{ij}]$ of the same size. Finally, the *adjoint matrix* of A is the matrix $\operatorname{adj} A = A_{\text{cof}}^T$.

Example 2.61. Compute the determinant, minors, cofactors, and adjoint matrices for $A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & -1 \\ 0 & 2 & 1 \end{bmatrix}$ and compute $A \operatorname{adj} A$.

Solution. The determinant is easily seen to be 2. Now for the matrix of minors:

$$M(A) = \begin{bmatrix} \begin{vmatrix} 0 & -1 \\ 2 & 1 \end{vmatrix} & \begin{vmatrix} 0 & -1 \\ 0 & 1 \end{vmatrix} & \begin{vmatrix} 0 & 0 \\ 0 & 2 \end{vmatrix} \\ \begin{vmatrix} 2 & 0 \\ 2 & 1 \end{vmatrix} & \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 0 & 2 \end{vmatrix} \\ \begin{vmatrix} 2 & 0 \\ 0 & -1 \end{vmatrix} & \begin{vmatrix} 1 & 0 \\ 0 & -1 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 0 & 0 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 1 & 2 \\ -2 & -1 & 0 \end{bmatrix}.$$

To get the matrix of cofactors, simply overlay $M(A)$ with the following "checkerboard" of $+/-$'s $\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$ to obtain the matrix $A_{\text{cof}} = \begin{bmatrix} 2 & 0 & 0 \\ -2 & 1 & -2 \\ -2 & 1 & 0 \end{bmatrix}$. Now transpose A_{cof} to obtain

$$\operatorname{adj} A = \begin{bmatrix} 2 & -2 & -2 \\ 0 & 1 & 1 \\ 0 & -2 & 0 \end{bmatrix}.$$

We check that

$$A \operatorname{adj} A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & -1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & -2 \\ 0 & 1 & 1 \\ 0 & -2 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = (\det A) I_3. \qquad \square$$

Of course, the example simply confirms equation (2.9) since this formula gives the (i, j)th entry of the product $(\operatorname{adj} A)A$. If we were to do determinants by row expansions, we would get a similar formula for the (i, j)th entry of $A \operatorname{adj} A$. We summarize this information in matrix notation as the following determinantal property:

Adjoint Formula

D8: For a square matrix A,

$$A \operatorname{adj} A = (\operatorname{adj} A)A = (\det A)I.$$

What does this have to do with inverses? We already know that A is invertible exactly when $\det A \neq 0$, so the answer is staring at us! Just divide the terms in D8 by $\det A$ to obtain an explicit formula for A^{-1}:

Inverse Formula

D9: For a square matrix A such that $\det A \neq 0$, .

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A.$$

Example 2.62. Compute the inverse of the matrix A of Example 2.61 by the inverse formula.

Solution. We already computed the adjoint matrix of A, and the determinant of A is just 2, so we have that

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A = \frac{1}{2} \begin{bmatrix} 2 & -2 & -2 \\ 0 & 1 & 1 \\ 0 & -2 & 0 \end{bmatrix}. \qquad \square$$

Example 2.63. Interpret the inverse formula for the 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

Solution. We have $M(A) = \begin{bmatrix} d & c \\ b & a \end{bmatrix}$, $A_{\text{cof}} = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$ and $\operatorname{adj} A = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$, so that the inverse formula becomes

$$A^{-1} = \frac{1}{\det A} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

This is exactly the same as the formula we obtained in Example 2.44. $\square$

Cramer's Rule

Thanks to the inverse formula, we can now find an explicit formula for solving linear systems with a nonsingular coefficient matrix. Here's how we proceed. To solve $A\mathbf{x} = \mathbf{b}$ we multiply both sides on the left by A^{-1} to obtain that $\mathbf{x} = A^{-1}\mathbf{b}$. Now use the inverse formula to obtain

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{1}{\det A} \operatorname{adj}(A)\mathbf{b}.$$

The explicit formula for the ith coordinate of $\mathbf{x}$ that comes from this fact is

$$x_i = \frac{1}{\det A} \sum_{j=1}^{n} A_{ji} b_j.$$

The summation term is exactly what we would obtain if we started with the determinant of the matrix B_i obtained from A by replacing the ith column of A by $\mathbf{b}$ and then expanding the determinant down the ith column. Therefore, we have arrived at the following rule:

Theorem 2.9. Cramer's Rule Let A be an invertible $n \times n$ matrix and $\mathbf{b}$ an $n \times 1$ column vector. Denote by B_i the matrix obtained from A by replacing the ith column of A by $\mathbf{b}$. Then the linear system $A\mathbf{x} = \mathbf{b}$ has unique solution $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where

$$x_i = \frac{\det B_i}{\det A}, \quad i = 1, 2, \ldots, n.$$

Example 2.64. Use Cramer's rule to solve the system

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ 4x_1 + 4x_2 &= 20. \end{aligned}$$

Solution. The coefficient matrix and right-hand-side vectors are

$$A = \begin{bmatrix} 2 & -1 \\ 4 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 20 \end{bmatrix},$$

so that $\det A = 8 - (-4) = 12$, and therefore,

$$x_1 = \frac{\begin{vmatrix} 1 & -1 \\ 20 & 4 \end{vmatrix}}{\begin{vmatrix} 2 & -1 \\ 4 & 4 \end{vmatrix}} = \frac{24}{12} = 2 \quad \text{and} \quad x_2 = \frac{\begin{vmatrix} 2 & 1 \\ 4 & 20 \end{vmatrix}}{\begin{vmatrix} 2 & -1 \\ 4 & 4 \end{vmatrix}} = \frac{36}{12} = 3. \qquad \square$$

Summary of Determinantal Laws

Here is a summary of the basic laws of determinants with laws D2–D4 stated in terms of elementary operations as multiplication by elementary matrices:

Laws of Determinants

Let A, B be $n \times n$ matrices.

D1: If A is upper triangular, $\det A$ is the product of all the diagonal elements of A.

D2: $\det(E_i(c)A) = c \cdot \det A$.

D3: $\det(E_{ij}A) = -\det A$.

D4: $\det(E_{ij}(s)A) = \det A$.

D5: The matrix A is invertible if and only if $\det A \neq 0$.

D6: $\det AB = \det A \det B$.

D7: $\det A^T = \det A$.

D8: $A \operatorname{adj} A = (\operatorname{adj} A)A = (\det A)I$.

D9: If $\det A \neq 0$, then $A^{-1} = \dfrac{1}{\det A} \operatorname{adj} A$.

Determinants of Some Block Matrices

This section began with a discussion of the 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and its determinant $D = ad - bc$. This raises an interesting question: Suppose that a matrix is blocked as $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Is there a formula for the determinant of M analogous to that of a 2×2 matrix? The answer is a qualified "yes", as the following theorem shows:

Theorem 2.10. Let $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, where A is $m \times m$, D is $n \times n$, B is $m \times n$ and C is $n \times m$. Then

(1) If $C = 0$ or $B = 0$, then $\det M = \det A \det D$.
(2) If $\det A \neq 0$, then $\det M = \det A \det \left(D - CA^{-1}B\right)$.
(3) If $\det D \neq 0$, then $\det M = \det D \det \left(A - BD^{-1}C\right)$.

Proof. To prove (1), assume first that $C = 0$. Block multiplication yields

$$M = \begin{bmatrix} A & B \\ 0 & D \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} A & B \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & B \\ 0 & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix}.$$

Note that expanding $\det \begin{bmatrix} I_m & 0 \\ 0 & D \end{bmatrix}$ down the first column yields $1 \cdot \det \begin{bmatrix} I_{m-1} & 0 \\ 0 & D \end{bmatrix}$ since the only nonzero entry of the first column is the first one. Repeating this argument $m-1$ times results in $|D|$. Similarly, expansion down the last column yields

$$\det \begin{bmatrix} A & 0 \\ 0 & I_n \end{bmatrix} = (-1)^{n+n} \det \begin{bmatrix} A & 0 \\ 0 & I_{n-1} \end{bmatrix}$$

and repeated application of this fact results in a value of $\det A$. The matrix $\begin{bmatrix} I & B \\ 0 & I \end{bmatrix}$ is clearly upper triangular, so D1 applies to it. Apply D6 to the product of matrices and we obtain

$$\det \begin{bmatrix} A & B \\ 0 & D \end{bmatrix} = \det \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} \cdot \det \begin{bmatrix} I & B \\ 0 & I \end{bmatrix} \cdot \det \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} = \det A \cdot 1 \cdot \det D.$$

The proof of the case $B = 0$ is similar and left as an exercise.

To prove (2) assume that $|A| \neq 0$ so that the matrix A is invertible by D9. Apply D6 and (1) to the factorization

$$\begin{bmatrix} A^{-1} & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & A^{-1}B \\ 0 & -CA^{-1}B + D \end{bmatrix}$$

to obtain

$$\begin{aligned} \det A^{-1} \det \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= (\det A)^{-1} \det M \\ &= \det \begin{bmatrix} I & A^{-1}B \\ 0 & -CA^{-1}B + D \end{bmatrix} \\ &= \det \left(D - CA^{-1}B \right), \end{aligned}$$

which proves (2).

The proof of (3) is similar and left as an exercise. □

*Verification of Some Determinantal Laws

D2: If B is obtained from A by multiplying one row of A by the scalar c, then $\det B = c \cdot \det A$.

To keep the notation simple, assume that the first row is multiplied by c, the proof being similar for other rows. Suppose we have established this for all determinants of size less than n (this is really another "proof by induction," which is how most of the following determinantal properties are established). For an $n \times n$ determinant we have

$$\begin{aligned} \det B &= \begin{vmatrix} c \cdot a_{11} & c \cdot a_{12} & \cdots & c \cdot a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\ &= c \cdot a_{11} \begin{vmatrix} a_{22} & a_{23} & \cdots & a_{2n} \\ a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ a_{n2} & a_{n3} & \cdots & a_{nn} \end{vmatrix} + \sum_{k=2}^{n} a_{k1}(-1)^{k+1} M_{k1}(B). \end{aligned}$$

But the minors $M_{k1}(B)$ all are smaller and have a common factor of c in the first row. Pull this factor out of every remaining term and we get that

$$\begin{vmatrix} c\cdot a_{11} & c\cdot a_{12} & \cdots & c\cdot a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = c\cdot \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}.$$

Thus, we have shown that property D2 holds for all matrices.

D3: If B is obtained from A by interchanging two rows of A, then $\det B = -\det A$.

To keep the notation simple, assume we switch the first and second rows. In the case of a 2×2 determinant, we get the negative of the original determinant (check this for yourself). Suppose we have established that the same is true for all matrices of size less than n. For an $n\times n$ determinant we have

$$\begin{aligned} \det B &= \begin{vmatrix} a_{21} & a_{22} & \cdots & a_{2n} \\ a_{11} & a_{12} & \cdots & a_{1n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\ &= a_{21}M_{11}(B) - a_{11}M_{21}(B) + \sum_{k=3}^{n} a_{k1}(-1)^{k+1}M_{k1}(B) \\ &= a_{21}M_{21}(A) - a_{11}M_{11}(A) + \sum_{k=3}^{n} a_{k1}(-1)^{k+1}M_{k1}(B). \end{aligned}$$

But all the determinants in the summation sign come from a submatrix of A with the first and second rows interchanged. Since they are smaller than n, each is just the negative of the corresponding minor of A. Notice that the first two terms are just the first two terms in the determinantal expansion of A, except that they are out of order and have an extra minus sign. Factor this minus sign out of every term and we have obtained D3. □

D4: If B is obtained from A by adding a multiple of one row of A to another row of A, then $\det B = \det A$.

Actually, it's a little easier to answer a slightly more general question: What happens if we replace a row of a determinant by that row plus some other row vector $\mathbf{r}$ (not necessarily a row of the determinant)? Again, simply for convenience of notation, we assume that the row in question is the first. The same argument works for any other row. Some notation: Let B be the matrix that we obtain from the $n\times n$ matrix A by adding the row vector $\mathbf{r} = [r_1, r_2, \ldots, r_n]$ to the first row and C the matrix obtained from A by replacing the first row by $\mathbf{r}$. The answer turns out to be that $|B| = |A| + |C|$. So we can say that the determinant function is "additive in each row." Let's see what happens in the one dimensional case:

$$|B| = |[a_{11} + r_1]| = a_{11} + r_1 = |[a_{11}]| + |[r_1]| = |A| + |C|.$$

Suppose we have established that the same is true for all matrices of size less than n and let A be $n \times n$. Then the minors $M_{k1}(B)$, with $k > 1$, are smaller than n, so the property holds for them. Hence, we have

$$\begin{aligned}
\det B &= \begin{vmatrix} a_{11} + r_1 & a_{12} + r_2 & \cdots & a_{1n} + r_n \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\
&= (a_{11} + r_1) M_{11}(A) + \sum_{k=2}^{n} a_{k1} (-1)^{k+1} M_{k1}(B) \\
&= (a_{11} + r_1) M_{11}(A) + \sum_{k=2}^{n} a_{k1} (-1)^{k+1} (M_{k1}(A) + M_{k1}(C)) \\
&= \sum_{k=1}^{n} a_{k1} (-1)^{k+1} M_{k1}(A) + r_1 M_{11}(C) + \sum_{k=2}^{n} a_{k1} (-1)^{k+1} M_{k1}(C) \\
&= \det A + \det C.
\end{aligned}$$

Now what about adding a multiple of one row to another in a determinant? For notational convenience, suppose we add s times the second row to the first. In the notation of the previous paragraph,

$$\det B = \begin{vmatrix} a_{11} + s \cdot a_{21} & a_{12} + s \cdot a_{22} & \cdots & a_{1n} + s \cdot a_{2n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

and

$$\det C = \begin{vmatrix} s \cdot a_{21} & s \cdot a_{22} & \cdots & s \cdot a_{2n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = s \cdot \begin{vmatrix} a_{21} & a_{22} & \cdots & a_{2n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = 0,$$

where we applied D2 to pull the common factor s from the first row and the result of Example 2.54 to get the determinant with repeated rows to be 0. But $|B| = |A| + |C|$. Hence, we have shown D4. □

D6: If matrices A, B are of the same size, then $\det AB = \det A \det B$.

The key here is that we now know that determinant calculation is intimately connected with elementary matrices, rank, and the reduced row echelon form. First let's reinterpret D2–D4 still one more time. First of all take $A = I$ in the discussion of the previous paragraph, and we see that

- $\det E_i(c) = c$
- $\det E_{ij} = -1$
- $\det E_{ij}(s) = 1$

Therefore, D2–D4 can be restated (yet again) as

- D2: $\det(E_i(c)A) = \det E_i(c) \cdot \det A$ (here $c \neq 0$.)
- D3: $\det(E_{ij}A) = \det E_{ij} \cdot \det A$
- D4: $\det(E_{ij}(s) = \det E_{ij}(s) \cdot \det A$

In summary: For any elementary matrix E and arbitrary matrix A of the same size, $\det(EA) = \det(E)\det(A)$.

Now let's consider this question: How does $\det(AB)$ relate to $\det(A)$ and $\det(B)$? If A is not invertible, rank $A < n$ by Theorem 2.6 and so rank $AB < n$ by Corollary 2.3. Therefore, $\det(AB) = 0 = \det A \cdot \det B$ in this case. Next suppose that A is invertible. Express it as a product of elementary matrices, say $A = E_1E_2 \cdots E_k$, and use our summary of D1–D3 to disassemble and reassemble the elementary factors:

$$\begin{aligned}\det(AB) &= \det(E_1E_2 \cdots E_kB) = (\det E_1 \det E_2 \cdots \det E_k) \det B \\ &= \det(E_1E_2 \cdots E_k) \det B = \det A \cdot \det B.\end{aligned}$$

Thus, we have shown that **D6** holds. □

D7: For all square matrices A, $\det A^T = \det A$.

Recall these facts about elementary matrices:

- $\det E_{ij}^T = \det E_{ij}$
- $\det E_i(c)^T = \det E_i(c)$
- $\det E_{ij}(c)^T = \det E_{ji}(c) = 1 = \det E_{ij}(c)$

Therefore, transposing does not affect determinants of elementary matrices. Now for the general case observe that since A and A^T are transposes of each other, one is invertible if and only if the other is by the Transpose/Inverse law. In particular, if both are singular, then $\det A^T = 0 = \det A$. On the other hand, if both are nonsingular, then write A as a product of elementary matrices, say $A = E_1E_2 \cdots E_k$, and obtain from the product law for transposes that $A^T = E_k^T E_{k-1}^T \ldots E_1^T$, so by D6

$$\begin{aligned}\det A^T &= \det E_k^T \det E_{k-1}^T \cdots \det E_1^T = \det E_k \det E_{k-1} \cdots \det E_1 \\ &= \det E_1 \det E_2 \cdots \det E_k = \det A.\end{aligned}$$ □

2.6 Exercises and Problems

Exercise 1. Compute all cofactors for these matrices.

(a) $\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1-\mathrm{i} \\ 0 & 1 \end{bmatrix}$

Exercise 2. Compute all minors for these matrices.

(a) $\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -3 & 0 \\ -2 & 1 & 0 \\ 0 & -2 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & \mathrm{i}+1 \\ \mathrm{i} & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 3 & 1 & -1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{bmatrix}$

Exercise 3. Compute these determinants. Which of the matrices represented are invertible?

(a) $\begin{vmatrix} 2 & -1 \\ 1 & 1 \end{vmatrix}$ (b) $\begin{vmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1+\mathrm{i} \end{vmatrix}$ (c) $\begin{vmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{vmatrix}$ (d) $\begin{vmatrix} 1 & -1 & 4 & 2 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 2 & 7 \\ -2 & 3 & 4 & 6 \end{vmatrix}$ (e) $\begin{vmatrix} -1 & -1 \\ 1 & 1-2\mathrm{i} \end{vmatrix}$

Exercise 4. Use determinants to determine which of these matrices are invertible.

(a) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ -2 & 3 & 4 & 6 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 0 & 1 & 3 \\ 1 & 1 & 2 & 7 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix}$ (e) $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$

Exercise 5. Verify by calculation that determinantal law D7 holds for the following choices of A.

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

Exercise 6. Let $A = B$ and verify by calculation that determinantal law D6 holds for the following choices of A.

(a) $\begin{bmatrix} -2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -1 & 1 \\ 1 & 2 & 0 \\ -1 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 3 \\ -1 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 2 & 7 \end{bmatrix}$

Exercise 7. Use determinants to find conditions on the parameters in these matrices under which the matrices are invertible.

(a) $\begin{bmatrix} a & 1 \\ ab & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & -1 \\ 1 & c & 1 \\ 0 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}$

Exercise 8. Find conditions on the parameters in these matrices under which the matrices are invertible.

(a) $\begin{bmatrix} a & b & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & b & a \\ 0 & 0 & -a & b \end{bmatrix}$ (b) $\begin{bmatrix} \lambda-1 & 0 & 0 \\ 1 & \lambda-2 & 1 \\ 3 & 1 & \lambda-1 \end{bmatrix}$ (c) $\lambda I_2 - \begin{bmatrix} 0 & 1 \\ -c_0 & -c_1 \end{bmatrix}$

Exercise 9. For each of the following matrices calculate the adjoint matrix and the product of the matrix and its adjoint.

(a) $\begin{bmatrix} 2 & 1 & 0 \\ -1 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 3 \\ -1 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 2 & 6 \end{bmatrix}$

Exercise 10. For each of the following matrices calculate the adjoint matrix and the product of the adjoint and the matrix.

(a) $\begin{bmatrix} -1 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -3 \end{bmatrix}$

Exercise 11. Find the inverse of following matrices by adjoints.

(a) $\begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & -2 & 2 \\ -1 & 2 & -1 \\ 1 & -3 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & \mathrm{i} \\ -2\mathrm{i} & 1 \end{bmatrix}$

Exercise 12. For each of the following matrices, find the inverse by superaugmented matrices and by adjoints.

(a) $\begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -1 & 3 \\ 2 & 2 & -4 \\ 1 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$

Exercise 13. Use Cramer's Rule to solve the following systems.

(a) $\begin{aligned} x - 3y &= 2 \\ 2x + y &= 11 \end{aligned}$ (b) $\begin{aligned} 2x_1 + x_2 &= b_1 \\ 2x_1 - x_2 &= b_2 \end{aligned}$ (c) $\begin{aligned} 3x_1 + x_3 &= 2 \\ 2x_1 + 2x_2 &= 1 \\ x_1 + x_2 + x_3 &= 6 \end{aligned}$

Exercise 14. Use Cramer's Rule to solve the following systems.

(a) $\begin{aligned} x + y + z &= 4 \\ 2x + 2y + 5z &= 11 \\ 4x + 6y + 8z &= 24 \end{aligned}$ (b) $\begin{aligned} x_1 - 2x_2 &= 2 \\ 2x_1 - x_2 &= 4 \end{aligned}$ (c) $\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ x_1 + 2x_2 &= 1 \\ x_1 - x_3 &= 2 \end{aligned}$

Exercise 15. Use Theorem 2.10 to compute determinants of the following matrices.

(a) $M = \begin{bmatrix} -1 & 1 & 1 & -1 & 2 \\ 0 & 1 & 2 & 5 & 3 \\ 3 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1+\mathrm{i} \\ 0 & 0 & 0 & 1-\mathrm{i} & -1 \end{bmatrix}$ (b) $M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 & 7 & 0 \\ -1 & 0 & 0 & 3 & 2 & -2 \end{bmatrix}$

Exercise 16. Use Theorem 2.10 to compute determinants of the following matrices.

(a) $M = \begin{bmatrix} -1 & 1 & 1 & -1 & 2 \\ 0 & 1 & 2 & 5 & 3 \\ 3 & 0 & 2 & 0 & 0 \\ 0 & -2 & 0 & 1 & 3 \\ 2 & 0 & 0 & 2 & 6 \end{bmatrix}$ (b) $M = \begin{bmatrix} -1 & 2 & 1 & -1 & 2 & 1 \\ -2 & 5 & 4 & 4 & 0 & 0 \\ 0 & 1 & 2 & 5 & 3 & -5 \\ 2 & 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 2 & 4 \\ 0 & 6 & 0 & -1 & 2 & 4 \end{bmatrix}$

Problem 17. Verify from definition that $\begin{vmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & e & f \\ 0 & 0 & g & h \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} \begin{vmatrix} e & f \\ g & h \end{vmatrix}$.

Problem 18. Confirm that the determinant of the matrix $A = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$ is -1. We can now assert without any further calculation that the inverse matrix of A has integer coefficients. Explain why in terms of laws of determinants.

Problem 19. Let

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{bmatrix}.$$

(V is a *Vandermonde* matrix.) Express $\det V$ as a product of factors $(x_j - x_k)$.

*Problem 20. Show that the determinant of the general Vandermonde matrix

$$V_n = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}.$$

is a product of factors $(x_j - x_k)$ with $j > k$.

Problem 21. Show by example that $\det A^* \neq \det A$ and prove that in general $\det A^* = \overline{\det A}$.

*Problem 22. Use a determinantal law to show that $\det(A) \det\left(A^{-1}\right) = 1$ if A is invertible.

Problem 23. Use the determinantal laws to show that any matrix with a row of zeros has zero determinant.

*Problem 24. If A is a 5×5 matrix, then in terms of $\det(A)$, what can we say about $\det(-2A)$? Explain and express a law about a general matrix cA, c a scalar, that contains your answer.

Problem 25. Let A be a skew-symmetric matrix, that is, $A^T = -A$. Show that if A has odd order n, i.e., A is $n \times n$, then A must be singular.

*Problem 26. Show that if

$$M = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$$

then $\det M = \det A \cdot \det D$.

Problem 27. Show that if $\det D \neq 0$ and

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

then $\det = \det D \det \left(A - BD^{-1}C\right)$.

*Problem 28. Let J_n be the $n \times n$ counteridentity, that is, J_n is a square matrix with ones along the counterdiagonal (the diagonal that starts in the lower left corner and ends in the upper right corner), and zeros elsewhere. Find a formula for $\det J_n$.

Problem 29. Show that the *companion matrix* of the polynomial $f(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1} + x^n$, which is defined to be

$$C(f) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -c_0 & -c_1 & \cdots & -c_{n-2} & -c_{n-1} \end{bmatrix},$$

is invertible if and only if $c_0 \neq 0$.

Problem 30. Prove that if real matrix A is invertible, then $\det(A^T A) > 0$.

Problem 31. Suppose that the square matrix A is singular. Prove that if the system $A\mathbf{x} = \mathbf{b}$ is consistent, then $(\operatorname{adj} A)\,\mathbf{b} = \mathbf{0}$.

Problem 32. Prove that if A is $n \times n$, then $\det(-A) = (-1)^n \det A$.

Problem 33. Let A and B be invertible matrices of the same size. Use determinantal law D9 to prove that $\operatorname{adj}\left(A^{-1}\right) = (\operatorname{adj}(A))^{-1}$ and $\operatorname{adj}(AB) = \operatorname{adj}(A) \cdot \operatorname{adj}(B)$.

2.7 *Tensor Products

How do we solve a system of equations in which the unknowns can be organized into a matrix X and the linear system in question is of the form

$$AX + XB = C, \tag{2.10}$$

where A, B, C are matrices? We call this equation the *Sylvester equation*. Such systems occur in a number of physical applications; for example, discretizing certain partial differential equations in order to solve them numerically can lead to such a system. Of course, we could simply expand each system laboriously. This direct approach offers us little insight as to the nature of the resulting system.

Sylvester Equation

We are going to develop a powerful "bookkeeping" method that will rearrange the variables of Sylvester's equation automatically. The first basic idea needed here is that of the tensor product of two matrices, which is defined as follows:

Definition 2.22. Tensor Product Let $A = [a_{ij}]$ be an $m \times p$ matrix and $B = [b_{ij}]$ an $n \times q$ matrix. Then the *tensor product* of A and B is the $mn \times pq$ matrix $A \otimes B$ that can be expressed in block form as

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1j}B & \cdots & a_{1p}B \\ a_{21}B & a_{22}B & \cdots & a_{2j}B & \cdots & a_{2p}B \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1}B & a_{i2}B & \cdots & a_{ij}B & \cdots & a_{ip}B \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mj}B & \cdots & a_{mp}B \end{bmatrix}.$$

Note 2.3. Some authors refer to the tensor product of matrices as the *Kronecker product* and reserve the term "tensor" for analogous operations on more abstract objects such as the vector spaces and linear operators discussed in Chapter 3.

Example 2.65. Let $A = \begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$. Exhibit $A \otimes B$, $B \otimes A$, and $I_2 \otimes A$ and conclude that $A \otimes B \neq B \otimes A$.

Solution. From the definition,

$$A \otimes B = \begin{bmatrix} 1B & 3B \\ 2B & 1B \end{bmatrix} = \begin{bmatrix} 4 & 12 \\ -1 & -3 \\ 8 & 4 \\ -2 & -1 \end{bmatrix}, \quad B \otimes A = \begin{bmatrix} 4A \\ -1A \end{bmatrix} = \begin{bmatrix} 4 & 12 \\ 8 & 4 \\ -1 & -3 \\ -2 & -1 \end{bmatrix},$$

$$\text{and} \quad I_2 \otimes A = \begin{bmatrix} 1A & 0A \\ 0A & 1A \end{bmatrix} = \begin{bmatrix} 1 & 3 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 2 & 1 \end{bmatrix}. \qquad \square$$

The other ingredient that we need to solve equation (2.10) is an operator that turns matrices into vectors. It is defined as follows.

Definition 2.23. Vec Operator Let A be an $m \times n$ matrix. Then the $mn \times 1$ vector $\operatorname{vec} A$ is obtained from A by stacking successive columns of A vertically, with the first column at the top and the last column of A at the bottom.

Example 2.66. Let $A = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 1 & 5 \end{bmatrix}$. Compute $\operatorname{vec} A$.

Solution. Stack the three columns to obtain $\operatorname{vec} A = [1, 2, 3, 1, 4, 5]^T$. □

The vec operator is linear ($\operatorname{vec}(aA + bB) = a \operatorname{vec} A + b \operatorname{vec} B$). We leave the proof, along with proofs of the following simple tensor facts, to the reader.

Theorem 2.11. Laws of Tensor Products Let A, B, C, D be suitably sized matrices. Then

(1) $(A + B) \otimes C = A \otimes C + B \otimes C$
(2) $A \otimes (B + C) = A \otimes B + A \otimes C$
(3) $(A \otimes B) \otimes C = A \otimes (B \otimes C)$
(4) $(A \otimes B)^T = A^T \otimes B^T$
(5) $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$
(6) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

The next theorem lays out the key bookkeeping relationship between tensor products and the vec operator.

Theorem 2.12. Bookkeeping Theorem If A, X, B are matrices conformable for multiplication, then

$$\operatorname{vec}(AXB) = \left(B^T \otimes A\right) \operatorname{vec} X.$$

Corollary 2.6. The following linear systems in the unknown X are equivalent.

(1) $A_1 X B_1 + A_2 X B_2 = C$
(2) $\left(\left(B_1^T \otimes A_1\right) + \left(B_2^T \otimes A_2\right)\right) \operatorname{vec} X = \operatorname{vec} C$

For Sylvester's equation, note that $AX + XB = AXI + IXB$.

The following is a very basic application of the tensor product. Suppose we wish to model a two-dimensional heat diffusion process on a flat plate that occupies the unit square in the xy-plane. We proceed as we did in the one-dimensional process described in Section 1.1. To fix ideas, we assume that the heat source is described by a function $f(x, y), 0 \le x \le 1, 0 \le y \le 1$, and that the temperature is held at 0 at the boundary of the unit square. Also, the conductivity coefficient is assumed to be the constant k. Cover the square with a uniformly spaced set of grid points $(x_i, y_j), 0 \le i, j \le n + 1$, called nodes, and assume that the spacing in each direction is a width $h = 1/(n + 1)$. Also

assume that the temperature function at the (i,j)th node is $u_{ij} = u(x_i, y_j)$ and that the source is $f_{ij} = f(x_i, y_j)$. Notice that the values of u on boundary grid points is set at 0. For example, $u_{01} = u_{20} = 0$. By balancing the heat flow in the horizontal and vertical directions, one arrives at a system of linear equations, one for each node, of the form

$$-u_{i-1,j} - u_{i+1,j} + 4u_{ij} - u_{i,j-1} - u_{i,j+1} = \frac{h^2}{k} f_{ij}, \quad i,j = 1,\ldots,n. \tag{2.11}$$

Observe that values of boundary nodes are zero, so these are not unknowns, which is why the indexing of the equations starts at 1 instead of 0. There are exactly as many equations as unknown grid point values. Each equation has a "molecule" associated with it that is obtained by circling the nodes that occur in the equation and connecting these circles. A picture of a few nodes is given in Figure 2.9.

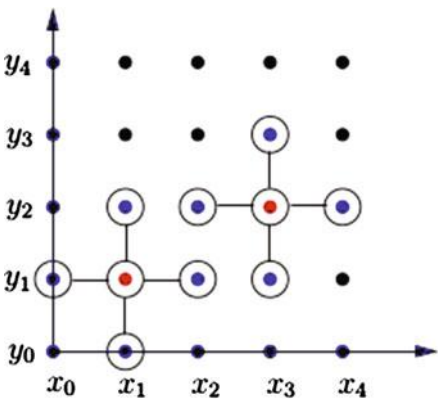


Fig. 2.9: Molecules for $(1,1)$th and $(3,2)$th grid points

Example 2.67. Set up and solve a system of equations for the two-dimensional heat diffusion problem described above.

Solution. Equation (2.11) gives us a system of n^2 equations in the n^2 unknowns u_{ij}, $i,j = 1,2,\ldots,n$. Rewrite equation (2.11) in the form

$$(-u_{i-1,j} + 2u_{ij} - u_{i+1,j}) + (-u_{i,j-1} + 2u_{ij} - u_{i,j+1}) = \frac{h^2}{k} f_{ij}.$$

Now form the $n \times n$ matrices

$$T_n = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}.$$

Set $U = [u_{ij}]$ and $F = [f_{ij}]$, and the system can be written in matrix form as

$$T_n U + U T_n = T_n U I_n + I_n U T_n = \frac{h^2}{k} F.$$

However, we can't as yet identify a coefficient matrix, which is where Corollary 2.6 comes in handy. Note that both I_n and T_n are symmetric and apply the corollary to obtain that the system has the form

$$(I_n \otimes T_n + T_n \otimes I_n) \operatorname{vec} U = \operatorname{vec} \frac{h^2}{k} F.$$

Now we have a coefficient matrix, and what's more, we have an automatic ordering of the doubly indexed variables u_{ij}, namely

$$u_{1,1}, u_{2,1}, \ldots, u_{n,1}, u_{1,2}, u_{2,2}, \ldots, u_{n,2}, \ldots, u_{1,n}, u_{2,n}, \ldots, u_{n,n}.$$

This is sometimes called the "row ordering," which refers to the rows of the nodes in Figure 2.9, and not the rows of the matrix U. □

Here is one more example of a problem in which tensor notation is an extremely helpful bookkeeper. This is a biological model that gives rise to an inverse theory problem. ("Here's the answer, what's the question?")

Example 2.68. Refer to Example 2.21, where a three-state insect (egg, juvenile, adult) is studied in stages spaced at intervals of two days. One might ask how the entries of the matrix were derived. Clearly, observation plays a role. Let us suppose that we have taken samples of the population at successive stages and recorded our estimates of the population state. Suppose we have estimates of states $\mathbf{x}^{(0)}$ through $\mathbf{x}^{(4)}$. How do we translate these observations into transition matrix entries?

Solution. We postulate that the correct transition matrix has the form

$$A = \begin{bmatrix} P_1 & 0 & F \\ G_1 & P_2 & 0 \\ 0 & G_2 & P_3 \end{bmatrix}.$$

Theoretically, we have the transition equation $\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}$ for $k = 0, 1, 2, 3$. Remember that this is an inverse problem, where the "answers," population states $\mathbf{x}^{(k)}$, are given, and the question "What is the transition matrix A?" is unknown. We could simply write out each transition equation and express the results as linear equations in the unknown entries of A. However, this is laborious and not practical for problems involving many states or larger amounts of data.

Here is a better idea: Assemble all of the transition equations into a single matrix equation by setting

$$M = \left[\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\right] = [m_{ij}] \quad \text{and} \quad N = \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}\right] = [n_{ij}].$$

The entire ensemble of transition equations becomes $AM = N$ with M and N known matrices and A the unknown. Here A is 3×3 and both M, N are 3×4. Next, write the transition equation as $I_3 AM = N$ and invoke the bookkeeping theorem to obtain the system

$$\operatorname{vec}(I_3 AM) = \left(M^T \otimes I_3\right) \operatorname{vec} A = \operatorname{vec} N.$$

This is a system of 12 equations in 9 unknowns. We can simplify it a bit by deleting the third, fourth, and eighth entries of $\operatorname{vec} A$ and the same columns of the coefficient matrix, since we know that the variables a_{31}, a_{12}, and a_{23} are zero. We thus end up with a system of 12 equations in 6 unknowns, which will determine the nonzero entries of A. □

2.7 Exercises and Problems

Exercise 1. Let $A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}$. Calculate the following.

(a) $A \otimes B$ (b)$B \otimes A$ (c)$A^{-1} \otimes B^{-1}$ (d)$(A \otimes B)^{-1}$

Exercise 2. Let $A = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 3 & -3 \\ 3 & 0 \end{bmatrix}$. Calculate the following.

(a) $A \otimes B$ (b)$B \otimes A$ (c)$A^T \otimes B^T$ (d)$(A \otimes B)^T$

Exercise 3. With A and B as in Exercise 1, $C = \begin{bmatrix} 2 & -1 \\ 1 & 0 \\ 1 & 3 \end{bmatrix}$, and $X = [x_{ij}]$ a 3×2 matrix of unknowns, use tensor products to determine the coefficient matrix of the linear system $AX + XB = C$ in matrix–vector form.

Exercise 4. Use the matrix A and methodology of Example 2.68 with $\mathbf{x}^{(0)} = (1, 2, 3)$, $\mathbf{x}^{(1)} = (0.9, 1.2, 3.6)$, and $\mathbf{x}^{(2)} = (1, 1.1, 3.4)$ to express the resulting system of equations in the six unknown nonzero entries of A in matrix–vector form.

*Problem 5. Verify parts (1) and (4) of Theorem 2.11.

Problem 6. Verify parts (5) and (6) of Theorem 2.11.

*Problem 7. Show that if A and B are square matrices of sizes m and n respectively, and if one of A and B is singular, then $|A \otimes B| = 0$.

Problem 8. Show that if A and B are square matrices of sizes m and n respectively, then $|A \otimes B| = |A|^n \, |B|^m$.

Problem 9. If heat is transported with a horizontal velocity v as well as diffused in Example 2.67, a new equation results at each node in the form

$$-u_{i-1,j} - u_{i+1,j} + 4u_{ij} - u_{i,j-1} - u_{i,j+1} - \frac{vh}{2k}\left(u_{i+1,j} - u_{i-1,j}\right) = \frac{h^2}{k} f_{ij}$$

for $i, j = 1, \ldots, n$. Vectorize the system and use tensor products to identify the coefficient matrix of this linear system.

*Problem 10. Prove the Bookkeeping Theorem (Theorem 2.12).

2.8 *Applications and Computational Notes

LU Factorization

Here is a problem: Suppose we want to solve a nonsingular linear system $Ax = b$ repeatedly, with different choices of b. A perfect example of this kind of situation is the heat flow problem Example 1.3, where the right-hand side is determined by the heat source term $f(x)$. Suppose that we need to experiment with different source terms. What happens if we use Gaussian or Gauss–Jordan elimination? Each time we carry out a complete calculation on the augmented matrix $\widetilde{A} = [A \,|\, b]$ we have to resolve the whole system. Yet, the main part of our work is the same: putting the part of $\widetilde{A}$ corresponding to the coefficient matrix A into reduced row echelon form. Changing the right-hand side has no effect on this work. What we want here is a way to somehow record our work on A, so that solving a new system involves very little additional work. This is exactly what the LU factorization is all about.

Definition 2.24. **LU Factorization** Let A be an $n \times n$ matrix. An LU factorization of A is a pair of $n \times n$ matrices L, U such that
(1) L is lower triangular.
(2) U is upper triangular.
(3) $A = LU$.

Even if we could find such beasts, what is so wonderful about them? The answer is that *triangular* systems $A\mathbf{x} = \mathbf{b}$ are easy to solve. For example, if A is upper triangular, we learned that the smart thing to do was to use the last equation to solve for the last variable, then the next-to-last equation for the next-to-last variable, etc. This is the secret of Gaussian elimination! But lower triangular systems are just as simple: Use the first equation to solve for the first variable, the second equation for the second variable, and so forth. Now suppose we want to solve $A\mathbf{x} = \mathbf{b}$ and we know that $A = LU$. The original system becomes $LU\mathbf{x} = \mathbf{b}$. Introduce an intermediate variable $\mathbf{y} = U\mathbf{x}$. Now perform these steps:

1. (Forward solve) Solve lower triangular system $L\mathbf{y} = \mathbf{b}$ for the variable $\mathbf{y}$.
2. (Back solve) Solve upper triangular system $U\mathbf{x} = \mathbf{y}$ for the variable $\mathbf{x}$.

This does it! Once we have the matrices L, U, we don't have to worry about right-hand sides, except for the small amount of work involved in solving two triangular systems. Notice that since A is assumed nonsingular, we have that if $A = LU$, then $\det A = \det L \det U \neq 0$. Therefore, neither triangular matrix L or U can have zeros on its diagonal. Thus, the forward and back solve steps can always be carried out to give a unique solution.

Example 2.69. You are given that

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -2 & 0 & -1 \\ 2 & 3 & -3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Use this fact to solve $Ax = b$ in the following cases:

(a) $\mathbf{b} = [1, 0, 1]^T$ (b) $\mathbf{b} = [-1, 2, 1]^T$

Solution. Set $\mathbf{x} = [x_1, x_2, x_3]^T$ and $\mathbf{y} = [y_1, y_2, y_3]^T$. For (a) forward solve

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

to get $y_1 = 1$, then $y_2 = 0 + 1y_1 = 1$, then $y_3 = 1 - 1y_1 - 2y_2 = -2$. Then back solve

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

to get $x_3 = -2/(-1) = 2$, then $x_2 = 1 + x_3 = 3$, then $x_1 = (1 - 1x_2)/2 = -1$.

For (b) forward solve

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$$

to get $y_1 = -1$, then $y_2 = 0 + 1y_1 = -1$, then $y_3 = 1 - 1y_1 - 2y_2 = 4$. Then back solve

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ 4 \end{bmatrix}$$

to get $x_3 = 4/(-1) = -4$, then $x_2 = 1 + x_3 = -3$, then $x_1 = (1 - 1x_2)/2 = 2$. □

Notice how simple the previous example was, if LU factorization is known. So how do we find such a factorization? In general, a nonsingular matrix may not have one. A good example is the matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. However, if Gaussian elimination can be performed on the matrix A *without row exchanges*, then such a factorization is really a by-product of Gaussian elimination. In this case let $[a_{ij}^{(k)}]$ be the matrix obtained from A after using the kth pivot to clear out entries below it (thus, $A = [a_{ij}^{(0)}]$). Remember that in Gaussian elimination we need only two types of elementary operations, namely row exchanges and adding a multiple of one row to another. Furthermore, the only elementary operations of the latter type that we use are of this form: $E_{ij}(-a_{ij}^{(k)}/a_{jj}^{(k)})$, where $[a_{ij}^{(k)}]$ is the matrix obtained from A from the various elementary operations up to this point.

Multipliers The numbers $m_{ij} = -a_{ij}^{(k)}/a_{jj}^{(k)}$, where $i > j$, are sometimes called *multipliers*. In the way of notation, let us call a triangular matrix a *unit* triangular matrix if its diagonal entries are all 1's.

Theorem 2.13. If Gaussian elimination is used without row exchanges on the nonsingular matrix A, resulting in the upper triangular matrix U, and if L is the unit lower triangular matrix whose entries below the diagonal are the negatives of the multipliers m_{ij}, then $A = LU$.

Proof. The proof of this theorem amounts to noticing that the product of all the elementary operations that reduces A to U is a unit lower triangular matrix $\widetilde{L}$ with the multipliers m_{ij} in the appropriate positions. Thus, $\widetilde{L}A = U$. To undo these operations, multiply by a matrix L with the negatives of the multipliers in the appropriate positions. This results in

$$L\widetilde{L}A = A = LU$$

as desired. □

The following example shows how one can write an efficient program to implement LU factorization. The idea is this: As we do Gaussian elimination, the U part of the factorization gradually appears in the upper parts of the transformed matrices $A^{(k)}$. Below the diagonal we replace nonzero entries with zeros, column by column. Instead of wasting this space, use it to store the negative of the multipliers in place of the element it zeros out. Of course, this storage part of the matrix should not be changed by subsequent elementary row operations. When we are finished with elimination, the diagonal and upper part of the resulting matrix is just U, and the strictly lower triangular part on the unit lower triangular matrix L is stored in the lower part of the matrix.

Example 2.70. Use the shorthand of the preceding discussion to compute an LU factorization for

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -2 & 0 & -1 \\ 2 & 3 & -3 \end{bmatrix}.$$

Solution. Proceed as in Gaussian elimination, but store negative multipliers:

$$\begin{bmatrix} \textcircled{2} & 1 & 0 \\ -2 & 0 & -1 \\ 2 & 3 & -3 \end{bmatrix} \xrightarrow[E_{31}(-1)]{E_{21}(1)} \begin{bmatrix} 2 & 1 & 0 \\ -1 & \textcircled{1} & -1 \\ 1 & 2 & -3 \end{bmatrix} \xrightarrow{E_{32}(-2)} \begin{bmatrix} 2 & 1 & 0 \\ -1 & 1 & -1 \\ 1 & 2 & -1 \end{bmatrix}.$$

Now we read off the results from the last matrix:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix}.$$

□

What can be said if row exchanges are required (for example, we might want to use a partial pivoting strategy)? Take the point of view that we could see our way to the end of Gaussian elimination and store the product P of all row-exchanging elementary operations that we use along the way. A product of such matrices is called a *permutation matrix*; such a matrix is invertible, since it is a product of invertible matrices. Thus, if we apply the correct permutation matrix P to A we obtain a matrix for which Gaussian elimination will succeed without further row exchanges. Consequently, we have a theorem that applies to all nonsingular matrices. Notice that it does not limit the usefulness of LU factorization since the linear system $Ax = b$ is equivalent to the system $PAx = Pb$. The following theorem could be called the "PLU factorization theorem."

Permutation Matrix

Theorem 2.14. If A is a nonsingular matrix, then there exists a permutation matrix P, upper triangular matrix U, and unit lower triangular matrix L such that $PA = LU$.

There are many other useful factorizations of matrices that numerical analysts have studied, e.g., LDU and Cholesky. We will stop at LU, but there is one last point we want to make. Recall that a "flop" in numerical linear algebra is a single addition or subtraction, or multiplication or division. The amount of work in finding the LU factorization is the same as Gaussian elimination itself, which is approximately $2n^3/3$ flops (see Section 1.5). The additional work of back and forward solving is about $2n^2$ flops. So the dominant amount of work is done by computing the factorization rather than the back and forward solving stages.

Efficiency of Determinants and Cramer's Rule in Computation

The truth of the matter is that Cramer's Rule and adjoints are good only for small matrices and theoretical arguments. For if you evaluate determinants in a straightforward way from the definition, the work in doing so is about $n \cdot n!$ flops for an $n \times n$ system. (Recall that a "flop" in numerical linear algebra is a single addition or subtraction, or multiplication or division.) For example, it is not hard to show that the operation of adding a multiple of one row vector of length n to another requires $2n$ flops. This number $n \cdot n!$ is vast when compared to the number $2n^3/3$ flops required for Gaussian elimination, even with "small" n, say $n = 10$. In this case we have $2 \cdot 10^3/3 \approx 667$, while $10 \cdot 10! = 36,288,000$.

Computational Efficiency of Determinants

On the other hand, there is a clever way to evaluate determinants that requires much less work than the definition: Use elementary row operations together with D2, D6, and the elementary operations that correspond to these rules to reduce the determinant to that of a triangular matrix. This requires about $2n^3/3$ flops. As a matter of fact, it is tantamount to Gaussian elimination. But to use Cramer's Rule,

you will have to calculate $n+1$ determinants. So why bother with Cramer's Rule on larger problems when it still will take about n times as much work as Gaussian elimination? A similar remark applies to computing adjoints instead of using Gauss–Jordan elimination on the superaugmented matrix of A.

Digital Signal Processing

As an introduction to the field of digital signal processing (DSP), let us reconsider the nonhomogeneous constant coefficient difference equation in a somewhat different format:

$$y_k = a_0 x_k + a_1 x_{k-1} + \cdots + a_m x_{k-m}, \quad k = m, m+1, m+2, \ldots, \tag{2.12}$$

where a_0 and a_m are nonzero. Rather than treating the x_k's as unknowns, view them as inputs and the resulting values y_k as outputs. More specifically, we think of the sequence $x_0, x_1, \ldots$ as a sampling of a continuous variable such as sound in a time domain or an image in a spatial domain. In this setting we think of equation (2.12) as a *linear digital filter* of length m and the resulting sequence of y_k's as the filtered data.

Digital Filter

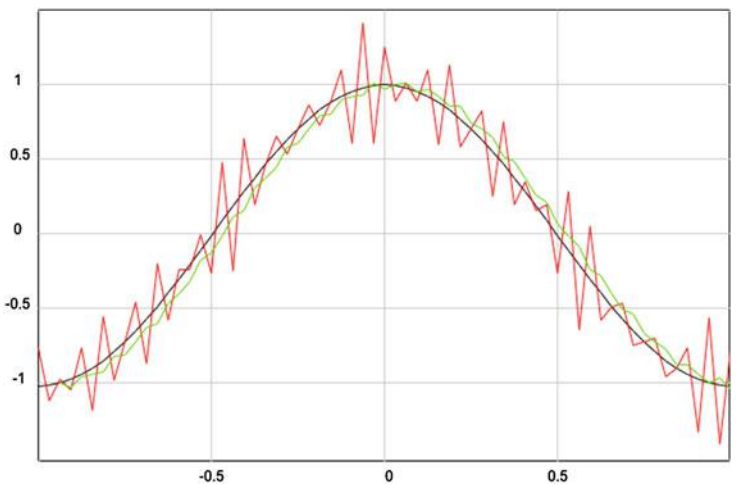


Fig. 2.10: Graph of data from Example 2.71: Exact (——), noisy (——) and filtered (——)

As a simple example, consider a continuous function $y = f(t)$ over the domain $-1 \leq t \leq 1$. Here the variable t could represent time or space. When you see a visual representation of this function (a graph of it) what you are really seeing is a discrete sampling of values of this continuous points at some resolution – a dot-to-dot so to speak.

Example 2.71. Consider the function $f(t) = \cos(\pi t)$, $-1 \le t \le 1$ which defines the exact signal that we want to sample. Suppose that what we actually sample is this signal plus noise, namely the function $g(t) = \cos(\pi t) + \frac{1}{5}\sin(24\pi t) + \frac{1}{4}\cos(30\pi t)$. Note that the signal $f(t)$ is the low frequency portion of $g(t)$ and the noise is the high frequency portion of $g(t)$. Suppose further that sampling is at the equally spaced points $t_k = -1 + \frac{2}{64}k$, $k = 0, 1, \ldots, 64$, yielding data points $x_k = g(t_k)$. We apply the following length two filter to the data:

$$y_k = \frac{1}{4}x_k + \frac{1}{2}x_{k-1} + \frac{1}{4}x_{k-2}, \quad k = 2, 3, \ldots, 64.$$

How effective is this filter in removing noise?

Solution. Rather than list the resulting numbers let's calculate and graph them. We shall interpret the number y_k as the filtered value of the noisy $x_k = g(t_k)$, $k = 2, 3, \ldots 64$ and therefore the approximation to $f(t_k)$ that results from this filtering. A graph of the exact data, noisy data and filtered data is given in Figure 2.10. Although it is somewhat crude (reliance on earlier values causes a slight forward shift in the filtered values), it appears to do a decent job of filtering out the noise in the sampled signal $g(t)$. □

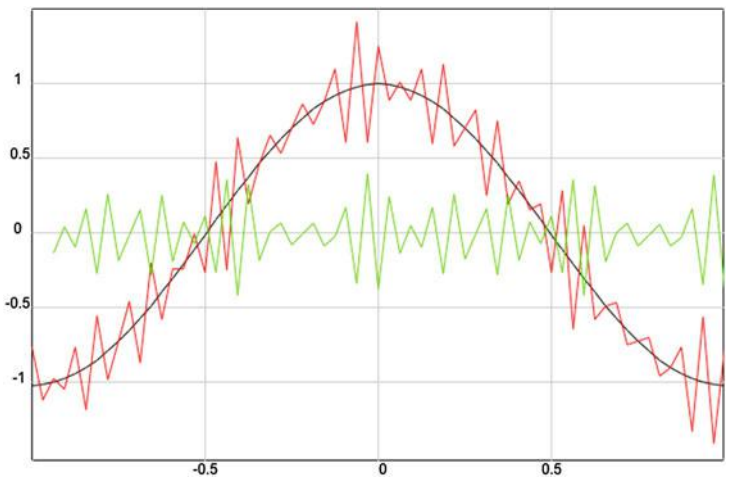


Fig. 2.11: Graph of data from Example 2.72: Exact (——), noisy (——) and filtered (——)

The filter of Example 2.71 is both *causal* (filtered data at a specific time only depends on samples from earlier times) and *low-pass* (filters out high frequency data but preserves low frequency data).

Causal and Low Pass Filter

Example 2.72. Apply the following length two filter to the data of Example 2.71:

$$y_k = -\frac{1}{4}x_k + \frac{1}{2}x_{k-1} - \frac{1}{4}x_{k-2}, \quad k = 2, 3, \ldots, 64.$$

What is the effect of this filter?

Solution. As in Example 2.71 we calculate and graph the resulting numbers. A graph of the true data, noisy data and filtered data is given in Figure 2.70. Evidently the effect of this filter is to filter out the low frequency portion of the sampled function $g(t)$, which is exactly the opposite of the lowpass filter of Example 2.71. □

Causal and High Pass Filter

The filter of Example 2.72 is both causal (filtered data at a specific time only depends on samples from earlier times) and *high-pass* (filters out low frequency data but preserves high frequency data). We will examine both low and high pass filters again in Chapters 4 and 6.

IsoRank

Consider this problem: Given two networks graphs, how can we compare them? Do they have similar subgraphs? If we are able to map one graph perfectly in an edge preserving fashion onto a copy of it contained in the other, we could say that we have perfect similarity between one graph and a subgraph of the other. But what if we can only find imperfect mappings of one into the other? How can we find a "best" match in some sense or other? This idea has many important applications in chemistry, network analysis, biology and bioinformatics among others. For example, one could compare certain gene sequences or PPIs (protein-protein interactions) between species (see [13] and [21] for more details.) Several different technologies have been developed to explore these problems, e.g., GeneRank, ProteinRank and IsoRank. All of these are ultimately special cases of PageRank. We shall introduce IsoRank by way a fairly simple example.

Example 2.73. What similarities can be found between the two graphs of Figure 2.12?

Solution. The two graphs are G_1 and G_2. Note that G_1 is extremely simple in structure and a visual inspection shows that G_1 is a subgraph of G_2 in several ways: It can be identified with the vertices A, B, D in pretty much any order, and similarly with the nodes B, C, D. However, the vertex E should not be identified with any part of the graph G_1. □

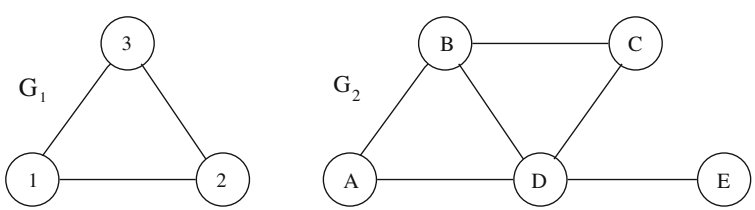


Fig. 2.12: Two graphs for comparison with IsoRank

The key question we want to answer here is: How can we design algorithms that can "see" similarities between portions of graphs such as we have observed in the previous example? To answer this question we shall use the following definition, which can be applied to a pair of graphs:

Definition 2.25. Tensor Product of Graphs Let G_i be graphs with vertex set V_i and edge set E_i, $i = 1, 2$. Then the tensor product of the two graphs is the graph $G = G_1 \times G_2$ with vertex set V and edge set E, where

$$V = \{(u, v) \mid u \in V_1 \text{ and } v \in V_2\}$$
$$E = \{\{\{u_1, v_1\}, \{u_2, v_2\}\} \mid \{u_1, u_2\} \in E_1 \text{ and } \{v_1, v_2\} \in E_2\}.$$

The idea behind this definition is that an edge between pairs of vertices in the two graphs will exist only if the corresponding vertices in each graph are themselves connected by an edge. Thus, it is a way of matching edges between the two graphs.

Next, we will assume that we have a surfing matrix for each graph. How do we construct a surfing matrix for their tensor product? In order to fix ideas, we need the PageRank matrices from the graphs of Figure 2.12. Note that there are no dangling nodes in any graph that is connected. Hence, there is no need for a correction vector for this problem. We can construct surfing matrices for each graph by inspection: Count the number of links n_j out of vertex v_j and give each target vertex v_i a probability $1/n_j$ of being reached from v_j. In other words, the (i, j)th entry of the surfing matrix is $1/n_j$. What results is that the surfing matrices $P = [p_{ij}]$ for G_1 and $Q = [q_{ij}]$ for G_2, where

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \text{ and } Q = \begin{bmatrix} 0 & \frac{1}{3} & 0 & \frac{1}{4} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 \\ 0 & 0 & 0 & \frac{1}{4} & 0 \end{bmatrix}.$$

Here the ordering of vertices is $1, 2, 3$, (also labeled as u_1, u_2, u_3, resp.) for G_1. For G_2 we use the ordering A, B, C, D, E (also labeled as v_1, v_2, v_3, v_4, v_5, resp.). The matter of ordering is significant because different orderings will result in different surfing matrices.

So how should we order the vertices of $G = G_1 \times G_2$ in our example? Two natural choices are to treat the sequence much like a double sum: For each vertex in the outer graph, cycle over the vertices in the inner graph. If we treat G_1 as inner and G_2 as outer, the ordering looks like

$$(u_1, v_1), (u_2, v_1), (u_3, v_1), (u_1, v_2), (u_2, v_2), (u_3, v_2), (u_1, v_3), (u_2, v_3), (u_3, v_3),$$
$$(u_1, v_4), (u_2, v_4), (u_3, v_4), (u_1, v_5), (u_2, v_5), (u_3, v_5),$$

while if we treat G_1 as outer and G_2 as inner, the ordering looks like

$$(u_1,v_1),(u_1,v_2),(u_1,v_3),(u_1,v_4),(u_1,v_5),(u_2,v_1),(u_2,v_2),(u_2,v_3),(u_2,v_4),$$
$$(u_2,v_5),(u_3,v_1),(u_3,v_2),(u_3,v_3),(u_3,v_4),(u_3,v_5).$$

A drawing of the graph G is too complicated to be helpful. Rather, we will find the surfing matrix with one of the ordering of vertices specified above. So the question to be answered is: Given a vertex (u_i,v_j) of G, what is the probability of it transitioning to the vertex (u_k,v_ℓ)? The answer is surprisingly simple: In order for this to happen, u_i must transition to u_k and v_j must transition to v_ℓ. However, these events are entirely independent of each other, so the probability of both happening is the product of probabilities of each happening, that is, $p_{ik}q_{j\ell}$.

Unfortunately, these indices alone are not sufficient to describe a surfing matrix for G. What is required is a specific ordering of the vertices of G. So let us consider the first ordering described above (G_1 inner, G_2 outer). The result is the following matrix:

$$S=\begin{bmatrix}
0&0&0&0&\frac16&\frac16&0&0&0&0&\frac18&\frac18&0&0&0\\
0&0&0&\frac16&0&\frac16&0&0&0&\frac18&0&\frac18&0&0&0\\
0&0&0&\frac16&\frac16&0&0&0&0&\frac18&\frac18&0&0&0&0\\
0&\frac14&\frac14&0&0&0&\frac14&\frac14&0&\frac18&\frac18&0&0&0&0\\
\frac14&0&\frac14&0&0&0&\frac14&0&\frac14&\frac18&0&\frac18&0&0&0\\
\frac14&\frac14&0&0&0&0&\frac14&\frac14&0&\frac18&\frac18&0&0&0&0\\
0&0&0&0&\frac16&\frac16&0&0&0&0&\frac18&\frac18&0&0&0\\
0&0&0&\frac16&0&\frac16&0&0&0&\frac18&0&\frac18&0&0&0\\
0&0&0&\frac16&\frac16&0&0&0&0&\frac18&\frac18&0&0&0&0\\
0&\frac14&\frac14&0&\frac16&\frac16&0&\frac14&\frac14&0&0&0&0&\frac12&\frac12\\
\frac14&0&\frac14&\frac16&0&\frac16&\frac14&0&\frac14&0&0&0&\frac12&0&\frac12\\
\frac14&\frac14&0&\frac16&\frac16&0&\frac14&\frac14&0&0&0&0&\frac12&\frac12&0\\
0&0&0&0&0&0&0&0&0&0&\frac18&\frac18&0&0&0\\
0&0&0&0&0&0&0&0&0&\frac18&0&\frac18&0&0&0\\
0&0&0&0&0&0&0&0&0&\frac18&\frac18&0&0&0&0
\end{bmatrix}$$

It would be impractical to construct all such surfing matrices on an ad hoc basis, so let us examine the entry formula $p_{ik}q_{j\ell}$ more closely. Note that our selected ordering of vertices of G occurs in successive blocks which means that the (row, column) indices of the blocks in S are (k,l), $k,\ell=1,2,3,4,5$. Within each block the (row, column) indices are (i,j), $i,j=1,2,3$ and $q_{j\ell}$ is fixed while p_{ij} ranges over the entries of P. Thus, the matrix S takes the form

$$S=\begin{bmatrix}
q_{1,1}P & q_{1,2}P & q_{1,3}P & q_{1,4}P & q_{1,5}P\\
q_{2,1}P & q_{2,2}P & q_{2,3}P & q_{2,4}P & q_{2,5}P\\
q_{3,1}P & q_{3,2}P & q_{3,3}P & q_{3,4}P & q_{3,5}P\\
q_{4,1}P & q_{4,2}P & q_{4,3}P & q_{4,4}P & q_{4,5}P\\
q_{5,1}P & q_{5,2}P & q_{5,3}P & q_{5,4}P & q_{5,5}P
\end{bmatrix}=Q\otimes P.$$

Had we chosen the second ordering (G_1 outer, G_2 inner) the resulting surfing matrix would have been $P\otimes Q$. These arguments are easily abstracted to yield the following theorem:

Theorem 2.15. Let G_1, G_2 be graphs with stochastic surfing matrices P, Q, resp. Then with a suitable block ordering of vertices, $P \otimes Q$ or $Q \otimes P$ is a surfing matrix for the graph $G_1 \times G_2$.

As an aside it is interesting to note that this implies that $P \otimes Q$ and $Q \otimes P$ are stochastic matrices. In fact this is true for any stochastic matrices P and Q, whether they are connected to any graph or not. We leave the proof of this as an exercise.

Returning to Example 2.73, we now have that $Q \otimes P$ is a surfing matrix for the tensor product $G = G_1 \times G_2$ so that we can apply the PageRank methodology to this graph. Set $\mathbf{e} = (1, 1, \ldots, 1)$ and define teleportation vector $\mathbf{v} = \mathbf{e}/15$ and teleportation parameter $\alpha = 0.85$. We can compute (with a suitable technology tool such as ALAMA calculator) the unique stationary vector $\mathbf{x}$ obtained by solving the system

$$(I - \alpha Q \otimes P)\,\mathbf{x} = (1 - \alpha)\mathbf{v}.$$

The G_1 inner, G_2 outer ordering suggests a very convenient way of displaying $\mathbf{x}$, namely to assemble the blocks of three into a 3×5 matrix with rows indexed by vertices of G_1 and columns by vertices of G_2. Here is the resulting matrix, rounded to three decimal places:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0.056 & 0.08 & 0.056 & 0.108 & 0.033 \\ 0.056 & 0.08 & 0.056 & 0.108 & 0.033 \\ 0.056 & 0.08 & 0.056 & 0.108 & 0.033 \end{bmatrix} \end{array}$$

What this table tells us is that the most likely (and therefore best) matching between vertices of G_1 and G_2 is to match any one of $1, 2, 3$ of G_1 with vertex D of G_2. This gives three possibilities. For example, let's match 2 to C. What remains in the table is

$$\begin{array}{c} \\ 1 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & E \end{array} \\ \begin{bmatrix} 0.056 & 0.08 & 0.056 & 0.033 \\ 0.056 & 0.08 & 0.056 & 0.033 \end{bmatrix} \end{array}.$$

This tells us that the best match for 1 or 3 is B, giving us a total of six possibilities. Match 1 to B and what remains in the table is

$$\begin{array}{c} \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} A & C & E \end{array} \\ \begin{bmatrix} 0.056 & 0.056 & 0.033 \end{bmatrix} \end{array}.$$

This tells us that the best match for 3 is A or C, which gives a total of 12 possibilities, namely, any permutation of B, D, A or of B, D, C. This is exactly what visual examination of the two graphs shows us.

We have introduced IsoRank in the context of graphs, but it can also work for digraphs with a few changes. First, for digraphs modify Definition 2.25 by

changing edges from unordered pairs $\{u, v\}$ to ordered pairs (u, v). Second, the construction of a suitable surfing matrix for both digraphs must be constructed from the adjacency matrices of these graphs via Theorem 2.7 and the use of correction vectors to handle dangling nodes.

2.8 Exercises and Problems

Exercise 1. Show that $L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix}$ and $U = \begin{bmatrix} 2 & -1 & 1 \\ 0 & 4 & -3 \\ 0 & 0 & -1 \end{bmatrix}$ is an LU factorization of $A = \begin{bmatrix} 2 & -1 & 1 \\ 2 & 3 & -2 \\ 4 & 2 & -2 \end{bmatrix}$.

Exercise 2. Show that $P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$, $L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}$ and $U = \begin{bmatrix} 4 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$ is a PLU factorization of $A = \begin{bmatrix} 0 & -1 & 1 \\ 2 & 3 & -2 \\ 4 & 2 & -2 \end{bmatrix}$.

Exercise 3. Use the LU factorization of Exercise 1 to solve $A\mathbf{x} = \mathbf{b}$, where (a) $\mathbf{b} = (6, -8, -4)$ (b) $\mathbf{b} = (2, -1, 2)$ (c) $\mathbf{b} = (1, 2, 4)$) (d) $\mathbf{b} = (1, 1, 1)$.

Exercise 4. Use the PLU factorization of Exercise 2 to solve $A\mathbf{x} = \mathbf{b}$, where (a) $\mathbf{b} = (3, 1, 4)$ (b) $\mathbf{b} = (2, -1, 3)$ (c) $\mathbf{b} = (1, 2, 0)$) (d) $\mathbf{b} = (1, 0, 0)$.

Exercise 5. Find an LU factorization of the matrix $A = \begin{bmatrix} 2 & 1 & 0 \\ -4 & -1 & -1 \\ 2 & 3 & -3 \end{bmatrix}$.

Exercise 6. Find a PLU factorization of the matrix $A = \begin{bmatrix} 2 & 1 & 3 \\ -4 & -2 & -1 \\ 2 & 3 & -3 \end{bmatrix}$.

*Problem 7. Show that if A is a nonsingular matrix with a zero $(1,1)$th entry, then A does not have an LU factorization.

*Problem 8. Repeat the IsoRank calculation for Example 2.73 with teleportation parameter $\alpha = 0.5$. Do you obtain the same embeddings as with $\alpha = 0.85$?

Problem 9. Apply the following digital filter to the noisy data of Example 2.71 and graph the results. Does it appear to be a low pass filter?

$$y_k = \frac{1}{2}x_k + \frac{1}{2}x_{k-1}, \quad k = 1, 2, \ldots, 33$$

Problem 10. Apply the following digital filter to the noisy data of Example 2.71 and graph the results. Does it appear to be a high pass filter?

$$y_k = \frac{1}{2}x_k - \frac{1}{2}x_{k-1}, \quad k = 1, 2, \ldots, 33$$

Problem 11. Show that the tensor product of any two stochastic matrices is itself stochastic.

2.9 *Projects and Reports

Project: LU Factorization

Write a program module that implements Theorem 2.14 using partial pivoting and implicit row exchanges. This means that space is allocated for the $n \times n$ matrix $A = [a[i, j]]$ and an array of row indices, say indx$[i]$. Initially, indx should consist of the integers $1, 2, \ldots, n$. Whenever two rows need to be exchanged, say the first and third, then the indices indx[1] and indx[3] are exchanged. References to array elements throughout the Gaussian elimination process should be indirect: Refer to the $(1, 4)$th entry of A as the element $a\,[\text{indx}[1], 4]$. This method of reference has the same effect as physically exchanging rows, but without the work. It also has the appealing feature that we can design the algorithm as though no row exchanges have taken place provided we replace the direct reference $a[i, j]$ by the indirect reference $a[\text{indx}[i], j]$. The module should return the lower/upper matrix in the format of Example 2.70 as well as the permuted array indx$[i]$. Effectively, this index array tells the user what the permutation matrix P is.

Use this module to implement an LU system solver module that uses the LU factorization to solve a general linear system. Also write a module that finds the inverse of an $n \times n$ matrix A by first using the LU factorization module, then making repeated use of the LU system solver to solve $A\mathbf{x}^{(i)} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the ith column of the identity. Then we will have

$$A^{-1} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}].$$

Be sure to document and test your code and report on the results.

Project: Markov Chains

Refer to Example 2.19 and Section 2.3 for background. Three automobile insurance firms compete for a fixed market of customers. Annual premiums are sold to these customers. Label the companies A, B, and C. You work for Company A, and your team of market analysts has done a survey that draws the following conclusions: In each of the past three years, the number of A customers switching to B is 20%, and to C is 30%. The number of B customers switching to A is 20%, and to C is 20%. The number of C customers switching to A is 30%, and to B is 10%. Those who do not switch continue to use their current company's insurance for the next year. Model this market as a Markov chain. Display the transition matrix for the model. Illustrate the workings of the model by showing what it would predict as the market shares three years from now if currently A, B, and C owned equal shares of the market.

The next part of your problem is as follows: Your team has tested two advertising campaigns in some smaller test markets and are confident that the first campaign will convince 20% of the B customers who would otherwise stay with B in a given year to switch to A. The second advertising campaign would convince 20% of the C customers who would otherwise stay with C in a given year to switch to A. Both campaigns have about equal costs and

would not change other customers' habits. Make a recommendation, based on your experiments with various possible initial state vectors for the market. Will these campaigns actually improve your company's market share? If so, which one do you recommend? Write up your recommendation in the form of a report, with supporting evidence. It's a good idea to hedge on your bets a little by pointing out limitations to your model and claims, so devote a few sentences to those points.

It would be a plus to carry the analysis further (your manager might appreciate that). For instance, you could turn the additional market share from, say B customers, into a variable and plot the long-term gain for your company against this variable. A manager could use this data to decide whether it was worthwhile to attempt gaining more customers from B.

Project: Affine Transforms in Real-Time Rendering

Refer to the examples in Section 2.3 for background. Graphics specialists find it important to distinguish between vector objects and point objects in three-dimensional space. They simultaneously manipulate these two kinds of objects with invertible linear operators, which they term *transforms*. To this end, they use the following clever ruse: Identify three-dimensional vectors and points in the usual way, that is, by their coordinates x_1, x_2, x_3. To distinguish between the two, embed them in the set of 4×1 vectors $\mathbf{x} = (x_1, x_2, x_3, x_4)$, called *homogeneous vectors*, with the understanding that if $x_4 = 0$, then $\mathbf{x}$ represents a three-dimensional vector object, and if $x_4 \neq 0$, then the vector represents a three-dimensional point with coordinates $\frac{x_1}{x_4}, \frac{x_2}{x_4}, \frac{x_3}{x_4}$.

Homogeneous Vector

Transforms (invertible linear operators) have the general form

$$T_M(\mathbf{x}) = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}.$$

If $m_{44} = 1$ and the remaining entries of the last row and column are zero, the transform is called a *homogeneous transform*. If $m_{44} = 1$ and the remaining entries of the last row are zero, the transform is called *affine*. If the transform matrix M takes the block form $M = \begin{bmatrix} I_3 & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$, the transform T_M is called a *translation* by the vector $\mathbf{t}$. All other operators are called *nonaffine*.

Homogeneous and Affine Transforms

In real-time rendering it is sometimes necessary to invert an affine transform. Computational efficiency is paramount in these calculations (after all, this is real time!). So your objective in this project is to design an algorithm that accomplishes this inversion with a minimum number of flops. Preface discussion of your algorithm with a description of affine transforms. Give a geometrical explanation of what homogeneous and translation transforms do to vectors and points. You might also find it helpful to show that every affine transform is the composition of a homogeneous and a translation transform.

Illustrate the algorithm with a few examples. Finally, discuss the stability of your algorithm. Could it be a problem? If so, how would you remedy it? See the discussion of roundoff error in Section 1.5.

Project: PageRank as Embedding Tool

Instructors: For simpler projects assign fewer of the tasks below.

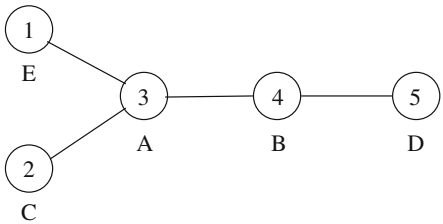


Fig. 2.13: Isomorphic relabeling of G_1: $1, 2, 3, 4, 5$ to G_2: A,B,C,D,E

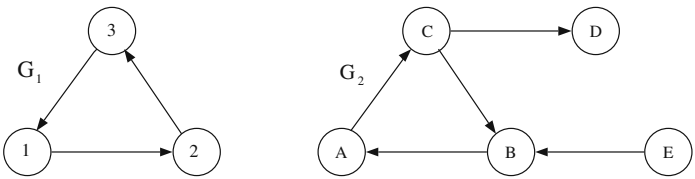


Fig. 2.14: Embedding examples

As we have seen, the notion of embedding one graph into another is a useful idea for some scientific studies. In this report you will test the basic idea of network embedding by using the variant IsoRank of the PageRank technique on three relatively simple examples. This project requires a technology tool for these calculations and the resulting output should be interpreted as in the discussion of the IsoRank technique following Example 2.73 of Section 2.8.

By an *isomorphism of graphs* we mean a one-to-one edge preserving map of vertices from one graph onto another. One can think of an isomorphism as simply a relabeling of the vertices of a graph. The first test is to provide an example of how well IsoRank can recognize isomorphisms. Consider the graph of Figure 2.13. Let G_1 be the graph with vertices $1, 2, 3, 4, 5$ in that order and G_2 the same graph with vertices A, B, C, D, E in that order. Apply IsoRank to these two graphs and discuss the validity of your results.

The next embedding test is to remove the edge connecting vertices B and C in Figure 2.12 and use IsoRank with teleportation vector $\mathbf{v} = \mathbf{e}/15$ and teleportation parameter $\alpha = 0.85$ to find the best matchings of the graph G_1 with the resulting graph G_2. List all possible mappings that are calculated.

The last embedding test is to use IsoRank with teleportation vector $\mathbf{v} = \mathbf{e}/15$ and teleportation parameter $\alpha = 0.85$, along with correction vector $\mathbf{u} = \mathbf{e}/5$ for G_2, to find the best matchings of the digraph G_1 with the digraph G_2 in Figure 2.14. List all possible mappings and discuss your calculations.

Report: Team Ranking in Sports

Instructors: For simpler projects assign fewer of the tasks below. Also, you could have students collect and use data from a sport of your choice.
Refer to Example 2.24 and Section 2.3 for background. As a sports analyst you are given the following data about a league of seven teams numbered 1–7, where the pair (j, k) represents a game in which team j defeated team k:

$$E = \{(1,2),(7,3),(2,4),(4,5),(3,2),(5,1),(6,1),(3,1),(7,2),(2,6),\\ (3,4),(7,4),(5,7),(6,4),(3,5),(5,6),(7,1),(5,2),(7,6),(1,4),(6,3)\}.$$

Based on these data you are to rank the teams. To this end, begin with the simplest method, ranking by win/loss record. Next, treat the data as defining a digraph. Begin this analysis by constructing the adjacency matrix of this digraph and drawing a picture of the digraph either by hand or using some software. Then rank the teams by using the following methods: First use the method of Example 2.26 to find a power ranking of each team. Then use the reverse PageRank idea of Example 2.47 to rank the the teams.

Next, suppose you are given additional information, namely, the game margins (winning score minus losing score) for each game. Following is a list of these margins matching the order of matches in the definition of E:

$$M = \{4,8,7,3,7,7,23,15,6,18,13,14,7,13,7,18,45,10,19,14,13\}\,.$$

In order to utilize these data examine your picture of the digraph and label each edge with the margin that matches it in M. You are now dealing with a weighted graph and one can construct a different sort of "adjacency matrix" by entering this margin in the (i, j)th entry according as team i defeated team j by that margin. Use this approach to calculate "power ranking".

As a last method of ranking, use PageRank on the reverse weighted digraph just as it is used with the unweighted digraph. Discuss your results, compare rankings and give reasons why you might prefer one over the other.

3

VECTOR SPACES

It is hard to overstate the importance of the idea of a vector space, a concept that has found application in mathematics, engineering, physics, chemistry, biology, the social sciences, and other areas. What we encounter is an abstraction of the idea of vector space that we studied in calculus or high school geometry. These "geometrical vectors" can easily be visualized. In this chapter, abstraction will come in two waves. The first wave, which could properly be called *generalization*, consists in generalizing the familiar ideas of geometrical vectors of calculus to vectors of size greater than three.

The second wave, *abstraction*, consists in abstracting the vector idea to entirely different kinds of objects. Abstraction can sometimes be difficult. For some, the study of abstract ideas is its own reward. For others, the natural reaction is to expect some payoff for the extra effort required to master abstraction. In the case of vector spaces we are happy to report that both kinds of students will be satisfied: Vector space theory really is a thing of beauty in itself and there is indeed a payoff for its study. It is a practical tool that enables us to understand phenomena that would otherwise escape our comprehension. Examples abound: The theory will be used in network analysis, for "best" solutions to an inconsistent system (least squares), for studying functions as systems of vectors, for establishing basic theory in linear programming, and to obtain new perspectives on our old friend $A\mathbf{x} = \mathbf{b}$.

3.1 Definitions and Basic Concepts

Generalization

We begin with the most concrete form of vector spaces, one that is closely in tune with what we learned when we were first introduced to two- and three-dimensional vectors using real numbers as scalars. However, we have seen that the complex numbers are a perfectly legitimate and useful field of numbers to work with. Therefore, our concept of a vector space must include

the selection of a field of scalars. The requirements for such a field are that it have binary operations of addition and multiplication that satisfy the usual arithmetic laws: Both operations are closed, commutative, and associative; have identities and satisfy distributive laws. And there exist additive inverses for all elements and multiplicative inverses for nonzero elements. Although other fields are possible, for our purposes the only fields of scalars are $\mathbb{F} = \mathbb{R}$ and $\mathbb{F} = \mathbb{C}$. Unless there is some indication to the contrary, the field of scalars will be assumed to be the default, the real numbers $\mathbb{R}$. However it should be noted that there are other fields of importance such as $\mathbb{Q}$, the field of rational numbers, or the finite field $\mathbb{F}_q$ of integers modulo p, where p is a prime number. The latter has significant applications in coding theory and cryptology.

A formal definition of vector space will come later. For now we describe a "vector space" over a field of scalars $\mathbb{F}$ as a nonempty set V of vectors of the same size, together with the binary operations of scalar multiplication and vector addition, subject to the following laws: For all vectors $\mathbf{u}, \mathbf{v} \in V$ and scalars $a \in \mathbb{F}$, (a) (Closure of vector addition) $\mathbf{u} + \mathbf{v} \in V$. (b) (Closure of scalar multiplication) $a\mathbf{v} \in V$. For vectors $\mathbf{u}, \mathbf{v}$, we define $-\mathbf{u} = (-1)\mathbf{u}$ and $\mathbf{u} - \mathbf{v} = \mathbf{u} + (-\mathbf{v})$.

Vector Negatives and Subtraction

Very simple examples are $\mathbb{R}^2$ and $\mathbb{R}^3$, which we discuss below. Another is any line through the origin in $\mathbb{R}^2$, which takes the form $V = \{c\,(x_0, y_0) \mid c \in \mathbb{R}\}$.

Geometrical vector spaces. We may have already seen the vector idea in geometry or calculus. In those contexts, a vector was supposed to represent a direction and a magnitude in two- or three-dimensional space, which is not the same thing as a point, that is, location in space. At first, one had to deal with these intuitive definitions until they could be turned into something more explicitly computational, namely the displacements of a vector in coordinate directions. This led to the following two vector spaces over the field of real numbers:

Geometrical Vectors

$$\mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}\,,$$
$$\mathbb{R}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{R}\}\,.$$

The distinction between vector spaces and points becomes a little hazy here. Once we have set up a coordinate system, we can identify each point in two- or three-dimensional space with its coordinates, which we write in the form of a tuple, i.e., a vector. The arithmetic of these two vector spaces is just the standard coordinatewise vector addition and scalar multiplication. One can visualize the direction represented by a vector (x, y) by drawing an arrow, i.e., directed line segment, from the origin of the coordinate system to the point with coordinates (x, y). The magnitude of this vector is the length of the arrow, which is just $\sqrt{x^2 + y^2}$. The arrows that we draw only *represent* the vector we are thinking of. More than one arrow could represent the same vector as in Figure 3.1. The definitions of vector arithmetic could be represented

geometrically too. For example, to get the sum of vectors $\mathbf{u}$ and $\mathbf{v}$, one places a representative of vector $\mathbf{u}$ in the plane, then places a representative of $\mathbf{v}$ whose tail is at the head of $\mathbf{v}$, and the vector $\mathbf{u}+\mathbf{v}$ is then represented by the third leg of this triangle, with base at the base of $\mathbf{u}$. To get a scalar multiple of a vector $\mathbf{w}$ one scales $\mathbf{w}$ in accordance with the coefficient. See Figure 3.1. Though instructive, this version of vector addition is not practical for calculations.

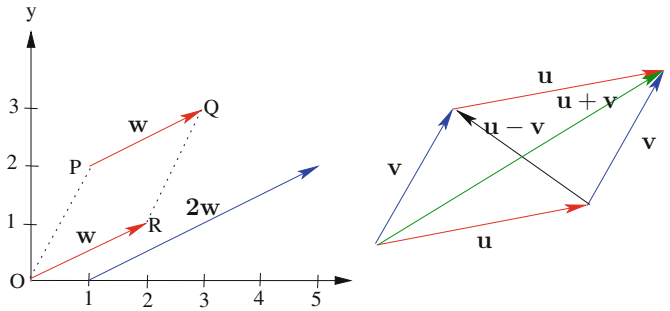


Fig. 3.1: Displacement vectors and graphical vector operations.

As a practical matter, it is also convenient to draw directed line segments connecting points; such a vector is called a *displacement vector*. For example, see Figure 3.1 for representatives of a displacement vector $\mathbf{w} = \overrightarrow{PQ}$ from the point P with coordinates $(1,2)$ to the point Q with coordinates $(3,3)$. One of the first nice outcomes of vector arithmetic is that this displacement vector can be deduced from a simple calculation,

Displacement and Position Vector

$$\mathbf{w} = (3,3) - (1,2) = (3-1, 3-2) = (2,1)\,.$$

A displacement vector of the form $\mathbf{w} = \overrightarrow{OR}$, where O is the origin, is called a *position vector*.

Geometrical vector spaces look a lot like the object we studied in Chapter 2 with the tuple notation as a shorthand for column vectors. The arithmetic of $\mathbb{R}^2$ and $\mathbb{R}^3$ is the same as the standard arithmetic for column vectors. Now, even though we can't draw real geometrical pictures of vectors with four or more coordinates, we have seen that larger vectors are useful in our search for solutions of linear systems. So the question presents itself, why stop at three? The answer is that we won't! We will use the familiar pictures of $\mathbb{R}^2$ and $\mathbb{R}^3$ to guide our intuition about vectors in higher-dimensional spaces, which we now define.

Definition 3.1. Standard Real Vector Space The *standard vector space of dimension n, where n is a positive integer, over the reals* is the set of vectors

$$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \mid x_1, x_2, \ldots, x_n \in \mathbb{R}\}$$

together with the standard vector addition and scalar multiplication. (Recall that $(x_1, x_2, \ldots, x_n)$ is shorthand for the column vector $[x_1, x_2, \ldots, x_n]^T$.)

We see immediately from the definition that the required closure properties of vector addition and scalar multiplication hold, so these really are vector spaces in the sense defined above. The standard real vector spaces are often called the real Euclidean vector spaces once the notion of a norm (a notion of length covered in the next chapter) is attached to them.

Homogeneous vector spaces. Graphics specialists and others find it important to distinguish between geometrical vectors and points (locations) in three-dimensional space. They want to be able to simultaneously manipulate these two kinds of objects, in particular, to do vector arithmetic and operator manipulation that reduces to the ordinary vector arithmetic when applied to geometrical vectors.

Here's the idea that neatly does the trick: Set up a coordinate system and identify geometrical vectors in the usual way, that is, by their coordinates x_1, x_2, x_3. Do the same with geometrical points. To distinguish between the two, embed them as vectors $\mathbf{x} = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ with the understanding that if $x_4 = 0$, then $\mathbf{x}$ represents a geometrical vector, and if $x_4 \neq 0$, then $\mathbf{x}$ represents a geometrical point. The vector $\mathbf{x}$ is called a *homogeneous vector* and $\mathbb{R}^4$ with the standard vector operations is called *homogeneous space*. If $x_4 \neq 0$, then $\mathbf{x}$ represents a point whose coordinates are $x_1/x_4, x_2/x_4, x_3/x_4$, and this point is said to be obtained from the vector $\mathbf{x}$ by *normalizing* the vector. Notice that the line through the origin that passes through the point $P = (x_1, x_2, x_3, 1)$ consists of vectors of the form (tx_1, tx_2, tx_3, t), where t is any real number. Conversely, any such nonzero vector is normalized $(tx_1/t, tx_2/t, tx_3/t, t/t) = P$. In this way, such lines through the origin correspond to points. (Readers who have seen projective spaces before may recognize this correspondence as identifying finite points in projective space with lines through the origin in $\mathbb{R}^4$. The ideas of homogeneous space actually originate in projective geometry.)

Homogeneous Vectors and Points

Now the standard vector arithmetic for $\mathbb{R}^4$ allows us to do arithmetic on geometrical vectors, for if $\mathbf{x} = (x_1, x_2, x_3, 0)$ and $\mathbf{y} = (y_1, y_2, y_3, 0)$ are such vectors, then as elements of $\mathbb{R}^4$ we have

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, x_2, x_3, 0) + (y_1, y_2, y_3, 0) = (x_1 + y_1, x_2 + y_2, x_3 + y_3, 0)\,, \\ c\mathbf{x} &= c\,(x_1, x_2, x_3, 0) = (cx_1, cx_2, cx_3, 0)\,, \end{aligned}$$

which result in geometrical vectors.

Example 3.1. Interpret the result of adding a point and vector in homogeneous space.

Solution. Notice that we can't add two points and obtain a point without some extra normalization; however, addition of a point $\mathbf{x} = (x_1, x_2, x_3, 1)$ and vector $\mathbf{y} = (y_1, y_2, y_3, 0)$ yields

$$\mathbf{x} + \mathbf{y} = (x_1, x_2, x_3, 1) + (y_1, y_2, y_3, 0) = (x_1 + y_1, x_2 + y_2, x_3 + y_3, 1)\,.$$

This has a rather elegant interpretation as the translation of the point $\mathbf{x}$ by the vector $\mathbf{y}$ to another point $\mathbf{x} + \mathbf{y}$. It reinforces the idea that geometrical vectors are simply displacements from one point to another. □

We can't draw pictures of $\mathbb{R}^4$, of course. But we can get an intuitive feeling for how homogenization works by moving down one dimension. Regard $\mathbb{R}^3$ as homogeneous space for the plane that consists of points $(x_1, x_2, 1)$. Figure 3.2 illustrates this idea.

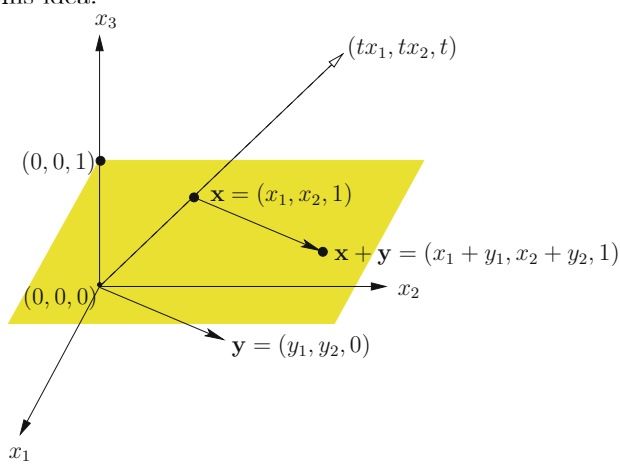


Fig. 3.2: Homogeneous space for planar points and vectors.

As in Chapter 2, we don't have to stop at the reals. For those situations in which we want to use complex numbers, we have the following vector spaces:

Definition 3.2. **Standard Complex Vector Space** The *standard vector space of dimension* n, where n is a positive integer, *over the complex numbers* is the set of vectors

$$\mathbb{C}^n = \{(x_1, x_2, \ldots, x_n) \mid x_1, x_2, \ldots, x_n \in \mathbb{C}\}$$

together with the standard vector addition and scalar multiplication.

The standard complex vector spaces are also sometimes called Euclidean spaces. It's rather difficult to draw honest spatial pictures of complex vectors. The space $\mathbb{C}^1$ isn't too bad: Complex numbers can be identified by points in

the complex plane. What about $\mathbb{C}^2$? Where can we put $(1+2\mathrm{i}, 3-\mathrm{i})$? It seems that we need four real coordinates, namely the real and imaginary parts of two independent complex numbers, to keep track of the point. This is too big to fit in real three-dimensional space, where we have only three independent coordinates. We don't let this technicality deter us. We can still draw fake vector pictures of elements of $\mathbb{C}^2$ to help our intuition, but do the algebra of vectors exactly from the definition.

Example 3.2. Find the displacement vector from the point P with coordinates $(1+2\mathrm{i}, 1-2\mathrm{i})$ to the point Q with coordinates $(3+\mathrm{i}, 2\mathrm{i})$.

Solution. We compute

$$\begin{aligned}\overrightarrow{PQ} &= (3+\mathrm{i}, 2\mathrm{i}) - (1+2\mathrm{i}, 1-2\mathrm{i}) \\ &= (3+\mathrm{i} - (1+2\mathrm{i}), 2\mathrm{i} - (1-2\mathrm{i})) \\ &= (2-\mathrm{i}, -1+4\mathrm{i}).\end{aligned}$$

□

Abstraction

We can see hints of a problem with the coordinate way of thinking about geometrical vectors. Suppose the vector in question represents a force. In one set of coordinates the force might have coordinates $(1, 0, 1)$. In another, it could have coordinates $(0, 1, 1)$. Yet the the force doesn't change, only its representation. This suggests an idea: Why not think about geometrical vectors as independent of any coordinate representation? From this perspective, geometrical vectors are really more abstract than the row or column vectors we have studied so far.

This line of thought leads us to consider an abstraction of our concept of vector space. First we have to identify the essential vector space properties, enough to make the resulting structure rich, but not so much that it is tied down to an overly specific form. We saw in Chapter 2 that many laws hold for the standard vector spaces. The essential laws were summarized in Section 2.1. These laws become the basis for our definition of an abstract vector space.

About notation: Just as in matrix arithmetic, for vectors $\mathbf{u}, \mathbf{v}$, we understand that $\mathbf{u} - \mathbf{v} = \mathbf{u} + (-\mathbf{v})$. We also suppress the dot $(\cdot)$ of scalar multiplication and usually write $a\mathbf{u}$ instead of $a \cdot \mathbf{u}$.

Abstract Vector Space An *(abstract) vector space* is a nonempty set V of elements called vectors, together with operations of vector addition $(+)$ and scalar multiplication $(\cdot)$, such that the following laws hold for all vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and scalars $a, b \in \mathbb{F}$:

(1) (Closure of vector addition) $\mathbf{u} + \mathbf{v} \in V$.
(2) (Commutativity of addition) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$.
(3) (Associativity of addition) $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$.
(4) (Additive identity) There exists an element $\mathbf{0} \in V$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u} = \mathbf{0} + \mathbf{u}$.
(5) (Additive inverse) There exists an element $-\mathbf{u} \in V$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0} = (-\mathbf{u}) + \mathbf{u}$.
(6) (Closure of scalar multiplication) $a \cdot \mathbf{u} \in V$.
(7) (Distributive law) $a \cdot (\mathbf{u} + \mathbf{v}) = a \cdot \mathbf{u} + a \cdot \mathbf{v}$.
(8) (Distributive law) $(a + b) \cdot \mathbf{u} = a \cdot \mathbf{u} + b \cdot \mathbf{u}$.
(9) (Associative law) $(ab) \cdot \mathbf{u} = a \cdot (b \cdot \mathbf{u})$.
(10) (Monoidal law) $1 \cdot \mathbf{u} = \mathbf{u}$.

Examples of these abstract vector spaces are the standard spaces just introduced, and these will be our main focus in this section. Yet, if we squint a bit, we can see vector spaces everywhere. There are other, entirely nonstandard examples, that make the abstraction worthwhile. Here are just a few such examples. Our first example is closely related to the standard spaces, though strictly speaking it is not one of them. It blurs the distinction between matrices and vectors in Chapter 2, since it makes matrices into "vectors" in the abstract sense of the preceding definition.

Example 3.3. Let $\mathbb{R}^{m,n}$ denote the set of all $m \times n$ matrices with real entries. Show that this set, with the standard matrix addition and scalar multiplication, forms a vector space.

Solution. We know that any two matrices of the same size can be added to yield a matrix of that size. Likewise, a scalar times a matrix yields a matrix of the same size. Thus, the operations of matrix addition and scalar multiplication are closed. Indeed, these laws and all the other vector space laws are summarized in the laws of matrix addition and scalar multiplication of page 70. □

Matrices as Vector Space

The next example is important in many areas of higher mathematics and is quite different from the standard vector spaces. Yet it is a perfectly legitimate vector space. All the same, at first it seems odd to think of functions as "vectors" even though this is meant in the abstract sense.

Example 3.4. Let $C[0, 1]$ denote the set of all real-valued functions that are continuous on the interval $[0, 1]$ and use the standard function addition and scalar multiplication for these functions. That is, for $f(x), g(x) \in C[0, 1]$ and real number c, we define the functions $f + g$ and cf by

$$(f+g)(x) = f(x) + g(x)$$
$$(cf)(x) = c(f(x)).$$

Show that $C[0,1]$ with these operations is a vector space.

Solution. We set $V = C[0,1]$ and check the vector space axioms for this V.

Function Space

For the rest of this example, we let f, g, h be arbitrary elements of V. We know from calculus that the sum of any two continuous functions is continuous and that any constant times a continuous function is also continuous. Therefore, the closure of addition and that of scalar multiplication hold. Now for all x such that $0 \leq x \leq 1$, we have from the definition and the commutative law of real number addition that

$$(f+g)(x) = f(x) + g(x) = g(x) + f(x) = (g+f)(x).$$

Since this holds for all x, we conclude that $f + g = g + f$, which is the commutative law of vector addition. Similarly,

$$\begin{aligned}((f+g)+h)(x) &= (f+g)(x) + h(x) = (f(x)+g(x)) + h(x) \\ &= f(x) + (g(x)+h(x)) = (f+(g+h))(x).\end{aligned}$$

Since this holds for all x, we conclude that $(f+g)+h = f+(g+h)$, which is the associative law for addition of vectors.

Next, if 0 denotes the constant function with value 0, then for any $f \in V$ we have that for all $0 \leq x \leq 1$,

$$(f+0)(x) = f(x) + 0 = f(x).$$

(We don't write the zero element of this vector space in boldface because it's customary not to write functions in bold.) Since this is true for all x we have that $f + 0 = f$, which establishes the additive identity law. Also, we define $(-f)(x) = -(f(x))$ so that for all $0 \leq x \leq 1$,

$$(f + (-f))(x) = f(x) - f(x) = 0,$$

from which we see that $f + (-f) = 0$. The additive inverse law follows. For the distributive laws note that for real numbers a, b and continuous functions $f, g \in V$, we have that for all $0 \leq x \leq 1$,

$$a(f+g)(x) = a(f(x)+g(x)) = af(x) + ag(x) = (af + ag)(x),$$

which proves the first distributive law. For the second distributive law, note that for all $0 \leq x \leq 1$,

$$((a+b)g)(x) = (a+b)g(x) = ag(x) + bg(x) = (ag + bg)(x),$$

and the second distributive law follows. For the scalar associative law, observe that for all $0 \leq x \leq 1$,

$$((ab)f)(x) = (ab)f(x) = a(bf(x)) = (a(bf))(x),$$

so that $(ab)f = a(bf)$, as required. Finally, we see that

$$(1f)(x) = 1f(x) = f(x),$$

from which we have the monoidal law $1f = f$. Thus, $C[0,1]$ with the prescribed operations is a vector space. □

The preceding example could have just as well been $C[a,b]$, the set of all continuous functions on the interval $a \leq x \leq b$, where $a < b$. Indeed, most of what we say about $C[0,1]$ is equally applicable to the more general space $C[a,b]$. We usually stick to the interval $0 \leq x \leq 1$ for simplicity. The next example is also based on the "functions as vectors" idea.

Example 3.5. One of the two sets $V = \{f(x) \in C[0,1] \mid f(1/2) = 0\}$ and $W = \{f(x) \in C[0,1] \mid f(1/2) = 1\}$, with the operations of function addition and scalar multiplication as in Example 3.4, forms a vector space over the reals, while the other does not. Determine which.

Solution. Notice that we don't have to check the commutativity of addition, associativity of addition, distributive laws, associative law, or monoidal law. The reason is that we already know from the previous example that these laws hold when the operations of the space $C[0,1]$ are applied to any elements of $C[0,1]$, whether they belong to V or W or not. So the only laws to be checked are the closure laws and the identity laws.

First let $f(x), g(x) \in V$ and let c be a scalar. By definition of the set V we have that $f(1/2) = 0$ and $g(1/2) = 0$. Add these equations together and we obtain

$$(f+g)(1/2) = f(1/2) + g(1/2) = 0 + 0 = 0.$$

It follows that V is closed under addition with these operations. Furthermore, if we multiply the identity $f(1/2) = 0$ by the real number c we obtain that

$$(cf)(1/2) = c \cdot f(1/2) = c \cdot 0 = 0.$$

It follows that V is closed under scalar multiplication. Now the zero function definitely belongs to V, since this function has value 0 at any argument. Therefore, V contains an additive identity element. Finally, we observe that the negative of a function $f(x) \in V$ is also an element of V, since

$$(-f)(1/2) = -1 \cdot f(1/2) = -1 \cdot 0 = 0.$$

Therefore, the set V with these operations satisfies all the vector space laws and is an (abstract) vector space in its own right.

When we examine the set W in a similar fashion, we run into a roadblock at the closure of addition law. If $f(x), g(x) \in W$, then by definition of the set W we have that $f(1/2) = 1$ and $g(1/2) = 1$. Add these equations together and we obtain

$$(f+g)(1/2) = f(1/2) + g(1/2) = 1 + 1 = 2.$$

This means that $f+g$ is not in W, so the closure of addition fails. We need go no further. If only one of the vector space axioms fails, then we do not have a vector space. Hence, W with these operations is not a vector space. □

There is a certain economy in this example. A number of laws did not need to be checked by virtue of the fact that the sets in question were subsets of existing vector spaces with the same vector operations. Here are two more examples that utilize this economy.

Example 3.6. Show that the set $\mathcal{P}_2$ of all polynomial functions of degree at most two with the standard function addition and scalar multiplication forms a vector space.

Solution. Polynomial functions are continuous functions. As in the preceding example, we don't have to check the commutativity of addition, associativity of addition, distributive laws, associative law, or monoidal law since we know that these laws hold for all continuous functions. Let $f, g \in \mathcal{P}_2$, say $f(x) = a_1 + b_1 x + c_1 x^2$ and $g(x) = a_2 + b_2 x + c_2 x^2$. Let c be any scalar. Then we have both

$$(f+g)(x) = f(x) + g(x) = (a_1 + a_2) + (b_1 + b_2)\, x + (c_1 + c_2)\, x^2 \in \mathcal{P}_2$$

and

$$(cf)(x) = cf(x) = c\left(a_1 + b_1 x + c_1 x^2\right) = ca_1 + cb_1 x + cc_1 x^2 \in \mathcal{P}_2.$$

Hence, $\mathcal{P}_2$ is closed under the operations of function addition and scalar multiplication. Furthermore, the zero function is a constant, hence a polynomial of degree at most two. Also, the negative of a polynomial of degree at most two is also a polynomial of degree at most two. So all of the laws for a vector space are satisfied and $\mathcal{P}_2$ is an (abstract) vector space. □

Example 3.7. Show that the set S_n of all $n \times n$ real symmetric matrices with the standard matrix addition and scalar multiplication form a vector space.

Solution. Just as in the preceding example, we don't have to check the commutativity of addition, associativity of addition, distributive laws, associative law, or monoidal law since we know that these laws hold for any matrices, symmetric or not. Now let $A, B \in S_n$. This means by definition that $A = A^T$ and $B = B^T$. Let c be any scalar. Then we have both

$$(A+B)^T = A^T + B^T = A + B$$

and

$$(cA)^T = cA^T = cA.$$

It follows that the set S_n is closed under the operations of matrix addition and scalar multiplication. Furthermore, the zero $n \times n$ matrix is clearly symmetric,

so the set S_n has an additive identity element. Finally, $(-A)^T = -A^T = -A$, so each element of S_n has an additive inverse as well. Therefore, all of the laws for a vector space are satisfied, so S_n together with these operations is an (abstract) vector space. □

One of the virtues of abstraction is that it allows us to cover many cases with one statement. For example, there are many simple facts that are deducible from the vector space laws alone. With the standard vector spaces, these facts seem transparently clear. For abstract spaces, the situation is not quite so obvious. Here are a few examples of what can be deduced from the definition.

Example 3.8. Let $\mathbf{v} \in V$, a vector space, and $\mathbf{0}$ the vector zero. Deduce *from the vector space properties alone* that $0\mathbf{v} = \mathbf{0}$.

Solution. Multiply both sides of the scalar identity $0+0=0$ on the right by the vector $\mathbf{v}$ to obtain that

$$(0+0)\mathbf{v} = 0\mathbf{v}.$$

Now use the distributive law to obtain

$$0\mathbf{v} + 0\mathbf{v} = 0\mathbf{v}.$$

Next add $-(0\mathbf{v})$ to both sides (remember, we don't know it's $\mathbf{0}$ yet), use the associative law of addition to regroup, and obtain that

$$0\mathbf{v} + (0\mathbf{v} + (-0\mathbf{v})) = 0\mathbf{v} + (-0\mathbf{v}).$$

Now use the additive inverse law to obtain that

$$0\mathbf{v} + \mathbf{0} = \mathbf{0}.$$

Finally, use the identity law to obtain

$$0\mathbf{v} = \mathbf{0},$$

which is what we wanted to show. □

Example 3.9. Show that the vector space V has only one zero element.

Solution. Suppose that both $\mathbf{0}$ and $\mathbf{0}_*$ act as zero elements in the vector space. Use the additive identity property of $\mathbf{0}$ to obtain that $\mathbf{0}_* + \mathbf{0} = \mathbf{0}_*$, while the additive identity property of $\mathbf{0}_*$ implies that $\mathbf{0} + \mathbf{0}_* = \mathbf{0}$. By the commutative law of addition, $\mathbf{0}_* + \mathbf{0} = \mathbf{0} + \mathbf{0}_*$. It follows that $\mathbf{0}_* = \mathbf{0}$, whence there can be only one zero element. □

There are several other such arithmetic facts that we want to identify, along with the one of this example. In the case of standard vectors, these facts are obvious, but for abstract vector spaces, they require a proof similar to the one we have just given. We leave these as exercises.

Laws of Vector Arithmetic
Let $\mathbf{v}$ be a vector in some vector space V and let c be any scalar. Then

(1) $0\mathbf{v} = \mathbf{0}$.
(2) $c\mathbf{0} = \mathbf{0}$.
(3) $(-c)\mathbf{v} = c(-\mathbf{v}) = -(c\mathbf{v})$.
(4) If $c\mathbf{v} = \mathbf{0}$, then $\mathbf{v} = \mathbf{0}$ or $c = 0$.
(5) A vector space has only one zero element.
(6) Every vector has only one additive inverse.

Linear Operators

We were introduced in Section 2.3 to the idea of a linear function in the context of standard vectors. Now that we have a notion of an abstract vector space, we can examine linearity in this larger setting. We have seen that some of our "vectors" can themselves be functions, as in the case of the vector space $C[0, 1]$ of continuous functions on the interval $[0, 1]$. In order to avoid confusion in cases like this, we prefer to designate linear functions by the term *linear operator*. Other common terms for this object are *linear mapping* and *linear transformation.*

Before giving the definition of linear operator, let us recall some notation that is associated with functions in general. We identify a function f with the notation $f : D \to T$, where D and T are the *domain* and *target* of the function, respectively. This means that for each x in the domain D, the value $f(x)$ is a uniquely determined element in the target T.

Domain, Range and Target

We want to emphasize at the outset that there is a difference here between the *target* of a function and its *range*. The *range* of the function f is defined as the subset of the target

$$\text{range}\,(f) = \{y \,|\, y = f(x) \text{ for some } x \in D\}\,,$$

which is just the set of all possible values of $f(x)$.

One-to-One and Onto Functions

A function is said to be *one-to-one* if, whenever $f(x) = f(y)$, then $x = y$. Also, a function is said to be *onto* if the range of f equals its target. For example, we can define a function $f : \mathbb{R} \to \mathbb{R}$ by the formula $f(x) = x^2$. It follows from our specification of f that the target of f is understood to be $\mathbb{R}$, while the range of f is the set of nonnegative real numbers. Therefore, f is not onto. Moreover, $f(-1) = f(1)$ and $-1 \neq 1$, so f is not one-to-one either.

A function that maps elements of one vector space into another, say $f : V \to W$, is sometimes called an *operator* or *transformation.* One of the simplest mappings of a vector space V is the *identity function* $\text{id}_V : V \to V$ given by $\text{id}_V(\mathbf{v}) = \mathbf{v}$, for all $\mathbf{v} \in V$. Here domain, range,

and target all agree. Of course, matters can become more complicated. For example, operator $f : \mathbb{R}^2 \to \mathbb{R}^3$ might be given by the formula

Identity Function

$$f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x^2 \\ xy \\ y^2 \end{bmatrix}.$$

Notice in this example that the target of f is $\mathbb{R}^3$, which is not the same as the range of f, since elements in the range have nonnegative first and third coordinates. From the point of view of linear algebra, this function lacks the essential feature that makes it really interesting, namely linearity.

Definition 3.3. Linear Operator A function $T : V \to W$ from the vector space V into the space W over the same field of scalars is called a *linear operator (mapping, transformation)* if for all vectors $\mathbf{u}, \mathbf{v} \in V$ and scalars c, d, we have

$$T(c\mathbf{u} + d\mathbf{v}) = cT(\mathbf{u}) + dT(\mathbf{v}).$$

By taking $c = d = 1$ in the definition, we see that a linear function T is *additive,* that is, $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$. Also, by taking $d = 0$ in the definition, we see that a linear function is *outative,* that is, $T(c\mathbf{u}) = cT(\mathbf{u})$. As a matter of fact, these two conditions imply the linearity property, and so are equivalent to it. We leave this fact as an exercise.

Additive and Outative Operator

An important special case of a linear operator is that in which the range of the operator is the field of scalars of the vector space. In this case, the operator is called a *linear functional.*

Linear Functional

By repeated application of the linearity definition, we can extend the linearity property to any linear combination of vectors, not just two terms. This means that for any scalars $c_1, c_2, \ldots, c_n$ and vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, we have

$$T(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n) = c_1T(\mathbf{v}_1) + c_2T(\mathbf{v}_2) + \cdots + c_nT(\mathbf{v}_n).$$

Example 3.10. Determine whether $T : \mathbb{R}^2 \to \mathbb{R}^3$ is a linear operator, where T is given by the formulas

(a) $T((x, y)) = (x^2, xy, y^2)$ or (b) $T((x, y)) = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$.

Solution. If T is given by (a) then we show by a simple example that T fails to be linear. Let us calculate

$$T((1, 0) + (0, 1)) = T((1, 1)) = (1, 1, 1),$$

while

$$T((1, 0)) + T((0, 1)) = (1, 0, 0) + (0, 0, 1) = (1, 0, 1).$$

These two are not equal, so T fails to satisfy the linearity property.

Next consider the operator T given as in (b). Write

$$A = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix},$$

and we see that the action of T is given by $T(\mathbf{v}) = A\mathbf{v}$. Now we have already seen in Section 2.3 that the operation of multiplication by a fixed matrix is a linear operator. □

Example 3.11. Let $\mathbf{t} = (t_1, t_2, t_3)$, $A = [a_{ij}]$ a 3×3 matrix and $M = \begin{bmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$. Show that the linear operator $T_M : \mathbb{R}^4 \to \mathbb{R}^4$ mapping homogeneous space into itself maps points to points and geometrical vectors to vectors.

Solution. Let $\mathbf{x} = (x_1, x_2, x_3, x_4) = (\mathbf{v}, x_4)$ with $\mathbf{v} = (x_1, x_2, x_3)$ and use block arithmetic to obtain that

$$T_M(\mathbf{x}) = \begin{bmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ x_4 \end{bmatrix} = \begin{bmatrix} A\mathbf{v} + x_4\mathbf{t} \\ x_4 \end{bmatrix}.$$

Thus, if $\mathbf{x}$ is a vector, which means $x_4 = 0$, then so is $T_M(\mathbf{x})$. Likewise, if $\mathbf{x}$ is a point, which means $x_4 = 1$, then so is $T_M(\mathbf{x})$. □

Recall that an operator $f : V \to W$ is said to be *invertible* if there is an operator $g : W \to V$ such that the composition of functions satisfies $f \circ g = \mathrm{id}_W$ and $g \circ f = \mathrm{id}_V$. In other words, $f(g(\mathbf{w})) = \mathbf{w}$ and $g(f(\mathbf{v})) = \mathbf{v}$ for all $\mathbf{w} \in W$ and $\mathbf{v} \in V$. We write $g = f^{-1}$ and call f^{-1} the inverse of f. One can show that for any operator f, linear or not, being invertible is equivalent to being both one-to-one and onto.

Invertible Operator

Example 3.12. Show that if $f : V \to W$ is an invertible linear operator on vector spaces, then f^{-1} is also a linear operator.

Solution. We need to show that for $\mathbf{u}, \mathbf{v} \in W$, the linearity property $f^{-1}(c\mathbf{u} + d\mathbf{v}) = cf^{-1}(\mathbf{u}) + df^{-1}(\mathbf{v})$ is valid. Let $\mathbf{w} = cf^{-1}(\mathbf{u}) + df^{-1}(\mathbf{v})$. Apply the function f to both sides and use the linearity of f to obtain that

$$f(\mathbf{w}) = f\left(cf^{-1}(\mathbf{u}) + df^{-1}(\mathbf{v})\right) = cf\left(f^{-1}(\mathbf{u})\right) + df\left(f^{-1}(\mathbf{v})\right) = c\mathbf{u} + d\mathbf{v}.$$

Apply f^{-1} to obtain that $\mathbf{w} = f^{-1}(f(\mathbf{w})) = f^{-1}(c\mathbf{u} + d\mathbf{v})$, which proves the linearity property. □

In Chapter 2 the following useful fact was shown, which we now restate for standard real vector spaces. It is also valid for standard complex spaces.

Theorem 3.1. Let A be an $m \times n$ matrix and define an operator $T_A : \mathbb{R}^n \to \mathbb{R}^m$ by the formula $T_A(\mathbf{v}) = A\mathbf{v}$, for all $\mathbf{v} \in \mathbb{R}^n$. Then T_A is a linear operator.

One can use this theorem and Example 3.12 to deduce the following fact, whose proof we leave as an exercise.

Corollary 3.1. Let A be an $n \times n$ matrix. The matrix operator T_A is invertible if and only if A is an invertible matrix.

Abstraction gives us a nice framework for certain key properties of mathematical objects, some of which we have seen before. For example, in calculus we were taught that differentiation has the "linearity property." Now we can express this assertion in a larger context: Let V be the space of differentiable functions and define an operator T on V by the rule $T(f(x)) = f'(x)$. Then T is a linear operator on the space V.

3.1 Exercises and Problems

In Exercises 1–2 the x-axis points east, y-axis north, and z-axis upward.

Exercise 1. Express the following geometric vectors as elements of $\mathbb{R}^3$.
(a) The displacement vector from the origin to the point P with coordinates $-2, 3, 1$.
(b) The displacement vector from the point P with coordinates $2, 1, 3$ to a location 3 units north, 4 units east, and 6 units upward.

Exercise 2. Express the following geometric points and vectors as elements of homogeneous space $\mathbb{R}^4$.
(a) The vectors of Exercise 1.
(b) The point situated 2 units upward, 4 units west, and -5 units north of the point with coordinates $1, 2, 0$.

In Exercises 3–10 determine whether the given set and operations define a vector space. If not, indicate which laws fail. Unless otherwise stated, the field of scalars is the default field $\mathbb{R}$.

Exercise 3. $V = \left\{ \begin{bmatrix} a & b \\ 0 & a+b \end{bmatrix} \mid a, b \in \mathbb{R} \right\}$ with the standard matrix addition and scalar multiplication.

Exercise 4. $V = \left\{ \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \mid a \in \mathbb{R} \right\}$ with the standard matrix addition and scalar multiplication.

Exercise 5. $V = \{[a, b, \bar{a}] \mid a, b \in \mathbb{C}\}$ with the standard matrix addition and scalar multiplication. In this example the scalar field is $\mathbb{C}$.

Exercise 6. V consists of all continuous functions $f(x)$ on the interval $[0, 1]$ such that $f(0) = 0$ with the standard function addition and scalar multiplication (see Example 3.4).

Exercise 7. $V = \mathbb{C}$ with the standard addition and scalar multiplication. In this example the scalar field is $\mathbb{R}$.

Exercise 8. $V = \left\{ z \,|\, z^2 = 0,\ z \in \mathbb{C} \right\}$ with the standard addition and scalar multiplication. In this example the scalar field is $\mathbb{C}$.

Exercise 9. V consists of all quadratic polynomial functions $f(x) = ax^2 + bx + c, a \neq 0$ with the standard function addition and scalar multiplication.

Exercise 10. V consists of all continuous functions $f(x)$ on the interval $[0, 1]$ such that $f(0) = f(1)$ with the standard function addition and scalar multiplication.

Exercise 11. V is the set of complex vectors $(z_1, z_2, z_3, 0)$ in space $\mathbb{C}^4$ with the standard vector addition and scalar multiplication.

Exercise 12. V is the set of points $(z_1, z_2, z_3, 1)$, $z_1, z_2, z_3 \in \mathbb{C}$, with scalar multiplication and vector addition given by $c\,(x_1, x_2, x_3, 1) = (cx_1, cx_2, cx_3, 1)$ and $(x_1, x_2, x_3, 1) + (y_1, y_2, y_3, 1) = (x_1 + y_1, x_2 + y_2, x_3 + y_3, 1)$.

Exercise 13. Determine which of the these formulas for $T : \mathbb{R}^3 \to \mathbb{R}^2$ is a linear operator. If so, write the operator as a matrix multiplication and determine whether the target of T equals its range. Here $\mathbf{x} = (x, y, z)$ and $T(\mathbf{x})$ follows.
(a) $(x, x + 2y - 4z)$ (b) $(x + y, xy)$ (c) (y, y) (d) $x\,(0, y)$ (e) $(\sin y, \cos z)$

Exercise 14. Repeat Exercise 13 for the following formulas for $T : \mathbb{R}^3 \to \mathbb{R}^3$.
(a) $(-y, z, -x)$ (b) $(x, y, 1)$ (c) $(y - x + z, 2x + z, 3x - y - z)$ (d) $\left(x^2, 0, z^2\right)$

Exercise 15. Let $V = C[0, 1]$ and define an operator $T : V \to V$ by the following formulas for $T(f)$ as a function of the variable x. Which of these operators is linear? If so, is the target V of the operator equal to its range?
(a) $f(1)x^2$ (b) $f^2(x)$ (c) $2f(x)$ (d) $\int_0^x f(s)\,ds$

Exercise 16. Let $V = \mathbb{R}^{2,2}$ and define an operator T with domain V by the following formulas for $T\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right)$. Which of these operators is linear?
(a) a_{22} (b) $\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$ (c) $\det A$ (d) $[a_{11}a_{22}, 0]$

Exercise 17. Is the identity operator on a vector space V, $\mathrm{id}_V : V \to V$ linear? Invertible? If so, specify its inverse.

Exercise 18. For arbitrary vector spaces U and V over the same scalars, is the zero operator $0_{U,V} : U \to V$ given by $0_{U,V}(\mathbf{v}) = \mathbf{0}$ linear? Invertible? If so, specify its inverse.

Exercise 19. A transform of homogeneous space is given by $M = \begin{bmatrix} I_3 & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ with $\mathbf{t} = (2, -1, 3)$. Calculate and describe in words the action of T_M on the point $\mathbf{x} = (x_1, x_2, x_3, 1)$. Find the inverse of this transform.

Exercise 20. A transform of homogeneous space is given by $M = \begin{bmatrix} A & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ with $\mathbf{t} = (2, -1, 3)$ and $A = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}$. Calculate and describe in words the action of T_M on the point $\mathbf{x} = (x_1, x_2, x_3, 1)$. Find the inverse of this transform. (See Example 2.18 in Chapter 2.)

*Problem 21. Use the definition of vector space to prove the vector law of arithmetic (2): $c\mathbf{0} = \mathbf{0}$.

Problem 22. Use the definition of vector space to prove the vector law of arithmetic (3): $(-c)\mathbf{v} = c(-\mathbf{v}) = -(c\mathbf{v})$.

Problem 23. Use the definition of vector space to prove the vector law of arithmetic (4): If $c\mathbf{v} = \mathbf{0}$, then $\mathbf{v} = \mathbf{0}$ or $c = 0$.

Problem 24. Let $\mathbf{u}, \mathbf{v} \in V$, where V is a vector space. Use the vector space laws to prove that the equation $\mathbf{x} + \mathbf{u} = \mathbf{v}$ has one and only one solution vector $\mathbf{x} \in V$, namely $\mathbf{x} = \mathbf{v} - \mathbf{u}$.

Problem 25. Let U and V be vector spaces over the same field of scalars and form the set $U \times V$ consisting of all ordered pairs $(\mathbf{u}, \mathbf{v})$ where $\mathbf{u} \in U$ and $\mathbf{v} \in V$. We can define an addition and scalar multiplication on these ordered pairs as follows:

$$(\mathbf{u}_1, \mathbf{v}_1) + (\mathbf{u}_2, \mathbf{v}_2) = (\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}_1 + \mathbf{v}_2),$$
$$c \cdot (\mathbf{u}_1, \mathbf{v}_1) = (c\mathbf{u}_1, c\mathbf{v}_1).$$

Verify that with these operations $U \times V$ becomes a vector space over the same field of scalars as U and V.

Problem 26. Show that for any vector space V, the identity function $\mathrm{id}_V : V \to V$ is a linear operator.

Problem 27. Let $T : \mathbb{R}^3 \to \mathcal{P}_2$ be given by $T((a, b, c)) = a + bx + cx^2$. Show that T is a linear operator whose range is $\mathcal{P}_2$.

Problem 28. Prove the remark following Definition 3.3: If a function $T : V \to W$ between vector spaces V and W is additive and outative, then it is linear.

*Problem 29. Prove Corollary 3.1.

*Problem 30. Transforms of homogeneous space are given by $M_1 = \begin{bmatrix} I_3 & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$, $\mathbf{t} = (t_1, t_2, t_3)$ and $M_2 = \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$, where A is an invertible 3×3 matrix. Show that the transform T_{M_1} (called a *translation transform*) and T_{M_2} (called a *homogeneous transform*) commute with each other, that is, $T_{M_1} \circ T_{M_2} = T_{M_2} \circ T_{M_1}$.

3.2 Subspaces

We now turn our attention to the concept of a *subspace*, which is a rich source for examples of vector spaces. It frequently happens that a certain vector space of interest is a subset of a larger, and possibly better understood, vector space, and that the vector operations are the same for both spaces. An example of this situation is given by the vector space V of Example 3.5, which is a subset of the larger vector space $C[0, 1]$ with both spaces sharing the same definitions of vector addition and scalar multiplication. Here is a precise formulation of the subspace idea.

Definition 3.4. **Subspace** A *subspace* of the vector space V is a subset W of V such that W, together with the binary operations it inherits from V, forms a vector space (over the same field of scalars as V) in its own right.

If W is a subset of the vector space V, we can apply the definition of vector space directly to the subset W to obtain the following very useful test.

Theorem 3.2. **Subspace Test** Let W be a subset of the vector space V. Then W is a subspace of V if and only if

(1) W contains the zero element of V.
(2) (Closure of addition) For all $\mathbf{u}, \mathbf{v} \in W$, $\mathbf{u} + \mathbf{v} \in W$.
(3) (Closure of scalar multiplication) For all $\mathbf{u} \in W$ and scalars c, $c\mathbf{u} \in W$.

Proof. Let W be a subspace of the vector space V. Then the closure of addition and scalar multiplication are automatically satisfied by the definition of vector space. For condition (1), we note that W must contain a zero element by definition of vector space. Let $\mathbf{0}_*$ be this element, so that $\mathbf{0}_* + \mathbf{0}_* = \mathbf{0}_*$. Add the negative of $\mathbf{0}_*$ (as an element of V) to both sides, cancel terms and we see that $\mathbf{0}_* = \mathbf{0}$, the zero of V. This shows that W satisfies condition (1).

Conversely, suppose that W is a subset of V satisfying the three conditions. Since the operations of W are those of the vector space V, and V is a vector space, most of the laws for W are automatic. Specifically, the laws of commutativity, associativity, distributivity, and the monoidal law hold for elements of W. The additive identity law follows from condition (1).

The only law that needs any work is the additive inverse law. Let $\mathbf{w} \in W$. By closure of scalar multiplication, $(-1)\mathbf{w}$ is in W. By the laws of vector arithmetic in the preceding section, this vector is simply $-\mathbf{w}$. This proves that every element of W has an additive inverse in W, which shows that W is a subspace of V. □

One notable point that comes out of the subspace test is that every subspace of V contains the zero vector. This is obviously not true of arbitrary subsets of V and serves to remind us that although every subspace is a subset

of V, not every subset is a subspace. Confusing the two is a common mistake, so much so that we issue the following caution:

Caution: Every subspace of a vector space is a subset, but not every subset is a subspace.

Example 3.13. Which of the following subsets of the standard vector space $V = \mathbb{R}^3$ are subspaces of V?

(a) $W_1 = \{(x, y, z) \,|\, x - 2y + z = 0\}$ (b) $W_2 = \{(x, y, z) \,|\, x, y, z \text{ are positive}\}$
(c) $W_3 = \{(0, 0, 0)\}$ (d) $W_4 = \{(x, y, z) \,|\, x^2 - y = 0\}$

Solution. (a) Take $\mathbf{w} = (0, 0, 0)$ and obtain that $0 - 2 \cdot 0 + 0 = 0$, so that $\mathbf{w} \in W_1$. Next, check closure of W_1 under addition. Let's name two general elements from W_1, say $\mathbf{u} = (x_1, y_1, z_1)$ and $\mathbf{v} = (x_2, y_2, z_2)$. Then we know from the definition of W_1 that

$$\begin{aligned} x_1 - 2y_1 + z_1 &= 0 \\ x_2 - 2y_2 + z_2 &= 0. \end{aligned}$$

We want to show that $\mathbf{u} + \mathbf{v} = (x_1 + x_2, y_1 + y_2, z_1 + z_2) \in W_1$. So add the two equations above and group terms to obtain

$$(x_1 + x_2) - 2(y_1 + y_2) + (z_1 + z_2) = 0.$$

This equation shows that the coordinates of $\mathbf{u} + \mathbf{v}$ fit the requirement for being an element of W_1, i.e., $\mathbf{u} + \mathbf{v} \in W_1$. Similarly, if c is a scalar then we can multiply the equation that says $\mathbf{u} \in W_1$, i.e., $x_1 - 2y_1 + z_1 = 0$, by c to obtain

$$(cx_1) - 2(cy_1) + (cz_1) = c0 = 0.$$

This shows that the coordinates of $c\mathbf{v}$ fit the requirement for being in W_1, i.e., $c\mathbf{u} \in W_1$. It follows that W_1 is closed under both addition and scalar multiplication, so it is a subspace of $\mathbb{R}^3$.

(b) This one is easy. Any subspace must contain the zero vector $(0, 0, 0)$. Clearly W_2 does not. Hence, it cannot be a subspace. Another way to see it is to notice that closure under scalar multiplication fails (try multiplying $(1, 1, 1)$ by -1).

(c) The only possible choice for arbitrary elements $\mathbf{u}, \mathbf{v}$, in this space is $\mathbf{u} = \mathbf{v} = (0, 0, 0)$. But then we see that W_3 obviously contains the zero vector and for any scalar c,

$$\begin{aligned} (0, 0, 0) + (0, 0, 0) &= (0, 0, 0), \\ c(0, 0, 0) &= (0, 0, 0). \end{aligned}$$

Therefore, W_3 is a subspace of V by the subspace test.

(d) First of all, $0^2 - 0 = 0$, which means that $(0, 0, 0) \in W_4$. Likewise we see that $(1, 1, 0) \in W_4$ as well. But $(1, 1, 0) + (1, 1, 0) = (2, 2, 0)$, which is not

an element of W_4 since $2^2 - 2 \neq 0$. Therefore, closure of addition fails and W_4 is not a subspace of V by the subspace test. □

Part (c) of this example highlights part of a simple fact about vector spaces. Every vector space V must have at least two subspaces, namely, $\{\mathbf{0}\}$, where $\mathbf{0}$ is the zero vector in V, and V itself. These are not terribly surprising subspaces, so they are commonly called the *trivial* subspaces.

Trivial Subspaces

Example 3.14. Show that the subset $\mathcal{P}[0,1]$ of $C[0,1]$ consisting of all polynomial functions on the interval $[0,1]$ is a subspace of $C[0,1]$ and that the subset $\mathcal{P}_n[0,1]$ consisting of all polynomials of degree at most n is a subspace of $\mathcal{P}[0,1]$.

Solution. Certainly, $\mathcal{P}[0,1]$ is a subset of $C[0,1]$, since every polynomial is continuous on the interval $[0,1]$ and $\mathcal{P}[0,1]$ contains the zero constant function, which is a polynomial function. Let f and g be two polynomial functions on the interval $[0,1]$, say

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n,$$
$$g(x) = b_0 + b_1 x + \cdots + b_n x^n,$$

where n is an integer equal to the maximum of the degrees of $f(x)$ and $g(x)$. Let c be any real number, and we see that

$$(f+g)(x) = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n,$$
$$(cf)(x) = ca_0 + ca_1 x + \cdots + ca_n x^n,$$

which shows that $\mathcal{P}[0,1]$ is closed under function addition and scalar multiplication. By the subspace test, $\mathcal{P}[0,1]$ is a subspace of $C[0,1]$. The equations above also show that the subset $\mathcal{P}_n[0,1]$ passes the subspace test, so it is a subspace of $\mathcal{P}[0,1]$. □

Example 3.15. Show that the set of all upper triangular matrices (see page 105) in the vector space $V = \mathbb{R}^{n,n}$ of $n \times n$ real matrices is a subspace of V.

Solution. Since the zero matrix is upper triangular, the subset W of all upper triangular matrices contains the zero element of V. Let $A = [a_{i,j}]$ and $B = [b_{i,j}]$ be any two matrices in W and let c be any scalar. By the definition of upper triangular, we must have $a_{i,j} = 0$ and $b_{i,j} = 0$ if $i > j$. However,

$$A + B = [a_{i,j} + b_{i,j}],$$
$$cA = [ca_{i,j}],$$

and for $i > j$ we have $a_{i,j} + b_{i,j} = 0 + 0 = 0$ and $ca_{i,j} = c0 = 0$, so that $A + B$ and cA are also upper triangular. It follows that W is a subspace of V by the subspace test. □

Linear Combinations There is an extremely useful type of subspace that requires the notion of a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in the vector space V: an expression of the form

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n,$$

where $c_1, c_2, \ldots, c_n$ are scalars. We can consider the set of all possible linear combinations of a list of vectors, which is what our next definition does.

Definition 3.5. Vector Span Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors in the vector space V. The *span* of these vectors, denoted by $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, is the subset of V consisting of all possible linear combinations of these vectors, i.e.,

$$\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} = \{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n \,|\, c_1, c_2, \ldots, c_n \text{ are scalars}\}$$

Caution: The scalars we are using really make a difference. For example, if $\mathbf{v}_1 = (1, 0)$ and $\mathbf{v}_2 = (0, 1)$ are viewed as elements of the real vector space $\mathbb{R}^2$, then

$$\begin{aligned}\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} &= \{c_1(1, 0) + c_2(0, 1) \,|\, c_1, c_2 \in \mathbb{R}\} \\ &= \{(c_1, c_2) \,|\, c_1, c_2 \in \mathbb{R}\} \\ &= \mathbb{R}^2.\end{aligned}$$

Similarly, if we view $\mathbf{v}_1$ and $\mathbf{v}_2$ as elements of the complex vector space $\mathbb{C}^2$, then we see that $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} = \mathbb{C}^2$. Now $\mathbb{R}^2$ consists of those elements of $\mathbb{C}^2$ whose coordinates have zero imaginary parts, so $\mathbb{R}^2$ is a sub*set* of $\mathbb{C}^2$; but these are definitely not equal sets. By the way, $\mathbb{R}^2$ is definitely not a sub*space* of $\mathbb{C}^2$ either, since the subset $\mathbb{R}^2$ is not closed under multiplication by complex scalars.

We should take note here that the definition of span would work perfectly well with infinite sets, as long as we understand that linear combinations in the definition would be finite and therefore not involve all the vectors in the span. A case in point is as follows: Consider the space $\mathcal{P}$ of all polynomial functions with the standard addition and scalar multiplication. It makes perfectly good sense to write

$$\mathcal{P} = \operatorname{span}\{1, x, x^2, x^3, \ldots, x^n, \ldots\},$$

since every polynomial is a *finite* linear combination of various monomials x^k.

Example 3.16. Interpret the following linear spans in $\mathbb{R}^3$ geometrically:

$$W_1 = \operatorname{span}\left\{\begin{bmatrix}1\\2\\1\end{bmatrix}\right\}, \qquad W_2 = \operatorname{span}\left\{\begin{bmatrix}1\\2\\1\end{bmatrix}, \begin{bmatrix}2\\0\\0\end{bmatrix}\right\}.$$

Solution. Elements of W_1 are simply scalar multiples of the single vector $(1, 2, 1)$. The set of all such multiples gives us a line through the origin $(0, 0, 0)$.

On the other hand, elements of W_2 give all possible linear combinations of two vectors $(1,2,1)$ and $(2,0,0)$. The locus of points generated by these combinations is a plane in $\mathbb{R}^3$ containing the origin, so it is determined by the points with coordinates $(0,0,0)$, $(1,2,1)$, and $(2,0,0)$. See Figure 3.3 for a picture of a portion of these spans. □

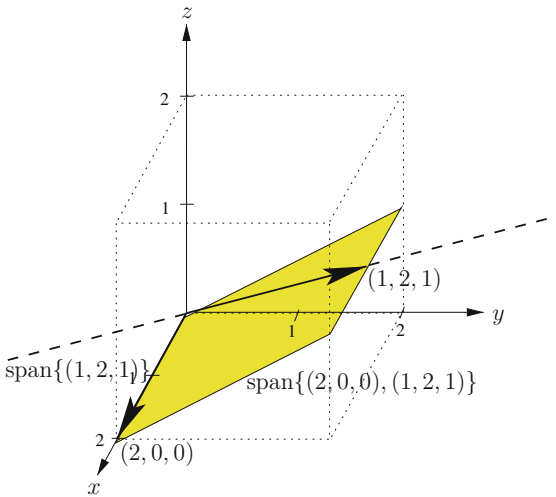


Fig. 3.3: Shaded portion of span $\{(2,0,0),(1,2,1)\}$ and dashed span $\{(1,2,1)\}$.

Spans are the premier examples of subspaces. In a certain sense, it can be said that *every* subspace is the span of some of its vectors. The following important fact is a very nice application of the subspace test.

Theorem 3.3. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors in the vector space V. Then $W = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a subspace of V.

Proof. First, we observe that the zero vector can be expressed as the linear combination $0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_n$, which is an element of W. Next, let c_i, d_i be any scalars and form general elements $\mathbf{u}, \mathbf{v} \in W$, say

$$\begin{aligned}\mathbf{u} &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n,\\ \mathbf{v} &= d_1\mathbf{v}_1 + d_2\mathbf{v}_2 + \cdots + d_n\mathbf{v}_n.\end{aligned}$$

Add these vectors and collect like terms to obtain

$$\mathbf{u} + \mathbf{v} = (c_1 + d_1)\mathbf{v}_1 + (c_2 + d_2)\mathbf{v}_2 + \cdots + (c_n + d_n)\mathbf{v}_n.$$

Thus, $\mathbf{u}+\mathbf{v}$ is also a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, so W is closed under vector addition. Finally, form the product $c\mathbf{u}$ to obtain

$$c\mathbf{u} = (cc_1)\mathbf{v}_1 + (cc_2)\mathbf{v}_2 + \cdots + (cc_n)\mathbf{v}_n,$$

which is again a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, so W is closed under scalar multiplication. By the subspace test, W is a subspace of V. □

If $W = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, we say that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a *spanning set* for the vector space W, and that W *is spanned by* the vectors $\mathbf{v}, \mathbf{v}_2, \ldots, \mathbf{v}_n$. There are a number of simple properties of spans that we will need from time to time. One of the most useful is this basic fact.

Spanning Set

Theorem 3.4. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors in the vector space V and let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ be vectors in $\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Then

$$\text{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\} \subseteq \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}.$$

Proof. Suppose that for each index $j = 1, 2, \ldots k$,

$$\mathbf{w}_j = c_{1j}\mathbf{v}_1 + c_{2j}\mathbf{v}_2 + \cdots + c_{nj}\mathbf{v}_n.$$

Write a linear combination of the $\mathbf{w}_j$'s by regrouping the coefficients of each $\mathbf{v}_k$ as

$$\begin{aligned} d_1\mathbf{w}_1 + d_2\mathbf{w}_2 + \cdots + d_k\mathbf{w}_k &= d_1(c_{11}\mathbf{v}_1 + c_{21}\mathbf{v}_2 + \cdots + c_{n1}\mathbf{v}_n) \\ &+ d_2(c_{12}\mathbf{v}_1 + c_{22}\mathbf{v}_2 + \cdots + c_{n2}\mathbf{v}_n) + \cdots + d_k(c_{1k}\mathbf{v}_1 + c_{2k}\mathbf{v}_2 + \cdots + c_{nk}\mathbf{v}_n) \\ &= \left(\sum_{j=1}^{k} d_j c_{1j}\right)\mathbf{v_1} + \left(\sum_{j=1}^{k} d_j c_{2j}\right)\mathbf{v_2} + \cdots + \left(\sum_{j=1}^{k} d_j c_{nj}\right)\mathbf{v_n}. \end{aligned}$$

It follows that each element of $\text{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\}$ belongs to the vector space $\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, as desired. □

Here is a simple application of this theorem: If $\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \ldots, \mathbf{v}_{i_k}$ is a subset of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, then

$$\text{span}\{\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \ldots, \mathbf{v}_{i_k}\} \subseteq \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}.$$

The reason is that for $j = 1, 2, \ldots, k$,

$$\mathbf{w}_j = \mathbf{v}_{i_j} = 0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 1\mathbf{v}_{i_j} + \cdots + 0\mathbf{v}_n,$$

so that the theorem applies to these vectors. Put another way, if we enlarge the list of spanning vectors, we enlarge the spanning set. However, we may not obtain a strictly larger spanning set, as the following example shows.

Example 3.17. Show that

$$\text{span}\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right\} = \text{span}\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}\right\}.$$

Why might one prefer the first spanning set?

Solution. Label vectors $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, and $\mathbf{v}_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Every element of $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ belongs to $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, since we can write $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + 0\mathbf{v}_3$. So we clearly have that $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} \subseteq \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. However, a little fiddling with numbers reveals this fact:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} = (-1)\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

In other words $\mathbf{v}_3 = -\mathbf{v}_1 + 2\mathbf{v}_2$. Therefore, any linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ can be written as

$$\begin{aligned} c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3(-\mathbf{v}_1 + 2\mathbf{v}_2) \\ &= (c_1 - c_3)\mathbf{v}_1 + (c_2 + 2c_3)\mathbf{v}_2. \end{aligned}$$

Thus, any element of $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ belongs to $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$, so the two spans are equal. This is an algebraic representation of the geometric fact that the three vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ belong to the same plane in $\mathbb{R}^2$ that is spanned by the two vectors $\mathbf{v}_1, \mathbf{v}_2$. It seems reasonable that we should prefer the spanning set $\mathbf{v}_1, \mathbf{v}_2$ to the set $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, since the former is smaller yet carries just as much information as the latter. As a matter of fact, we would get the same span if we used $\mathbf{v}_1, \mathbf{v}_3$ or $\mathbf{v}_2, \mathbf{v}_3$. The spanning set $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ has "redundant" vectors in it. □

As another application of vector spans, let's consider the problem of determining all subspaces of the vector space $\mathbb{R}^2$, the plane, from a geometrical perspective. First, we have the trivial subspaces $\{(0,0)\}$ and $\mathbb{R}^2$. Next, consider the subspace $V = \operatorname{span}\{\mathbf{v}\}$, where $\mathbf{v} \neq \mathbf{0}$. It's easy to see that the set of all multiples of $\mathbf{v}$ constitutes a straight line through the origin. Finally, consider the subspace $V = \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$, where $\mathbf{w} \notin \operatorname{span}\{\mathbf{v}\}$. We can see that any point in the plane can be a corner of a parallelogram with edges that are multiples of $\mathbf{v}$ and $\mathbf{w}$. Hence, $V = \mathbb{R}^2$. Consequently, the only subspaces of $\mathbb{R}^2$ are $\{(0,0)\}$, $\mathbb{R}^2$, and lines through the origin. In a similar fashion, you can convince yourself that the only subspaces of $\mathbb{R}^3$ are $\{(0,0,0)\}$, lines through the origin, planes through the origin, and $\mathbb{R}^3$.

3.2 Exercises and Problems

In Exercises 1–10, determine whether the subset W is a subspace of the vector space V.

Exercise 1. $V = \mathbb{R}^3$ and $W = \{(a, b, a - b + 1) \mid a, b \in \mathbb{R}\}$.

Exercise 2. $V = \mathbb{R}^3$ and $W = \{(a, 0, a - b) \mid a, b \in \mathbb{R}\}$.

Exercise 3. $V = \mathbb{R}^3$ and $W = \{(a, b, c) \mid 2a - b + c = 0\}$.

Exercise 4. $V = \mathbb{R}^{2,3}$ and $W = \left\{ \begin{bmatrix} a & b & 0 \\ b & a & 0 \end{bmatrix} \mid a, b \in \mathbb{R} \right\}$.

Exercise 5. $V = C[0,1]$ and $W = \{f(x) \in C[0,1] \,|\, f(1) + f(1/2) = 0\}$.

Exercise 6. $V = C[0,1]$ and $W = \{f(x) \in C[0,1] \,|\, f(1) \leq 0\}$.

Exercise 7. $V = \mathbb{R}^{n,n}$ and W is the set of all invertible matrices in V.

Exercise 8. $V = \mathbb{R}^{2,2}$ and W is the set of all matrices $A = \begin{bmatrix} a & b \\ -b & c \end{bmatrix}$, for some scalars a, b, c. (Such matrices are called *skew-symmetric* since $A^T = -A$.)

Exercise 9. V is the subset of geometrical vectors $(x_1, x_2, x_3, 0)$ in homogeneous space $W = \mathbb{R}^4$ with the standard vector addition and scalar multiplication.

Exercise 10. V is the subset of geometrical points $(x_1, x_2, x_3, 1)$ in homogeneous space $W = \mathbb{R}^4$ with vector addition and scalar multiplication defined by $(x_1, x_2, x_3, 1) + (y_1, y_2, y_3, 1) = (x_1 + y_1, x_2 + y_2, x_3 + y_3, 1)$ and $c\,(x_1, x_2, x_3, 1) = (cx_1, cx_2, cx_3, 1)$.

Exercise 11. Show that $\operatorname{span}\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}\right\} = \operatorname{span}\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}-2\\1\end{bmatrix}\right\}$.

Exercise 12. Show that $\operatorname{span}\left\{\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}1\\1\\1\end{bmatrix}\right\} = \operatorname{span}\left\{\begin{bmatrix}0\\-1\\-1\end{bmatrix}, \begin{bmatrix}1\\2\\2\end{bmatrix}\right\}$.

Exercise 13. Which of the following spans equal the space $\mathcal{P}_2$ of polynomial functions of degree at most 2? Justify your answers.

(a) $\operatorname{span}\left\{1, 1+x, x^2\right\}$ (b) $\operatorname{span}\left\{x, 4x - 2x^2, x^2\right\}$

(c) $\operatorname{span}\left\{1 + x + x^2, 1 + x, 3\right\}$ (d) $\operatorname{span}\left\{1 - x^2, 1\right\}$

Exercise 14. Which of the following spans equal the space $\mathbb{R}^2$? Justify your answers.

(a) $\operatorname{span}\{(1,0), (-1,-1)\}$ (b) $\operatorname{span}\{(1,2), (2,4)\}$

(c) $\operatorname{span}\{(1,0), (0,0), (0,-1)\}$ (d) $\operatorname{span}\{(-1,-2), (-1,-1)\}$

Exercise 15. Expand the following subsets of vector spaces into spanning sets of the space with the fewest possible additional elements. Justify your answer.

(a) $\{(1,2), (-2,-4)\} \subseteq \mathbb{R}^2$ (b) $\left\{x^3 + 1, x - 1, x^2 - 1, x\right\} \subseteq \mathcal{P}_3$

Exercise 16. Expand the following subsets of vector spaces into spanning sets of the space with the fewest possible additional elements. Justify your answer.

(a) $\{(1,0,1,0), (2,0,2,1), (0,1,0,1)\} \subseteq \mathbb{R}^4$ (b) $\left\{\begin{bmatrix}2&1\\0&1\end{bmatrix}, \begin{bmatrix}2&0\\1&1\end{bmatrix}\right\} \subseteq \mathbb{R}^{2,2}$

Exercise 17. Let $\mathbf{u} = (2,-1,1)$, $\mathbf{v} = (0,1,1)$, and $\mathbf{w} = (2,1,3)$. Show that $\operatorname{span}\{\mathbf{u}+\mathbf{w}, \mathbf{v}-\mathbf{w}\} \subseteq \operatorname{span}\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ and determine whether or not these spans are actually equal.

Exercise 18. Find two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$ such that if $\mathbf{u} = (1, -1, 1)$, then $\mathbb{R}^3 = \text{span}\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$.

*Problem 19. Let U and V be subspaces of W. Use the subspace test to prove the following.

(a) The set intersection $U \cap V$ is a subspace of W.

(b) The sum of the spaces, $U + V = \{u + v \,|\, u \in U \text{ and } v \in V\}$, is a subspace of W.

(c) The set union $U \cup V$ is not a subspace of W unless one of U or V is contained in the other.

Problem 20. Let V and W be subspaces of $\mathbb{R}^3$ given by

$$V = \{(x, y, z) \,|\, x = y = z \in \mathbb{R}\} \text{ and } W = \{(x, y, 0) \,|\, x, y \in \mathbb{R}\}.$$

Show that $V + W = \mathbb{R}^3$ and $V \cap W = \{\mathbf{0}\}$.

*Problem 21. Prove that if $V = \mathbb{R}^{n,n}$, then the set of all diagonal matrices is a subspace of V.

*Problem 22. Let V be the space of 2×2 matrices and associate with each $A \in V$ the vector $\text{vec}(A) \in \mathbb{R}^4$ obtained from A by stacking the columns of A underneath each other. (For example, $\text{vec}\left(\begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}\right) = (1, -1, 2, 1)$.) Show the following.

(a) The vec operator establishes a one-to-one correspondence between matrices in V and vectors in $\mathbb{R}^4$.

(b) The vec operator, $\text{vec} : \mathbb{R}^{2,2} \to \mathbb{R}^4$, is a linear operator.

Problem 23. You will need a technology tool for this exercise. Use the matrix

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

and the vec operator of the preceding exercise to turn powers of A into vectors. Then use your technology tool to find a spanning set (or basis, which is a special spanning set) for subspaces $V_k = \text{span}\left\{A^0, A^1, \ldots, A^k\right\}$, $k = 1, 2, 3, 4, 5, 6$. Based on this evidence, how many matrices will be required for a span of V_k? (Remember that $A^0 = I$.)

Problem 24. Show that the set $C^1[0, 1]$ of continuous functions that have a continuous derivative on the interval $[0, 1]$ is a subspace of the vector space $C[0, 1]$.

3.3 Linear Combinations

We have seen in Section 3.2 that linear combinations give us a rich source of subspaces for a vector space. In this section we will take a closer look at linear combinations. But first let's clarify the difference between list

Lists and Sets

and sets: *Lists* involve an ordering of elements (they can just as well be called finite sequences), while *sets* don't really imply any ordering of elements. Thus, every list of vectors, e.g., $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, gives rise to a unique set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. A different list $\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_2$ may define the same set $\{\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_2\} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. Lists can have repeats in them, while sets don't. For instance, the list $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_1$ defines the set $\{\mathbf{v}_1, \mathbf{v}_2\}$. The default meaning of the terminology "the vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$" is "the list of vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$," although occasionally it means a set or even both. For example, the definitions below work perfectly well for either sets or lists.

Linear Dependence

Let's make precise the idea of redundant vectors encountered in Example 3.17.

Definition 3.6. ***Redundant Vector*** The vector $\mathbf{v}_i$ is *redundant* in the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ if the linear span $\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ does not change when $\mathbf{v}_i$ is removed.

An easy example of a redundant vector is the zero vector, which is clearly redundant in any set or list containing it.

Example 3.18. Which vectors are redundant in the set consisting of $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\mathbf{v}_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$?

Solution. As in Example 3.17, we notice that $\mathbf{v}_3 = (-1)\mathbf{v}_1 + 2\mathbf{v}_2$. Thus, any linear combination involving $\mathbf{v}_3$ can be expressed in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$. Therefore, $\mathbf{v}_3$ is redundant in the list $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. But there is more going on here. Let's write the equation above in a form that doesn't single out any one vector:

$$0 = (-1)\mathbf{v}_1 + 2\mathbf{v}_2 + (-1)\mathbf{v}_3.$$

Now we see that we could solve for *any* of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ in terms of the remaining two vectors. Therefore, each of these vectors is redundant in the set. However, this doesn't mean that we can discard all three and get the same linear span. This is obviously false. What we can do is discard any *one* of them, then start over and examine the remaining set for redundant vectors. □

This example shows that what really counts for redundancy is that the vector in question occurs with a nonzero coefficient in a linear combination that equals 0. This situation warrants a name:

Definition 3.7. **Linearly Dependent or Independent Vectors** The vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are said to be *linearly dependent* if there exist scalars $c_1, c_2, \ldots, c_n$, not all zero, such that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}. \tag{3.1}$$

Otherwise, the vectors are called *linearly independent*.

This fact inspires some notation: We will call a linear combination *trivial* if every coefficient is zero; otherwise it is *nontrivial*. We say that a linear combination has *value zero* if it sums to zero. Thus, linear dependence is equivalent to the existence of a nontrivial linear combination with value zero. Just as with redundancy, linear dependence or independence is a property of the list or set in question, *not* of the individual vectors. Here is the key connection between linear dependence and redundancy.

Trivial and Nontrivial Linear Combination

Theorem 3.5. Redundancy Test The list of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of a vector space has redundant vectors if and only if it is linearly dependent, in which case the redundant vectors are those that occur with nonzero coefficient in some linear combination with value zero.

Proof. Observe that if (3.1) holds and some scalar, say c_1, is nonzero, then we can use the equation to solve for $\mathbf{v}_1$ as a linear combination of the remaining vectors to obtain

$$\mathbf{v}_1 = \frac{-1}{c_1}\left(c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + \cdots + c_n\mathbf{v}_n\right).$$

Thus, we see that any linear combination involving $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ can be expressed using only $\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n$. It follows that

$$\operatorname{span}\{\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n\} = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}.$$

Conversely, if these spans are equal then $\mathbf{v}_1$ belongs to the left-hand side, so there are scalars $d_2, d_3, \ldots, d_n$ such that

$$\mathbf{v}_1 = d_2\mathbf{v}_2 + d_3\mathbf{v}_3 + \cdots + d_n\mathbf{v}_n.$$

Now bring all terms to the right-hand side and obtain the nontrivial linear combination

$$-\mathbf{v}_1 + d_2\mathbf{v}_2 + d_3\mathbf{v}_3 + \cdots + d_n\mathbf{v}_n = \mathbf{0}.$$

All of this works equally well for any index other than 1, so the theorem is proved. □

It is instructive to examine the simple case of two vectors $\mathbf{v}_1, \mathbf{v}_2$. What does it mean to say that these vectors are linearly dependent? Simply that one of the vectors can be expressed in terms of the other, in other words, that each vector is a scalar multiple of the other. However, matters are more complex when we proceed to three or more vectors, a point that is often overlooked. So we issue a warning here.

Caution: If we know that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is linearly dependent, it does *not* necessarily imply that one of these vectors is a multiple of one of the others unless $n = 2$. In general, all we can say is that one of these vectors is a linear combination of the others.

Example 3.19. Which of the following lists of vectors have redundant vectors, i.e., are linearly dependent?

(a) $\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}1\\-1\\-2\end{bmatrix}$ (b) $\begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}$ (c) $\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}$

Solution. Let's try to see the big picture. Consider the vectors in each list to be designated as $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. Define matrix $A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$ and vector $\mathbf{c} = (c_1, c_2, c_3)$. Then the general linear combination can be written as

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix} = A\,\mathbf{c}.$$

This is the key idea of "linear combination as matrix–vector multiplication" that we saw in Theorem 2.1. Now we see that a nontrivial linear combination with value zero amounts to a nontrivial solution to the homogeneous equation $A\mathbf{c} = \mathbf{0}$. We know how to find these! In case (a) we have that

$$\begin{bmatrix}1&0&1\\1&1&-1\\0&1&-2\end{bmatrix} \xrightarrow[E_{21}(-1)]{} \begin{bmatrix}1&0&1\\0&1&-2\\0&1&-2\end{bmatrix} \xrightarrow[E_{32}(-1)]{} \begin{bmatrix}1&0&1\\0&1&-2\\0&0&0\end{bmatrix},$$

so that the solutions to the homogeneous system are $\mathbf{c} = (-c_3, 2c_3, c_3) = c_3(-1, 2, 1)$. Take $c_3 = 1$ and we have that

$$-1\mathbf{v}_1 + 2\mathbf{v}_2 + 1\mathbf{v}_3 = 0,$$

which shows that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is a linearly dependent list of vectors.

We'll solve (b) by a different method, which can be applied to any set of n vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$. Notice that

Determinant Test for Linear Independence

$$\det\begin{bmatrix}0&1&0\\1&1&1\\0&0&1\end{bmatrix} = -1\det\begin{bmatrix}1&0\\0&1\end{bmatrix} = -1.$$

It follows that A is nonsingular, so the only solution to the system $A\mathbf{c} = 0$ is $\mathbf{c} = 0$. Since every linear combination of the columns of A takes the form $A\mathbf{c}$, the vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ must be linearly independent.

Finally, we see by inspection in (c) that since $\mathbf{v}_3$ is a repeat of $\mathbf{v}_1$, we have that

$$\mathbf{v}_1 + 0\mathbf{v}_2 - \mathbf{v}_3 = \mathbf{0}.$$

Thus, this list of vectors is linearly dependent. Notice, by the way, that not every coefficient c_i has to be nonzero. □

Example 3.20. Show that any list of vectors that contains the zero vector is linearly dependent.

Solution. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be such a list and suppose that for some index j, $\mathbf{v}_j = 0$. Examine the following linear combination:

$$0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 1\mathbf{v}_j + \cdots + 0\mathbf{v}_n = \mathbf{0}.$$

This linear combination of value zero is nontrivial because the coefficient of the vector $\mathbf{v}_j$ is 1. Therefore, this list is linearly dependent by the definition of dependence. □

The Basis Idea

We are now ready for one of the big ideas of vector space theory, the notion of a basis. We already know what a spanning set for a vector space V is. This is a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ such that $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. However, we saw back in Example 3.17 that some spanning sets are better than others because they are more economical. We know that a set of vectors has no redundant vectors in it if and only if it is linearly independent. This observation is the inspiration for the following definition.

Definition 3.8. **Basis of Vector Space** A *basis* for the vector space V is a spanning set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ that is a linearly independent set.

Basis As Minimal Spanning Set

We should take note here that we could have just as well defined a basis as a minimal spanning set, by which we mean a spanning set such that any proper subset is not a spanning set. The proof that this is equivalent to our definition of basis is left as an exercise.

Ordered Basis

Usually we think of a basis as a set of vectors and the order in which we list them is convenient but not important. Occasionally, ordering is important. In such a situation we speak of an *ordered basis* of $\mathbf{v}$, by which we mean a spanning list of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ that is a linearly independent list.

Example 3.21. Which subsets of $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} = \left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}\right\}$ yield bases of the vector space $\mathbb{R}^2$?

Solution. These are just the vectors of Example 3.17 and Example 3.18. Referring back to that example, we saw that

$$-\mathbf{v}_1 + 2\mathbf{v}_2 - \mathbf{v}_3 = 0,$$

which told us that we could remove any one of these vectors and get the same span. Moreover, we saw that these three vectors span $\mathbb{R}^2$, so the same is true of any two of them. Clearly, a single vector cannot span $\mathbb{R}^2$, since the span of a single vector is a line through the origin. Therefore, the subsets $\{\mathbf{v}_1, \mathbf{v}_2\}$, $\{\mathbf{v}_2, \mathbf{v}_3\}$, and $\{\mathbf{v}_1, \mathbf{v}_3\}$ are all bases of $\mathbb{R}^2$. □

Example 3.22. Which subsets of $\{1+x,\ x+x^2,\ 1,\ x\}$ yield bases of the vector space $\mathcal{P}_2$ of all polynomials of degree at most two?

Solution. Any linear combination of $1+x$, 1, and x yields a linear polynomial, so cannot equal x^2. Hence, $x+x^2$ must be in the basis. On the other hand, any element of the set $\{x, 1+x, 1\}$ can be expressed as a combination of the other two, so is redundant in the set. Discard redundant vectors from this set and we obtain three candidates for bases of $\mathcal{P}_2$: $\{x+x^2, 1+x, 1\}$, $\{x+x^2, x, 1\}$, and $\{x+x^2, x, 1+x\}$. It's easy to see that the span of any one of these sets contains 1, x, and x^2, so is a spanning set for $\mathcal{P}_2$. We leave it to the reader to check that each set contains no redundant vectors, hence is linearly independent. Therefore, each of these sets forms a basis of $\mathcal{P}_2$. □

An extremely important generic type of basis is provided by the columns of the identity matrix. For future reference, we establish this notation. The *standard basis* of $\mathbb{R}^n$ or $\mathbb{C}^n$ is the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_j$ is the jth column of the identity matrix I_n.

Standard Basis

Example 3.23. Let V be the standard vector space $\mathbb{R}^n$ or $\mathbb{C}^n$. Verify that the standard basis really is a basis of this vector space.

Solution. Let $\mathbf{v} = (c_1, c_2, \ldots, c_n)$ be a vector from V so that $c_1, c_2, \ldots, c_n$ are scalars of the appropriate type. Now we have

$$\begin{aligned}\mathbf{v} &= \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + c_n \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \\ &= c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_n\mathbf{e}_n.\end{aligned}$$

This equation tells us two things: First, every vector in V is a linear combination of the $\mathbf{e}_j$'s, so $V = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. Second, if some linear combination of vectors has value zero, then each scalar coefficient of the combination is 0. Therefore, these vectors are linearly independent. Therefore, the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is a basis of V. □

Coordinates

In the case of the standard basis $\mathbf{e}_1, \mathbf{e}_2, , \mathbf{e}_3$ of $\mathbb{R}^3$ we know that it is very easy to write out any other vector $\mathbf{v} = (c_1, c_2, c_3)$ in terms of the standard basis:

$$\mathbf{v} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + c_3\mathbf{e}_3.$$

We call the scalars c_1, c_2, c_3 the *coordinates* of the vector $\mathbf{v}$. Up to this point, this is the only sense in which we have used the term "coordinates." We can

see that these coordinates are strongly tied to the standard basis. Yet $\mathbb{R}^3$ has many bases. Is there a corresponding notion of "coordinates" relative to other bases? The answer is a definite yes, thanks to the following fact.

Theorem 3.6. Uniqueness of Coordinates Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be a basis of the vector space V. Then every $\mathbf{v} \in V$ can be expressed uniquely as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, up to order of terms.

Proof. To see this, note first that since $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, there exist scalars $c_1, c_2, \ldots, c_n$ such that

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n.$$

Suppose that we could also write

$$\mathbf{v} = d_1\mathbf{v}_1 + d_2\mathbf{v}_2 + \cdots + d_n\mathbf{v}_n.$$

Subtract these two equations and obtain

$$\mathbf{0} = (c_1 - d_1)\mathbf{v}_1 + (c_2 - d_2)\mathbf{v}_2 + \cdots + (c_n - d_n)\mathbf{v}_n.$$

However, a basis is a linearly independent set, so it follows that each coefficient of this equation is zero, whence $c_j = d_j$, for $j = 1, 2, \ldots, n$. □

In view of this fact, we may speak of *coordinates of a vector relative to a basis*. Here is the notation that we employ:

Definition 3.9. Vector Coordinates and Coordinate Vector If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis B of the vector space V and $\mathbf{v} \in V$ with $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$, then the scalars $c_1, c_2, \ldots, c_n$ are called the coordinates of $\mathbf{v}$ *with respect to the basis* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. The coordinate vector of $\mathbf{v}$ with respect to B is $[\mathbf{v}]_B = (c_1, c_2, \ldots, c_n)$.

Standard Coordinates

As we have noted, coordinates of a vector with respect to the standard basis are what we have referred to as "coordinates" so far in this text. Perhaps we should call these the *standard coordinates* of a vector, but we will usually stick to the convention that an unqualified reference to a vector's coordinates assumes that we mean standard coordinates unless otherwise stated. Normally, vectors in $\mathbb{R}^n$ are given explicitly in terms of their standard coordinates, so these are trivial to identify. Coordinates with respect to other bases are fairly easy to calculate if we have enough information about the structure of the vector space.

Example 3.24. The following vectors form a basis of $\mathcal{P}_2$: $B = \{x + x^2, 1 + x, 1\}$ (see Example 3.22). Find the coordinate vector of $p(x) = 2 - 2x - x^2$ with respect to this basis.

Solution. The coordinates are c_1, c_2, c_3, where

$$2 - 2x - x^2 = c_1\left(x + x^2\right) + c_2\,(1 + x) + c_3 \cdot 1 = (c_2 + c_3) + (c_1 + c_2)\,x + c_1 x^2.$$

We note here that the order in which we list the basis elements matters for the coordinates. Now we simply equate coefficients of like powers of x to obtain that $c_2 + c_3 = 2$, $c_1 + c_2 = -2$, and $c_1 = -1$. It follows that $c_2 = -2 - c_1 = -1$ and that $c_3 = 2 - c_2 = 3$. Thus, $[p\,(x)]_B = (-1, -1, 3)$. Incidentally, we note here that the order in which we list the basis elements matters for the coordinates. □

Example 3.25. The following vectors form a basis B of $\mathbb{R}^3$: $\mathbf{v}_1 = (1, 1, 0)$, $\mathbf{v}_2 = (0, 2, 2)$, and $\mathbf{v}_3 = (1, 0, 1)$. Find the coordinate vector of $\mathbf{v} = (2, 1, 5)$ with respect to this basis.

Solution. Notice that the basis $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ was given in terms of standard coordinates. Begin by writing

$$\begin{aligned}\mathbf{v} = \begin{bmatrix} 2 \\ 1 \\ 5 \end{bmatrix} &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 \\ &= [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix},\end{aligned}$$

where the coordinates c_1, c_2, c_3 of $\mathbf{v}$ relative to the basis $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are to be determined. This is a straightforward system of equations with coefficient matrix $A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$ and right-hand side $\mathbf{v}$. It follows that the solution we want is given by

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}.$$

This shows us that

$$\mathbf{v} = -1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} + 3 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

It does *not* prove that $\mathbf{v} = (-1, 1, 3)$, which is plainly false. Only in the case of the standard basis can we expect that a vector actually equals its vector of coordinates with respect to the basis. What we have is that the coordinate vector of $\mathbf{v}$ with respect to basis B is $[\mathbf{v}]_B = (-1, 1, 3)$. □

In general, vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ are linearly independent if and only if the system $A\mathbf{c} = 0$ has only the trivial solution, where $A = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$. This in turn is equivalent to the matrix A being of full column rank n.

(See Theorem 2.6, where we see that these are equivalent conditions for a matrix to be invertible). We can see how this idea can be extended, and doing so tells us something remarkable. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ be a basis of $V = \mathbb{R}^n$ and form the $n \times k$ matrix $A = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$. By the same reasoning as in the example, for any $\mathbf{b} \in V$ there is a unique solution to the system $A\mathbf{x} = \mathbf{b}$. In view of Theorem 1.5 we see that A has full column rank k. Therefore, $k \leq n$. On the other hand, we can take $\mathbf{b}$ to be any one of the standard basis vectors $\mathbf{e}_j$, $j = 1, 2, \ldots, n$, solve the resulting systems, and stack the solution vectors together to obtain a solution to the system $AX = I_n$. From our rank inequalities, we see that

Dimension Theorem for $\mathbb{R}^n$

$$n = \operatorname{rank} I_n = \operatorname{rank} AX \leq \operatorname{rank} A = k.$$

What this shows is that $k = n$, that is, every basis of $\mathbb{R}^n$ has exactly n elements in it, which would justify calling n the *dimension* of the space $\mathbb{R}^n$. Amazing! Does this idea extend to abstract vector spaces? Indeed it does, and we shall return to this issue in Section 3.5. Among other things, we have shown the following handy fact, which gives us yet one more characterization of invertible matrices to add to Theorem 2.6.

Theorem 3.7. An $n \times n$ real matrix A is invertible if and only if its columns are linearly independent, in which case they form a basis of $\mathbb{R}^n$.

Here is a problem that comes to us straight from analytical geometry (classification of conics) and shows how the matrix and coordinate tools we have developed can shed light on geometrical problems.

Example 3.26. Suppose we want to understand the character of the graph of the curve $x^2 - xy + y^2 - 6 = 0$. It is suggested to us that if we execute a change of variables by rotating the xy-axis by $\pi/4$ to get a new $x'y'$-axis, the graph will become more intelligible. OK, we do it. The algebraic connection between the coordinate pairs x, y and x', y' representing the same point in the plane and resulting from a rotation of θ can be worked out using a bit of trigonometry (which we omit) to yield

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta \\ y' &= -x\sin\theta + y\cos\theta. \end{aligned}$$

Use matrix methods to formulate these equations and execute the change of variables.

Solution. First, we write the change of variable equations in matrix form as

Givens and Rotation Matrix

$$\mathbf{x}' = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = G(\theta)\,\mathbf{x}.$$

(Such a matrix $G(\theta)$ is often referred to as a *Givens* matrix.) This matrix isn't exactly what we need for substitution into our curve equation. Rather, we need x, y explicitly. That's easy enough. Simply invert $G(\theta)$ to obtain the rotation matrix $R(\theta)$ as

$$G(\theta)^{-1} = R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Therefore, $\mathbf{x} = G(\theta)^{-1}\mathbf{x}' = R(\theta)\mathbf{x}'$. Now observe that the original equation can be put in the form (as in Example 2.37)

$$\begin{aligned} x^2 - xy + y^2 - 6 &= \mathbf{x}^T \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \mathbf{x} - 6 \\ &= (\mathbf{x}')^T \mathbf{R}(\theta)^T \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \mathbf{R}(\theta)\mathbf{x}' - 6. \end{aligned}$$

We leave it as an exercise to check that with $\theta = \pi/4$, so that $\cos\theta = 1/\sqrt{2} = \sin\theta$, the equation reduces to $\frac{1}{2}(x'^2 + 3y'^2) - 6 = 0$ or equivalently

$$\frac{x'^2}{12} + \frac{y'^2}{4} = 1.$$

This curve is an ellipse with semimajor axis of length $2\sqrt{3}$ and semiminor axis of length 2. With respect to the $x'y'$-axes, this ellipse is in standard form. For a graph of the ellipse, see Figure 3.4. □

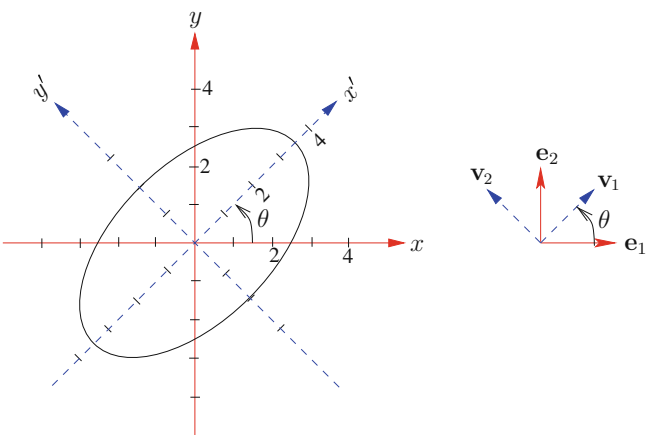


Fig. 3.4: Change of variables and the curve $x^2 - xy + y^2 - 6 = 0$.

The change of variables we have just seen can be interpreted as a *change of coordinates* in the following sense: The variables x and y are just the standard coordinates (with respect to the standard basis $B = \{\mathbf{e}_1, \mathbf{e}_2\}$) of a general vector

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} 0 \\ 1 \end{bmatrix} = x\mathbf{e}_1 + y\mathbf{e}_2.$$

The meaning of the variables x' and y' becomes clear when we set $\mathbf{x}' = (x', y')$ and write the matrix equation $\mathbf{x} = R(\theta)\mathbf{x}'$ out in detail as a linear combination of the columns of $R(\theta)$:

$$\mathbf{x} = R(\theta)\mathbf{x}' = x' \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + y' \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} = x'\mathbf{v}_1 + y'\mathbf{v}_2.$$

Thus, the numbers x' and y' are just the coordinates of the vector $\mathbf{x}$ with respect to a new basis $C = \{\mathbf{v}_1, \mathbf{v}_2\}$ of $\mathbb{R}^2$. This basis consists of unit vectors in the direction of the x' and y' axes. See Figure 3.4 for a picture of the two bases. For these reasons, the matrix $R(\theta)$ is sometimes called a *change of coordinates* matrix.

The matrix $R(\theta)$ is also called a *change of basis* matrix, due to the fact that the coordinate equation above is equivalent to $[\mathbf{x}]_B = R(\theta)[\mathbf{x}]_C$. Thus, $R(\theta)$ shows us how to change from the basis $C = \{\mathbf{v}_1, \mathbf{v}_2\}$ to the standard basis $B = \{\mathbf{e}_1, \mathbf{e}_2\}$. What makes a change of basis desirable is that sometimes a problem looks a lot easier if we look at it using a basis other than the standard one, such as in our example.

From a change of coordinates perspective, the vectors $\mathbf{x}$ and $\mathbf{x}'$ simply represent different coordinates for the same point and are connected by way of the formula $\mathbf{x} = R(\theta)\mathbf{x}'$. This is to be contrasted with the use of the rotation matrix $R(\theta)$ in Example 2.18. In that example we have only one coordinate system — the standard one — and we move a vector $\mathbf{x}$ by way of a rotation of θ in the counterclockwise direction to a new vector $\mathbf{y}$. This defined a linear operator, and the connection between the two vectors is that $\mathbf{y} = R(\theta)\mathbf{x} = T_{R(\theta)}(\mathbf{x})$.

Change of Basis Matrix

In general, a change of basis matrix from basis C to basis B of vector space V is a matrix P such that for any vector $\mathbf{v} \in V$, $[\mathbf{v}]_B = P[\mathbf{v}]_C$. These matrices are treated in more detail in Section 3.7. However, we will record this simple fact about change of basis matrices.

Theorem 3.8. Change of Basis Formula If $V = \mathbb{R}^n$, B is the standard basis, and $C = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ any other basis, then the change of basis matrix from basis C to B is $P = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$.

Proof. To see this, note first that for any $\mathbf{v} \in V$, we have $\mathbf{v} = [\mathbf{v}]_B$ since B is the standard basis. Let

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n,$$

so that $c_1, c_2, \ldots, c_n$ are the coordinates of $\mathbf{v}$ relative to C, i.e., Then

$$\mathbf{v} = [\mathbf{v}]_B = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n][c_1, c_2, \ldots, c_n]^T = P[\mathbf{v}]_C,$$

which shows that P is the change of basis matrix from B to C. □

3.3 Exercises and Problems

Exercise 1. Find the redundant vectors, if any, in the following lists.
(a) $(1,0,1)$, $(1,-1,1)$ (b) $(1,2,1)$, $(2,1,1)$, $(3,3,2)$, $(2,0,1)$
(c) $(1,0,-1)$, $(1,1,0)$, $(1,-1,-2)$ (d) $(0,1,-1)$, $(1,0,0)$, $(-1,1,3)$

Exercise 2. Find the redundant vectors, if any, in the following lists.
(a) x, $5x$ (b) 2, $2-x$, x^2, $1+x^2$
(c) $1+x$, $1+x^2$, $1+x+x^2$ (d) $x-1$, x^2-1, $x+1$

Exercise 3. Which of the following sets are linearly independent in $V=\mathcal{P}_3$? If not linearly independent, which vectors are redundant in the lists?
(a) $1, x, x^2, x^3$ (b) $1+x, 1+x^2, 1+x^3$
(c) $1-x^2, 1+x, 1-x-2x^2$ (d) $x^2-x^3, x, -x+x^2+3x^3$

Exercise 4. Which of the following sets are linearly independent in $V=\mathbb{R}^4$? If not linearly independent, which vectors are redundant in the lists?
(a) $(1,-1,0,1)$, $(-2,2,1,1)$ (b) $(1,1,0,0)$, $(1,0,1,0)$, $(1,0,0,1)$, $(-1,1,-2,0)$
(c) $(0,1,-1,2)$, $(0,1,3,4)$, $(0,2,2,6)$ (d) $(1,1,1,1)$, $(0,2,0,0)$, $(0,2,1,1)$

Exercise 5. Find the coordinates of $\mathbf{v}$ with respect to the following bases:
(a) $\mathbf{v}=(-1,1)$, basis $(2,1)$, $(2,-1)$ of $\mathbb{R}^2$.
(b) $\mathbf{v}=2+x^2$, basis $1+x$, $x+x^2$, $1-x$ of $\mathcal{P}_2$.
(c) $\mathbf{v}=\begin{bmatrix} a & b \\ b & c \end{bmatrix}$, basis $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ of the space of real symmetric 2×2 matrices.
(d) $\mathbf{v}=(1,2)$, basis $(2+\mathrm{i},1)$, $(-1,\mathrm{i})$ of $\mathbb{C}^2$.

Exercise 6. Find the coordinate vector of $\mathbf{v}$ with respect to the following bases:
(a) $\mathbf{v}=(0,1,2)$, basis $(2,0,1)$, $(-1,1,0)$, $(0,1,1)$ of $\mathbb{R}^3$.
(b)$\mathbf{v}=\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$, basis $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ of the space of upper triangular 2×2 matrices.
(c) $\mathbf{v}=(1,\mathrm{i},\mathrm{i})$, basis $(1,1,0)$, $(0,1,1)$, $(0,0,\mathrm{i})$ of $\mathbb{C}^3$.
(d) $\mathbf{v}=4$, basis $1+2x$, $1-x$ of $\mathcal{P}_1$.

Exercise 7. Let $\mathbf{u}_1=(1,0,1)$ and $\mathbf{u}_2=(1,-1,1)$.
(a) Determine whether $\mathbf{v}=(2,1,2)$ belongs to the space span $\{\mathbf{u}_1,\mathbf{u}_2\}$.
(b) Find a basis of $\mathbb{R}^3$ that contains $\mathbf{u}_1$ and $\mathbf{u}_2$.

Exercise 8. Let $\mathbf{u}_1=1-x+x^2$ and $\mathbf{u}_2=x+2x^2$.
(a) Determine whether $\mathbf{v}=4-7x-x^2$ belongs to the space span $\{\mathbf{u}_1,\mathbf{u}_2\}$.
(b) Find a basis of $\mathcal{P}_2$ that contains $\mathbf{u}_1$ and $\mathbf{u}_2$.

Exercise 9. If $\mathbf{v}_2 + 2\mathbf{v}_3 = \mathbf{0}$, find all subsets of the vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ that could form a minimal spanning set of span $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$.

Exercise 10. If $2\mathbf{v}_1 + \mathbf{v}_3 + \mathbf{v}_4 = \mathbf{0}$ and $\mathbf{v}_2 + \mathbf{v}_3 = \mathbf{0}$, find all subsets of the vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ that could form a minimal spanning set of span $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$.

Exercise 11. For what values of the parameter c is the set of vectors $(1, 1, c)$, $(2, c, 4)$, $(3c + 1, 3, -4)$ in $\mathbb{R}^3$ linearly independent?

Exercise 12. For what values of the parameter λ is the set of vectors $\left(1, \lambda^2, 1, 2\right)$, $(2, \lambda, 4, 8)$, $(0, 0, 1, 2)$ in $\mathbb{R}^4$ linearly dependent?

Exercise 13. Let e_{ij} be a matrix with a one in the (i, j)th entry and zeros elsewhere. Which 2×2 matrices e_{ij} can be added to the set below to form a basis of $\mathbb{R}^{2,2}$?

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

Exercise 14. Which 2×2 matrices e_{ij} can be added to the set below to form a basis of $\mathbb{R}^{2,2}$?

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Exercise 15. The Wronskian of smooth functions $f(x), g(x), h(x)$ is defined as

$$W(f, g, h)(x) = \det \begin{bmatrix} f(x) & g(x) & h(x) \\ f'(x) & g'(x) & h'(x) \\ f''(x) & g''(x) & h''(x) \end{bmatrix}.$$

(A similar definition can be made for any number of functions.) Calculate the Wronskians of the polynomial functions of Exercise 2 (c) and (d). What does Problem 25 tell you about these calculations?

Exercise 16. Show that the functions e^x, x^3, and $\sin(x)$ are linearly independent in $C[0, 1]$ in two ways:

(a) Use Problem 25.

(b) Assume that a linear combination with value zero exists and evaluate it at various points to obtain conditions on the coefficients.

Exercise 17. Let $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ and $A = \begin{bmatrix} 1 & \frac{-1}{2} \\ \frac{-1}{2} & 1 \end{bmatrix}$. Calculate $R(\theta)^T A R(\theta)$ in the case that $\theta = \pi/4$.

Exercise 18. Use matrix methods as in Example 3.26 to express the equation of the curve $11x^2 + 10\sqrt{3}xy + y^2 - 16 = 0$ in new variables x', y' obtained by rotating the xy-axis by an angle of $\pi/6$.

Problem 19. Show that for $m \times n$ matrix A and vector $\mathbf{b} \in \mathbb{R}^m$, if the vectors $\mathbf{u_j} \in \mathbb{R}^n$, $j = 1, 2, \ldots, N$, solve the equation $A\mathbf{x} = \mathbf{b}$, then so does the linear *convex combination* $\mathbf{w} = \sum_{j=1}^{N} \alpha_j \mathbf{u}_j$, where all $\alpha_j \geq 0$ and $\sum_{j=1}^{N} \alpha_j = 1$.

Problem 20. Let $V = \mathbb{R}^{n,n}$ be the vector space of real $n \times n$ matrices and let $A, B \in \mathbb{R}^{n,n}$ be such that both are nonzero matrices, A is nilpotent (some power of A is zero), and B is idempotent ($B^2 = B$). Show that the subspace $W = \text{span}\{A, B\}$ cannot be spanned by a single element of W.

Problem 21. Show that a basis is a minimal spanning set and conversely.

Problem 22. Let V be a vector space whose only subspaces are $\{\mathbf{0}\}$ and V. Show that V is the span of a single vector.

*Problem 23. Prove that a list of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ with repeated vectors in it is linearly dependent.

Problem 24. Suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are linearly independent elements of $\mathbb{R}^n$ and $A = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$. Show that $\text{rank}\, A = k$.

*Problem 25. Show that smooth functions $f(x), g(x), h(x)$ are linearly dependent if and only if for all x, $W(f, g, h)(x) = 0$.

Problem 26. Show that a linear operator $T : V \to W$ maps a linearly dependent set $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ to linearly dependent set $T(\mathbf{v}_1), T(\mathbf{v}_2), \ldots, T(\mathbf{v}_n)$, but if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent, $T(\mathbf{v}_1), T(\mathbf{v}_2), \ldots, T(\mathbf{v}_n)$ need not be linearly independent (give a specific counterexample).

*Problem 27. Suppose that a linear change of variables from old coordinates x_1', x_2' to new coordinates x_1, x_2 is given by the equations

$$\begin{aligned} x_1 &= p_{11}x_1' + p_{12}x_2', \\ x_2 &= p_{21}x_1' + p_{22}x_2', \end{aligned}$$

where the 2×2 change of basis matrix $P = [p_{ij}]$ is invertible. Show that if a linear matrix multiplication function $T_A : \mathbb{R}^2 \to \mathbb{R}^2$ is given in new coordinates by

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = T_A\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = T_A(\mathbf{x}) = A\mathbf{x},$$

where $A = [a_{ij}]$ is any 2×2 matrix, then it is given by $\mathbf{y}' = P^{-1}AP\mathbf{x}' = T_{P^{-1}AP}(\mathbf{x}')$ in old coordinates.

3.4 Subspaces Associated with Matrices and Operators

Certain subspaces are a rich source of information about the behavior of a matrix or a linear operator. We define and explore the properties of these subspaces in this section.

Subspaces Defined by Matrices

There are three very useful subspaces that can be associated with a matrix A. Understanding these subspaces is a great aid in vector space calculations that might have nothing to do with matrices per se, such as the determination of a minimal spanning set for a vector space. Each definition below is followed by an illustration using the following example matrix:

$$A = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 0 & 1 & 2 & 1 \end{bmatrix}. \tag{3.2}$$

We make the default assumption that the scalars are the real numbers, but the definitions we will give can be stated just as easily for the complex numbers.

Caution: Do not confuse any of the spaces defined below with the matrix A itself. They are objects that are derived from the matrix, but do not even uniquely determine the matrix A.

Definition 3.10. **Column Space** The *column space* of the $m \times n$ matrix A is the subspace $\mathcal{C}(A)$ of $\mathbb{R}^m$ spanned by the columns of A.

Example 3.27. Describe the column space of the matrix A in equation (3.2).

Solution. Here we have that $\mathcal{C}(A) \subseteq \mathbb{R}^2$. Also

$$\mathcal{C}(A) = \text{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

Technically, this describes the column space in question, but we can do better. We saw in Example 3.17 that the vector $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ was really redundant since it is a linear combination of the first two vectors. We also see that

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} = -2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

so that Theorem 3.4 shows us that

$$\mathcal{C}(A) = \text{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}.$$

This description is much better, in that it exhibits a *basis* of $\mathcal{C}(A)$. It also shows that not all the columns of the matrix A are really needed to span the entire subspace $\mathcal{C}(A)$. □

Definition 3.11. Row Space The *row space* of the $m \times n$ matrix A is the subspace $\mathcal{R}(A)$ of $\mathbb{R}^n$ spanned by the transposes of the rows of A.

The "transpose" part of the preceding definition seems a bit odd. Why would we want rows to look like columns? It's a technicality, but later it will be convenient for us to have the row spaces live inside a $\mathbb{R}^n$ instead of an $(\mathbb{R}^n)^T$. Remember, we had to make a choice about $\mathbb{R}^n$ consisting of rows or columns. Just to make the elements of a row space look like rows, we can always adhere to the tuple notation instead of matrix notation. We gain one convenience: $\mathcal{R}(A) = \mathcal{C}(A^T)$, so that whatever we understand about column spaces can be applied to row spaces.

Example 3.28. Describe the row space of A in equation (3.2).

Solution. We have from the definition that

$$\mathcal{R}(A) = \text{span}\{(1,1,1,-1),(0,1,2,1)\} \subseteq \mathbb{R}^4.$$

Now it's easy to see that neither one of these vectors can be expressed as a multiple of the other (if we had $c(1,1,1,-1) = (0,1,2,1)$, then read the first coordinates and obtain $c = 0$), so that span is given as economically as we can do, that is, the two vectors listed constitute a *basis* of $\mathcal{R}(A)$. □

Definition 3.12. Null Space The *null space* of the $m \times n$ matrix A is the subset $\mathcal{N}(A)$ of $\mathbb{R}^n$ defined to be

$$\mathcal{N}(A) = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\}.$$

Observe that $\mathcal{N}(A)$ is the solution set to the homogeneous linear system $A\mathbf{x} = \mathbf{0}$. This means that null spaces are really very familiar. We were computing these solution sets way back in Chapter 1. We didn't call them subspaces at the time. Here is an application of this concept. Let A be a square matrix. We know that A is invertible exactly when the system $A\mathbf{x} = \mathbf{0}$ has only the trivial solution (see Theorem 2.6). Now we can add one more equivalent condition to the long list of equivalences for invertibility: A is invertible exactly if $\mathcal{N}(A) = \{\mathbf{0}\}$. We next justify the subspace property implied by the term "null space."

Example 3.29. Use the subspace test to verify that $\mathcal{N}(A)$ is a subspace of $\mathbb{R}^n$.

Solution. Since $A\mathbf{0} = \mathbf{0}$, the zero vector is in $\mathcal{N}(A)$. Now let c be a scalar and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ arbitrary elements of $\mathcal{N}(A)$. By definition, $A\mathbf{u} = \mathbf{0}$ and $A\mathbf{v} = \mathbf{0}$. Add these two equations to obtain that

$$\mathbf{0} = \mathbf{0} + \mathbf{0} = A\mathbf{u} + A\mathbf{v} = A(\mathbf{u} + \mathbf{v}).$$

Therefore, $\mathbf{u} + \mathbf{v} \in \mathcal{N}(A)$. Next multiply the equation $A\mathbf{u} = \mathbf{0}$ by the scalar c to obtain

$$\mathbf{0} = c\mathbf{0} = c(A\mathbf{u}) = A(c\mathbf{u}).$$

Thus, we see from that definition that $c\mathbf{u} \in \mathcal{N}(A)$. The subspace test implies that $\mathcal{N}(A)$ is a subspace of $\mathbb{R}^n$. □

Example 3.30. Describe the null space of the matrix A of equation (3.2).

Solution. Proceed as in Section 1.4. We find the reduced row echelon form of A, identify the free variables, and solve for the bound variables using the implied zero right-hand side and solution vector $x = [x_1, x_2, x_3, x_4]^T$:

$$\begin{bmatrix} 1 & 1 & 1 & -1 \\ 0 & 1 & 2 & 1 \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 1 \end{bmatrix}.$$

Pivots are in the first and second columns, so it follows that x_3 and x_4 are free, x_1 and x_2 are bound, and

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

Let's write out the form of a general solution in terms of the free variables as a combination of x_3 times some vector plus x_4 times another vector:

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

We have seen this clever trick before in Example 2.6. Remember that free variables can take on arbitrary values, so we see that the general solution to the homogeneous system has the form of an arbitrary linear combination of the two vectors on the right. In other words,

$$\mathcal{N}(A) = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ 0 \\ 1 \end{bmatrix} \right\} \subseteq \mathbb{R}^4.$$

Neither of these vectors is a multiple of the other, so this is as economical an expression for $\mathcal{N}(A)$ as we can hope for. In other words, we have exhibited a minimal spanning set, that is, a *basis* of $\mathcal{N}(A)$. □

The following example relates null spaces to the idea of a limiting state for a Markov chain as discussed in Example 2.20. Recall that in that example we observed that the sequence of state vectors $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, appeared to converge to a steady-state vector $\mathbf{x}$, no matter what the initial (probability distribution) state vector $\mathbf{x}^{(0)}$. We will call a stochastic

matrix (Markov chain transition matrix) A that has this property *stable*. Null spaces can tell us something about such matrices. In Chapter 5 we will apply this concept to general discrete dynamical systems.

Stable Stochastic Matrix

Example 3.31. Suppose that a Markov chain has an stable transition matrix $A = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix}$. Determine the steady-state vector for the Markov chain

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}.$$

Solution. We reason as follows: Since the limit of the state vectors $\mathbf{x}^{(k)}$ is $\mathbf{x}$, and the state vectors are related by the formula

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)},$$

we can take the limits of both sides of this matrix equation and obtain that $\mathbf{x} = A\mathbf{x}$. Therefore,

$$\mathbf{0} = \mathbf{x} - A\mathbf{x} = I\mathbf{x} - A\mathbf{x} = (I - A)\mathbf{x}.$$

It follows that $\mathbf{x} \in \mathcal{N}(I - A)$. Now

$$I - A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.3 & -0.4 \\ -0.3 & 0.4 \end{bmatrix}.$$

Calculate the null space by Gauss–Jordan elimination:

$$\begin{bmatrix} 0.3 & -0.4 \\ -0.3 & 0.4 \end{bmatrix} \xrightarrow[E_1(1/0.3)]{E_{21}(1)} \begin{bmatrix} 1 & -4/3 \\ 0 & 0 \end{bmatrix}.$$

Therefore, the null space of $I - A$ is spanned by the single vector $(4/3, 1)$. In particular, any multiple of this vector qualifies as a possible limiting vector. If we want a limiting vector whose entries are nonnegative and sum to 1 (which is required for states in a Markov chain), then the only choice is the vector resulting from dividing $(4/3, 1)$ by the sum of its coordinates to obtain

$$(3/7)(4/3, 1) = (4/7, 3/7) \approx (0.57143, 0.42857).$$

Interestingly enough, this is the vector that was calculated on page 91. □

Caution: We have no guarantee that the transition matrix A of the preceding example is actually stable. We have only experimental evidence so far. We will *prove* stability using eigenvalue ideas in Chapter 5.

Here is a way of thinking about $\mathcal{C}(A)$. The key is the "linear combination as matrix–vector multiplication" idea that was first introduced

Column Space as Matrix Products in Example 2.9 and formalized in Theorem 2.1. Recall that it asserts that if matrix A has columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$, i.e., $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$, and if $\mathbf{x} = [x_1, x_2, ..., x_n]^T$, then

$$A\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n. \tag{3.3}$$

This equation shows that the column space of the matrix A can be thought of as the set of all possible matrix products $A\mathbf{x}$, i.e.,

$$\mathcal{C}(A) = \{A\mathbf{x} \,|\, \mathbf{x} \in \mathbb{R}^n\}.$$

An insight that follows from these observations: The linear combination of columns of A with coefficients from the vector $\mathbf{x}$ is zero exactly when $\mathbf{x} \in \mathcal{N}(A)$. Thus, we can use null space calculations to identify redundant vectors in a set of column vectors, as in the next example.

Example 3.32. Find all possible linear combinations with value zero of the columns of matrix A of equation (3.2) and use this information to find a basis of $\mathcal{C}(A)$.

Solution. As in Example 3.30 we find the reduced row echelon form of A, identify the free variables, and solve for the bound variables using the implied zero right-hand side. The result is a solution vector $\mathbf{x} = (x_1, x_2, x_3, x_4) = (x_3 + 2x_4, -2x_3 - x_4, x_3, x_4)$. Write $A = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4]$, and we see that the linear combinations of A are just

$$\mathbf{0} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3 + x_4\mathbf{a}_4 = (x_3 + 2x_4)\,\mathbf{a}_1 - (2x_3 + x_4)\,\mathbf{a}_2 + x_3\mathbf{a}_3 + x_4\mathbf{a}_4.$$

Here we think of x_3 and x_4 as free variables. Take $x_3 = 1$ and $x_4 = 0$, and we obtain $\mathbf{0} = \mathbf{a}_1 - 2\mathbf{a}_2 + \mathbf{a}_3$, so that $\mathbf{a}_3$ is a linear combination of $\mathbf{a}_1$ and $\mathbf{a}_2$. Similarly, take $x_3 = 0$ and $x_4 = 1$, and we obtain $\mathbf{0} = 2\mathbf{a}_1 - \mathbf{a}_2 + \mathbf{a}_4$, so that $\mathbf{a}_4$ is a linear combination of $\mathbf{a}_1$ and $\mathbf{a}_2$. Hence, $\mathcal{C}(A) = \text{span}\{\mathbf{a}_1, \mathbf{a}_2\} = \text{span}\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right\}$, the same conclusion we reached by trial and error in Example 3.27. □

Subspaces Defined by a Linear Operator

Suppose we are given a linear operator $T : V \to W$. We immediately have three spaces we can associate with the operator, namely, the domain V, target W, and range $T(V) = \{\mathbf{y} \,|\, \mathbf{y} = T(\mathbf{x}) \text{ for some } \mathbf{x} \in V\}$ of the operator. The domain and range are vector spaces by definition of linear operator. That the range is a vector space is a nice application of the subspace test.

Example 3.33. Show that if $T : V \to W$ is a linear operator, then range(T) is a subspace of W.

Solution. Apply the subspace test. First, we observe that range(T) contains $T(0)$. We leave it as an exercise for the reader to check that $T(0)$ is the zero element of W. Next let $\mathbf{y}$ and $\mathbf{z}$ be in range(T), say $\mathbf{y} = T(\mathbf{u})$ and $\mathbf{z} = T(\mathbf{v})$. We show closure of range(T) under addition: by the linearity property of T,

$$\mathbf{y} + \mathbf{z} = T(\mathbf{u}) + T(\mathbf{v}) = T(\mathbf{u} + \mathbf{v}) \in \text{range}(T),$$

where the latter term belongs to range(T) by the definition of image. Finally, we show closure under scalar multiplication: Let c be a scalar, and we obtain from the linearity property of T that

$$c\mathbf{y} = cT(\mathbf{u}) = T(c\mathbf{u}) \in \text{range}(T),$$

where the latter term belongs to range(T) by the definition of range. Thus, the subspace test shows that range(T) is a subspace of W. □

Here is another space that has proven to be very useful in understanding the nature of a linear operator.

Definition 3.13. **Kernel of Operator** The *kernel* of the linear operator $T : V \to W$ is the subspace of V defined by

$$\ker(T) = \{\mathbf{x} \in V \,|\, T(\mathbf{x}) = \mathbf{0}\}.$$

The definition claims that the kernel is a subspace and not merely a subset of the domain. This is true, and a proof of this fact is left to the exercises. In fact, we have been computing kernels since the beginning of the text. To see this, suppose that the linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is given by matrix multiplication, that is, $T(\mathbf{x}) = T_A(\mathbf{x}) = A\mathbf{x}$, for all $\mathbf{x} \in \mathbb{R}^n$. Then

$$\ker(T) = \{\mathbf{x} \in \mathbb{R}^n \,|\, T_A(\mathbf{x}) = \mathbf{0}\} = \{\mathbf{x} \in \mathbb{R}^n \,|\, A\mathbf{x} = \mathbf{0}\} = \mathcal{N}(A).$$

In other words, for matrix operators kernels are the same thing as null spaces.

Here is one very nice application of kernels. Suppose we are interested in knowing whether an operator $T : V \to W$ is one-to-one, i.e., whether the equation $T(\mathbf{u}) = T(\mathbf{v})$ always implies that $\mathbf{u} = \mathbf{v}$. For general functions this is a nontrivial question. If, for example, $V = W = \mathbb{R}$, then we could graph the function T and try to determine whether a horizontal line cut the graph twice. But for *linear* operators, the answer is very simple:

Theorem 3.9. The linear operator $T : V \to W$ is one-to-one if and only if $\ker(T) = \{\mathbf{0}\}$.

Proof. If T is one-to-one, then only one element can map to $\mathbf{0}$ under T. Thus, $\ker(T)$ can consist of only one element. However, $\ker(T)$ contains the zero vector since it is a subspace of the domain of T. Therefore, $\ker(T) = \{\mathbf{0}\}$.

Conversely, suppose that $\ker(T) = \{\mathbf{0}\}$. If $\mathbf{u}$ and $\mathbf{v}$ are such that $T(\mathbf{u}) = T(\mathbf{v})$, then subtract terms and use the linearity of T to obtain that

$$0 = T(\mathbf{u}) - T(\mathbf{v}) = T(\mathbf{u}) + (-1)T(\mathbf{v}) = T(\mathbf{u} - \mathbf{v}).$$

It follows that $\mathbf{u} - \mathbf{v} \in \ker(T) = \{\mathbf{0}\}$. Therefore, $\mathbf{u} - \mathbf{v} = \mathbf{0}$ and so $\mathbf{u} = \mathbf{v}$. □

Before we leave the topic of one-to-one linear mappings, let's digest its significance in a very concrete case. The space $\mathcal{P}_2 = \text{span}\{1, x, x^2\}$ of polynomials of degree at most 2 has a basis of three elements, like $\mathbb{R}^3$, and it seems very reasonable to think that $\mathcal{P}_2$ is "just like" $\mathbb{R}^3$ in that a polynomial $p(x) = a + bx + cx^2$ is uniquely described by its vector of coefficients $(a, b, c) \in \mathbb{R}^3$, and corresponding polynomials and vectors add and scalar multiply in a corresponding way. Here is the precise version of these musings: Define an operator $T : \mathcal{P}_2 \to \mathbb{R}^3$ by the formula $T(a + bx + cx^2) = (a, b, c)$. One can check that T is linear, the range of T is its target, $\mathbb{R}^3$, and $\ker(T) = 0$. By Theorem 3.9 the function T is one-to-one. Hence, it describes a one-to-one correspondence between elements of $\mathcal{P}_2$ and elements of $\mathbb{R}^3$ such that sums and scalar products in one space correspond to the corresponding sums and scalar products in the other. In plain words, this means we can get one of the vector spaces from the other simply by relabeling elements of one of the spaces. So, in a very real sense, they are "the same thing." More generally, whenever there is a one-to-one linear mapping of one vector space onto another, we say that the two vector spaces are *isomorphic*, which is a fancy way of saying that they are the same, up to a relabeling of elements. The mapping T itself is called an *isomorphism*. Actually, we have already encountered isomorphisms in the form of invertible linear operators. The following theorem, whose proof we leave as an exercise, explains the connection between these ideas.

Isomorphism and Isomorphic Vector Spaces

Theorem 3.10. The linear operator $T : V \to W$ is an isomorphism if and only if T is an invertible linear operator.

In summary, there are four important subspaces associated with a linear operator $T : V \to W$, the domain, target, kernel, and range. In symbols:

$$\begin{aligned} \text{domain}(T) &= V \\ \text{target}(T) &= W \\ \ker(T) &= \{\mathbf{v} \in V \,|\, T(\mathbf{v}) = 0\} \\ \text{range}(T) &= \{T(\mathbf{v}) \,|\, \mathbf{v} \in V\}. \end{aligned}$$

There are important connections between these subspaces and those associated with a matrix. Let A be an $m \times n$ matrix and $T_A : \mathbb{R}^n \to \mathbb{R}^m$ the corresponding matrix operator defined by multiplication by A. We have

$$\begin{aligned} \text{domain}(T_A) &= \mathbb{R}^n \\ \text{target}(T_A) &= \mathbb{R}^m \\ \ker(T_A) &= \mathcal{N}(A) \\ \text{range}(T_A) &= \mathcal{C}(A). \end{aligned}$$

The proofs of these are left to the exercises. One last example of subspaces associated with a linear operator $T : V \to W$ is really a whole family of subspaces. Suppose that U is a subspace of the domain V. Then we define the *image of U under T* to be the set

$$T(U) = \{T(u) \,|\, u \in U\}\,.$$

One can show that $T\,(U)$ is always a subspace of range (T). We leave the proof of this fact as an exercise. What this says is that a linear operator maps subspaces of its domain into subspaces of its range.

3.4 Exercises and Problems

Exercise 1. Find bases for null spaces of the following matrices.

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

Exercise 2. Find bases for null spaces of the following matrices.

(a) $\begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 4 \\ -1 & -2 \\ 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 3 & 1 & 1 \\ 0 & 0 & 0 \\ 6 & 2 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & -1 & \mathrm{i} \\ 2 & -2 & 2-\mathrm{i} \end{bmatrix}$

Exercise 3. Find bases for the column spaces of the matrices in Exercise 1.

Exercise 4. Find bases for the column spaces of the matrices in Exercise 2.

Exercise 5. Find bases for the row spaces of the matrices in Exercise 1.

Exercise 6. Find bases for the row spaces of the matrices in Exercise 2.

Exercise 7. For the following matrices find the null space of $I - A$ and find state vectors with nonnegative entries that sum to 1 in the null space, if any. Are these matrices stable (yes/no)?

(a) $A = \begin{bmatrix} 0.5 & 0 & 1 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 \end{bmatrix}$ (b) $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Exercise 8. Find the null space of $I - A$ and find state vectors (nonnegative entries that sum to 1) in the null space, if any, for the matrix $A = \begin{bmatrix} 1 & 0 & 1/3 \\ 0 & 1 & 1/3 \\ 0 & 0 & 1/3 \end{bmatrix}$. Is this matrix stable? Explain your answer.

Exercise 9. For each of the following linear operators, find the kernel and range of the operator. Is the operator one-to-one? onto?

(a) $T : \mathbb{R}^3 \to \mathbb{R}^3$ and $T((x_1, x_2, x_3)) = \begin{bmatrix} x_1 - 2x_2 + x_3 \\ x_1 + x_2 + x_3 \\ 2x_1 - x_2 + 2x_3 \end{bmatrix}$

(b) $T : \mathcal{P}_2 \to \mathbb{R}$ and $T(p(x)) = p(1)$

Exercise 10. For each of the following linear operators, find the kernel and range of the operator. Is the operator one-to-one? onto?

(a) $T : \mathcal{P}_2 \to \mathcal{P}_3$ and $T\left(a + bx + cx^2\right) = ax + bx^2/2 + cx^3/3$.

(b) $T : \mathbb{R}^3 \to \mathbb{R}^2$ and $T((x_1, x_2, x_3)) = \begin{bmatrix} 2x_2 \\ 3x_3 \end{bmatrix}$

Exercise 11. The linear operator $T : V \to \mathbb{R}^2$ is such that $T(\mathbf{v}_1) = (-1, 1)$, $T(\mathbf{v}_2) = (1, 1)$, and $T(\mathbf{v}_3) = (2, 0)$, where $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is a basis of V. Compute $\ker T$ and $\operatorname{range} T$. Is T one-to-one? onto? an isomorphism? *(Hint:* For the kernel calculation use Theorem 3.6 and find conditions on coefficients such that $T(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3) = \mathbf{0}$.)

Exercise 12. Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ be a basis of the vector space V and let the linear operator $T : V \to \mathbb{R}^3$ be such that $T(\mathbf{v}_1) = (0, 1, 1)$, $T(\mathbf{v}_2) = (1, 1, 0)$, and $T(\mathbf{v}_3) = (-1, 0, 1)$. Compute $\ker T$ and $\operatorname{range} T$. Is T one-to-one? onto? an isomorphism?

Problem 13. Let $T_A : \mathbb{R}^n \to \mathbb{R}^m$ be the matrix multiplication operator given by the $m \times n$ matrix A. Show that $\ker T_A = \mathcal{N}(A)$ and $\operatorname{range} T_A = \mathcal{C}(A)$.

Problem 14. Prove that if T is a linear operator, then for all $\mathbf{u}, \mathbf{v}$ in the domain of T and scalars c and d, we have $T(c\mathbf{u} - d\mathbf{v}) = cT(\mathbf{u}) - dT(\mathbf{v})$.

*Problem 15. Show that if $T : V \to W$ is a linear operator, then $T(\mathbf{0}) = \mathbf{0}$.

Problem 16. Show that if $T : V \to W$ is a linear operator, then the kernel of T is a subspace of V.

*Problem 17. Let the function $T : \mathbb{R}^3 \to \mathcal{P}_2$ be given by

$$T([c_1, c_2, c_3]^T) = c_1 x + c_2(x - 1) + c_3 x^2.$$

Show that T is an isomorphism of vector spaces.

Problem 18. Let $T : V \to W$ be a linear operator and U a subspace of V. Show that the image of U, $T(U) = \{T(\mathbf{v}) \,|\, \mathbf{v} \in U\}$, is a subspace of W.

*Problem 19. Prove that if A is a nilpotent matrix then $\mathcal{N}(A) \neq \{\mathbf{0}\}$ and $\mathcal{N}(I - A) = \{\mathbf{0}\}$.

Problem 20. Let V be a vector space over the reals with basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Show that the linear operator $T : \mathbb{R}^n \to V$ given by

$$T((c_1, c_2, \ldots, c_n)) = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

is an isomorphism of vector spaces.

Problem 21. Let A be an $m \times n$ matrix with $m \leq n$. Show that every subset of m columns of A is linearly independent if and only if every $m \times m$ submatrix B of A satisfies $\det B \neq 0$.

*Problem 22. Given $A \in \mathbb{R}^{m,n}$ and $\mathbf{b} \in \mathbb{R}^m$, show that $\mathbf{b}^T A$ is a linear combination of the rows of A.

3.5 Bases and Dimension

We have used the word "dimension" many times already, without really making the word precise. Intuitively, it makes sense when we say that $\mathbb{R}^2$ is "two-dimensional" or that $\mathbb{R}^3$ is "three-dimensional," for we reason that it takes two coordinate numbers to determine a vector in $\mathbb{R}^2$ and three for a vector in $\mathbb{R}^3$. What can we say about general vector spaces? Is there some number that is a measure of the size of the vector space? We answer these questions in this section. In the familiar cases of geometrical vector spaces, the answers will confirm our intuition.

The Basis Theorem

We know that the standard vector spaces always have a basis: The standard basis. What about subspaces of a standard space? Or, for that matter, abstract vector spaces? It turns out that the answer in all cases is yes, but we will be satisfied to answer the question for a special class of abstract vector spaces. The following concept turns out to be helpful.

Definition 3.14. **Finite-Dimensional Vector Space** The vector space V is called *finite-dimensional* if V has a finite spanning set.

Examples of finite-dimensional vector spaces are the standard spaces $\mathbb{R}^n$ and $\mathbb{C}^n$. As a matter of fact, we will see shortly that every subspace of a finite-dimensional vector space is finite-dimensional, and this includes most of the vector spaces we have studied so far. However, some very important vector spaces are *not* finite-dimensional, and accordingly, we call them *infinite-dimensional* spaces. Here is an example.

Example 3.34. Show that the space of all polynomial functions $\mathcal{P}$ is not a finite-dimensional space, while the subspaces $\mathcal{P}_n$ are finite-dimensional.

Solution. If $\mathcal{P}$ were a finite-dimensional space, then there would be a finite spanning set of polynomials $p_1(x), p_2(x), \ldots, p_m(x)$ for $\mathcal{P}$. This means that any other polynomial could be expressed as a linear combination of these polynomials. Let m be the maximum of all the degrees of the polynomials $p_j(x)$. Notice that any linear combination of polynomials of degree at most m must itself be a polynomial of degree at most m. (Remember that polynomial multiplication plays no part here, only addition and scalar multiplication.) Therefore, it is not possible to express the polynomial $q(x) = x^{m+1}$ as a linear combination of these polynomials, which means that they cannot be a basis. Hence, the space $\mathcal{P}$ has no finite spanning set.

On the other hand, it is obvious that the polynomial

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$

is a linear combination of the monomials $1, x, \ldots, x^n$ from which it follows that $\mathcal{P}_n$ is a finite-dimensional space. □

Here is the first basic result about these spaces. It is simply a formalization of what we have already done with preceding examples.

Theorem 3.11. **Basis Theorem** Every finite-dimensional vector space has a basis.

Proof. To see this, suppose that V is a finite-dimensional vector space with

$$V = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}.$$

Now if the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ has a redundant vector in it, discard it and obtain a smaller spanning set of V. Continue discarding vectors until you reach a spanning set for V that has no redundant vectors in it. (Since you start with a finite set, this can't go on indefinitely.) By the redundancy test, this spanning set must be linearly independent. Hence, it is a basis of V. □

The Dimension Theorem

No doubt you have already noticed that every basis of the vector space $\mathbb{R}^2$ must have exactly two elements in it. Similarly, one can reason geometrically that any basis of $\mathbb{R}^3$ must consist of exactly three elements. These numbers somehow measure the "size" of the space in terms of the degrees of freedom (number of coordinates) one needs to describe a general vector in the space. The dimension theorem asserts that this number can be unambiguously defined. As a matter of fact, the discussion on Page 214 shows that every basis of $\mathbb{R}^n$ has exactly n elements. Our next stop: arbitrary finite-dimensional vector spaces. Along the way, we need a very handy theorem that is sometimes called the *Steinitz substitution principle*. This principle is a mouthful to swallow, so we will precede its statement with an example that illustrates its basic idea.

Example 3.35. Let $\mathbf{w}_1 = (1,-1,0)$, $\mathbf{w}_2 = (0,-1,1)$, $\mathbf{v}_1 = (0,1,0)$, $\mathbf{v}_2 = (1,1,0)$, and $\mathbf{v}_3 = (0,1,1)$. Then $\mathbf{w}_1, \mathbf{w}_2$ form a linearly independent set and $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ form a basis of $V = \mathbb{R}^3$ (assume this). Show how to substitute both $\mathbf{w}_1$ and $\mathbf{w}_2$ into the set $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ while substituting out some of the $\mathbf{v}_j$'s and at the same time retaining the basis property of the set.

Solution. Since $\mathbb{R}^3 = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, we can express $\mathbf{w}_1$ as a linear combination of these vectors. We have a formal procedure for finding such combinations, but in this case we don't have to work too hard. A little trial and error shows that

$$\mathbf{w}_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = -2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = -2\mathbf{v}_1 + 1\mathbf{v}_2 + 0\mathbf{v}_3,$$

so that $1\mathbf{w}_1 + 2\mathbf{v}_1 - \mathbf{v}_2 - 0\mathbf{v}_3 = 0$. It follows that $\mathbf{v}_1$ or $\mathbf{v}_2$ is redundant in the set $\mathbf{w}_1, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. So discard, say, $\mathbf{v}_2$, and obtain a spanning set $\mathbf{w}_1, \mathbf{v}_1, \mathbf{v}_3$. In fact, it is actually a basis of V since two vectors can span only a plane. Now start over: Express $\mathbf{w}_2$ as a linear combination of this new basis. Again, a little trial and error shows that

$$\mathbf{w}_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} = -2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = 0\mathbf{w}_1 - 2\mathbf{v}_1 + 1\mathbf{v}_3.$$

Therefore, $\mathbf{v}_1$ or $\mathbf{v}_3$ is redundant in the set $\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_1, \mathbf{v}_3$. So discard, say, $\mathbf{v}_3$, and obtain a spanning set $\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}_1$. Again, this set is actually a basis of V since two vectors can span only a plane; and this is the kind of set we were looking for. □

Theorem 3.12. Steinitz Substitution Principle Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ be a linearly independent set in the space V and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be a basis of V. Then $r \leq n$ and we may substitute all of the $\mathbf{w}_i$'s for r of the $\mathbf{v}_j$'s in such a way that the resulting set of vectors is still a basis of V.

Proof. Let's do the substituting one step at a time. Start at $k = 0$. Now suppose that $k < r$ and that we have relabeled the remaining $\mathbf{v}_i$'s so that

$$V = \text{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s\}$$

with $k + s = n$ and $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$ is a basis of V.

We show how to substitute the next vector $\mathbf{w}_{k+1}$ into the basis and remove exactly one $\mathbf{v}_j$. We know that $\mathbf{w}_{k+1}$ is expressible *uniquely* as a linear combination of elements of the basis $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$ by Theorem 3.6. Also, there have to be some $\mathbf{v}_i$'s left in such a combination if $k < r$, for otherwise the set of vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ would not be linearly independent. Relabel the $\mathbf{v}_j$ again so that $b_s \neq 0$ in the unique expression

$$\mathbf{w}_{k+1} = a_1\mathbf{w}_1 + a_2\mathbf{w}_2 + \cdots + a_k\mathbf{w}_k + b_1\mathbf{v}_1 + b_2\mathbf{v}_2 \cdots + b_s\mathbf{v}_s$$

for $\mathbf{w}_{k+1}$. Thus, we can solve this equation to express $\mathbf{v}_s$ *uniquely* in terms of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k+1}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{s-1}$; otherwise the expression for $\mathbf{w}_{k+1}$ is not unique. It follows that $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k+1}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{s-1}$ are also linearly independent, else the expression for $\mathbf{v}_s$ is not unique. From these expressions we see that

$$\text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{v}_1, \ldots, \mathbf{v}_s\} = \text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_{k+1}, \mathbf{v}_1, \ldots, \mathbf{v}_{s-1}\}.$$

Hence, we have accomplished the substitution of $\mathbf{w}_{k+1}$ into the basis by removal of a single $\mathbf{v}_j$ and preserved the equality $n = k+s = (k+1)+(s-1)$. Continue this process until $k = r$ and we obtain the desired basis of V. □

Here is an important application of the Steinitz substitution principle.

Corollary 3.2. Every linearly independent set in a finite-dimensional vector space can be expanded to a basis of the space.

Proof. Suppose that $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ is a linearly independent set in V and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis of V. Apply the Steinitz substitution principle to this linearly independent set and basis to obtain a basis of V that includes $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$. □

Next, the dimension theorem is an easy consequence of Steinitz substitution, which has done the hard work for us.

Theorem 3.13. **Dimension Theorem** Let V be a finite-dimensional vector space. Then any two bases of V have the same number of elements, which is called the dimension of the vector space and denoted by $\dim V$.

Proof. Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be two bases of V. Apply the Steinitz substitution principle to the linearly independent set $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ and the basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ to obtain that $r \leq n$. Now reverse the roles of these two sets in the substitution principle to obtain the reverse inequality $n \leq r$. We conclude that $r = n$, as desired. □

Remember that a vector space always carries a field of scalars with it. If we are concerned about that field we could specify it explicitly as part of the dimension notation. For instance, we could write

$$\dim \mathbb{R}^n = \dim_{\mathbb{R}} \mathbb{R}^n \text{ or } \dim \mathbb{C}^n = \dim_{\mathbb{C}} \mathbb{C}^n.$$

Usually, the field of scalars is clear from context and we don't need the subscript notation.

As a first application of the dimension theorem, let's dispose of the standard spaces. We already know from Example 3.23 that these vector spaces have a basis consisting of n elements, namely the standard basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

According to the dimension theorem, this is all we need to specify the dimension of these spaces.

Corollary 3.3. The standard spaces satisfy $\dim \mathbb{R}^n = n$ and $\dim \mathbb{C}^n = n$.

There is one more question we want to answer. How do dimensions of a finite-dimensional vector space V and a subspace W of V relate to each other? At the outset, we don't even know whether W is finite-dimensional. Our intuition tells us that subspaces should have smaller dimension. Sure enough, our intuition is right this time!

Corollary 3.4. If W is a subspace of the finite-dimensional vector space V, then W is also finite-dimensional and $\dim W \leq \dim V$ with equality if and only if $V = W$.

Proof. Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ be a linearly independent set in W and suppose that $\dim V = n$. According to the Steinitz substitution principle, $r \leq n$. Now if the span of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ were smaller than W, then we could find a vector $\mathbf{w}_{r+1}$ in W but not in $\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r\}$. The new set $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r, \mathbf{w}_{r+1}$ would also be linearly independent (we leave this fact as an exercise) and $r+1 \leq n$. Since we cannot continue adding vectors indefinitely, we have to conclude that at some point we obtain a basis $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s$ for W. So W is finite-dimensional and furthermore, $s \leq n$, so we conclude that $\dim W \leq \dim V$. Finally, if we had equality, then a basis of W would be the same size as a basis of V. However, Steinitz substitution ensures that any linearly independent set can be expanded to a basis of V. It follows that this basis for W is also a basis for V, whence $W = V$. □

If U and V are subspaces of the vector space W, then the sum of these subspaces, $U+V = \{u+v \mid u \in U \text{ and } v \in V\}$, is also a subspace of W. These corollaries can be used to show how to calculate the dimension of $U+V$.

Corollary 3.5. If U and V are subspaces of the finite-dimensional vector space W, then $\dim(U+V) = \dim U + \dim V - \dim U \cap V$.

Proof. Corollary 3.4 shows that U, V, and $U \cap V$ are all finite-dimensional, say $\dim U = m$ and $\dim V = n$. Since $U \cap V$ is also a subspace of both U and V, $U \cap V$ has a basis, say $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$, with $r \leq m$ and $r \leq n$. Apply Corollary 3.2 to this basis to expand this basis to bases $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_s$ of U and $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t$ of V. Then $r+s = m$ and $r+t = n$. We leave it as a exercise to verify that $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_s, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t$ is a basis of $U+V$. Thus

$$\dim(U+V) = r+s+t = m+n-r = \dim U + \dim V - \dim U \cap V. \quad \square$$

A particularly nice special case of subspace sums is the case in which $U \cap V = \{\mathbf{0}\}$, which implies that $\dim(U+V) = \dim U + \dim V$:

Definition 3.15. Direct Sum and Summands If U and V are subspaces of the vector space W such that $U \cap V = \{\mathbf{0}\}$, then the subspace $U + V$ is called a *direct sum of subspaces* and denoted by $U \oplus V$. In this case U and V are called *summands* of $U \oplus V$.

Note 3.1. Direct sums as defined above could be called *internal* direct sums to distinguish them from what is called an *external* direct sum of two vector spaces: This direct sum is really the direct product $U \times V$ of two spaces with coordinate-wise arithmetic on elements of the product.

In the special case that $W = U \oplus V$ we say that V is a *complement* of U in W. Given a subspace U of W, a complement V of U in W is easy to manufacture thanks to Corollary 3.4:

Corollary 3.6. Complementary Subspace If U is a subspace of the finite-dimensional vector space W, then U has a complement V in W.

Proof. If $\dim W = n$, then $\dim U = r \leq n$ by Corollary 3.4. Furthermore, a basis $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ of U can be expanded to a basis $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ of W by the Steinitz substitution principle. Let $V = \text{span}\{\mathbf{w}_{r+1}, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$. Then $W = U + V$. If $\mathbf{0} \neq \mathbf{v} \in U \cap V$, then $\mathbf{v}$ could be expressed as a nontrivial linear combination of both $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ and $\mathbf{w}_{r+1}, \mathbf{w}_2, \ldots, \mathbf{w}_n$. These combinations would be equal, so bringing all terms to one side of the equation would yield a nontrivial combination of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ that sums to $\mathbf{0}$. This contradicts linear independence of this basis. Hence, $W = U \oplus V$. □

For a linear operator $T : V \to V$, where V is an n-dimensional vector space, two important subspaces associated with T are $\ker(T) = \{\mathbf{v} \in V \,|\, T(\mathbf{v}) = 0\}$ and $\text{range}(T) = \{T(\mathbf{v}) \,|\, \mathbf{v} \in V\}$. These subspaces have a special quality: They are T-*invariant*, i.e., T maps them into themselves. We will see in the next section that in the case of a matrix operator the sum of the dimensions of these subspaces is n. Does this mean that V is the direct sum of these subspaces? Exercise 13 of this section shows us that this is not the case. However, something can be salvaged:

Invariant Subspace

Theorem 3.14. If $T : V \to V$, where V is an n-dimensional vector space and T a linear operator, then $V = \ker(T^n) \oplus \text{range}(T^n)$ and both of these summands are T-invariant.

Proof. Define $U_j = \ker(T^j)$ and $W_j = \text{range}(T^j)$, $j = 0, 1, 2, \ldots$. We leave it as an exercise to show that for all j, $U_j \subseteq U_{j+1}$ and $W_j \supseteq W_{j+1}$. If at any point $W_j = W_{j+1}$, then since $W_{j+1} = T(W_j)$, all subsequent W_k are equal to W_j. Now every strict containment $W_j \supset W_{j+1}$ decreases dimension by at least 1 and $W_0 = V$ has dimension n; it follows that $W_m = W_n$ for all $m > n$. Thus, $T(W_n) = W_n$ and W_n is T-invariant. Likewise, if we have only strict inclusions $U_j \subset U_{j+1}$ for for $j < n$, then each containment increases

dimension by at least 1, so we would have $U_n = V = U_m$ for all $m > n$. On the other hand, if $U_j = U_{j+1}$ for some $j < n$, then $T^{j+1}(\mathbf{v}) = \mathbf{0}$ implies that $T^j(\mathbf{v}) = \mathbf{0}$. Therefore, , if $T^{j+2}(\mathbf{v}) = T^{j+1}(T(\mathbf{v})) = \mathbf{0}$, then $T^j(T(\mathbf{v})) = \mathbf{0}$, i.e., $T^{j+1}(\mathbf{v}) = \mathbf{0}$ and hence $U_{j+1} = U_{j+2}$. Continuing in this fashion we see that $U_{j+k} = U_j$ for all $k > 0$. So in all cases $U_n = U_{n+k}$ for all $k > 0$. From definition we see that $T(U_n) \subseteq U_{n-1}$, so U_n is also T-invariant.

Suppose that $\mathbf{v} \in \ker(T^n) \cap \text{range}(T^n) = U_n \cap W_n$. To show that $\mathbf{v}$ must be $\mathbf{0}$, first note that $T(W_n) = W_{n+1} = W_n$. We leave it as an exercise to show that any linear operator that maps a finite dimensional space onto itself must be an isomorphism. It follows that the operator T restricted to the space W_n is an isomorphism. In particular, if $\mathbf{0} \neq \mathbf{v} \in U_n \cap W_n$, then $T(\mathbf{v}) \neq \mathbf{0}$ and this argument can be applied repeatedly to show that $T^k(\mathbf{v}) \neq \mathbf{0}$ for any k. Yet $\mathbf{v} \in \ker(T^n)$, so we must have $T^n(\mathbf{v}) = \mathbf{0}$, a contradiction. Hence, $U_n \cap W_n = \{\mathbf{0}\}$.

Next, let $\mathbf{v}$ be any element of V. Certainly, $T^n(\mathbf{v}) \in W_n$. Note that since T restricted to W_n is an isomorphism, so is any power of T. Thus, there exists $\mathbf{w} \in W_n$ such that $T^n(\mathbf{v}) = T^{2n}(\mathbf{w})$, so that

$$T^n(\mathbf{v} - T^n(\mathbf{w})) = T^n(\mathbf{v}) - T^{2n}(\mathbf{w}) = T^n(\mathbf{v}) - T^n(\mathbf{v}) = \mathbf{0}.$$

Therefore, $\mathbf{v} - T^n(\mathbf{w}) = \mathbf{u} \in U_n$ and $T^n(\mathbf{w}) \in W_n$, so $\mathbf{v} = \mathbf{u} + T^n(\mathbf{w}) \in U_n + V_n$, which proves the theorem. □

As another application of the dimension theory developed in this section, consider the problem of determining all solutions to the homogeneous difference equation

$$a_m y_{k+m} + a_{m-1} y_{k+m-1} + \cdots + a_1 y_{k+1} + a_0 y_k = 0, \quad k = 0, 1, 2, \ldots, \tag{3.4}$$

where $a_0, a_1, \ldots, a_m$ are constant real coefficients with $a_0 \neq 0$, $a_m \neq 0$ and $y_0, y_1, \ldots, y_{m-1}$ are real initial values. Let S be the set of all solutions $\{y_k\}_{k=0}^{\infty}$ of (3.4). We leave it as an exercise to show that S is a vector space over $\mathbb{R}$ with the obvious operations of addition and scalar multiplication: $\{y_k\} + \{z_k\} = \{y_k + z_k\}$ and $c\{y_k\} = \{cy_k\}$.

Example 3.36. Determine the dimension of S, the space of all solutions to equation (3.4).

Solution. Determining the dimension of this space is accomplished by considering the linear map $T : S \to \mathbb{R}^m$ defined by $T(\{y_k\}) = (y_0, y_1, \ldots, y_{m-1})$. This mapping is onto and one-to-one. It is therefore an isomorphism of vector spaces, so these spaces have the same dimension, namely m. □

Example 3.37. Find a formula for all solutions to the linear difference equation $2y_{k+2} - 3y_{k+1} - 2y_k = 0$.

Solution. We see from the preceding example that the dimension of this solution space is 2. Moreover, by finding the roots the characteristic polynomial equation $2x^2 - 3x - 2 = 0$ of this difference equation we saw in Example 2.28 (page 99) that two particular solutions to this difference equation are

$y_k = 2^k$ and $y_k = \left(-\frac{1}{2}\right)^k$, $k = 0, 1, 2, \ldots$. These solutions are linearly independent since one is not a multiple of the other. Therefore, they form a basis for the solution space and hence a general formula for all solutions $\{y_k\}_{k=0}^{\infty}$ to the difference equation is given by

$$y_k = c_1 2^k + c_2 \left(-\frac{1}{2}\right)^k, \; k = 0, 1, 2, \ldots,$$

where c_1 and c_2 are arbitrary constants. □

3.5 Exercises and Problems

Exercise 1. Find all possible subsets of the following sets of vectors that form a basis of $\mathbb{R}^3$.

(a) $(1,0,1), (1,-1,1)$ (b) $(1,2,1), (2,1,1), (3,4,1), (2,0,1)$

(c) $(2,-3,1), (4,-2,-3), (0,-4,5), (1,0,0), (0,0,0)$

Exercise 2. Find all possible subsets of the following sets of vectors that form a basis of $\mathbb{R}^{2,2}$.

(a) $\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$

Exercise 3. Let $V = \mathbb{R}^3$ and $\mathbf{w}_1 = (2,1,0)$, $\mathbf{v}_1 = (1,3,1)$, $\mathbf{v}_2 = (4,2,0)$. The set $\mathbf{v}_1, \mathbf{v}_2$ is linearly independent in V. Determine which $\mathbf{v}_j$'s could be replaced by $\mathbf{w}_1$ while retaining the linear independence of the resulting set.

Exercise 4. Let $V = \mathbb{R}^3$ and $\mathbf{w}_1 = (0,1,0)$, $\mathbf{w}_2 = (1,1,1)$, $\mathbf{v}_1 = (1,3,1)$, $\mathbf{v}_2 = (2,-1,1)$, $\mathbf{v}_3 = (1,0,1)$. The set $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is a basis of V. Determine which $\mathbf{v}_j$'s could be replaced by $\mathbf{w}_1$, and which $\mathbf{v}_j$'s could be replaced by both $\mathbf{w}_1$ and $\mathbf{w}_2$, while retaining the basis property.

Exercise 5. Let $V = C[0,1]$ and $\mathbf{w}_1 = \sin^2 x$, $\mathbf{w}_2 = \cos x$, $\mathbf{v}_1 = \sin x$, $\mathbf{v}_2 = \cos^2 x$, $\mathbf{v}_3 = 1$. The set $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is linearly independent in V. Determine which $\mathbf{v}_j$'s could be replaced by $\mathbf{w}_1$, and which $\mathbf{v}_j$'s could be replaced by both $\mathbf{w}_1$ and $\mathbf{w}_2$, while retaining the linear independence of the resulting set.

Exercise 6. Let $V = \mathcal{P}_2$ and $\mathbf{w}_1 = x$, $\mathbf{w}_2 = x^2$, $\mathbf{v}_1 = 1 - x$, $\mathbf{v}_2 = 2 + x$, $\mathbf{v}_3 = 1 + x^2$. The set $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is a basis of V. Determine which $\mathbf{v}_j$'s could be replaced by $\mathbf{w}_1$, and which $\mathbf{v}_j$'s could be replaced by both $\mathbf{w}_1$ and $\mathbf{w}_2$, while retaining the basis property.

Exercise 7. Let $\mathbf{w}_1 = (0,1,1)$. Expand $\{\mathbf{w}_1\}$ to a basis of $\mathbb{R}^3$.

Exercise 8. Let $\mathbf{w}_1 = x + 1$. Expand $\{\mathbf{w}_1\}$ to a basis of $\mathcal{P}_2$.

Exercise 9. Find two complements of the subspace spanned by $\mathbf{w} = (1, 1, 1)$ in $\mathbb{R}^3$.

Exercise 10. Find two complements of the subspace spanned by $\mathbf{w} = x^2 + 1$ in $\mathcal{P}_2$.

Exercise 11. Assume that $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} \subseteq V$, where V is a vector space of dimension n. Answer True/False to the following:
(a) If S is a basis of V then $k = n$.
(b) If S spans V then $k \leq n$.
(c) If S is linearly independent then $k \leq n$.
(d) If S is linearly independent and $k = n$ then S spans V.
(e) If S spans V and $k = n$ then S is a basis for V.
(f) If A is a 5 by 5 matrix and $\det A = 2$, then the first 4 columns of A span a 4 dimensional subspace of $\mathbb{R}^5$.

Exercise 12. Assume that V is a vector space of dimension n and $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} \subseteq V$. Answer True/False to the following:
(a) S is either a basis or contains redundant vectors.
(b) A linearly independent set contains no redundant vectors.
(c) If $V = \operatorname{span}\{\mathbf{v}_2, \mathbf{v}_3\}$ and $\dim V = 2$, then $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a linearly dependent set.
(d) A set of vectors containing the zero vector is a linearly independent set.
(e) Every vector space is finite-dimensional.
(f) The set of vectors $[\mathrm{i}, 0]^{\mathrm{T}}, [0, \mathrm{i}]^{\mathrm{T}}, [1, \mathrm{i}]^{\mathrm{T}}$ in $\mathbb{C}^2$ contains redundant vectors.

Exercise 13. Let $A = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. Determine if $\ker T_A + \operatorname{range} T_A$ is direct and confirm Theorem 3.14 for the operator $T = T_A$ and space $V = \mathbb{R}^3$.

Exercise 14. Repeat Exercise 13 with $A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}$.

Problem 15. Show that the set S of Example 3.36 is a vector space.

*Problem 16. Show that the mapping T of Example 3.36 is a linear operator.

Problem 17. Let $V = \{\mathbf{0}\}$, a vector space with a single element. Explain why the element $\mathbf{0}$ is *not* a basis of V and the dimension of V must be 0.

*Problem 18. Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ be linearly independent vectors in the vector space W. Show that if $\mathbf{w} \in W$ and $\mathbf{w} \notin \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r\}$, then $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r, \mathbf{w}$ is a linearly independent set.

*Problem 19. Let $e_{i,j}$ be the $m \times n$ matrix with a unit in the (i,j)th entry and zeros elsewhere. Show that $\{e_{i,j} \mid i = 1, \ldots, m,\ j = 1, \ldots, n\}$ is a basis of the vector space $\mathbb{R}^{m,n}$.

*Problem 20. Complete the proof of Corollary 3.5.

Problem 21. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be bases of U and V, respectively, where U and V are subspaces of the vector space W. Show by example that if $U \cap V \neq \{\mathbf{0}\}$, then their union need not be a basis if $U + V$.

*Problem 22. Determine the dimension of the subspace of $\mathbb{R}^{3,3}$ consisting of all symmetric matrices by exhibiting a basis.

Problem 23. Let U be the subspace of $W = \mathbb{R}^{3,3}$ consisting of all symmetric matrices and V the subspace of all skew-symmetric matrices.
(a) Show that $U + V = U \oplus V$.
(b) Use Problems 19, 22 and Corollary 3.5 to calculate $\dim V$.

Problem 24. Show that the functions $1, x, x^2, \ldots, x^n$ form a basis for the space $\mathcal{P}_n$ of polynomials of degree at most n.

Problem 25. Show that $C[0,1]$ is an infinite-dimensional space.

Problem 26. Let $T : V \to W$ be a linear operator such that $\operatorname{range} T = W$ and $\ker T = \{0\}$. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be a basis of V. Show that the image of these vectors, $T(\mathbf{v}_1), T(\mathbf{v}_2), \ldots, T(\mathbf{v}_n)$, is a basis of W.

*Problem 27. Let $p(x) = c_0 + c_1 x + \cdots + c_m x^m$ be a polynomial and A an $n \times n$ matrix. Use the result of Problem 19 to show that there exists a polynomial $p(x)$ of degree at most n^2 for which $p(A) = 0$. (Aside: This estimate is actually much too pessimistic. The Cayley–Hamilton theorem in Chapter 5 shows that n works in place of n^2.)

Problem 28. Show that a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in the vector space V is a basis if and only if it has no redundant vectors and $\dim V \leq n$.

Problem 29. Let $T : V \to W$ be a linear operator where V is a finite-dimensional space and U is a subspace of V. Show that $\dim T(U) \leq \dim U$.

Problem 30. Show that if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a spanning set for the vector space V and the subset $\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \ldots, \mathbf{v}_{i_k}$ is linearly independent, then this set can be expanded to a basis of V using only elements of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$.

Problem 31. Let $T : V \to V$ be a linear operator and define $U_j = \ker(T^j)$ and $W_j = \operatorname{range}(T^j)$, $j = 1, 2, \ldots$. Show that for all j, $U_j \subseteq U_{j+1}$ and $W_j \supseteq W_{j+1}$.

Problem 32. Show that if $T : V \to V$ is a linear operator such that $T(V) = V$ and V is finite dimensional, then T is an isomorphism.

*Problem 33. Verify Theorem 3.14 for the linear operator $T : \mathcal{P}_2 \to \mathcal{P}_2$ given by

$$T\left(c_0 + c_1 x + c_2 x^2\right) = 2c_0 - c_2 + 3c_3 + (2c_0 - 3c_1 + 4c_2)\, x^2.$$

3.6 Linear Systems Revisited

We now have some very powerful tools for understanding the nature of solution sets of the standard linear system $A\mathbf{x} = \mathbf{b}$. This understanding will help us design practical computational methods for finding dimension and bases for vector spaces and other problems as well.

The first business at hand is to describe solution sets of inhomogeneous systems. Recall that every homogeneous system is consistent since it has the trivial solution. Inhomogeneous systems are another matter. We already have one criterion, namely that the rank of augmented matrix and coefficient matrix of the system must agree. Here is one more way to view the consistency of such a system in the language of vector spaces.

Theorem 3.15. Consistency in Terms of Column Space The linear system $A\mathbf{x} = \mathbf{b}$ of m equations in n unknowns is consistent if and only if $\mathbf{b} \in \mathcal{C}(A)$.

Proof. The key to this fact is Theorem 2.1, which says that the vector $A\mathbf{x}$ is a linear combination of the columns of A with the entries of $\mathbf{x}$ as scalar coefficients. Therefore, to say that $A\mathbf{x} = \mathbf{b}$ has a solution is simply to say that some linear combination of columns of A adds up to $\mathbf{b}$, i.e., $\mathbf{b} \in \mathcal{C}(A)$. □

Inclusion in a Span

The next example shows how to to determine whether a vector belongs to a subspace specified by a spanning set of standard vectors.

Example 3.38. One of the following vectors belongs to the space V spanned by $\mathbf{v}_1 = (1, 1, 3, 3)$, $\mathbf{v}_2 = (0, 2, 2, 4)$, and $\mathbf{v}_3 = (1, 0, 2, 1)$. The vectors in question are $\mathbf{u} = (2, 1, 5, 4)$ and $\mathbf{w} = (1, 0, 0, 0)$. Which and why?

Solution. Theorem 3.15 tells us that if $A = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$, then we need only determine whether the systems $A\mathbf{x} = \mathbf{u}$ and $A\mathbf{x} = \mathbf{w}$ are consistent. In the interests of efficiency, we may as well do both at once by forming the augmented matrix for both right-hand sides at once as

$$[A \,|\, \mathbf{u} \,|\, \mathbf{w}] = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 \\ 1 & 2 & 0 & 1 & 0 \\ 3 & 2 & 2 & 5 & 0 \\ 3 & 4 & 1 & 4 & 0 \end{bmatrix} \text{ with reduced row echelon form } \begin{bmatrix} 1 & 0 & 1 & 2 & 0 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Observe that there is a pivot in the fifth column but not in the fourth column. This tells us that the system with augmented matrix $[A \,|\, \mathbf{u}]$ is consistent, but the system with augmented matrix $[A \,|\, \mathbf{w}]$ is not consistent. Therefore, $\mathbf{u} \in \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, but $\mathbf{w} \notin \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. As a matter of fact, the reduced row echelon form of $[A \,|\, \mathbf{u}]$ tells us what linear combinations will work, namely

$$\mathbf{u} = (2 - c_3)\mathbf{v}_1 + \frac{1}{2}(c_3 - 1)\mathbf{v}_2 + c_3\mathbf{v}_3,$$

where c_3 is an arbitrary scalar. The reason for nonuniqueness of the coordinates of $\mathbf{u}$ is that the vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are not linearly independent. □

The next item of business is a description of the solution space itself, given that it is not empty. We already have a pretty good conceptual model for the solution of a homogeneous system $A\mathbf{x} = 0$. Remember that this is just the null space, $\mathcal{N}(A)$, of the matrix A. In fact, the definition of $\mathcal{N}(A)$ is the set of vectors $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$. The important point here is that we proved that $\mathcal{N}(A)$ really is a subspace of the appropriate n-dimensional standard space $\mathbb{R}^n$ or $\mathbb{C}^n$. As such we can really picture it when n is 2 or 3: $\mathcal{N}(A)$ is either the origin, a line through the origin, a plane through the origin, or in the case $A = 0$, all of $\mathbb{R}^3$. What can we say about an inhomogeneous system? Here is a handy way of understanding these solution sets.

Theorem 3.16. Form of General Solution Suppose the system $A\mathbf{x} = \mathbf{b}$ is consistent with a particular solution $\mathbf{x}_*$. Then the general solution $\mathbf{x}$ to this system can be described by the equation

$$\mathbf{x} = \mathbf{x}_* + \mathbf{z},$$

where $\mathbf{z}$ runs over all elements of $\mathcal{N}(A)$.

Proof. On the one hand, suppose we are given a vector of the form $\mathbf{x} = \mathbf{x}_* + \mathbf{z}$, where $A\mathbf{x}_* = \mathbf{b}$ and $\mathbf{z} \in \mathcal{N}(A)$. Then

$$A\mathbf{x} = A(\mathbf{x}_* + \mathbf{z}) = A\mathbf{x}_* + A\mathbf{z} = \mathbf{b} + \mathbf{0} = \mathbf{b}.$$

Thus, $\mathbf{x}$ is a solution to the system. Conversely, suppose we are given any solution $\mathbf{x}$ to the system and that $\mathbf{x}_*$ is a particular solution to the system. Then

$$A(\mathbf{x} - \mathbf{x}_*) = A\mathbf{x} - A\mathbf{x}_* = \mathbf{b} - \mathbf{b} = \mathbf{0}.$$

Thus, $\mathbf{x} - \mathbf{x}_* = \mathbf{z} \in \mathcal{N}(A)$, so that $\mathbf{x}$ has the required form $\mathbf{x}_* + \mathbf{z}$. □

This is really a pretty fact, so let's be clear about what it is telling us. It says that the solution space to a consistent system, as a set, can be described as the set of all translates of elements in the null space of A by some fixed vector. Such a set is sometimes called an *affine set* or a *flat*. When n is 2 or 3 this says that the solution set is either a single point, a line or a plane—*not* necessarily through the origin!

Example 3.39. Describe geometrically the solution sets to the system

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

Solution. First solve the system, which has augmented matrix

$$\begin{bmatrix} 1 & 2 & 0 & 3 \\ 1 & 1 & 1 & 3 \end{bmatrix} \xrightarrow{E_{21}(-1)} \begin{bmatrix} 1 & 2 & 0 & 3 \\ 0 & -1 & 1 & 0 \end{bmatrix} \xrightarrow[E_2(-1)]{E_{12}(2)} \begin{bmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & -1 & 0 \end{bmatrix}.$$

The general solution to the system is given in terms of the free variable z, which we will relabel as $z = t$ to obtain

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

We may recognize this from calculus as a parametric representation of a line in three-dimensional space $\mathbb{R}^3$. This this line does not pass through the origin since $z = 0$ forces $x = 3$. So the solution set is not a subspace of $\mathbb{R}^3$. □

Now we turn to another computational matter. How do we find bases of vector spaces that are prescribed by a spanning set? How do we find the linear dependencies in a spanning set or implement the Steinitz substitution principle in a practical way? We have all the tools we need now to solve these problems. Let's begin with the question of finding a basis. We are going to solve this problem in two ways. Each has its own merits. First we examine the row space approach. We require two simple facts.

Theorem 3.17. Let A be any matrix and E an elementary matrix. Then

$$\mathcal{R}(A) = \mathcal{R}(EA).$$

Proof. Suppose the rows of A are the vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$, so that we have $\mathcal{R}(A) = \text{span}\left\{\mathbf{r}_1^T, \mathbf{r}_2^T, \ldots, \mathbf{r}_n^T\right\}$. If $E = E_{ij}$, then the effect of multiplication by E is to switch the ith and jth rows, so the rows of EA are simply the rows of A in a different order. Hence, $\mathcal{R}(A) = \mathcal{R}(EA)$ in this case. If $E = E_i(a)$, with a a nonzero scalar, then the effect of multiplication by E is to replace the ith row by a nonzero multiple of itself. Clearly, this doesn't change the span of the rows either. To simplify notation, consider the case $E = E_{12}(a)$. Then the first row $\mathbf{r}_1$ is replaced by $\mathbf{r}_1 + a\mathbf{r}_2$, so that any combination of the rows of EA is expressible as a linear combination of the rows of A. Conversely, since $\mathbf{r}_1 = \mathbf{r}_1 + a\mathbf{r}_2 - a\mathbf{r}_2$, we see that any combination of $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$ can be expressed in terms of the rows of EA. This proves the theorem. □

Theorem 3.18. If the matrix R is in a reduced row form, then the transposes of the nonzero rows of R form a basis of $\mathcal{R}(R)$.

Proof. Suppose the rows of R are $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$, so that we have $\mathcal{R}(R) = \text{span}\left\{\mathbf{r}_1^T, \mathbf{r}_2^T, \ldots, \mathbf{r}_k^T\right\}$, where the first k rows of R are nonzero and the remaining rows are zero rows. Then the nonzero rows span $\mathcal{R}(R)$. In order for these vectors to form a basis, they must also be a linearly independent set. If some linear combination of these vectors has value zero, say

$$\mathbf{0} = c_1\mathbf{r}_1 + \cdots + c_k\mathbf{r}_k,$$

we examine the first coordinate of this linear combination, corresponding to the column in which the first pivot appears. In that column $\mathbf{r}_1$ has a nonzero

coordinate value and all other $\mathbf{r}_j$ have a value of zero. Therefore, the linear combination above yields that $c_1 = 0$. Repeat this argument for each index and we obtain that all $c_i = 0$. Hence, the nonzero rows must be linearly independent. Therefore, transposes of these vectors form a basis of $\mathcal{R}(R)$. □

These theorems are the foundations for the following algorithm for finding a basis for a vector space.

Row Space Algorithm
Given $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\} \subseteq \mathbb{R}^n$ or $\mathbb{C}^n$.

(1) Form the $m \times n$ matrix A whose rows are $\mathbf{v}_1^T, \mathbf{v}_2^T, \ldots, \mathbf{v}_m^T$.
(2) Find a reduced row form R of A.
(3) List the nonzero rows of R. Their transposes form a basis of V.

Example 3.40. Let the vector space V be spanned by vectors $\mathbf{v}_1 = (1, 1, 3, 3)$, $\mathbf{v}_2 = (0, 2, 2, 4)$, $\mathbf{v}_3 = (1, 0, 2, 1)$, and $\mathbf{v}_4 = (2, 1, 5, 4)$. Find a basis of V by the row space algorithm.

Solution. Form the matrix whose rows are the $\mathbf{v}_j$'s and find its reduced row echelon form:

$$A = \begin{bmatrix} 1 & 1 & 3 & 3 \\ 0 & 2 & 2 & 4 \\ 1 & 0 & 2 & 1 \\ 2 & 1 & 5 & 4 \end{bmatrix} \xrightarrow[\substack{E_{31}(-1) \\ E_{41}(-2) \\ E_2(1/2)}]{} \begin{bmatrix} 1 & 1 & 3 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & -1 & -1 & -2 \\ 0 & -1 & -1 & -2 \end{bmatrix} \xrightarrow[\substack{E_{32}(1) \\ E_{42}(1) \\ E_{12}(-1)}]{} \begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

From this we see that the vectors $(1, 0, 2, 1)$ and $(0, 1, 1, 2)$ form a basis for the row space of A. □

This algorithm for computing a basis does more than find a basis: It formalizes an idea we encountered in Section 3.4 that determines when linear combinations have value zero.

Theorem 3.19. Let A be a matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$. Suppose the indices of the nonpivot columns in the reduced row echelon form of A are $i_1, i_2, \ldots, i_k$. Then every linear combination of value zero,

$$\mathbf{0} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n,$$

of the columns of A is uniquely determined by the values of $c_{i_1}, c_{i_2}, \ldots, c_{i_k}$. In particular, if these coefficients are 0, then all the other coefficients must be 0.

Proof. Express the linear combination in the form

$$\mathbf{0} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]\mathbf{c} = A\mathbf{c},$$

where $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ and $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$. In other words, the column $\mathbf{c}$ of coefficients is in the null space of A. Every solution $\mathbf{c}$ to this system is

uniquely specified as follows: Assign arbitrary values to the free variables, then use the rows of the reduced row echelon form of A to solve for each bound variable. This is exactly what we wanted to show. □

Corollary 3.7. If A is a matrix with columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ and $j_1, j_2, \ldots, j_r$ the indices of the pivot columns of the reduced row echelon form of A, then the columns $\mathbf{a}_{j_1}, \mathbf{a}_{j_2}, \ldots, \mathbf{a}_{j_r}$ form a basis of $\mathcal{C}(A)$.

Proof. Theorem 3.19 implies that the columns of A corresponding to pivot columns in the reduced row echelon form of A must be themselves a linearly independent set. Moreover the proof shows that we can express any column corresponding to a nonpivot column in terms of columns corresponding to pivot columns by setting the free variable corresponding to the nonpivot column to 1, and all other free variables to 0. Therefore, the columns of A corresponding to pivot columns form a basis of $\mathcal{C}(A)$. □

This corollary justifies the following algorithm for finding a basis for a vector space.

Column Space Algorithm
Given $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq \mathbb{R}^m$ or $\mathbb{C}^m$:

(1) Form the $m \times n$ matrix A whose columns are $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$.
(2) Find a reduced row form R of A.
(3) List the columns of A that correspond to pivot columns of R. These form a basis of V.

Caution: It is not the columns (or the rows) of the reduced row echelon form matrix R that yield the basis vectors for V. In fact, if E is an elementary matrix, in general we have $\mathcal{C}(A) \neq \mathcal{C}(EA)$.

Example 3.41. Let the vector space V be spanned by vectors $\mathbf{v}_1 = (1, 1, 3, 3)$, $\mathbf{v}_2 = (0, 2, 2, 4)$, $\mathbf{v}_3 = (1, 0, 2, 1)$, and $\mathbf{v}_4 = (2, 1, 5, 4)$. Find a basis of V by the column space algorithm.

Solution. Form the matrix A whose columns are these vectors and reduce the matrix to its reduced row echelon form:

$$\begin{bmatrix} 1 & 0 & 1 & 2 \\ 1 & 2 & 0 & 1 \\ 3 & 2 & 2 & 5 \\ 3 & 4 & 1 & 4 \end{bmatrix} \xrightarrow[E_{31}(-3)]{E_{21}(-1)} \begin{bmatrix} 1 & 0 & 1 & 2 \\ 0 & 2 & -1 & -1 \\ 0 & 2 & -1 & -1 \\ 0 & 4 & -2 & -2 \end{bmatrix} \xrightarrow[E_{42}(-2)]{E_{32}(-1)} \begin{bmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We can see from this calculation that the first and second columns will be pivot columns, while the third and fourth will not be. According to the column space algorithm, $\mathcal{C}(A)$ is a two-dimensional space with the first two columns $\mathbf{v}_1 = (1, 1, 3, 3)$ and $\mathbf{v}_2 = (0, 2, 2, 4)$ of A for a basis. □

Note 3.2. While any reduced row form suffices, what is gained by the reduced row echelon form is the ability to use Theorem 3.19 to determine linear combinations of value zero easily.

Consider Example 3.41. From the first two rows of the reduced row echelon form of A we see that if $\mathbf{c} = (c_1, c_2, c_3, c_4)$ and $A\mathbf{c} = 0$, then

$$c_1 = -(c_3 + 2c_4),$$
$$c_2 = \frac{1}{2}(c_3 + c_4),$$

and c_3, c_4 are free. Hence, the general linear combination with value zero is

$$-(c_3 + 2c_4)\mathbf{v}_1 + \frac{1}{2}(c_3 + c_4)\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4 = \mathbf{0}.$$

For example, take $c_3 = 0$ and $c_4 = 1$ to obtain

$$-2\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2 + 0\mathbf{v}_3 + 1\mathbf{v}_4 = \mathbf{0},$$

so that $\mathbf{v}_4 = 2\mathbf{v}_1 - \frac{1}{2}\mathbf{v}_2$. A similar calculation with $c_3 = 1$ and $c_4 = 0$ shows that $\mathbf{v}_3 = \mathbf{v}_1 - \frac{1}{2}\mathbf{v}_2$.

Finally, we consider the problem of finding a basis for a null space. Actually, we have already dealt with this problem in an earlier example (Example 3.30), but now we will justify what we did there.

Theorem 3.20. Let A be an $m \times n$ matrix such that $\operatorname{rank} A = r$. Suppose the general solution to the homogeneous equation $A\mathbf{x} = 0$ with $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is written in the form

$$\mathbf{x} = x_{i_1}\mathbf{w}_1 + x_{i_2}\mathbf{w}_2 + \cdots + x_{i_{n-r}}\mathbf{w}_{n-r},$$

where $x_{i_1}, x_{i_2}, \ldots, x_{i_{n-r}}$ are the free variables. Then $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n-r}$ form a basis of $\mathcal{N}(A)$.

Proof. The vector $\mathbf{x} = 0$ occurs precisely when all the free variables are set equal to 0, for the bound variables are linear combinations of the free variables. This means that the only linear combinations with value zero of the vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n-r}$ are those for which all the coefficients $x_{i_1}, x_{i_2}, \ldots, x_{i_{n-r}}$ are 0. Hence, these vectors are linearly independent. They span $\mathcal{N}(A)$ since every element $\mathbf{x} \in \mathcal{N}(A)$ is a linear combination of them. Therefore, $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n-r}$ form a basis of $\mathcal{N}(A)$. □

The formula in Theorem 3.20 shows that each of the vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n-r}$ is recovered from the general solution by setting one free variable to one and the others to zero. It also shows that the following algorithm is valid.

Null Space Algorithm
Given an $m \times n$ matrix A.

(1) Compute the reduced row echelon form R of A.
(2) Use R to find the general solution to the homogeneous system $A\mathbf{x} = 0$.
(3) Write the general solution $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to the homogeneous system in the form

$$\mathbf{x} = x_{i_1}\mathbf{w}_1 + x_{i_2}\mathbf{w}_2 + \cdots + x_{i_{n-r}}\mathbf{w}_{n-r},$$

where $x_{i_1}, x_{i_2}, \ldots, x_{i_{n-r}}$ are the free variables.
(4) List the vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n-r}$. These form a basis of $\mathcal{N}(A)$.

Example 3.42. Find a basis for the null space of the matrix A in Example 3.41 by the null space algorithm.

Solution. From the Example the reduced row echelon form of A is

$$R = \begin{bmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The variables x_3 and x_4 are free, while x_1 and x_2 are bound. Hence, the general solution of $A\mathbf{x} = 0$ can be written as

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

which becomes, in vector notation,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = x_3 \begin{bmatrix} -1 \\ 1/2 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -2 \\ 1/2 \\ 0 \\ 1 \end{bmatrix}.$$

Hence, $\mathbf{w}_1 = (-1, 1/2, 1, 0)$ and $\mathbf{w}_1 = (-2, 1/2, 0, 1)$ form a basis of $\mathcal{N}(A)$. □

A summary of key dimensional facts that we have learned in this section:

Theorem 3.21. **Rank Theorem** Let A be an $m \times n$ matrix such that $\operatorname{rank} A = r$. Then

(1) $\dim \mathcal{C}(A) = r$
(2) $\dim \mathcal{R}(A) = r$
(3) $\dim \mathcal{N}(A) = n - r$

The following example offers insight into the nature of rank-one matrices.

Example 3.43. Show that every rank-one matrix can be expressed as an outer product of vectors.

Solution. Let A be an $m \times n$ rank-one matrix. Let the rows of A be $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m$. Since $\dim \mathcal{R}(A) = 1$, the row space of A is spanned by a single row of A, say the kth one. Hence, there are constants $c_1, c_2, \ldots, c_m$ such that $\mathbf{r}_j = c_j \mathbf{r}_k$, $k = 1, \ldots, m$. Let $\mathbf{c} = [c_1, c_2, \ldots, c_m]^T$ and $\mathbf{d} = \mathbf{r}_k^T$, and it follows that A and the outer product of $\mathbf{c}$ and $\mathbf{d}$, $\mathbf{c}\mathbf{d}^T$, have the same rows, hence are equal. □

Rank-One Matrix as Outer Product

3.6 Exercises and Problems

Exercise 1. Use the fact that B is a reduced row form of A to find bases for the row and column spaces of A with no calculations, and null space with minimum calculations, where $A = \begin{bmatrix} 3 & 5 & -1 & 5 & 1 \\ 1 & 2 & -1 & 2 & 0 \\ 2 & 3 & 0 & 3 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 & 3 & 0 & 2 \\ 0 & 1 & -2 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$.

Exercise 2. Let $A = \begin{bmatrix} 3 & 1 & -2 & 0 & 1 & 2 & 1 \\ 1 & 1 & 0 & -1 & 1 & 2 & 2 \\ 3 & 2 & -1 & 1 & 1 & 8 & 9 \\ 0 & 2 & 2 & -1 & 1 & 6 & 8 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 0 & -2 & 0 & 0 & -4 & -6 \\ 0 & 2 & 2 & 0 & 0 & 4 & 6 \\ 0 & 0 & 0 & -2 & 2 & 4 & 4 \\ 0 & 0 & 0 & 0 & 1 & 6 & 7 \end{bmatrix}$, and repeat Exercise 1.

Exercise 3. Find two bases for the space spanned by each of the following sets of vectors by using the row space algorithm and column space algorithm with the reduced row echelon form.
(a) $(0, -1, 1)$, $(2, 1, 1)$ in $\mathbb{R}^3$.
(b) $(2, -1, 1)$, $(2, 0, 1)$, $(-4, 2, -2)$ in $\mathbb{R}^3$.
(c) $(1, -1)$, $(2, 2)$, $(-1, 2)$, $(2, 0)$ in $\mathbb{R}^2$.
(d) $1+x^2$, $-2-x+3x^2$, $5+x$, $4+4x^2$ in $\mathcal{P}_2$. (*Hint:* See the discussion following Theorem 3.9 of Section 3.4 for a way of thinking of polynomials as vectors.)

Exercise 4. Find two bases for each of the following sets of vectors by using the row space algorithm and the column space algorithm.
(a) $(1, -1)$, $(1, 1)$, $(2, 0)$ in $\mathbb{R}^2$.
(b) $(2, 2, -4)$, $(-4, -4, 8)$ in $\mathbb{R}^3$.
(c) $(1, 0, 0)$, $(1 + \mathrm{i}, 2, 2 - \mathrm{i})$, $(-1, 0, \mathrm{i})$ in $\mathbb{C}^3$.
(d) $1 + 2x + 2x^3$, $-2 - 5x + 5x^2 + 6x^3$, $-x + 5x^2 + 6x^3$, $x - 5x^2 + 4x^3$ in $\mathcal{P}_3$.

Exercise 5. Find bases for the row, column, and null space of each of the following matrices.

(a) $[2, 0, -1]$ (b) $\begin{bmatrix} 1 & 2 & 0 & 0 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 3 & 6 & 2 & 2 & 3 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 2 & 0 & 4 & 0 \\ 1 & 3 & 5 & 2 & 1 \\ 2 & 3 & -5 & 10 & 0 \\ 2 & 4 & 0 & 8 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & -3 & -1 \\ 0 & 2 & 0 \\ 2 & 4 & 1 \end{bmatrix}$

Exercise 6. Find bases for the row, column, and null space of the following.

(a) $\begin{bmatrix} 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 \\ 2 & 4 & 1 & 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 4 & 0 & -4 & 0 & -2 \\ 2 & 4 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 2 & 1 & 5 \\ 1 & 2 & 0 & -2 & 0 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & \mathrm{i} & 0 \\ 1 & 2 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 & 0 & 0 \\ 3 & 6 & 2 & 2 \end{bmatrix}$

Exercise 7. Find all possible linear combinations with value zero of the following sets of vectors and the dimension of the space spanned by them.
(a) $(0,1,1)$, $(2,0,1)$, $(2,2,3)$, $(0,2,2)$ in $\mathbb{R}^3$.
(b) x, x^2+x, x^2-x in $\mathcal{P}_2$.
(c) $(1,1,2,2)$, $(0,2,0,2)$, $(1,0,2,1)$, $(2,1,4,4)$ in $\mathbb{R}^4$.

Exercise 8. Repeat Exercise 7 for the following sets of vectors.
(a) $(1,1,3,3)$, $(0,2,2,4)$, $(1,0,2,1)$, $(2,1,5,4)$ in $\mathbb{R}^4$.
(b) $1+x$, $1+x-x^2$, $1+x+x^2$, $x-x^2$, $1+2x$ in $\mathcal{P}_2$.
(c) $\cos(2x)$, $\sin^2 x$, $\cos^2 x$, 2 in $C[0,1]$.

Exercise 9. Let $A = \begin{bmatrix} 5 & 2 & -1 \\ 3 & 1 & 0 \\ -1 & 0 & -1 \end{bmatrix}$, $B = \begin{bmatrix} 4 & -3 \\ -2 & 3 \\ 1 & -2 \end{bmatrix}$, $U = \mathcal{C}(A)$, and $V = \mathcal{C}(B)$.
(a) Compute $\dim U$ and $\dim V$.
(b) Use the column algorithm on the matrix $[A, B]$ to compute $\dim(U+V)$.
(c) Use Corollary 3.5 of Section 3.5 to determine $\dim U \cap V$.

Exercise 10. Repeat Exercise 9 with $A = \begin{bmatrix} 4 & 3 & 5 \\ 5 & 4 & 3 \\ 2 & 1 & 9 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 & 3 \\ -2 & -1 & -4 \\ 7 & 5 & 17 \end{bmatrix}$.

Exercise 11. Find a basis of $U \cap V$ in Exercise 9. (*Hint:* Solve the system $\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathbf{0}$ and use the fact that any nonzero solution will give an element in the intersection, namely $A\mathbf{x}$ or $B\mathbf{y}$. Now just look for the right number of linearly independent elements in the intersection.)

Exercise 12. Find a basis of $U \cap V$ in Exercise 10.

Exercise 13. Let $A = \begin{bmatrix} 0 & 1 & 0 & 1 & 2 \\ 1 & 0 & 2 & 1 & 2 \\ 2 & 2 & 4 & 4 & 8 \end{bmatrix}$. Use the column space algorithm on the matrix $[A\ I]$ to find a basis B of $\mathcal{C}(A)$ and to expand it to a basis of $\mathbb{R}^3$.

Exercise 14. Use the isomorphism $T : \mathcal{P}_3 \to \mathbb{R}^4$ given by $T(a+bx+cx^2+dx^3) = (a,b,c,d)$ to find a basis B of

$$V = \operatorname{span}\left\{1-x+2x^2+2x^3, 2x+3x^3, 2-2x+4x^2+4x^3, 2-6x+4x^2-2x^3\right\}$$

and expand it to a basis of $\mathcal{P}_3$ using the method of Exercise 13.

*Problem 15. Suppose that the linear system $A\mathbf{x} = \mathbf{b}$ is a consistent system of equations, where A is an $m \times n$ matrix and $\mathbf{x} = [x_1, \ldots, x_n]^T$. Prove that if the set of columns of A has redundant vectors in it, then the system has more than one solution.

Problem 16. Use Theorem 3.17 and properties of invertible matrices to show that if P and Q are invertible and PAQ is defined, then $\operatorname{rank} PAQ = \operatorname{rank} A$.

*Problem 17. Let A be an $m \times n$ matrix of rank r. Suppose that there exists a vector $\mathbf{b} \in R^m$ such that the system $A\mathbf{x} = \mathbf{b}$ is inconsistent. Use the consistency and rank theorems of this section to deduce that the system $A^T\mathbf{y} = \mathbf{0}$ must have nontrivial solutions.

Problem 18. Use the rank theorem and Problem 16 to prove that if P and Q are invertible and PAQ is defined, then $\dim \mathcal{N}(PAQ) = \dim \mathcal{N}(A)$.

3.7 *Change of Basis and Linear Operators

How much information do we need to uniquely identify an operator? For a general operator the answer is a lot! Specifically, we don't really know everything about it until we know how to find its value at every possible argument. This is an infinite amount of information. Yet we know that in some circumstances we can do better. For example, to know a polynomial function completely, we need only a finite amount of data, namely the coefficients of the polynomial. We have already seen that linear operators are special. Are they described by a finite amount of data? The answer is a resounding yes in the situation in which the domain and target are finite-dimensional.

Let's begin with some notation. We will indicate that $T : V \to W$ is a linear operator, B is a basis of V, and C is a basis of W with the notation

$$T : V_B \to W_C \text{ or } V_B \xrightarrow{T} W_C.$$

Now let $\mathbf{v} \in V$, $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ and $C = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$. We know that there exists a unique set of scalars, the coordinates $c_1, c_2, \ldots, c_n$ of $\mathbf{v}$ with respect to this basis, such that

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n.$$

Thus, by linearity of T we see that

$$T(\mathbf{v}) = T(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n) = c_1T(\mathbf{v}_1) + c_2T(\mathbf{v}_2) + \cdots + c_nT(\mathbf{v}_n).$$

It follows that we know everything about the linear operator T if we know the vectors $T(\mathbf{v}_1), T(\mathbf{v}_2), \ldots, T(\mathbf{v}_n)$.

Now go a step further. Each vector $T(\mathbf{v}_j)$ can be expressed uniquely as a linear combination of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$, namely

$$T(\mathbf{v}_j) = a_{1,j}\mathbf{w}_1 + a_{2,j}\mathbf{w}_2 + \cdots + a_{m,j}\mathbf{w}_m. \tag{3.5}$$

In other words, the scalars $a_{1,j}, a_{2,j}, \ldots, a_{m,j}$ are the coordinates of $T(\mathbf{v}_j)$ with respect to the basis $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$. Stack these in columns and we now have the $m \times n$ matrix $A = [a_{i,j}]$, which contains everything we need to know in order to compute $T(\mathbf{v})$. In fact, with the above terminology, we have

$$\begin{aligned} T(\mathbf{v}) &= c_1 T(\mathbf{v}_1) + c_2 T(\mathbf{v}_2) + \cdots + c_n T(\mathbf{v}_n) \\ &= c_1 \left(a_{1,1}\mathbf{w}_1 + a_{2,1}\mathbf{w}_2 + \cdots + a_{m,1}\mathbf{w}_m\right) + \\ &\qquad \cdots + c_n(a_{1,n}\mathbf{w}_1 + a_{2,n}\mathbf{w}_2 + \cdots + a_{m,n}\mathbf{w}_m) \\ &= (a_{1,1}c_1 + a_{1,2}c_2 + \cdots + a_{1,n}c_n)\mathbf{w}_1 + \\ &\qquad \cdots + (a_{m,1}c_1 + a_{m,2}c_2 + \cdots + a_{m,n}c_n)\,\mathbf{w}_m. \end{aligned}$$

Look closely and we see that the coefficients of these vectors are themselves coordinates of a matrix product, namely the matrix A times the column vector of coordinates of $\mathbf{v}$ with respect to the chosen basis of V. The result of this matrix multiplication is a column vector whose entries are the coordinates of $T(\mathbf{v})$ relative to the chosen basis of W. So in a certain sense, computing the value of a linear operator amounts to no more than multiplying a (coordinate) vector by the matrix A. Thus, we make the following definition.

Definition 3.16. Matrix of Linear Operator The *matrix of the linear operator* $T : V_B \to W_C$ *relative to the bases* B and C is the matrix $[T]_{C,B} = [a_{i,j}]$ whose entries are specified by equation (3.5). In the case that $B = C$, we simply write $[T]_B$.

Recall that we denote the coordinate vector of a vector $\mathbf{v}$ with respect to a basis B by $[\mathbf{v}]_B$. Then the above calculation of $T(\mathbf{v})$ can be stated succinctly in matrix/vector terms as

$$[T(\mathbf{v})]_C = [T]_{C,B}\,[\mathbf{v}]_B\,. \tag{3.6}$$

Standard Matrix of Linear Operator

This equation has a very interesting application to the standard spaces. Recall that a matrix operator is a linear operator $T_A : \mathbb{R}^n \to \mathbb{R}^m$ defined by the formula $T_A(\mathbf{x}) = A\mathbf{x}$, where A is an $m \times n$ matrix. It turns out that *every* linear operator on the standard vector spaces is a matrix operator. The matrix A for which $T = T_A$ is called the *standard matrix* of T.

Theorem 3.22. Linear Operator on Standard Spaces Is Matrix Operator If $T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator, B and C the standard bases for $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and $A = [T]_{C,B}$, then $T = T_A$.

Proof. The proof is straightforward: For vectors $\mathbf{x}$, $\mathbf{y} = \mathbf{T}(\mathbf{x})$ in standard spaces with standard bases B, C, we have $\mathbf{x} = [\mathbf{x}]_B$ and $\mathbf{y} = [\mathbf{y}]_C$. Therefore,

$$T(\mathbf{x}) = \mathbf{y} = [\mathbf{y}]_C = [T(\mathbf{x})]_C = [T]_{C,B}[\mathbf{x}]_B = [T]_{C,B}\mathbf{x} = A\mathbf{x},$$

which proves the theorem. □

Even in the case of an operator as simple as the identity function $\mathrm{id}_V(\mathbf{v}) = \mathbf{v}$, the matrix of a linear operator can be useful and interesting.

Definition 3.17. Change of Basis Matrix Let $\mathrm{id}_V : V_B \to V_C$ be the identity function of V. Then the matrix $[\mathrm{id}_V]_{C,B}$ is called the *change of basis* matrix from the basis B to the basis C.

This definition and equation (3.6) show us that for any vector $\mathbf{v} \in V$,

$$[\mathbf{v}]_C = [\mathrm{id}_V(\mathbf{v})]_C = [\mathrm{id}_V]_{C,B}[\mathbf{v}]_B. \tag{3.7}$$

This allows us to change an expression from one involving basis C to one involving basis B by replacing terms $[\mathbf{v}]_C$ by $[\mathrm{id}_V]_{B,C}[\mathbf{v}]_B$ (which is how we used change of basis in Example 3.26.) Also note that if C is a standard basis, to obtain the change of basis matrix from B to C one simply forms the matrix that has the vectors of basis B listed as its columns.

Example 3.44. Let $V = \mathbb{R}^2$. What is the change of basis matrix from the basis $B = \left\{\mathbf{v}_1 = \begin{bmatrix}1\\2\end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix}-1\\1\end{bmatrix}\right\}$ to the standard basis $C = \{\mathbf{e}_1, \mathbf{e}_2\}$?

Solution. We see that

$$\begin{aligned}\mathrm{id}_V(\mathbf{v}_1) &= \mathbf{v}_1 = 1\,\mathbf{e}_1 + 2\,\mathbf{e}_2\\ \mathrm{id}_V(\mathbf{v}_2) &= \mathbf{v}_2 = -1\,\mathbf{e}_1 + 1\,\mathbf{e}_2.\end{aligned}$$

Compare these equations to (3.5) and we see that the change of basis matrix is

$$[\mathrm{id}_V]_{C,B} = \begin{bmatrix}1 & -1\\ 2 & 1\end{bmatrix}.$$

As predicted, we only have to form the matrix that has the vectors of B listed as its columns. Compare this to the discussion following Example 3.26. □

Next, suppose that $S : U \to V$ and $T : V \to W$ are linear operators. Can we relate the matrices of S, T and the function composition of these operators, $T \circ S$? The answer to this question is a very fundamental fact.

Theorem 3.23. Matrix of Operator Composition If $U_B \xrightarrow{S} V_C \xrightarrow{T} W_D$, then $[T \circ S]_{D,B} = [T]_{D,C}[S]_{C,B}$.

Proof. Let $\mathbf{u} \in U$ and set $\mathbf{v} = S(\mathbf{u})$. With the notation of equation (3.6) we have that $[(T \circ S)(\mathbf{u})]_D = [T \circ S]_{D,B}[\mathbf{u}]_B$ and by definition of function composition that $(T \circ S)(\mathbf{u}) = T(S(\mathbf{u})) = T(\mathbf{v})$. Therefore,

$$[T \circ S]_{D,B} [\mathbf{u}]_B = [(T \circ S)(\mathbf{u})]_D = [T(S(\mathbf{u}))]_D = [T(\mathbf{v})]_D .$$

On the other hand, equation (3.6) also implies that $[T(\mathbf{v})]_D = [T]_{D,C} [\mathbf{v}]_C$ and $[S(\mathbf{u})]_C = [S]_{C,B} [\mathbf{u}]_B$. Hence, we deduce that

$$[T \circ S]_{D,B} [\mathbf{u}]_B = [T]_{D,C} [\mathbf{v}]_C = [T]_{D,C} [S]_{C,B} [\mathbf{u}]_B . \tag{3.8}$$

If we choose $\mathbf{u}$ such that $\mathbf{e}_j = [\mathbf{u}]_B$, where $\mathbf{e}_j$ is the jth standard vector, then we obtain from equation (3.8) that the jth columns of left- and right-hand side agree for all j. Hence, the matrices themselves agree, which is what we wanted to show. □

Corollary 3.8. If the finite-dimensional vector space V has bases B and C and $T : V_B \to V_C$ is an invertible linear operator, then $[T]_{C,B}^{-1} = \left[T^{-1}\right]_{B,C}$.

Proof. Apply Theorem 4.11 to the composition $T^{-1} \circ T = \mathrm{id}_V$, where $T : V_B \to V_C$ and $T^{-1} : V_C \to V_B$ and deduce that $\left[T^{-1}\right]_{B,C} [T]_{C,B} = I$ from which we obtain $\left[T^{-1}\right]_{B,C} = [T]_{C,B}^{-1}$. □

We can now also see exactly what happens when we make a change of basis in the domain and target of a linear operator and recalculate the matrix of the operator. Specifically, suppose that $T : V \to W$ and that B, B' are bases of V and C, C' are bases of W. Let P and Q be the change of basis matrices from B to B' and C to C', respectively. From Corollary 3.8 we obtain that P^{-1} is the change of basis matrix from B' to B. Identify a matrix with its operator action by multiplication, and we have a chain of operators

Operator Matrix Under Change of Bases

$$V_{B'} \overset{\mathrm{id}_V}{\to} V_B \overset{T}{\to} W_C \overset{\mathrm{id}_W}{\to} W_{C'}.$$

Application of Theorem 3.23 shows that

$$[T]_{C',B'} = [\mathrm{id}_W]_{C',C} [T]_{C,B} [\mathrm{id}_V]_{B,B'} = Q[T]_{C,B} P^{-1}.$$

We have just obtained a very important insight into the matrix of a linear transformation. Here is the form it takes for the standard spaces.

Corollary 3.9. Change of Basis for Matrix Operator Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear operator, B a basis of $\mathbb{R}^n$, and C a basis of $\mathbb{R}^m$. Let P and Q be the change of basis matrices from the bases B and C to the standard bases, respectively. If A is the matrix of T with respect to the standard bases and $M = [T]_{C,B}$ the matrix of T with respect to the bases B and C, then

$$A = QMP^{-1}.$$

Example 3.45. Given the linear operator $T : \mathbb{R}^4 \to \mathbb{R}^2$ by the rule

$$T(x_1, x_2, x_3, x_4) = \begin{bmatrix} x_1 + 3x_2 - x_3 \\ 2x_1 + x_2 - x_4 \end{bmatrix},$$

find the standard matrix of T.

Solution. We see that

$$T(\mathbf{e}_1) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad T(\mathbf{e}_2) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad T(\mathbf{e}_3) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad T(\mathbf{e}_4) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Since the standard coordinate vector of a standard vector is itself, we have

$$[T] = \begin{bmatrix} 1 & 3 & -1 & 0 \\ 2 & 1 & 0 & -1 \end{bmatrix}. \qquad \square$$

Example 3.46. With T as in the previous example, find the matrix of T with respect to the domain basis $B = \{(1,0,0,0), (1,1,0,0), (1,0,1,0), (1,0,0,1)\}$ and range basis $C = \{(1,2), (-1,1)\}$

Solution. Let A be the matrix of the previous example, so it represents the standard matrix of T. Let B' and C' be the standard bases for the domain and target of T. Then we have

$$A = [T] = [T]_{C',B'}.$$

Further, we have only to stack columns of B and C to obtain change of basis matrices from these bases to the standard bases B' and C':

$$P = [\mathrm{id}_{\mathbb{R}^4}]_{B',B} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } Q = [\mathrm{id}_{\mathbb{R}^2}]_{C',C} = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}.$$

Now apply Corollary 3.9 to obtain that

$$\begin{aligned} [T]_{C,B} &= Q^{-1}\,[T]_{C',B'}\,P = Q^{-1}AP \\ &= \frac{1}{3}\begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 3 & -1 & 0 \\ 2 & 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \frac{1}{3}\begin{bmatrix} 3 & 7 & 2 & 2 \\ 0 & -5 & 2 & -1 \end{bmatrix}. \end{aligned}$$

$\square$

3.7 Exercises and Problems

Exercise 1. Find the standard matrix, kernel, and range of the linear operator $T : \mathbb{R}^3 \to \mathbb{R}^3$ given by $T((x, y, z)) = (x + 2y, x - y, y + z)$.

Exercise 2. Find the standard matrix, kernel, and range of the linear operator $T : \mathbb{R}^4 \to \mathbb{R}^2$ given by $T((x_1, x_2, x_3, x_4)) = (x_2 - x_4 + 3x_3, 3x_2 - x_4 + x_3)$.

Exercise 3. Bases $B = \{(1, 1), (1, -1)\} = \{\mathbf{u}_1, \mathbf{u}_2\}$ and $B' = \{(2, 0), (3, 1)\} = \{\mathbf{u}'_1, \mathbf{u}'_2\}$ of $\mathbb{R}^2$ are given.
(a) Find the change of basis to the standard basis from each of these bases.
(b) Use (a) to compute the change of basis matrix from B to B' by applying Corollary 3.9 to $T = \mathrm{id}_{\mathbb{R}^2}$.
(c) Suppose that $\mathbf{w} = 3\mathbf{u}_1 + 4\mathbf{u}_2$ and use (b) to express $\mathbf{w}$ as a linear combination of $\mathbf{u}'_1$ and $\mathbf{u}'_2$.

Exercise 4. Given bases $B = \{(0, 1, 1), (1, 0, 1), (1, 0, -1)\} = \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and $B' = \{(0, 0, -1), (0, 3, 1), (2, 0, 0)\} = \{\mathbf{u}'_1, \mathbf{u}'_2, \mathbf{u}'_3\}$ of $\mathbb{R}^3$, find the change of basis matrix from B to B' and use it to express $\mathbf{w}' = 2\mathbf{u}'_1 + \mathbf{u}'_2 - 2\mathbf{u}'_3$ in terms of $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$.

Exercise 5. Find the matrix of the operator $T_A : \mathbb{R}^3 \to \mathbb{R}^2$, where $A = \begin{bmatrix} 2 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$, with respect to the bases $B = \{(1, 0, 1), (1, -1, 0)), (0, 0, 2)\}$ and $C = \{(3, 4), (4, -3)\}$.

Exercise 6. Find the matrix of the operator $T : \mathcal{P}_3 \to \mathcal{P}_2$, where T is given by $T\left(a + bx + cx^2 + dx^3\right) = b + 2cx + 3dx^2$, with respect to the bases $B = \left\{1, x, x^2, x^3\right\}$ and $C = \left\{1, x, 2x^2 - 1\right\}$.

Problem 7. Two $n \times n$ matrices A and B are called *similar* if there exists an invertible matrix P such that $B = P^{-1}AP$. Use Corollary 3.9 to show that similar matrices A and B are both matrices of the same linear operator, namely T_A, with respect to different bases.

Problem 8. Show that a change of basis matrix from one orthonormal basis to another is an orthogonal matrix. Use this to simplify the change of basis formula of Corollary 3.9 in the case that C is an orthonormal basis.

*Problem 9. Define the *determinant* of a linear operator $T : V \to V$ to be the determinant of $[T]_B$, where B is any basis of the finite-dimensional vector space V. Show that this definition is independent of the basis B.

Problem 10. Let λ be a scalar and A, B similar $n \times n$ matrices, i.e., for some invertible matrix P, $B = P^{-1}AP$. Show that $\dim \mathcal{N}(\lambda I - A) = \dim \mathcal{N}(\lambda I - B)$.

3.8 *Introduction to Linear Programming

Basic Ideas

The term "linear programming" does not refer to programming in the computer sense but rather the more general idea of a program as a plan. As mathematics goes, linear programming is a relatively recent development: Early full formulations of the problem were given in the late 1930s by mathematician Leonid Kantorovich, Dutch American mathematician and economist T. C. Koopmans, and economist Wassily Leontief. However, it was George Danzig's development in 1947 of the simplex method for solving linear programming problems that turned linear programming into a practical tool for all sorts of industrial and operations research problems.

In this section we will utilize some useful notation regarding real vectors and matrices. If A and B are matrices or vectors of the same size, we interpret the statement $A \leq B$ to mean that entries from corresponding coordinates in A and B satisfy the same inequality. This also applies to the symbols $<$, $\geq$, $>$ and so forth. Thus, we can make statements like "$[1, 0, 3] \leq [2, 1, 3]$" and "$\begin{bmatrix} 4 & 1 \\ -2 & 2 \end{bmatrix} > \begin{bmatrix} 2 & 0 \\ -3 & 0 \end{bmatrix}$".

Matrix Vector Inequalities

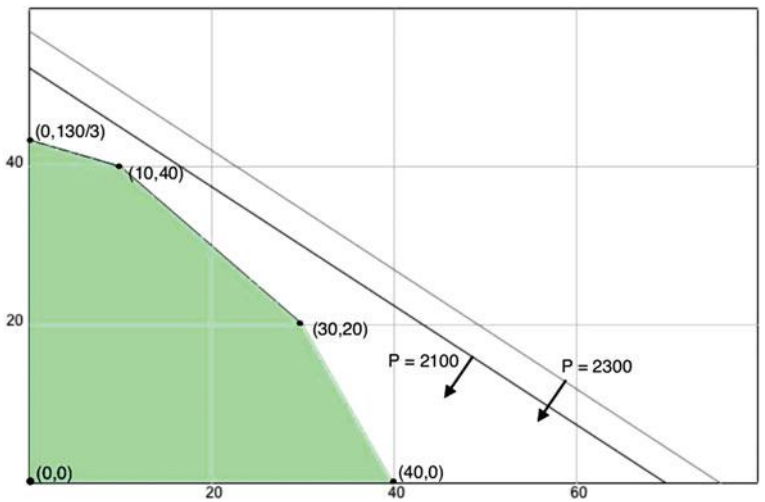


Fig. 3.5: Feasible set and profit lines for Example 3.47.

In industrial and economic practice, optimization (finding either minimum or maximum possible value of a quantity subject to certain constraints) is an important process that could involve hundreds or even thousands of variables. Linear programming is designed to handle certain types of optimization problems. To illustrate some of the ideas behind linear programming, we consider the following (highly simplified) examples of manufacturing and diet problems.

Objective Function

The first version of optimizing is to maximize a quantity which is described by a linear function called the *objective function*.

Example 3.47. Suppose that a company produces two vacuum cleaners, upright (V_1) and canister (V_2), for which it makes 30 and 40 dollars of profit per item, respectively. Production requires time on each of three assembly lines, A_1, A_2, and A_3. Currently, the number of available hours that could be spent in the production of these items by each assembly line is given in the table below, along with product rates for products by units:

	Hours/product		
	V_1	V_2	Available
A_1	2	1	80
Lines A_2	1	3	130
A_3	2	2	100

Describe the problem of maximizing profit from utilization of the available times from these assembly lines in terms of ordinary and matrix/vector inequalities.

Solution. Let x_j be the total number of product V_j produced, $j = 1, 2$. The objective function of this problem is the profit function which can be written as $P = 30x_1 + 40x_2$. Given these numbers of products, the problem is to maximize P subject to the time constraints on the three assembly lines and products:

$$\begin{aligned} 2x_1 + x_2 &\leq 80 \\ x_1 + 3x_2 &\leq 130 \\ 2x_1 + 2x_2 &\leq 100 \\ x_j &\geq 0, \ j = 1, 2. \end{aligned}$$

In terms of matrices and vectors, set

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \ \mathbf{c} = \begin{bmatrix} 30 \\ 40 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} 80 \\ 130 \\ 100 \end{bmatrix}, \ A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \\ 2 & 2 \end{bmatrix}$$

and we can express the problem concisely as this *linear program*:

Maximize $P = \mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$. □

Max Linear Program

In the max linear program think of the objective function $P = \mathbf{c}^T\mathbf{x}$ as a function of the variable $\mathbf{x}$. The next version of optimizing is to minimize a quantity.

Example 3.48. Suppose that a food service must provide a diet that satisfies certain nutritional requirements. In particular, suppose that a combination

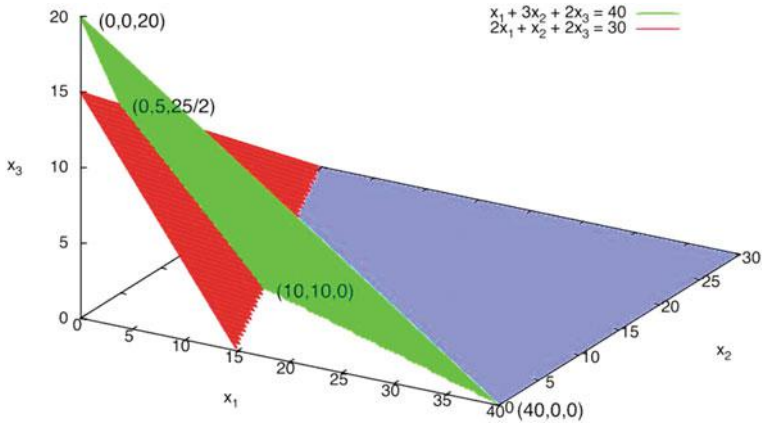


Fig. 3.6: Feasible set for Example 3.48 is above planes in the first octant.

from three different food items, F_1, F_2 and F_3, must supply certain minimum number of milligrams of vitamins C (V_1) and B complex (V_2). The unit costs of the food items, amounts of each vitamin contained in each food item and minimum requirements are given in this table:

		F_1	F_2	F_3	Minimum Requirement
Nutrients	V_1	2	1	2	30
	V_2	1	3	2	40
	Unit cost	80	130	100	

Describe the problem of minimizing cost of this diet while satisfying the minimum nutritional requirements in terms of ordinary and matrix/vector inequalities.

Solution. Let x_j be the number of units of food source F_j required in the diet, $j = 1, 2, 3$. The objective function here is the cost function which can be written as $C = 80x_1 + 130x_2 + 100x_3$. Given these units of food items, the problem is to minimize C subject to the constraints on the minimum requirements of the two vitamins:

$$\begin{aligned} 2x_1 + x_2 + 2x_3 &\geq 30 \\ x_1 + 3x_2 + 2x_3 &\geq 40 \\ x_j &\geq 0, \ j = 1, 2, 3. \end{aligned}$$

In terms of matrices and vectors, set

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \ \mathbf{c} = \begin{bmatrix} 80 \\ 130 \\ 100 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} 30 \\ 40 \end{bmatrix}, \ A = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 3 & 2 \end{bmatrix}$$

and we can express the problem concisely as this *linear program*:

Minimize $C = \mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} \geq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$. □

Min Linear Program

Notice that minimizing $C = \mathbf{c}^T\mathbf{x}$ is equivalent to maximizing $P = -\mathbf{c}^T\mathbf{x}$ and that the constraints $A\mathbf{x} \geq \mathbf{b}$ are the same as constraints $-A\mathbf{x} \leq -\mathbf{b}$. Hence, every min linear program can be converted to a max linear program (and vice versa.)

So how do we solve these problems? Before describing a general procedure we shall attack the problem of Example 3.47 with a geometrical approach. First, some terminology: A vector $\mathbf{x}$ satisfying all the constraints of a linear programming problem is called a *feasible solution*, and the set of all such vectors is the *feasible set* for the problem. Figure 3.5 illustrates the feasible set for the following example.

Feasible Solution and Set

Example 3.49. Find the solution to the linear program of Example 3.47 using a geometric method.

Solution. First we identify the lines bounding the inequalities of the problem as (1): $2x_1+x_2 = 80$, (2): $x_1+3x_2 = 130$ and (3): $2x_1+2x_2 = 100$. Examine Figure 3.5 where we see that the point $(0, 130/3)$ is at the intersection of the positive x_2 axis with line (2), point $(10, 40)$ at the intersection of line (2) and line (3), point $(30, 20)$ is at the intersection of lines (3) and (1), and $(40, 0)$ at the intersection of line (1) and the x_1-axis. The feasible set for this problem lies under these lines, above the x_1-axis and to the right of the x_2-axis.

To visualize the solution, think of P as controlling the placement of the lines of constant slope $P = 30x_1 + 45x_2$. At $P = 2300$ and $P = 2100$ these lines lie outside the feasible set. But as we steadily decrease P, it is clear that first contact with the feasible set will be at an upper corner of this polyhedron. Visual inspection shows that the first point of contact is at $(10, 40)$. Hence, with the given constraints the largest possible value of the profit P is $P = 30{\cdot}10+40{\cdot}40 = 1900$, and it occurs with production of $x_1 = 10$ upright vacuum cleaners and $x_2 = 40$ canister vacuum cleaners. To confirm our answer, check that $P = 1733.33$ at the corner $(0, 130/3)$, $P = 0$ at $(0, 0)$, $P = 1700$ at $(30, 20)$ and $P = 1200$ at $(40, 0)$. So $P = 1900$ is optimal and $\mathbf{x} = (10, 40)$ is an optimal solution to the problem. □

Example 3.50. Find the solution to the linear program of Example 3.48 using a geometric method.

Solution. Matters are a bit more difficult to visualize in Example 3.48. The feasible set lies in the first octant, bounded by the coordinate planes $x_j = 0$, $j = 1, 2, 3$, and lies above the planes $2x_1 + x_2 + 2x_3 = 30$ and $x_1 + 3x_2 + 2x_3 = 40$. Visual inspection for the lowest point of contact of a plane of the form $C = 80x_1 + 130x_2 + 100x_3$ with a corner of the feasible set is a bit difficult here. We see from Figure 3.6 and calculating intersections with one or two coordinate values set to zero that this set has four corners: $(0, 0, 20)$, $\left(0, 5, \frac{25}{2}\right)$, $(10, 10, 0)$ and $(40, 0, 0)$. We could try to visualize the planes $C = 80x_1+130x_2+100x_3$ moving towards the feasible set as C increases

from zero, but it's simpler to check the values of C at each corner: $C = 2000$ at $(0, 0, 20)$, $C = 1900$ at $\left(0, 5, \frac{25}{2}\right)$, $C = 2100$ at $(10, 10, 0)$ and $C = 3200$ at $(40, 0, 0)$. Hence, the optimal minimum value of the cost is $C = 1900$ and this occurs with a diet using 0 units of F_1, 5 units of F_2 and 12.5 units of F_3. So $\left(0, 5, \frac{25}{2}\right)$ is an optimal solution to this problem. □

Although the geometrical methods which we used here give us an intuitive understanding of solutions to linear programming problems, they are impractical for higher dimensional problems. What is needed is a systematic algebraic approach to solutions, a system that deals with linear equations rather than complicated inequalities. The first step that we shall take is the observation that the inequalities of our examples can be turned into equalities – if we are willing to accept more nonnegative variables.

Example 3.51. Convert the constraints of Examples 3.47 and 3.48 into linear equations by introducing additional nonnegative variables.

Solution. Any inequality of the form $a \leq b$ can be converted to an equality by introducing a new variable $x \geq 0$ taking up the slack on the left side by satisfying $a + x = b$. Thus, in the case of Example 3.47 we need three new *slack variables* x_3, x_4, x_5 satisfying the linear system

Slack Variable

$$\begin{aligned} 2x_1 + x_2 + x_3 &= 80 \\ x_1 + 3x_2 + x_4 &= 130 \\ 2x_1 + 2x_2 + x_5 &= 100 \\ x_j \geq 0,\ j &= 1, 2, 3, 4, 5. \end{aligned}$$

Similarly any inequality of the form $a \geq b$ can be converted to an equality by introducing a new variable $x \geq 0$ removing the surplus on the left side by satisfying $a - x = b$. Thus, in the case of Example 3.47 we need two new *surplus variables* x_4, x_5 satisfying the linear system

Surplus Variable

$$\begin{aligned} 2x_1 + x_2 + 2x_3 - x_4 &= 30 \\ x_1 + 3x_2 + 2x_3 - x_5 &= 40 \\ x_j \geq 0,\ j &= 1, 2, 3, 4, 5. \end{aligned}$$

□

The previous example reveals a key idea: The max and min linear programs of Examples 3.47 and 3.48 can be expressed in a common format provided we make a minor adjustment to the objective function: Expand it with zero coefficients for all new variables introduced into the problem. This will not affect the optimal value of the objective function since artificial variables make zero contribution to it. In the following, the term "optimize" is understood to mean either "maximize" or "minimize". This is the common format, which we will call a linear program in *standard form*:

Given an $m \times p$ matrix B of rank m, vectors $\mathbf{d} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^p$,

optimize the objective function $\mathbf{c}^T\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^p$ subject to the constraints

Standard Form

$$\begin{aligned} B\mathbf{x} &= \mathbf{d} \\ \mathbf{x} &\geq \mathbf{0} \end{aligned} \tag{3.9}$$

Any $\mathbf{x} \in \mathbb{R}^p$ that solves this problem is called an optimal solution. The standard form of a linear program has some substantial advantages over the min or max linear program form. For one it can accommodate mixed inequalities in a single problem. It can even allow for incorporating equalities into a linear programming problem.

Optimal Solution

Example 3.52. Express the linear programming problems of Examples 3.47 and 3.48 in standard form.

Solution. We use the formats of Example 3.51. In the case of Example 3.47 set

$\mathbf{x} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{pmatrix}$, $\mathbf{c} = (30, 40, 0, 0, 0)$, $\mathbf{d} = (80, 130, 100)$ and

$$B = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 \\ 1 & 3 & 0 & 1 & 0 \\ 2 & 2 & 0 & 0 & 1 \end{bmatrix}.$$

For Example 3.48 set

$\mathbf{x} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{pmatrix}$, $\mathbf{c} = (80, 130, 100, 0, 0)$, $\mathbf{d} = (30, 40)$ and

$$B = \begin{bmatrix} 2 & 1 & 2 & -1 & 0 \\ 1 & 3 & 2 & 0 & -1 \end{bmatrix}.$$

□

A key idea in linear programming is the notion of a basic solution to a linear program in standard form.

Definition 3.18. **Basic Solution and Variables** Given a linear system $B\mathbf{x} = \mathbf{d}$ where $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p]$ is $m \times p$ of rank $m \leq p$ and a set of m linearly independent columns of A, say $\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \ldots, \mathbf{b}_{i_m}$, the *basic solution* defined by this set is the vector $\mathbf{x} \in \mathbb{R}^p$ such that the coordinates of $\mathbf{x}$ corresponding to these columns satisfy the equation $[\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \ldots, \mathbf{b}_{i_m}]\,[x_{i_1}, x_{i_2}, \ldots, x_{i_m}]^T = \mathbf{d}$ and all other coordinates of $\mathbf{x}$ are zero. The coordinates $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$ are called the *basic variables* of this basic solution.

Note that a set m of linearly independent columns of B is actually a basis of $\mathcal{C}(B)$ since this space has dimension m. Moreover, the $m \times m$ matrix $C = [\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \ldots, \mathbf{b}_{i_m}]$ is invertible since its rank is m. Therefore, the basic solution $\mathbf{x}$ of this definition is uniquely defined. However, in the context of a linear program in the standard form (3.9), a basic solution $\mathbf{x}$ will not be useful unless it is also feasible, i.e., $\mathbf{x} \geq \mathbf{0}$. Such solutions turn out to be key tools in the simplex method of solving linear programming problems which we will discuss shortly. Indeed, finding an initial basic feasible solution is crucial for the simplex method.

Example 3.53. Find basic feasible solutions to the standard forms of Examples 3.47 and 3.48.

Solution. We use the standard forms of Example 3.52. In the case of Example 3.47 the relevant system is

$$B\mathbf{x} = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 \\ 1 & 3 & 0 & 1 & 0 \\ 2 & 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 80 \\ 130 \\ 100 \end{bmatrix} = \mathbf{d}.$$

Here a basic feasible solution is obvious: Use the last three columns of A as the basis and we obtain that $\mathbf{x} = (0, 0, 80, 130, 100)$ is a basic feasible solution. This example also highlights the advantage of having the columns of a basic feasible solution be the columns of an identity matrix in some order: We can read off the value of this coordinate of the basic variable from the entries of the right-hand side vector $\mathbf{b}$.

In the case of Example 3.48 the relevant system is

$$B\mathbf{x} = \begin{bmatrix} 2 & 1 & 2 & -1 & 0 \\ 1 & 3 & 2 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 30 \\ 40 \end{bmatrix} = \mathbf{d}.$$

Here the choice of the last two columns of B will not do the job. It does supply a *basic* solution, namely $\mathbf{x} = (0, 0, 0, -30, -40)$, but this solution is clearly not feasible. One could easily find a basic feasible solution by trial and error (or consulting Figure 3.6), but we want a more systematic algebraic approach that works or more general problems. What we will do is introduce two new *artificial* variables, x_6 and x_7, which we require to be nonnegative. The revised system has this enlarged coefficient matrix $B_e = \begin{bmatrix} 2 & 1 & 2 & -1 & 0 & 1 & 0 \\ 1 & 3 & 2 & 0 & -1 & 0 & 1 \end{bmatrix}$ and augmented matrix $[B_e \,|\, \mathbf{d}] = \begin{bmatrix} 2 & 1 & 2 & -1 & 0 & 1 & 0 & 30 \\ 1 & 3 & 2 & 0 & -1 & 0 & 1 & 40 \end{bmatrix}$ with basic variables $x_6 = 30$ and $x_7 = 40$, all others nonbasic. Next we will perform row operations on the system so as to replace x_6 or x_7 by x_1, x_2, or x_3 as basic, but we must proceed carefully. If we decide to make x_3 basic, we must choose a pivot in the third column to solve for the value of x_3 by way of row operations. Looking ahead, we see that using the $(2, 3)$th entry as a pivot would be a bad idea, since it would cause x_6 to have a negative value of -10 and hence our basic solution would no longer be feasible. So use the $(1, 3)$th entry as pivot as follows:

Artificial Variables

$$[B_e \,|\, \mathbf{d}] = \begin{bmatrix} 2 & 1 & \textcircled{2} & -1 & 0 & 1 & 0 & 30 \\ 1 & 3 & 2 & 0 & -1 & 0 & 1 & 40 \end{bmatrix} \xrightarrow[E_{21}(-2)]{E_1\left(\frac{1}{2}\right)} \begin{bmatrix} 1 & \frac{1}{2} & 1 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & 15 \\ -1 & 2 & 0 & 1 & -1 & -1 & 1 & 10 \end{bmatrix}.$$

Now x_3 and x_7 are basic. If we decide to make x_2 basic, looking ahead shows us that the $(1, 2)$th entry of the second column is a bad choice since it will

lead to the other basic variable x_7 having a negative value of -50. So use the $(2,2)$th entry for pivot as follows:

$$\left[\begin{array}{cccccccc} 1 & \frac{1}{2} & 1 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & 15 \\ -1 & \textcircled{2} & 0 & 1 & -1 & -1 & 1 & 10 \end{array}\right] \begin{array}{c} \longrightarrow \\ E_2\left(\frac{1}{2}\right) \\ E_{12}\left(-\frac{1}{2}\right) \end{array} \left[\begin{array}{cccccccc} \frac{5}{4} & 0 & 1 & -\frac{3}{4} & \frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & \frac{25}{2} \\ -\frac{1}{2} & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 5 \end{array}\right].$$

At this point we can discard the non-basic artificial variables x_6 and x_7 since they have value zero and contribute zero to the objective function. We have found a basic feasible solution to the original problem, namely $\mathbf{x} = \left(0, 5, \frac{25}{2}, 0, 0\right)$. Notice that the point $\left(0, 5, \frac{25}{2}\right)$ appears as a corner point to the feasible set in Figure 3.6. □

There are many other difficulties besides that of the previous example that one can encounter in solving a linear program. For example, the problem may have no solution because the feasible set is too large (e.g., maximize $P = 2x_1$ subject to constraint $x_1 \geq 3$) or may be empty (minimize $P = 2x_1$ subject to constraints $x_1 \geq 2$ and $x_1 \leq 1$) or may have multiple solutions (maximize $P = x_1 + x_2$ subject to constraints $x_1 + x_2 = 2$, $x_1, x_2 \geq 0$). Or the search for an optimal solution may lead to *cycling*, i.e., going around in a loop of successive feasible solutions with no change in objective function. There are other difficulties as well, but the following key theorem assures us that a search for basic feasible variables will not fail to yield results, if there is a solution at all.

Theorem 3.24. **Fundamental Theorem of Linear Programming** If a linear program in standard form has a feasible solution, then it has a basic feasible solution, and if it has an optimal feasible solution, then it has an optimal basic feasible solution.

Proof. We are given that $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p]$ is $m \times p$ of rank $m \leq p$. We may assume that the columns are all nonzero since any zero column contributes nothing to the problem and can be deleted. Suppose first that the linear program has a feasible solution $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. Relabel the column indices so that nonzero coefficients come first, say $x_j > 0$ for $1 \leq j \leq q \leq p$ and $x_j = 0$ for $q < j \leq p$. What results is that

$$B\mathbf{x} = x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + \cdots + x_q\mathbf{b}_q = \mathbf{d}. \tag{3.10}$$

Next, suppose that $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q$ are linearly dependent, say

$$y_1\mathbf{b}_1 + y_2\mathbf{b}_2 + \cdots + y_q\mathbf{b}_q = \mathbf{0}. \tag{3.11}$$

Multiply (3.11) by $-\alpha$ and add it to (3.10) to obtain the equation

$$(x_1 - \alpha y_1)\,\mathbf{b}_1 + (x_2 - \alpha y_2)\,\mathbf{b}_2 + \cdots + (x_q - \alpha y_q)\,\mathbf{b}_q = \mathbf{d}. \tag{3.12}$$

For $|\alpha|$ sufficiently small all of the coefficients $z_j = x_j - \alpha y_j$ are clearly positive since all the x_j's are. Since at least one $y_j \neq 0$, as we increase $|\alpha|$ we will reach a

point where one or more of the x_j's vanishes while the others are still positive, say at $\alpha = \alpha_0$. Discard the corresponding columns and what results is a new linear combination of fewer than p columns of A with all positive coefficients and summing to $\mathbf{b}$.

Next, suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ is an optimal feasible solution. Again, relabel the column indices so that nonzero coefficients come first, say $x_j > 0$ for $1 \leq j \leq q \leq p$ and $x_j = 0$ for $q < j \leq p$. With the notation of equation (3.12) set $\mathbf{y} = (y_1, y_2, \ldots, y_p)$ where $y_j = 0$ for $q < j \leq p$. Then we have

$$P = \mathbf{c}^T(\mathbf{x} - \alpha\mathbf{y}) = \mathbf{c}^T\mathbf{x} - \alpha\mathbf{c}^T\mathbf{y}.$$

As we have noted, for all α with $|\alpha|$ sufficiently small, $\mathbf{x} - \alpha\mathbf{y}$ is a feasible solution to the problem. If $\mathbf{c}^T\mathbf{y}$ were nonzero, then a suitable choice of small positive or negative α would increase or decrease the value of the objective function P, so it cannot be optimal maximal or minimal. Therefore, $\mathbf{c}^T\mathbf{y} = 0$. So choosing $\mathbf{y}$ as in the first paragraph will not affect the optimal value of P.

In summary, what we have shown is that if the columns $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q$ of (3.10) are linearly dependent, then they can be replaced by a smaller set of columns for which $\mathbf{d}$ is a linear combination of them with positive coefficients. Moreover, the vector formed by using these positive coefficients along with and zero coefficients for columns of A outside this set is a feasible vector which is optimal if the feasible solution $\mathbf{x}$ is optimal. Hence, we can remove vectors from (3.10) one at a time until we have reached a linearly independent set of $q \leq p$ columns of A which yield an equation of the form

$$z_{i_1}\mathbf{b}_{i_1} + z_{i_2}\mathbf{b}_{i_2} + \cdots + z_{i_q}\mathbf{b}_{i_q} = \mathbf{d}, \tag{3.13}$$

where all coefficients $z_j > 0$. Since the columns of A are a spanning set for the m dimensional vector space $V = \mathcal{C}(B)$, the linearly independent set $\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \ldots, \mathbf{b}_{i_q}$ can be expanded to a basis $\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \ldots, \mathbf{b}_{i_m}$ of V using only additional columns of B (see Exercise 30 of Section 3.5). Set $z_k = 0$ for for any index k other than i_j, $1 \leq j \leq q$, so that $[\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \ldots, \mathbf{b}_{i_m}]\,[z_{i_1}, z_{i_2}, \ldots, z_{i_m}]^T = \mathbf{d}$. Then the resulting solution $z = (z_1, z_2, \ldots, z_n)$ to the linear program is basic feasible. Moreover, if the original solution $\mathbf{x}$ is optimal, so is $\mathbf{z}$. □

The Simplex Method

We've already seen some basic steps of the simplex method in Example 3.53. In general, the simplex method for a problem in standard form consists of applying elementary row operations to a matrix which we call the standard augmented matrix: If the problem is to optimize $\mathbf{c}^T\mathbf{x}$ subject to the constraints $B\mathbf{x} = \mathbf{d}$ and $\mathbf{x} \geq \mathbf{0}$, then the standard augmented matrix is

Standard Augmented Matrix

$\tilde{B} = \begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}^T & 0 \end{bmatrix}$. The simplex method uses the last row to guide us to an optimal value

of the objective function that will appear in the lower right corner of the standard augmented matrix. The goal is to find an optimal basic feasible solution since Theorem 3.24 tells us that if there is an optimal feasible solution then there is an optimal basic feasible solution.

We think of the last row as representing the equation $P - \mathbf{c}^T\mathbf{x} = u$, where the values of P and u are defined by the current basic feasible solution. The variable P does not warrant an extra column because it does not appear in any earlier equation and will not be affected by subsequent operations on this augmented matrix. The only admissible elementary row operations are any elementary operations on the first m rows of $\tilde{B}$ and the elementary operations of adding a multiple of one of the first m rows to the last row. We use these operations to solve for the values of the basic variables by converting the columns corresponding to these variables to the columns of the identity matrix in some order. Thus, we can read off the value of basic variables as entries of the last column of $\tilde{B}$ (nonbasic variables are always set to zero). A key result is that these operations essentially do not change the outcome in the sense that they give us a new optimization problem with exactly the same solutions as the original:

Admissible Operations

Theorem 3.25. If admissible elementary operations are applied to a standard augmented matrix of the form $\begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}^T & u \end{bmatrix}$, where B is $m \times p$ of rank $m \leq p$, then the resulting matrix is a standard augmented matrix for an equivalent problem.

Proof. We leave it as an exercise to show that all possible admissible transformations of the system are accounted for by a matrix of the form $\begin{bmatrix} E & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix}$, where E is an $m \times m$ product of elementary row operations and $\mathbf{v} \in \mathbb{R}^p$. Such operations result in a new standard form, namely

$$\tilde{B}_t = \begin{bmatrix} E & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix} \begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}^T & u \end{bmatrix} = \begin{bmatrix} EB & E\mathbf{d} \\ \mathbf{v}^T B - \mathbf{c}^T & \mathbf{v}^T\mathbf{d} + u \end{bmatrix}.$$

So the problem of optimizing $\mathbf{c}^T\mathbf{x} + u$ subject to $B\mathbf{x} = \mathbf{d}$ is transformed into the problem of optimizing $\left(\mathbf{c}^T - \mathbf{v}^T B\right)\mathbf{x} + \mathbf{v}^T\mathbf{d} + u$ subject to $EB\mathbf{x} = E\mathbf{d}$. However the constraints are equivalent to B$\mathbf{x} = \mathbf{d}$ since E is invertible, and the new objective function is

$$\left(\mathbf{c}^T - \mathbf{v}^T B\right)\mathbf{x} + \mathbf{v}^T\mathbf{d} + u = \mathbf{c}^T\mathbf{x} - \mathbf{v}^T B\mathbf{x} + \mathbf{v}^T\mathbf{d} + u = \mathbf{c}^T\mathbf{x} + u$$

since B$\mathbf{x} = \mathbf{d}$. Thus, the problem is unchanged by these elementary operations and we can think of $\tilde{B}_t$ as the augmented standard matrix of a new problem that is completely equivalent to the original problem. □

Corollary 3.10. Suppose that admissible elementary operations are applied to the standard augmented matrix $\begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}^T & u \end{bmatrix}$ of a maximization problem to yield the matrix $\tilde{B}_f = \begin{bmatrix} B_f & \mathbf{d}_f \\ -\mathbf{c}_f^T & u_f \end{bmatrix}$, where B_f is $m \times p$ of rank $m \leq p$, and $\mathbf{c}_f \leq \mathbf{0}$. If $\mathbf{x}_*$ is a basic feasible solution for $\tilde{B}_f$ such that the columns of $\tilde{B}_f$ corresponding to the basic variables of $\mathbf{x}_*$ are columns of I_{m+1}, then $\mathbf{x}_*$ is an optimal feasible solution for the problem and $\mathbf{c}_f^T\mathbf{x}_* + u_f$ is the optimal value of the objective function of the problem.

Proof. By Theorem 3.25 we may as well assume that the problem in question is defined by $\tilde{B}_f$. In this case the objective function is $\mathbf{c}_f^T\mathbf{x} + u_f$, where coefficients of the objective function corresponding to basic variables of $\mathbf{x}_*$ are zero and all others are nonpositive. Any other feasible solution $\mathbf{x}$ may differ from $\mathbf{x}_*$ in a basic coordinate, which does not change the value of the objective function, or in a nonbasic coordinate which will produce no improvement in the objective function since the corresponding coordinate of $\mathbf{c}_f$ is nonpositive. Hence, no improvement in the value of the objective function is possible and $\mathbf{x}_*$ is an optimal feasible solution to the problem. □

Here is a general description of how this applies to a linear program in standard form: First, we must be able to find an initial basic feasible solution. In the case of a max linear program with system of inequalities $A\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{b} \geq \mathbf{0}$, convert the constraints $A\mathbf{x} \leq \mathbf{b}$ to equalities by adding nonnegative slack variables so that the new system has coefficient matrix $B = \begin{bmatrix} A & I_m \end{bmatrix}$, vector of unknowns $\mathbf{x}_e = (x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m})$, objective vector $\mathbf{c}_e = \begin{bmatrix} \mathbf{c}^T, \mathbf{0}_m^T \end{bmatrix}$ and standard augmented matrix

$$\tilde{B} = \begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}_e^T & u \end{bmatrix} = \begin{bmatrix} A & I & \mathbf{b} \\ -\mathbf{c}^T & \mathbf{0}^T & 0 \end{bmatrix}.$$

Thus, we have an immediate basic feasible solution $\mathbf{x}_e = (0, \ldots, 0, \mathrm{b}_1, \ldots, \mathrm{b}_m)$. In other cases we have to use different approaches, as in the second part of Example 3.53, but in all cases we should begin with a basic feasible solution to an optimization problem with standard augmented matrix. Of course if none exists then the problem has no feasible solution by Theorem 3.24.

Once we have put our problem into standard form and found an initial basic feasible solution, our goal in the simplex method for maximization is to end up with a standard augmented matrix of the form $\tilde{B} = \begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}^T & u \end{bmatrix}$ with $\mathbf{c}^T \leq \mathbf{0}$ so that no further improvement in the objective function is possible by way of basic feasible solutions. To this end we use the last row to find a nonbasic variable x_j to trade into the current basic solution because it would improve the value of the objective function, and we use elementary operations to convert its corresponding column to a unit column, i.e., a column of the identity matrix. We leave it as an exercise to verify that if such a state

is achieved, the current basic feasible solution is indeed an optimal feasible solution. The following examples illustrates the simplex procedure.

Example 3.54. Solve the max linear program of Example 3.47 using the simplex method.

Solution. In this case the standard augmented system looks like

$$\tilde{A}_e = \begin{bmatrix} A & I & \mathbf{b} \\ -\mathbf{c}^T & \mathbf{0}^T & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1\,0\,0 & 80 \\ 1 & 3 & 0\,1\,0 & 130 \\ 2 & 2 & 0\,0\,1 & 100 \\ -30 & -40 & 0\,0\,0 & 0 \end{bmatrix}.$$

Use the last three columns of A as a basis of $\mathcal{C}(A)$ and we have that $\mathbf{x} = (0, 0, 80, 130, 100)$ is the initial basic feasible solution. The last row indicates that if x_1 were to become basic by trading it for a current basic variable, the increment in P would be $30x_1$, while using x_2 would give an increment of $40x_2$. We choose the greatest increase in P, so x_2 will become basic. To decide which entry in the second column should be a pivot, divide each entry in the last column (other than the last row) by its corresponding (nonzero) entry in the second column to obtain the numbers $80/1 = 80$, $130/3 = 46\frac{1}{3}$ and $100/2 = 50$. These are the numbers that would result from making that entry a pivot. The second row has the smallest positive entry, so using the $(2, 2)$th entry as pivot will not cause any basic variable to become negative. Now perform admissible elementary row operations to obtain

$$\begin{bmatrix} 2 & 1 & 1\,0\,0 & 80 \\ 1 & \textcircled{3} & 0\,1\,0 & 130 \\ 2 & 2 & 0\,0\,1 & 100 \\ -30 & -40 & 0\,0\,0 & 0 \end{bmatrix} \xrightarrow[\begin{array}{c} E_{12}(-1) \\ E_{32}(-2) \\ E_{24}(40) \end{array}]{E_2\left(\frac{1}{3}\right)} \begin{bmatrix} \frac{5}{3} & 0\,1 & -\frac{1}{3} & 0 & \frac{110}{3} \\ \frac{1}{3} & 1\,0 & \frac{1}{3} & 0 & \frac{130}{3} \\ \frac{4}{3} & 0\,0 & -\frac{2}{3} & 1 & \frac{40}{3} \\ -\frac{50}{3} & 0\,0 & \frac{40}{3} & 0 & \frac{5200}{3} \end{bmatrix}.$$

We see from this matrix that in terms of the current values of the basic variable, the last row says that $P - \frac{50}{3}x_1 + \frac{40}{3}x_4 = \frac{5200}{3}$ with the value contributed by the current values of the basic variables in the lower right corner. So improvement in the objective function could be made by making x_1 basic. Again, to find which entry in the first column to use as a pivot, divide each entry in the last column (other than the last row) by its corresponding (nonzero) entry in column one to obtain the numbers $110/5 = 22$, 130 and $40/4 = 10$. Using the smallest positive ratio selects the $(3, 1)$th entry as pivot and we obtain

$$\begin{bmatrix} \frac{5}{3} & 0\,1 & -\frac{1}{3} & 0 & \frac{110}{3} \\ \frac{1}{3} & 1\,0 & \frac{1}{3} & 0 & \frac{130}{3} \\ \textcircled{\frac{4}{3}} & 0\,0 & -\frac{2}{3} & 1 & \frac{40}{3} \\ -\frac{50}{3} & 0\,0 & \frac{40}{3} & 0 & \frac{5200}{3} \end{bmatrix} \xrightarrow[\begin{array}{c} E_{34}\left(\frac{50}{3}\right) \\ E_{23}\left(-\frac{1}{3}\right) \\ E_{13}\left(-\frac{5}{3}\right) \end{array}]{E_3\left(\frac{3}{4}\right)} \begin{bmatrix} 0\,0\,1 & \frac{1}{2} & -\frac{5}{4} & 20 \\ 0\,1\,0 & \frac{1}{2} & -\frac{1}{4} & 40 \\ 1\,0\,0 & -\frac{1}{2} & \frac{3}{4} & 10 \\ 0\,0\,0 & 5 & \frac{25}{2} & 1900 \end{bmatrix}. \tag{3.14}$$

From this matrix we can read off the solution to our problem: Basic feasible solution $\mathbf{x} = (10, 40, 20, 0, 0)$ yields an optimal value of P. Read the last row of this augmented matrix as

$$P - 0 \cdot x_1 - 0 \cdot x_2 - 0 \cdot x_3 + 5 \cdot x_4 + \frac{25}{2} x_5 = 1900.$$

Since $x_4 = x_5 = 0$, we obtain $P = 1900$. Of course we are mainly interested in the values of x_1 and x_2 since they are the only variables that contribute to the equivalent original form of the objective function, namely

$$P = 30x_1 + 40x_2 + 0 \cdot x_3 + 0 \cdot x_4 + 0 \cdot x_5 = 30 \cdot 10 + 40 \cdot 40 = 1900.$$

□

The simplex solution to Example 3.47 agrees with the geometrical solution that we found in Example 3.49. Notice that at each step we were able to increase the current value of the objective function by a positive amount. If one of the basic variables had a value of zero then a new basic variable replacing it would also have zero value and we might cycle through the basic variables without increasing the objective function and worse, return to a basic variable we had already traded out. This situation mandates a name: A basic feasible solution for which one or more of the basic variables has a value of zero is called a *degenerate* basic feasible solution, otherwise it is called a *nondegenerate* basic feasible solution. More generally, we have the following definition:

Nondegenerate Solution

Definition 3.19. **Nondegenerate Problem** A linear programming problem in standard form $B\mathbf{x} = \mathbf{d}$ with $m \times p$ system matrix B is *nondegenerate* if any subset of m columns of the augmented matrix $\left[\, B \;\; \mathbf{d} \,\right]$ is linearly independent.

We leave it as an exercise to prove that if a linear program in standard form is nondegenerate, then every basic feasible solution to it is nondegenerate. Thus, cycling cannot occur. It should be noted that degeneracy is relatively rare, but the simplex methods (with suitable modifications) and results of this section can be applied without the assumption of nondegeneracy. There is another roadblock to the pivot selection procedure for a maximization problem outlined in the previous example:

Caution: If there are no positive entries in the column corresponding to the variable that we wish to trade into the current basic feasible solution due to negative last row entry in that column, then P can be made arbitrarily large and there is no solution to the original linear program.

For let's suppose that we have re-indexed the variables so that the basic variables are last, the column of the variable we wish to make basic is first and that we have converted to columns corresponding to the basic variables to unit columns. Then with a suitable reordering of the basic variables we can express the current augmented system matrix in the form $\begin{bmatrix} A & I & \mathbf{d} \\ -\mathbf{c}^T & \mathbf{0}^T & u \end{bmatrix}$. Here $\mathbf{d} \geq \mathbf{0}$ since the current solution is basic feasible and $c_1 > 0$ since we expect it to improve the objective function by making x_1 positive. Also $A = [a_{i,j}] = [\mathbf{a}_1, \mathbf{a}_2 \ldots, \mathbf{a}_n]$ with $\mathbf{a}_1 \leq \mathbf{0}$ and $I = [\mathbf{e}_1, \mathbf{e}_2 \ldots, \mathbf{e}_m]$, so the system of constraint equations is

$$x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n + x_{n+1}\mathbf{e}_1 + \cdots + x_{n+m}\mathbf{e}_m = \mathbf{d}$$

with objective function as

$$P = c_1x_1 + \cdots + c_nx_n + 0 \cdot x_{n+1} + \cdots 0 \cdot x_{n+m} + u.$$

Keep $x_j = 0$, $2 \le j \le n$ but allow for x_1 to be possibly positive and the system becomes

$$(x_{n+1} + x_1a_{1,1})\,\mathbf{e}_1 + \cdots + (x_{n+m} + x_1a_{m,1})\,\mathbf{e}_m = \mathbf{d}.$$

Thus, x_1 can be made arbitrarily large as long as the equations $x_{n+j} + x_1a_{j,1} = d_j$, $j = 1, \ldots, m$. are maintained. The new values of $x_{n+j} = d_j - x_1a_{j,1}$ are always nonnegative since $a_{j,1} \le 0$, so $\left(x_1, 0, \ldots, 0, x_{n+1} + x_1a_{1,1}, \ldots, x_{n+m} + x_1a_{m,1}\right)$ is a feasible solution for which the value of the objective function is $P = c_1x_1 + u$. Hence, P can be made arbitrarily large in the feasible set.

Next we consider how the simplex method works on the min linear program of Example 3.48.

Example 3.55. Solve the min linear program of Example 3.48 using the simplex method.

Solution. First recast the optimization problem of minimizing $C = 80x_1 + 130x_2 + 100x_3$ to the equivalent problem of maximizing $P = -C = -80x_1 - 130x_2 - 100x_3$. Next, follow the lead of Example 3.53 in introducing two additional artificial variables to account for the absence of an obvious choice for initial basic variables. We use the outcome of this calculation, which was to identify first admissible pivot in the $(1,3)$th position and second admissible pivot in the $(2,1)$th position. So append the row corresponding to the value of the objective function and complete the appropriate row operations:

$$\begin{bmatrix} 2 & 1 & \textcircled{2} & -1 & 0 & 30 \\ 1 & 3 & 2 & 0 & -1 & 40 \\ 80 & 130 & 100 & 0 & 0 & 0 \end{bmatrix} \xrightarrow[E_{21}(-2)\\ E_{31}(-100)]{E_1\left(\frac{1}{2}\right)} \begin{bmatrix} 1 & \frac{1}{2} & 1 & -\frac{1}{2} & 0 & 15 \\ -1 & 2 & 0 & 1 & -1 & 10 \\ -20 & 80 & 0 & 50 & 0 & -1500 \end{bmatrix}$$

$$\begin{bmatrix} 1 & \frac{1}{2} & 1 & -\frac{1}{2} & 0 & 15 \\ -1 & \textcircled{2} & 0 & 1 & -1 & 10 \\ -20 & 80 & 0 & 50 & 0 & -1500 \end{bmatrix} \xrightarrow[E_{12}\left(-\frac{1}{2}\right)\\ E_{32}(-80)]{E_2\left(\frac{1}{2}\right)} \begin{bmatrix} \frac{5}{4} & 0 & 1 & -\frac{3}{4} & \frac{1}{4} & \frac{25}{2} \\ -\frac{1}{2} & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 5 \\ 20 & 0 & 0 & 10 & 40 & -1900 \end{bmatrix}. \quad (3.15)$$

Thus, we have found a minimum of $C = -P = 1900$ at the basic feasible solution $\left(0, 5, \frac{25}{2}, 0, 0\right)$ which yields the solution $\mathbf{x} = \left(0, 5, \frac{25}{2}\right)$ to the original problem. This is the same solution found by the geometrical method of Example 3.50. □

If we look carefully at the last row of the final matrix in (3.15) we see something very curious, namely that the solution to the max linear program of Example 3.48 is sitting in that row under the surplus variable columns: $\mathbf{x} = (10, 40)$. Similarly, the last row of the final matrix in (3.14) contains the

solution $\mathbf{x} = \left(0, 5, \frac{25}{2}\right)$ to the min linear program of Example 3.48 under the surplus variable columns. What is the connection? These two problems have a very close connection which is codified in the following definition.

Definition 3.20. Primal and Dual Problems The dual problem for the primal problem of maximizing $P = \mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$ is the problem of minimizing $P = \mathbf{b}^T\mathbf{y}$ subject to the constraints $A^T\mathbf{y} \geq \mathbf{c}$ and $\mathbf{y} \geq \mathbf{0}$.

There is a symmetry here: If the dual problem is converted to a max linear program, then we leave it as an exercise to show that the dual of the dual is the primal. Another basic theorem of linear programming is the following connection between a primal linear program and its dual.

Theorem 3.26. Equivalence of Primal and Dual Suppose that the primal problem of maximizing $P = \mathbf{c}^T\mathbf{x}$ subject to the constraints $A\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$ and its dual are nondegenerate. If the primal has a feasible solution $\mathbf{x}$ and the dual has a feasible solution $\mathbf{y}$, then $\mathbf{c}^T\mathbf{x} \leq \mathbf{b}^T\mathbf{y}$, both have optimal solutions $\mathbf{x}_*$ and $\mathbf{y}_*$ and $\mathbf{c}^T\mathbf{x}_* = \mathbf{b}^T\mathbf{y}_*$.

Proof. Suppose first that the primal has feasible solution $\mathbf{x}$ and the dual has feasible solution $\mathbf{y}$. We leave it as an exercise to show that multiplying by nonnegative vectors preserves vector inequalities. Since $\mathbf{y} \geq \mathbf{0}$, we have that $\mathbf{y}^T A\mathbf{x} \leq \mathbf{y}^T\mathbf{b}$ and likewise, since $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{y}^T A \geq \mathbf{c}^T$, we have that $\mathbf{y}^T A\mathbf{x} \geq \mathbf{c}^T\mathbf{x}$. Put these together and we conclude that $\mathbf{c}^T\mathbf{x} \leq \mathbf{b}^T\mathbf{y}$. This implies that there is an upper bound to values of the primal problem and a lower bound to values of its dual.

Next, notice that there are only a finite number of basic feasible solutions to this problem: Nondegeneracy implies that any m columns of the augmented system matrix $\left[\, A_e \;\; \mathbf{b} \,\right]$ are linearly independent so a linear combination of them that equals $\mathbf{b}$ has uniquely determined coefficients. It follows that there is only a finite number of positive improvements in the objective function achieved by replacing one basic feasible solution by another. Hence, there is a smallest positive improvement, say α. Since the primal is nondegenerate, we see that replacing any basic feasible solution by another that whose column has a negative last entry and ensuring that the this column is a unit column by admissible elementary operations ensures that the right-hand side consists of values of the basic variables, hence has positive entries. Therefore, any such change improves the solution by a positive amount greater than or equal to α. Since the objective function is bounded above, it follows that after a finite number of simplex steps must terminate with basic feasible solution $\mathbf{x}_*^e$ and there can be no further improvement via basic feasible solutions. Thus, the objective function of the last standard matrix $\tilde{B} = \begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}_e^T & u \end{bmatrix}$ has $\mathbf{c}_e \leq \mathbf{0}$. It follows from Corollary 3.10 that $\mathbf{c}_e^T\mathbf{x}_*^e + u$ is the optimal value of the original objective function and $\mathbf{x}_*^e$ is an optimal feasible solution to the initial problem

with standard matrix $\begin{bmatrix} A & I & \mathbf{b} \\ -\mathbf{c}^T & \mathbf{0}^T & 0 \end{bmatrix}$. Thus, the optimal value $\mathbf{c}^T\mathbf{x}_*^e + 0$ for the primal problem is achieved by the feasible solution $\mathbf{x}_* \in \mathbb{R}^m$ defined as the first m coordinates of $\mathbf{x}_*^e$.

Finally, let $\begin{bmatrix} E & \mathbf{0} \\ y_*^T & 1 \end{bmatrix}$ be the composition of all elementary operations used to reach the final matrix and we see that

$$\begin{bmatrix} E & \mathbf{0} \\ \mathbf{y}_*^T & 1 \end{bmatrix} \begin{bmatrix} A & I & \mathbf{b} \\ -\mathbf{c}^T & \mathbf{0}^T & 0 \end{bmatrix} = \begin{bmatrix} EA & EI & E\mathbf{b} \\ \mathbf{y}_*^T A - \mathbf{c}^T & \mathbf{y}_*^T & \mathbf{y}_*^T\mathbf{b} \end{bmatrix}$$

where $\mathbf{y}_*^T A - \mathbf{c}^T \geq \mathbf{0}$, $\mathbf{y}_*^T \geq \mathbf{0}$ and $u = \mathbf{y}_*^T\mathbf{b}$ is the maximum value of the objective function $\mathbf{c}^T\mathbf{x}$. It follows from definition of the dual that $\mathbf{y}_*$ is a feasible solution to the dual. From the first paragraph above we see that if $\mathbf{y}$ is any other feasible solution to the dual, then

$$\mathbf{y}_*^T\mathbf{b} = u = \mathbf{c}^T\mathbf{x}^* \leq \mathbf{y}^T\mathbf{b} = \mathbf{b}^T\mathbf{y}.$$

Hence, u must also be the minimum value of the objective function $\mathbf{b}^T\mathbf{y}$ for the dual problem and this minimum is achieved by the vector $\mathbf{y}_*$. □

The last paragraph of the preceding proof shows why the solution to Example 3.48 appeared in the final form of the augmented matrix for Example 3.47. In our last example we put the simplex method to work on a slightly nonstandard max linear program.

Example 3.56. Use the simplex method to solve the following linear program:

Maximize $P = x_1 + 4x_2$ subject to constraints:

$$\begin{aligned} x_1 + 3x_2 &\leq 45 \\ 2x_1 + x_2 &\leq 40 \\ x_1 + x_2 &\geq 18 \\ x_j &\geq 0,\ j = 1, 2. \end{aligned}$$

Solution. First we follow the lead of Example 3.53: Correct the lack of a basic feasible variable in third equation by appending an artificial variable x_6 to obtain the system

$$\begin{aligned} x_1 + 3x_2 + x_3 &= 45 \\ 2x_1 + x_2 + x_4 &= 40 \\ x_1 + x_2 - x_5 + x_6 &= 18 \\ x_j &\geq 0,\ j = 1, 2, \ldots, 6. \end{aligned}$$

Next construct the augmented standard matrix and use admissible elementary operations to bring the simplex method to conclusion on this problem. We leave the calculations as an exercise. The result from beginning to end is as follows:

$$\left[\begin{array}{cccccc|c} 1 & \textcircled{3} & 1 & 0 & 0 & 0 & 45 \\ 2 & 1 & 0 & 1 & 0 & 0 & 40 \\ 1 & 1 & 0 & 0 & -1 & 1 & 18 \\ -1 & -4 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \xrightarrow[\cdots]{} \left[\begin{array}{cccccc|c} \frac{1}{3} & 1 & \frac{1}{3} & 0 & 0 & 0 & 15 \\ \frac{5}{3} & 0 & -\frac{1}{3} & 1 & 0 & 0 & 25 \\ \frac{2}{3} & 0 & -\frac{1}{3} & 0 & -1 & 1 & 3 \\ \frac{1}{3} & 0 & \frac{4}{3} & 0 & 0 & 0 & 60 \end{array}\right]. \tag{3.16}$$

So the optimal solution appears to be $x_2 = 15$, $x_4 = 25$, $x_6 = 3$ and $x_1 = x_3 = x_5 = 0$. But there's a problem: This is not a solution to the original problem since it yields that $x_1 + x_2 = 15 < 18$. So what went wrong? Is Theorem 3.25 contradicted? The answer lies in understanding the difference between slack and surplus versus artificial variables. The former yield equations entirely equivalent to the original inequalities. The latter does not. In other words, the system equalities that we created are not equivalent to the original problem.

One solution is to follow the lead of Example 3.53 and make the variable x_6 nonbasic by substituting it with another variable. In more complex cases this may not be a good strategy. We shall work around this difficulty with a more general procedure as follows: First try to minimize the artificial variable x_6 to a value of zero, using only admissible row operations that remain admissible for the original problem. Then delete its column from the problem. To this end, we want to minimize x_6 to a value of zero, i.e., maximize $-x_6$ to zero. So use this as a temporary objective function and add the additional row to our augmented system matrix. The first step is to make the column for x_6 equal to an identity column. Then we follow the usual simplex procedure for this objective function while also updating the original objective function, details of which are left as an exercise. The result is

$$\left[\begin{array}{cccccc|c} 1 & 3 & 1 & 0 & 0 & 0 & 45 \\ 2 & 1 & 0 & 1 & 0 & 0 & 40 \\ 1 & 1 & 0 & 0 & -1 & \textcircled{1} & 18 \\ -1 & -4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \xrightarrow[\cdots]{E_{53}(-1)} \left[\begin{array}{cccccc|c} 0 & 1 & \frac{1}{2} & 0 & \frac{1}{2} & -\frac{1}{2} & \frac{27}{2} \\ 0 & 0 & \frac{1}{2} & 1 & \frac{5}{2} & -\frac{5}{2} & \frac{35}{2} \\ 1 & 0 & -\frac{1}{2} & 0 & -\frac{3}{2} & \frac{3}{2} & \frac{9}{2} \\ 0 & 0 & \frac{3}{2} & 0 & \frac{1}{2} & -\frac{1}{2} & \frac{117}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right]. \tag{3.17}$$

Variable x_6 is no longer basic, so delete it and the temporary objective function of the last row to obtain an augmented system matrix equivalent to the original problem, namely,

$$\left[\begin{array}{cccccc|c} 0 & 1 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{27}{2} \\ 0 & 0 & \frac{1}{2} & 1 & \frac{5}{2} & \frac{35}{2} \\ 1 & 0 & -\frac{1}{2} & 0 & -\frac{3}{2} & \frac{9}{2} \\ 0 & 0 & \frac{3}{2} & 0 & \frac{1}{2} & \frac{117}{2} \end{array}\right].$$

As it turns out, there is no additional work to be done, so we read off the solution to the problem: $x_1 = \frac{9}{2}$, $x_2 = \frac{27}{2}$, $x_4 = \frac{35}{2}$ and $x_3 = x_5 = 0$ with an optimal value of $P = 0 \cdot x_3 + \frac{117}{2} = \frac{117}{2}$. We leave it as an exercise to verify that this is the same as the geometrical solution $x_1 = \frac{9}{2}$, $x_2 = \frac{27}{2}$, $P = x_1 + 4x_2 = \frac{9}{2} + 4 \cdot \frac{27}{2} = \frac{117}{2}$ of this problem. □

We have only scratched the surface of the extensive topic of linear programming and there are many excellent textbooks, e.g., Hillier and Lieberman [16],

Bertsimas and Tsitsiklis [4] and others, that cover all aspects of this subject in great detail.

3.8 Exercises and Problems

Exercise 1. Express the following problem in standard form: Minimize $C = x_1 + 2x_1 + x_3$ subject to the constraints $x_1 + x_2 \geq 4$, $x_1 + x_3 \leq 6$, $x_1 \geq 1$, $x_2, x_3 \geq 0$.

Exercise 2. Express the following problem in standard form: Maximize $P = x_1 - x_2 + x_3$ subject to the constraints $x_1 - 2x_2 + x_3 \leq 3$, $x_1 + x_3 \leq 4$, $x_1, x_2, x_3 \geq 0$.

Exercise 3. Solve the following problem using both geometric and simplex methods: Maximize $P = x_1 + 2x_2$ subject to constraints $x_1 + x_2 \leq 6$, $-x_1 + x_2 \leq 2$, $x_1 \leq 5$, $x_1, x_2 \geq 0$.

Exercise 4. Solve the following problem using the simplex method: Maximize $P = 4x_1 - x_2 + 3x_3$ subject to constraints $x_1 + x_2 - x_3 \leq 2$, $x_1 + x_2 + x_3 \leq 6$, $x_1, x_2, x_3 \geq 0$.

Exercise 5. Solve the problem of Example 3.56 by the geometric method.

Exercise 6. Solve the problem of Exercise 4 by the geometric method.

Exercise 7. Use the simplex method to show that the problem of maximizing $P = 3x_1 + x_2$ subject to the constraints $-x_1 + x_2 \leq 4$, $-x_1 + 2x_2 \leq 10$, $x_1, x_2 \geq 0$, has unbounded objective function and show that the dual problem has no feasible solution.

Exercise 8. Use the simplex method to show that the problem of maximizing $P = x_1 + x_2 + x_3$ subject to the constraints $-x_1 + x_2 \leq 20$, $-x_1 - x_2 + 2x_3 \leq 10$, $x_1, x_2, x_3 \geq 0$, has unbounded objective function and show that the dual problem has no feasible solution.

Exercise 9. A company produces four different tool kits, K_1, K_2, K_3, K_4, parts of which are created at site S_1, then finished and assembled into kits at site S_2. The hours required for these kits at each site, as well as the profit per kit and maximum currently available hours at each site are detailed in this table:

		K_1	K_2	K_3	K_4	Available Hours
Sites	S_1	2	1	3	2	900
	S_2	3	2	2	2	1200
	Profit per kit	35	20	40	25	

Express the problem of maximizing profit subject to these constraints as a linear programming problem in standard form and solve it.

Exercise 10. The firm Grain Associates manages two granaries G_1 and G_2 which currently have available 9 and 15 tons of grain, resp. The firm has contracted with three flour mills M_1, M_2 and M_3 to supply $8, 12$ and 4 tons of grain, resp., so that supply balances demand. Total costs of transport in hundreds of dollars per ton are given in this table.

		Destination		
		M_1	M_2	M_3
Source	G_1	2	1	3
	G_2	3	2	2

Express the problem of minimizing transport costs subject to these constraints as a linear programming problem in standard form and solve it. (Note: The standard form is not full row rank, so one equation can be eliminated.)

Exercise 11. Use the simplex method to solve the problem of minimizing $C = 6x_1 + x_2 + 4x_3$ subject to the constraints $x_1 + x_2 + 2x_3 \leq 40$, $x_1 + x_2 \geq 10$, $x_1, x_2, x_3 \geq 0$.

Exercise 12. Use graphical and simplex method to solve the problem of minimizing $C = 10x_1 + 4x_2$ subject to the constraints $x_1 + 2x_2 \leq 8$, $x_1 + x_2 \leq 13$, $2x_1 - x_2 \geq 11$, $2x_1 + x_2 \geq 13$, $x_1, x_2 \geq 0$. (Hint: Try minimizing the sum of the two artificial variables.)

Exercise 13. Use graphical method, then duality and the simplex method to solve the problem of minimizing $C = 3x_1 + 2x_2$ subject to constraints $3x_1 + x_2 \geq 6$, $x_1 + 2x_2 \geq 6$, $x_1, x_2 \geq 0$.

Exercise 14. Use graphical method on the dual, then duality and the simplex method to solve the problem of minimizing $C = 2x_1 + 3x_2 + 6x_3 + 5x_4$ subject to constraints $-x_1 + x_3 + x_4 \geq 1$, $x_1 + x_2 + x_3 \geq 2$ and $x_j \geq 0$, $j = 1, 2, 3, 4$.

Problem 15. Show that if $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ with $\mathbf{u} \leq \mathbf{v}$ and $\mathbf{w} \geq \mathbf{0}$, then $\mathbf{w}^T\mathbf{u} \leq \mathbf{w}^T\mathbf{v}$.

*Problem 16. Show that if B is $m \times p$, $\mathbf{d} \in \mathbb{R}^m$, $\tilde{B} = \begin{bmatrix} B & \mathbf{d} \\ -\mathbf{c}^T & u \end{bmatrix}$ and a sequence of arbitrary elementary operations on the first m rows of $\tilde{B}$ is applied, along with elementary operations of adding multiples of the first m rows to the last row, then the sequence of operations on $\tilde{B}$ is accomplished by a matrix of the form $\begin{bmatrix} E & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix}$, where E is an $m \times m$ product of elementary row operations and $\mathbf{v} \in \mathbb{R}^p$.

Problem 17. Given the problem of optimizing $\mathbf{c}^T\mathbf{x}$ subject to the constraints $B\mathbf{x} = \mathbf{d}$ and $\mathbf{x} \geq \mathbf{0}$, and feasible solutions $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{c}^T\mathbf{x} < \mathbf{c}^T\mathbf{y}$, show that for any convex combination of the form $\mathbf{z} = (1 - \alpha)\mathbf{x} + \alpha\mathbf{y}$, $0 \leq \alpha \leq 1$, $\mathbf{z}$ is feasible and $\mathbf{c}^T\mathbf{x} \leq \mathbf{c}^T\mathbf{z} \leq \mathbf{c}^T\mathbf{y}$ with strict inequality if $0 < \alpha < 1$.

*Problem 18. Show that if the problem of optimizing $\mathbf{c}^T\mathbf{x}$ subject to the constraints $B\mathbf{x} = \mathbf{d}$ and $\mathbf{x} \geq \mathbf{0}$ has two optimal feasible solutions then it has infinitely many such solutions.

Problem 19. Show that if a linear program in standard form is nondegenerate, then every basic feasible solution to it is nondegenerate.

Problem 20. Convert the dual of the primal problem of maximizing $\mathbf{c}^T\mathbf{x}$ subject to $\mathbf{x} \geq \mathbf{0}$ and $A\mathbf{x} \leq \mathbf{b}$ to a max linear program and use this to show that the dual of the dual is the primal problem.

Problem 21. Use Theorem 3.26 to show that if a primal problem has unbounded objective function then the dual has no feasible solution.

3.9 *Applications and Computational Notes

Spaces Associated with a Directed Graph

There are significant practical applications of vector space theory to modeling with digraphs. In addition to the adjacency matrix of Section 2.3, lurking in the background is another matrix that describes all the data necessary to construct a digraph or graph.

Definition 3.21. The *incidence matrix* of a graph or digraph has rows indexed by its vertices and columns by its edges in some specific order. If the edge (i, j) is in the digraph, then the column corresponding to this edge has -1 in its ith row and $+1$ in its jth row. In the case of a graph, if the edge $\{i, j\}$ is in the graph, then column corresponding to this edge has $+1$ in its ith and jth rows. All other entries are 0.

For example, the incidence matrices A of the graph of Figure 2.4 and B of the digraph of Figure 2.5 are given by

$$A = \begin{bmatrix} 1&1&1&0&0&0&0&0 \\ 1&0&0&1&1&0&0&0 \\ 0&0&0&1&0&1&1&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&1&1&1 \\ 0&0&1&0&1&0&0&1 \end{bmatrix} \text{ and } B = \begin{bmatrix} -1&-1&1&0&0 \\ 1&0&0&-1&1 \\ 0&0&-1&1&0 \\ 0&1&0&0&-1 \end{bmatrix}.$$

Definition 3.22. **Loops** A *loop* in a digraph is a directed walk that starts and ends at the same node.

For example, the sequence $(1,2),(2,3),(3,1)$ is a loop L in the digraph of Figure 2.5. So is $(1,2),(2,3),(3,1),(1,2),(2,3),(3,1)$, which could reasonably the thought of as $2L$.

Each column of the incidence matrix defines an edge. Thus, linear combinations of these columns with integer coefficients, say $\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_8\mathbf{v}_8 = A\mathbf{c}$, represent a listing of edges, possibly with repeats. When will such a combination represent a directed loop? Consider such a linear combination with defining vector of coefficients $\mathbf{c} = (c_1, \ldots, c_8)$. Here's the key idea: Examine this combination locally, that is, at each vertex. There we expect the total number of "in-arrows" (-1's) to be exactly canceled by the total number of "out-arrows" ($+1$'s). In other words, each coordinate of $\mathbf{v}$ should be 0 and so $\mathbf{c} \in \mathcal{N}(A)$. Thus, the description of all possible loops amounts to the description of a subspace of $\mathbb{R}^8$. Of course, one has to distinguish between "algebraic" loops (any c_i's) and directed loops (all $c_i \geq 0$).

3.10 *Projects and Reports

Projects: Modeling with Directed Graphs

Instructors: Formulate projects by selecting one or more from the following items. These items are listed roughly in order of difficulty.

Project Descriptions: These projects introduce more applications of digraphs as mathematical modeling tools. You are given that the digraph G has vertex set $V = \{1,2,3,4,5,6\}$ and edge set

$$E = \{(1,2),(2,3),(3,4),(4,2),(1,4),(3,1),(3,6),(6,3),(4,5),(5,6)\}\,.$$

Address the following points regarding G.

1. (a) Draw a picture of this digraph. You may leave space in your report and draw this by hand, or if you prefer, you may use the computer drawing applications available to you on your system.

(b) Exhibit the incidence matrix A of this digraph and find a basis for $\mathcal{N}(A)$ using its reduced row echelon form. Some of the basis elements may be algebraic but not directed loops. Use this basis to find a basis of directed loops (e.g., non-directed basis element $\mathbf{c}_1$ might be replaced by directed $\mathbf{c}_1 + \mathbf{c}_2$).

2. Think of the digraph as representing an electrical circuit where an edge represents some electrical object like a resistor or capacitor. Each node represents the circuit space between these objects. and we can attach a potential value to each node, say the potentials are $x_1, \ldots, x_6$. The potential difference across an edge is the potential value of head minus tail. Kirchhoff's second law of electrical circuits says that the sum of potential differences around a circuit loop must be zero. Assume and use the fact (p. 422) that $A\mathbf{x} = \mathbf{b}$ implies that for all $\mathbf{y} \in \mathcal{N}(A^T)$, $\mathbf{y}^T\mathbf{b} = 0$ to find conditions that a vector $\mathbf{b}$ must satisfy in order for it to be a vector of potential differences for some potential distribution on the vertices.

3. Assume that across each edge of a circuit a current flows. Thus, we can assign to each edge a "weight," namely the current flow along the edge. This is an example of a *weighted* digraph. However, not just any set of current weights will do, since Kirchhoff's first law of circuits says that the total flow of current in and out of any node should be 0. Use this law to find a matrix condition that must be satisfied by the currents and solve it to exhibit some current flows.

4. Think of the digraph as representing a directed communications network. Here loops determine which nodes have bidirectional communication since any two nodes of a loop can only communicate with each other by way of a loop. By examining only a basis of directed loops how could you determine which nodes in the network can communicate with each other?

5. Think of vertices of the digraph as representing airports and edges representing flight connections between airports for Gamma Airlines. Suppose further that for each connection there is a maximum number of daily flights that will be allowed by the destination airport from an origin airport and that, in the order that the edges in E are listed above, these limits are

$$M = \{4, 3, 8, 7, 2, 6, 7, 10, 5, 8\}\,.$$

Now suppose that Gamma wants to maximize the flow of flights into airport 1 and out of airport 6. Count inflows into an airport as positive and outflows as negative. Assume that the net in/outflow of Gamma flights at each airport 1 to 5 is zero, while the net inflow of such flights into airport 1 matches the net outflow from 6.

(a) Describe the problem of maximizing this inflow to airport 1 as a linear programming problem and express it in a standard form (block matrices are helpful.) Note that the appropriate variables are all outflows from one airport to another, i.e., along edges, together with the net inflow into airport 1.

(b) Solve the problem of part (a). Also solve the reverse problem: Maximize inflow into airport 6 and matching outflow from 1. Explain and justify your answers.

6. With the same limits on allowable flights into airports as in item 5, suppose that Gamma Airlines wants to determine an allocation of planes that will maximize their profits, given the following constraints: (1) Airports 1 and 6 have repair facilities for their planes, so no limit is placed on the inflow or outflow of their planes other than the airport limits. (2) Flights through airports 2-5 of Gamma planes are pass through, i.e., inflow and outflow must match. (3) Gamma has 32 planes available for this network of airports. (4) The profits per flight in thousands are, in the order that the edges in E are listed above,

$$P = \{5, 6, 7, 9, 10, 8, 9, 5, 6, 10\}\,.$$

(a) Set this problem up as a linear programming problem in standard form. Clearly identify the variables and explain how the constraints follow.

(b) Solve this problem explicitly and specify the operations taken to do so. Example 3.56 is instructive for this problem, so be aware of it. Use a technology tool that allows you to use elementary operations (ALAMA calculator has this capability).

4

GEOMETRICAL ASPECTS OF STANDARD SPACES

The standard vector spaces have many important extra features that we have largely ignored up to this point. These extra features made it possible to do sophisticated calculations in the spaces and enhance our insight into vector spaces by appealing to geometry. For example, in the geometrical spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ that were studied in algebra and calculus, it was possible to compute the length of a vector and angles between vectors. These are visual concepts that feel very comfortable to us. In this chapter we generalize these ideas to standard vector spaces and their subspaces. We will abstract these ideas to general vector spaces in Chapter 6.

4.1 Standard Norm and Inner Product

The Norm Idea

Consider this problem. How do we formulate precisely the idea of a sequence of vectors $\mathbf{u}_i$ converging to a limit vector $\mathbf{u}$, i.e.,

$$\lim_{n\to\infty} \mathbf{u}_n = \mathbf{u},$$

in standard spaces? A reasonable answer is to mean that the distance between the vectors should tend to 0 as $n \to \infty$. By *distance* we mean the length of the difference. So what we need is some idea about the *length,* i.e., *norm*, of a vector. We have seen such an idea in the geometrical spaces $\mathbb{R}^2$ and $\mathbb{R}^3$. There are different ways to measure length. We shall begin with the most standard method. It is one of the outcomes of geometry and the Pythagorean theorem. As with standard spaces, there is no compelling reason to stop at geometrical dimensions of two or three, so here is the general definition.

Definition 4.1. Standard Real Vector Norm Let $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n$. The *(standard) norm* of $\mathbf{u}$ is the nonnegative real number

$$\|\mathbf{u}\| = \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2}.$$

Example 4.1. Compute the norms of the vectors $\mathbf{u} = (1, -1, 3)$ and $\mathbf{v} = [2, -1, 0, 4, 2]^T$.

Solution. From the definition,

$$\|\mathbf{u}\| = \sqrt{1^2 + (-1)^2 + 3^2} = \sqrt{11} \approx 3.3166,$$

and

$$\|\mathbf{v}\| = \sqrt{2^2 + (-1)^2 + 0^2 + 4^2 + 2^2} = \sqrt{25} = 5. \qquad \square$$

Even though we can't really "see" the five-dimensional vector $\mathbf{v}$ of this example, it is interesting to note that calculating its length is just as routine as calculating the length of the three-dimensional vector $\mathbf{u}$. What about complex vectors? Shouldn't there be an analogous definition for such objects? The answer is yes, but we have to be a little careful. We can't use the same definition that we did for real vectors. Consider the vector $\mathbf{x} = (1, 1 + \mathrm{i})$. The sum of the squares of the coordinates is just

$$1^2 + (1 + \mathrm{i})^2 = 1 + 1 + 2\mathrm{i} - 1 = 1 + 2\mathrm{i}.$$

This isn't good. We don't want "length" to be measured in complex numbers. The fix is very simple. We already have a way of measuring the length of a complex number z, namely the absolute value $|z|$, so length squared is $|z|^2$. That is the inspiration for the following definition, which is entirely consistent with our first definition when applied to real vectors:

Definition 4.2. Standard Complex Vector Norm Let $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \mathbb{C}^n$. The *(standard) norm* of $\mathbf{u}$ is the nonnegative real number

$$\|\mathbf{u}\| = \sqrt{|u_1|^2 + |u_2|^2 + \cdots + |uz|^2}.$$

Notice that $|z|^2 = \overline{z}z$. (Remember that if $z = a + b\mathrm{i}$, then $\overline{z} = a - b\mathrm{i}$ and $\overline{z}z = a^2 + b^2 = |z|^2$.) Therefore,

$$\|\mathbf{u}\| = \sqrt{\overline{u}_1 u_1 + \overline{u}_2 u_2 + \cdots + \overline{u}_n u_n}.$$

Example 4.2. Compute the norms of the vectors $\mathbf{u} = (1, 1 + \mathrm{i})$ and $\mathbf{v} = (2, -1, \mathrm{i}, 3 - 2\mathrm{i})$.

Solution. From the definition,

$$\|\mathbf{u}\| = \sqrt{1^2 + (1-\mathrm{i})\,(1+\mathrm{i})} = \sqrt{1+1+1} \approx 1.7321$$

and

$$\begin{aligned}\|\mathbf{v}\| &= \sqrt{2^2 + (-1)^2 + (-\mathrm{i})\mathrm{i} + (3+2\mathrm{i})(3-2\mathrm{i})} \\ &= \sqrt{4+1+1+9+4} = \sqrt{19} \approx 4.3589.\end{aligned}$$

□

Here are the essential properties of the norm concept:

Basic Norm Laws

Let c be a scalar and $\mathbf{u}, \mathbf{v} \in V$ where the vector space V has the standard norm $\|\ \|$. Then the following hold.

(1) $\|\,\mathbf{u}\,\| \geq 0$ with $\|\,\mathbf{u}\,\| = 0$ if and only if $\mathbf{u} = \mathbf{0}$.
(2) $\|\,c\mathbf{u}\,\| = |\,c\,|\ \|\,\mathbf{u}\,\|$.
(3) $\|\,\mathbf{u} + \mathbf{v}\,\| \leq \|\,\mathbf{u}\,\| + \|\,\mathbf{v}\,\|$. (Triangle Inequality)

That (1) is true is immediate from the definition of $\|\mathbf{u}\|$ as a sum of the lengths squared of the coordinates of $\mathbf{u}$. This sum is zero exactly when each term is zero. Condition (2) is fairly straightforward too. Suppose $\mathbf{u} = (z_1, z_2, \ldots, z_n)$, so that

$$\begin{aligned}\|c\mathbf{u}\| &= \sqrt{(\overline{cu}_1)cu_1 + (\overline{cu}_2)cu_2 + \cdots + (\overline{cu}_n)cu_n} \\ &= \sqrt{(\overline{c}c)(\overline{u}_1u_1 + \overline{u}_2u_2 + \cdots + \overline{u}_nu_n)} \\ &= \sqrt{|c|^2}\sqrt{\overline{u}_1u_1 + \overline{u}_2u_2 + \cdots + \overline{u}_nu_n} \\ &= |c| \cdot \|\mathbf{u}\|.\end{aligned}$$

The triangle inequality (which gets its name from the triangle with representatives of the vectors $\mathbf{u}, \mathbf{v}, \mathbf{u}+\mathbf{v}$ as its sides) can be proved easily in two- or three-dimensional geometrical space by appealing to the fact that the sum of lengths of any two sides of a triangle is greater that the length of the third side. A justification for higher dimensions is a nontrivial bit of algebra that we postpone until after the introduction of inner products below.

First we consider a few applications of the norm concept. We say that two vectors *determine the same direction* if one is a positive multiple of the other and *determine opposite directions* ifone is a negative multiple of the other. The first application is the idea of "normalizing" a vector. This means finding a *unit vector*, which means a *vector of length* 1, that has the same direction as the vector. This process is sometimes called "normalization."The following simple fact shows us how to do it.

Unit Vectors

Theorem 4.1. Let $\mathbf{u}$ be a nonzero vector. Then the vector

$$\mathbf{w} = \frac{1}{\|\mathbf{u}\|}\mathbf{u}$$

is a unit vector in the same direction as $\mathbf{u}$.

Proof. Since $\|\mathbf{u}\|$ is positive, we see immediately that $\mathbf{w}$ and $\mathbf{u}$ determine the same direction. Now check the length of $\mathbf{w}$ by using the basic norm law 2 to obtain that

$$\|\mathbf{w}\| = \left\| \frac{1}{\|\mathbf{u}\|}\mathbf{u} \right\| = \left| \frac{1}{\|\mathbf{u}\|} \right| \|\mathbf{u}\| = \frac{\|\mathbf{u}\|}{\|\mathbf{u}\|} = 1.$$

Hence, $\mathbf{w}$ is a unit vector, as desired. □

Example 4.3. Use the normalization procedure to find unit vectors in the directions of vectors $\mathbf{u} = (2, -1, 0, 4)$ and $\mathbf{v} = (-4, 2, 0, -8)$. Conclude that these vectors determine opposite directions.

Solution. Let us find a unit vector in the same direction of each vector. We have norms

$$\|\mathbf{u}\| = \sqrt{2^2 + (-1)^2 + 0^2 + 4^2} = \sqrt{21}$$

and

$$\|\mathbf{v}\| = \sqrt{-4^2 + (2)^2 + +0^2 + (-8)^2} = \sqrt{84} = 2\sqrt{21}.$$

It follows that unit vectors in the directions of $\mathbf{u}$ and $\mathbf{v}$, respectively, are

$$\mathbf{w_1} = (2, -1, 0, 4)/\sqrt{21},$$
$$\mathbf{w_2} = (-4, 2, 0, -8)/(2\sqrt{21}) = -(2, -1, 0, 4)/\sqrt{21} = -\mathbf{w_1}.$$

Therefore, $\mathbf{u}$ and $\mathbf{v}$ determine opposite directions. □

Example 4.4. Find a unit vector in the direction of the vector $\mathbf{v} = (2 + \mathrm{i}, 3)$.

Solution. We have

$$\|\mathbf{u}\| = \sqrt{2^2 + 1^2 + 3^2} = \sqrt{14}.$$

It follows that a unit vector in the direction of $\mathbf{v}$ is

$$\mathbf{w} = \frac{1}{\sqrt{14}}(2 + \mathrm{i}, 3).$$ □

In order to work the next example we must express the idea of vector convergence of a sequence $\mathbf{u}_1, \mathbf{u}_2, \ldots$ to the vector $\mathbf{u}$ in a sensible way. The norm idea makes this straightforward: to say that the $\mathbf{u}_n$'s approach the vector $\mathbf{u}$ should mean that the distance between $\mathbf{u}$ and $\mathbf{u}_n$ goes to 0 as $n \to \infty$. But norm measures distance. Therefore, the correct definition is as follows:

Definition 4.3. Convergence of Vectors Let $\mathbf{u}_1, \mathbf{u}_2, \ldots$ be a sequence of vectors in the vector space V and $\mathbf{u}$ also a vector in V. We say that the sequence *converges* to $\mathbf{u}$ and write

$$\lim_{n\to\infty} \mathbf{u}_n = \mathbf{u}$$

if the sequence of real numbers $\|\mathbf{u}_n - \mathbf{u}\|$ converges to 0, i.e.,

$$\lim_{n\to\infty} \|\mathbf{u}_n - \mathbf{u}\| = 0.$$

Example 4.5. Use the norm concept to justify the statement that

$$\lim_{n\to\infty} \mathbf{u}_n = \mathbf{u},$$

where $\mathbf{u}_n = \left(1 + 1/n^2, 1/(n^2+1), \sin n/n\right)$ and $\mathbf{u} = (1, 0, 0)$.

Solution. In our case we have

$$\mathbf{u}_n - \mathbf{u} = \begin{bmatrix} 1 + 1/n^2 \\ 1/(n^2+1) \\ \sin n/n \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/n^2 \\ 1/(n^2+1) \\ \sin n/n \end{bmatrix},$$

so

$$\|\mathbf{u}_n - \mathbf{u}\| = \sqrt{\left(\frac{1}{n^2}\right)^2 + \left(\frac{1}{(n^2+1)}\right)^2 + \left(\frac{\sin n}{n}\right)^2} \underset{n\to\infty}{\to} \sqrt{0+0+0} = 0,$$

which is what we wanted to show. □

The Inner Product Idea

In addition to the norm concept we have another fundamental tool in our arsenal when we tackle two- and three-dimensional geometrical vectors. This tool is the so-called *dot product* (or *inner product*) of two vectors. It has many handy applications, but the most powerful of these is the ability to determine the angle between two vectors. In fact, some authors use this idea to define dot products as follows: let θ be the angle between representatives of the vectors $\mathbf{u}$ and $\mathbf{v}$ (see Figure 4.1.) The dot product of $\mathbf{u}$ and $\mathbf{v}$ is defined to be the quantity $\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta$. The law of cosines, with this definition and the notation of Figure 4.1, can be stated as

$$\|\mathbf{v} - \mathbf{u}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta \quad \textbf{Law of Cosines}$$

With a bit of algebra one can use the law of cosines and this definition of dot product to derive a very convenient form for inner products; for example, if $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$, then

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + u_3 v_3. \tag{4.1}$$

This makes the calculation of dot products vastly easier since we don't have to use any trigonometry to compute it. A particularly nice application is that we can determine $\cos\theta$ quite easily from the dot product, namely

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}. \tag{4.2}$$

It is useful to try to extend these geometrical ideas to higher dimensions even if we can't literally use trigonometry and the like. So what we do is reverse the sequence of ideas we've discussed and take equation (4.1) as the prototype for our next definition. As with norms, we are going to have to distinguish carefully between the cases of real and complex scalars. First we focus on the more common case of real coefficients.

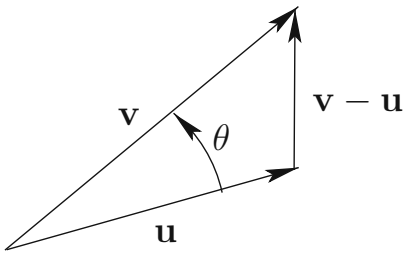


Fig. 4.1: Angle θ between vectors $\mathbf{u}$ and $\mathbf{v}$.

Definition 4.4. **Real Dot Product** Let $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ be vectors in $\mathbb{R}^n$. The *(standard) inner product*, also called the *dot product* of $\mathbf{u}$ and $\mathbf{v}$, is the real number

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n.$$

We can see from the first form of this definition where the term "inner product" came from. Recall from Section 2.4 that the matrix product $\mathbf{u}^T \mathbf{v}$ is called the inner product of these two vectors.

Example 4.6. Compute the dot product of the vectors $\mathbf{u} = (1, -1, 3, 2)$ and $\mathbf{v} = (2, -1, 0, 4)$ in $\mathbb{R}^4$.

Solution. From the definition,

$$\mathbf{u} \cdot \mathbf{v} = 1 \cdot 2 + (-1) \cdot (-1) + 3 \cdot 0 + 2 \cdot 4 = 11. \qquad \square$$

There is a wonderful connection between the standard inner product and the standard norm for vectors that is immediately evident from the definitions. Here it is:

$$\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}. \tag{4.3}$$

Thus, computing norms amounts to an inner product calculation followed by a square root. Actually, we can even avoid the square root and put the equation in the form

$$\|\mathbf{u}\|^2 = \mathbf{u} \cdot \mathbf{u}.$$

We say that the standard norm is *induced* by the standard inner product. We would like this property to carry over to complex vectors. Now we have to be a bit careful. In general, the quantity $\mathbf{u}^T\mathbf{u}$ may not even be a real number, or may be negative. This means that $\sqrt{\mathbf{u}^T\mathbf{u}}$ could be complex, which doesn't seem like a good idea for measuring "length." So how can we avoid this problem? Recall that when we introduced transposes, we also introduced conjugate transposes and remarked that for complex vectors, this is a more natural tool than the transpose. Now we can back up that remark! Recall the definition for complex norm: for $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \mathbb{C}^n$, the *norm* of $\mathbf{u}$ is the nonnegative real number

$$\|\mathbf{u}\| = \sqrt{\overline{u}_1 u_1 + \overline{u}_2 u_2 + \cdots + \overline{u}_n u_n} = \sqrt{\mathbf{u}^*\mathbf{u}}.$$

Therefore, in our definition of complex "dot products" we had better replace transposes by conjugate transposes. This inspires the following definition:

Definition 4.5. **Complex Dot Product** Let $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ be vectors in $\mathbb{C}^n$. The *(standard) inner product*, also called the *dot product* of $\mathbf{u}$ and $\mathbf{v}$, is the complex number

$$\mathbf{u} \cdot \mathbf{v} = \overline{u}_1 v_1 + \overline{u}_2 v_2 + \cdots + \overline{u}_n v_n = \mathbf{u}^*\mathbf{v}.$$

(Be aware that some authors prefer to put the conjugate sign on the second term in this definition.) With this definition we still have the close connection given above in equation (4.3) between norm and standard inner product of complex vectors.

Example 4.7. Compute the dot product of the vectors $\mathbf{u} = (1 + 2\mathrm{i}, \mathrm{i}, 1)$ and $\mathbf{v} = (\mathrm{i}, -1 - \mathrm{i}, 0)$ in $\mathbb{C}^3$.

Solution. Simply apply the definition:

$$\mathbf{u} \cdot \mathbf{v} = \overline{(1 + 2\mathrm{i})}\mathrm{i} + \overline{\mathrm{i}}(-1 - \mathrm{i}) + 1 \cdot 0 = (1 - 2\mathrm{i})\mathrm{i} - \mathrm{i}(-1 - \mathrm{i}) = 1 + 2\mathrm{i}. \quad \square$$

What are the essential defining properties of these standard inner products? It turns out that we can answer the question for both real and complex inner products at once. However, we should bear in mind that in most cases we will be dealing with real dot products, and in such cases all the dot products in question are real numbers, so that any reference to a complex conjugate can be omitted.

Basic Inner Product Laws
Let c be a scalar and $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, where V is a vector space with the standard inner product. Then the following hold:

(1) $\mathbf{u} \cdot \mathbf{u} \geq 0$ with $\mathbf{u} \cdot \mathbf{u} = 0$ if and only if $\mathbf{u} = \mathbf{0}$.
(2) $\mathbf{u} \cdot \mathbf{v} = \overline{\mathbf{v} \cdot \mathbf{u}}$.
(3) $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$.
(4) $\mathbf{u} \cdot (c\mathbf{v}) = c(\mathbf{u} \cdot \mathbf{v})$.

That (1) is true is immediate from the fact that $\mathbf{u} \cdot \mathbf{u} = \mathbf{u}^*\mathbf{u}$ is a sum of the lengths squared of the coordinates of $\mathbf{u}$. This sum is zero exactly when each term is zero. Condition (2) follows from this line of calculation:

$$\overline{\mathbf{v} \cdot \mathbf{u}} = \overline{\mathbf{v}^*\mathbf{u}} = \left(\overline{\mathbf{v}^*\mathbf{u}}\right)^T = (\mathbf{v}^*\mathbf{u})^* = \mathbf{u}^*\mathbf{v} = \mathbf{u} \cdot \mathbf{v}.$$

One point that stands out in this calculation is the following:

Caution: A key difference between real and complex inner products is in the commutative law $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$, which holds for real vectors but *not* for complex vectors, where instead $\mathbf{u} \cdot \mathbf{v} = \overline{\mathbf{v} \cdot \mathbf{u}}$.

Conditions (3) and (4) are similarly verified and left to the exercises. We can also use (4) to prove this fact for real vectors:

$$(c\mathbf{u}) \cdot \mathbf{v} = \mathbf{v} \cdot (c\mathbf{u}) = c(\mathbf{v} \cdot \mathbf{u}) = c(\mathbf{u} \cdot \mathbf{v}).$$

If we are dealing with complex dot products, matters are a bit trickier. One can show then that

$$(c\mathbf{u}) \cdot \mathbf{v} = \overline{c}(\mathbf{u} \cdot \mathbf{v}),$$

so we don't quite have the symmetry that we have for real products.

The Cross Product Idea

We complete this discussion of vector arithmetic with a tool that can be used only in three dimensions, the cross product of vectors. It has more sophisticated relatives, called *wedge products*, that operate in higher-dimensional spaces; this is an advanced topic in multilinear algebra that we shall not pursue. Unlike the dot product, cross products transform vectors into vectors.

In the traditional style of three-dimensional vector analysis, we use the symbols $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ to represent the standard basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ of $\mathbb{R}^3$. Here is the definition of cross product along with a handy determinant mnemonic.

Definition 4.6. **Cross Product of Vectors** Let $\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k}$ and $\mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ be vectors in $\mathbb{R}^3$. The cross product $\mathbf{u} \times \mathbf{v}$ of these vectors is defined to be the vector in $\mathbb{R}^3$ given by

$$\mathbf{u} \times \mathbf{v} = (u_2v_3 - u_3v_2)\,\mathbf{i} + (u_3v_1 - u_1v_3)\,\mathbf{j} + (u_1v_2 - u_2v_1)\,\mathbf{k} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}.$$

Strictly speaking, the "determinant" of this definition is not a determinant in the usual sense. However, formal calculations with it are perfectly valid and provide us with useful insights. For example:

(1) Vectors $\mathbf{u}$ and $\mathbf{v}$ are parallel if and only if $\mathbf{u} \times \mathbf{v} = \mathbf{0}$, since a determinant with one row a multiple of another is zero. In particular, $\mathbf{u} \times \mathbf{u} = \mathbf{0}$.

(2) $\mathbf{w} \cdot \mathbf{u} \times \mathbf{v} = \begin{vmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$, since the result of dotting $\mathbf{w} = u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k}$ with $\mathbf{u} \times \mathbf{v}$ using the first form of the definition of cross product is equal to this determinant. (Note: parentheses are not needed since the only interpretation of $\mathbf{w} \cdot \mathbf{u} \times \mathbf{v}$ that makes sense is $\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$.)

(3) $\mathbf{u} \cdot \mathbf{u} \times \mathbf{v} = 0$ and $\mathbf{v} \cdot \mathbf{u} \times \mathbf{v} = 0$, since a determinant with repeated rows is zero.

(4) $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$, since interchanging two rows of a determinant changes its sign.

(5) $\mathbf{i} \times \mathbf{j} = \mathbf{k}$, $\mathbf{j} \times \mathbf{k} = \mathbf{i}$, $\mathbf{k} \times \mathbf{i} = \mathbf{j}$, as a direct calculation with the definition shows. Thus, the products follow a circular pattern, with the product of any successive two yielding the next vector in the loop $\mathbf{i} \to \mathbf{j} \to \mathbf{k} \to \mathbf{i}$.

Example 4.8. Confirm by direct calculation that $\mathbf{u} \cdot \mathbf{u} \times \mathbf{v} = 0$ and $\mathbf{v} \cdot \mathbf{u} \times \mathbf{v} = 0$ if $\mathbf{u} = (2, -1, 3)$ and $\mathbf{v} = (1, 1, 0)$.

Solution. We calculate that

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 2 & -1 & 3 \\ 1 & 1 & 0 \end{vmatrix} = (-1 \cdot 0 - 1 \cdot 3)\,\mathbf{i} - (2 \cdot 0 - 3 \cdot 1)\,\mathbf{j} + (2 \cdot 1 - (-1)\,1)\,\mathbf{k}$$
$$= -3\mathbf{i} + 3\mathbf{j} + 3\mathbf{k}.$$

Thus

$$\mathbf{u} \cdot \mathbf{u} \times \mathbf{v} = (2\mathbf{i} - 1\mathbf{j} + 3\mathbf{k}) \cdot (-3\mathbf{i} + 3\mathbf{j} + 3\mathbf{k}) = -6 - 3 + 9 = 0$$
$$\mathbf{v} \cdot \mathbf{u} \times \mathbf{v} = (\mathbf{i} + \mathbf{j}) \cdot (-3\mathbf{i} + 3\mathbf{j} + 3\mathbf{k}) = -3 + 3 = 0.$$

□

Here is a summary of some of the basic laws of cross products:

Basic Cross Product Laws
Let $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^3$ and $c \in \mathbb{R}$. Then

(1) $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$.
(2) $(c\mathbf{u}) \times \mathbf{v} = c\,(\mathbf{u} \times \mathbf{v}) = \mathbf{u} \times (c\mathbf{v})$.
(3) $\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = \mathbf{u} \times \mathbf{v} + \mathbf{u} \times \mathbf{w}$.
(4) $(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = \mathbf{u} \times \mathbf{w} + \mathbf{v} \times \mathbf{w}$.
(5) (Scalar triple product) $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \mathbf{u} \times \mathbf{v} \cdot \mathbf{w}$.
(6) (Vector triple product) $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\,\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\,\mathbf{w}$.
(7) $\|\mathbf{u} \times \mathbf{v}\|^2 = \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - (\mathbf{u} \cdot \mathbf{v})^2$.
(8) $\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{u}\| \|\mathbf{v}\| |\sin\theta|$, where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$.

Items (1)–(7) can be verified directly from the definition of cross product and properties of the dot product, while item (8) follows from (7), equation (4.2) and the definition of dot product. Note that (8) has an interesting geometrical interpretation for vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$, namely that $\|\mathbf{u} \times \mathbf{v}\|$ is the area of the parallelogram with adjacent sides represented by the vectors $\mathbf{u}$ and $\mathbf{v}$.

4.1 Exercises and Problems

Exercise 1. For the following pairs of vectors, calculate $\mathbf{u} \cdot \mathbf{v}$, $\|\mathbf{u}\|$, and $\|\mathbf{v}\|$.
(a) $(3, -5)$, $(2, 4)$ (b) $(1, 1, 2)$, $(2, -1, 3)$ (c) $(2, 1, -2, -1)$, $(3, 0, 1, -4)$
(d) $(1 + 2\mathrm{i}, 2 + \mathrm{i})$, $(4 + 3\mathrm{i}, 1)$ (e) $(3, 1, 2, -4)$, $(2, 0, 1, 1)$ (f) $(2, 2, -2)$, $(2, 1, 5)$

Exercise 2. For the following pairs of vectors, calculate $\mathbf{u} \cdot \mathbf{v}$ and unit vectors in the direction of $\mathbf{u}$ and $\mathbf{v}$.
(a) $(4, -2, 2)$, $(1, 3, 2)$ (b) $(1, 1)$, $(2, -2)$ (c) $(4, 0, 1, 2 - 3\mathrm{i})$, $(1, 1 - 2\mathrm{i}, 1, \mathrm{i})$
(d) $(\mathrm{i}, -\mathrm{i})$, $(3\mathrm{i}, 1)$ (e) $(1, -1, 1, -1)$, $(2, 2, 1, 1)$ (f) $(4, 1, 2)$, $(1, 0, 0)$

Exercise 3. Let θ be the angle between the following pairs of real vectors and compute $\cos\theta$ using dot products.
(a) $(2, -5)$, $(4, 2)$ (b) $(3, 4)$, $(4, -3)$ (c) $(1, 1, 2)$, $(2, -1, 3)$ (d) $\mathbf{j} + \mathbf{k}$, $2\mathbf{i} + \mathbf{k}$

Exercise 4. Compute an angle θ between the following pairs of real vectors.
(a) $(4, 5)$, $(-4, 4)$ (b) $\mathbf{i} - 5\mathbf{j}$, $\mathbf{i} + \mathbf{k}$ (c) $(4, 0, 2)$, $(1, 1, 1)$

Exercise 5. Compute the cross product of the vector pairs in Exercise 4. (Express two-dimensional vectors in terms of $\mathbf{i}$ and $\mathbf{j}$ first.)

Exercise 6. Compute $\sin\theta$, where θ is the angle between the following pairs of real vectors, using cross products.
(a) $3\mathbf{i} - 5\mathbf{j}$, $2\mathbf{i} + 4\mathbf{j}$ (b) $3\mathbf{i} - 5\mathbf{j} + 2\mathbf{k}$, $2\mathbf{i} - 4\mathbf{k}$ (c) $(-4, 2, 4)$, $(4, 1, -5)$

Exercise 7. Let $c = 3$, $\mathbf{u} = (4, -1, 2, 3)$, and $\mathbf{v} = (-2, 2, -2, 2)$. Verify that the four basic norm laws hold for these vectors and scalars.

Exercise 8. Let $c = 2$, $\mathbf{u} = (-3, 2, 1)$, $\mathbf{v} = (4, 2, -3)$, and $\mathbf{w} = (1, -2, 1)$. Verify the four basic inner product laws for these vectors and scalars.

Exercise 9. Let $c = -2$, $\mathbf{u} = (0, 2, 1)$, $\mathbf{v} = (4, 0, -3)$, and $\mathbf{w} = (1, -2, 1)$. Verify cross product laws (1)–(4) for these vectors and scalars.

Exercise 10. Let $\mathbf{u} = (1, 2, 2)$, $\mathbf{v} = (0, 2, -3)$, and $\mathbf{w} = (1, 0, 1)$. Verify cross product laws (5)–(7) for these vectors and scalars.

Exercise 11. Let $\mathbf{u}, \mathbf{v} \in \mathbb{C}^3$ be given by $\mathbf{u} = (\mathrm{i}, 2, 1 - \mathrm{i})$ and $\mathbf{v} = (2 + 3\mathrm{i}, 2, -1)$. Verify the *parallelogram equality*

$$2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 = \|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2$$

with these vectors.

Exercise 12. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ be given by $\mathbf{u} = (1, 2, 1)$ and $\mathbf{v} = (-1, 2, -3)$. Compute $\|\mathbf{u}\|^2$, $\|\mathbf{v}\|^2$ and $\|\mathbf{u} + \mathbf{v}\|^2$. What does this tell you about these vectors?

Exercise 13. Verify that $\mathbf{u}_n = [2/n, (1 + n^2)/(2n^2 + 3n + 5)]^T$, $n = 1, 2, \ldots$, converges to a limit vector $\mathbf{u}$ by using the norm definition of vector limit.

Exercise 14. Let $\mathbf{u}_n = [\mathrm{i}, \left(\mathrm{n}^2\mathrm{i} + 1\right) / \left((\mathrm{ni})^2 + \mathrm{n}\right)]$, $n = 1, 2, \ldots$, and verify that $\mathbf{u}_n$ converges to a limit vector $\mathbf{u}$.

*Problem 15. Show that for real vectors $\mathbf{u}$, $\mathbf{v}$ and real number c one has

$$(c\mathbf{u}) \cdot \mathbf{v} = \mathbf{v} \cdot (c\mathbf{u}) = c(\mathbf{v} \cdot \mathbf{u}) = c(\mathbf{u} \cdot \mathbf{v}).$$

Problem 16. Prove this basic norm law: $\|\mathbf{u}\| \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$.

*Problem 17. Let $A = \mathbf{u}\mathbf{u}^T$ with $\mathbf{u} \in \mathbb{R}^n$. Derive a formula for A^m, m a positive integer, in terms of $\mathbf{u}$.

Problem 18. Show that if $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ (or $\mathbb{C}^n$) and c is a scalar, then
(a) $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$ (b) $\mathbf{u} \cdot (c\mathbf{v}) = c(\mathbf{u} \cdot \mathbf{v})$

Problem 19. Show from the definition that if $\lim_{n\to\infty} \mathbf{u}_n = \mathbf{u}$, where $\mathbf{u}_n = (x_n, y_n) \in \mathbb{R}^2$ and $\mathbf{u} = (x, y)$, then $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} y_n = y$.

*Problem 20. Prove that if $\mathbf{v}$ is a vector and c is a positive real, then normalizing $\mathbf{v}$ and normalizing $c\mathbf{v}$ yield the same unit vector. How are the normalized vectors related if c is negative?

Problem 21. Show that if A is a real $n \times n$ matrix and $\mathbf{u}, \mathbf{v}$ are vectors in $\mathbb{R}^n$, then $\left(A^T\mathbf{u}\right) \cdot \mathbf{v} = \mathbf{u} \cdot (A\mathbf{v})$.

*Problem 22. Show that $|\,\|\mathbf{u}\| - \|\mathbf{v}\|\,| \leq \|\mathbf{u}\|$ for any two vectors $\mathbf{u}, \mathbf{v}$ in the same space.

Problem 23. Verify the scalar triple product law $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \mathbf{u} \times \mathbf{v} \cdot \mathbf{w}$.

*Problem 24. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$. Use the law of cosines for the triangle determined by $\mathbf{u}$ and $\mathbf{v}$ in terms of the coordinates of of $\mathbf{u}$ and $\mathbf{v}$ to verify equation (4.2).

Problem 25. Verify the parallelogram equality (see Exercise 11) for vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$.

Problem 26. Verify item (7) of the Basic Cross Product Laws.

*Problem 27. Use item (7) of the Basic Cross Product Laws to verify item (8).

4.2 Applications of Norms and Vector Products

Projections and Angles

Now that we have dot products under our belts we can tackle geometrical issues such as angles between vectors in higher dimensions. For the matter of angles, we will stick to real vector spaces, though we could do it for complex vector spaces with a little extra work. What we would like to do is take equation (4.2) as the *definition* of the angle between two vectors. There's one slight problem: how do we know that it will give a quantity that could be a cosine? After all, cosines take on only values between -1 and 1. We could use some help and the Cauchy–Bunyakovsky–Schwarz inequality (CBS for short) is just what we need:

Theorem 4.2. CBS Inequality For vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$,

$$|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\| .$$

Proof. Let c be an arbitrary real number and compute the nonnegative quantity

$$\begin{aligned} f(c) &= \|\mathbf{u} + c\mathbf{v}\|^2 \\ &= (\mathbf{u} + c\mathbf{v}) \cdot (\mathbf{u} + c\mathbf{v}) \\ &= \mathbf{u} \cdot \mathbf{u} + \mathbf{u} \cdot (c\mathbf{v}) + (c\mathbf{v}) \cdot \mathbf{u} + (c\mathbf{v}) \cdot (c\mathbf{v}) \\ &= \|\mathbf{u}\|^2 + 2c(\mathbf{u} \cdot \mathbf{v}) + c^2 \|\mathbf{v}\|^2 \\ &= \left(\|\mathbf{u}\|^2 - \left(\frac{(\mathbf{u} \cdot \mathbf{v})}{\|\mathbf{v}\|} \right)^2 \right) + \left(\frac{(\mathbf{u} \cdot \mathbf{v})}{\|\mathbf{v}\|} + \|\mathbf{v}\| \, c \right)^2 . \end{aligned}$$

The function $f(c)$ is a quadratic in the variable c with nonnegative values, whose low point occurs where the squared term is zero, i.e.,

$$c = \frac{-(\mathbf{u} \cdot \mathbf{v})}{\|\mathbf{v}\|^2}.$$

Evaluate f at this point to get that

$$0 \leq \|\mathbf{u}\|^2 - \frac{(\mathbf{u} \cdot \mathbf{v})^2}{\|\mathbf{v}\|^2}.$$

Now add $(\mathbf{u} \cdot \mathbf{v})^2 / \|\mathbf{v}\|^2$ to both sides of the inequality and multiply by $\|\mathbf{v}\|^2$ to obtain that

$$(\mathbf{u} \cdot \mathbf{v})^2 \leq \|\mathbf{u}\|^2 \, \|\mathbf{v}\|^2 .$$

Take square roots, use the fact that $|x| = \sqrt{x^2}$ for real numbers x, and the desired inequality follows. □

This inequality has a number of useful applications. Because of it we can articulate a definition of angle between vectors. Note that there is a certain ambiguity in discussing the angle between vectors, since more than one angle works. It is the cosine of these angles that is actually unique.

Definition 4.7. Angle Between Vectors For nonzero vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ we define the *angle* between $\mathbf{u}$ and $\mathbf{v}$ to be any angle θ satisfying

$$\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}.$$

Thanks to the CBS inequality, we know that $|\mathbf{u}\cdot\mathbf{v}|\,/(\|\mathbf{u}\|\,\|\mathbf{v}\|) \leq 1$, so that this formula for $\cos\theta$ makes sense.

Example 4.9. Find the angle between the vectors $\mathbf{u} = (1,1,0,1)$ and $\mathbf{v} = (1,1,1,1)$ in $\mathbb{R}^4$.

Solution. We have that

$$\cos\theta = \frac{(1,1,0,1)\cdot(1,1,1,1)}{\|(1,1,0,1)\|\,\|(1,1,1,1)\|} = \frac{3}{2\sqrt{3}} = \frac{\sqrt{3}}{2}.$$

Hence, we can take $\theta = \pi/6$. □

Example 4.10. Use the laws of inner products and the CBS inequality to verify the triangle inequality for vectors $\mathbf{u}$ and $\mathbf{v}$. What happens to this inequality if we also know that $\mathbf{u}\cdot\mathbf{v} = 0$?

Solution. Here the trick is to avoid square roots. Square both sides of equation (4.3) to obtain that

$$\begin{aligned}
\|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) \\
&= \mathbf{u}\cdot\mathbf{u} + \mathbf{u}\cdot\mathbf{v} + \mathbf{v}\cdot\mathbf{u} + \mathbf{v}\cdot\mathbf{v} \\
&= \|\mathbf{u}\|^2 + 2(\mathbf{u}\cdot\mathbf{v}) + \|\mathbf{v}\|^2 \\
&\leq \|\mathbf{u}\|^2 + 2\,|\mathbf{u}\cdot\mathbf{v}| + \|\mathbf{v}\|^2 \\
&\leq \|\mathbf{u}\|^2 + 2\,\|\mathbf{u}\|\,\|\mathbf{v}\| + \|\mathbf{v}\|^2 \\
&= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2,
\end{aligned}$$

where the last inequality follows from the CBS inequality. If $\mathbf{u}\cdot\mathbf{v} = 0$, then the third equality yields the Pythagorean theorem. □

We have just seen a very important case of angles between vectors that warrants its own name. Recall from geometry that two vectors are *perpendicular* or *orthogonal* if the angle between them is $\pi/2$. Since $\cos\pi/2 = 0$, we see that this amounts to the equation $\mathbf{u}\cdot\mathbf{v} = 0$. Now we can extend the perpendicularity idea to arbitrary vectors, including complex vectors.

Definition 4.8. Orthogonal Vectors Two vectors $\mathbf{u}$ and $\mathbf{v}$ in an inner product space are *orthogonal* if $\mathbf{u} \cdot \mathbf{v} = 0$. In this case we write $\mathbf{u} \perp \mathbf{v}$.

In the case that one of the vectors is the zero vector, we have the little oddity that the zero vector is orthogonal to every other vector, since the dot product is always 0 in this case. Some authors require that $\mathbf{u}$ and $\mathbf{v}$ be nonzero as part of the definition. It's a minor point and we won't worry about it. When $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, i.e., $\mathbf{u} \cdot \mathbf{v} = 0$, we see from the third equality in the derivation of CBS above that

Pythagorean Theorem

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2,$$

which is really the Pythagorean theorem for vectors in $\mathbb{R}^n$.

Example 4.11. Determine whether the following pairs of vectors are orthogonal.

(a) $\mathbf{u} = (2, -1, 3, 1)$ and $\mathbf{v} = (1, 2, 1, -2)$
(b) $\mathbf{u} = (1 + \mathrm{i}, 2)$ and $\mathbf{v} = (-2\mathrm{i}, 1 + \mathrm{i})$.

Solution. For (a) we calculate

$$\mathbf{u} \cdot \mathbf{v} = 2 \cdot 1 + (-1)2 + 3 \cdot 1 + 1(-2) = 1,$$

so that $\mathbf{u}$ is not orthogonal to $\mathbf{v}$. For (b) we calculate

$$\mathbf{u} \cdot \mathbf{v} = (1 - \mathrm{i})(-2\mathrm{i}) + 2(1 + \mathrm{i}) = -2\mathrm{i} - 2 + 2 + 2\mathrm{i} = 0,$$

so that $\mathbf{u}$ is orthogonal to $\mathbf{v}$ in this case. □

The next example illustrates a handy little trick well worth remembering.

Example 4.12. Find a vector orthogonal to the vector (a, b) in $\mathbb{R}^2$ or $\mathbb{C}^2$.

Solution. Simply interchange coordinates, conjugate them (this does nothing if the entries are real), and insert a minus sign in front of one of the coordinates, say the first. We obtain $(-\overline{b}, \overline{a})$. Now check that

$$(a, b) \cdot (-\overline{b}, \overline{a}) = \overline{a}(-\overline{b}) + \overline{b}\overline{a} = 0.$$ □

Parallel Vectors

By *parallel vectors* we mean two vectors that are nonzero scalar multiples of each other. Notice that parallel vectors may determine the same or opposite directions. Our next application of the dot product relates back to a fact that we learned in geometry: given two nonzero vectors in the plane, it is always possible to resolve one of them into a sum of a vector parallel to the other and a vector orthogonal to the other (see Figure 4.2). The parallel component is called the projection of one vector along the other. This idea is useful, for example, in physics problems where we want to resolve a force into orthogonal components. As a matter of fact,

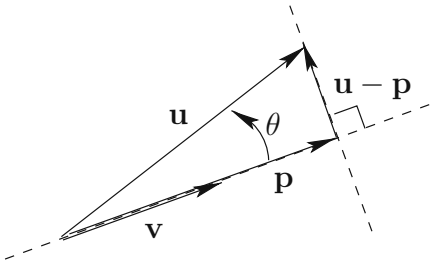


Fig. 4.2: Angle between vectors $\mathbf{u}$ and $\mathbf{v}$, projection $\mathbf{p}$ of $\mathbf{u}$ along $\mathbf{v}$ and $(\mathbf{u}-\mathbf{p}) \perp \mathbf{v}$.

we can develop this same idea in arbitrary standard vector spaces. That is the content of the following useful fact.

Theorem 4.3. Projection Formula for Vectors Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in a vector space with $\mathbf{v} \neq \mathbf{0}$. Let

$$\mathbf{p} = \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{v} \cdot \mathbf{v}} \mathbf{v} \quad \text{and} \quad \mathbf{q} = \mathbf{u} - \mathbf{p}.$$

Then $\mathbf{p}$ is parallel to $\mathbf{v}$, $\mathbf{q}$ is orthogonal to $\mathbf{v}$, and $\mathbf{u} = \mathbf{p} + \mathbf{q}$.

Proof. Let $\mathbf{p} = c\mathbf{v}$, an arbitrary multiple of $\mathbf{v}$. Then $\mathbf{p}$ is automatically parallel to $\mathbf{v}$. Impose the constraint that $\mathbf{q} = \mathbf{u} - \mathbf{p}$ be orthogonal to $\mathbf{v}$. This means, by definition, that

$$0 = \mathbf{v} \cdot \mathbf{q} = \mathbf{v} \cdot (\mathbf{u} - \mathbf{p}) = \mathbf{v} \cdot \mathbf{u} - \mathbf{v} \cdot (c\mathbf{v}).$$

Add $\mathbf{v} \cdot (c\mathbf{v})$ to both sides and pull the scalar c outside the dot product to obtain that

$$c(\mathbf{v} \cdot \mathbf{v}) = \mathbf{v} \cdot \mathbf{u}$$

and therefore

$$c = \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{v} \cdot \mathbf{v}}.$$

So for this choice of c, $\mathbf{q}$ is orthogonal to $\mathbf{p}$. Clearly, $\mathbf{u} = \mathbf{p} + \mathbf{u} - \mathbf{p}$, so the proof is complete. □

It is customary to call the vector $\mathbf{p}$ of this theorem the *(parallel) projection of* $\mathbf{u}$ *along* $\mathbf{v}$. As above, we write

Projection Vector

$$\text{proj}_{\mathbf{v}} \mathbf{u} = \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{v} \cdot \mathbf{v}} \mathbf{v}.$$

The projection of one vector along another is itself a vector quantity.

A scalar quantity that is frequently associated with these calculations is the *component* of $\mathbf{u}$ along $\mathbf{v}$. It is defined as

Component of Vector

$$\operatorname{comp}_{\mathbf{v}} \mathbf{u} = \frac{\mathbf{v} \cdot \mathbf{u}}{\|\mathbf{v}\|}.$$

The connection between these two quantities is that

$$\operatorname{proj}_{\mathbf{v}} \mathbf{u} = \frac{\operatorname{comp}_{\mathbf{v}} \mathbf{u}}{\|\mathbf{v}\|}\mathbf{v}.$$

Notice that $\mathbf{v}/\|\mathbf{v}\|$ is a unit vector in the same direction as $\mathbf{v}$. Therefore, $\operatorname{comp}_{\mathbf{v}} \mathbf{u}$ is the signed magnitude of the projection of $\mathbf{u}$ along $\mathbf{v}$ and will be negative if the angle between $\mathbf{u}$ and $\mathbf{v}$ exceeds $\pi/2$.

Orthogonal Projection

The vector $\mathbf{q}$ of Theorem 4.3 that is orthogonal to $\mathbf{v}$ also has a name: the *orthogonal projection* of $\mathbf{u}$ to $\mathbf{v}$. We write

$$\operatorname{orth}_{\mathbf{v}} \mathbf{u} = \mathbf{u} - \operatorname{proj}_{\mathbf{v}} \mathbf{u}.$$

Note, however, that the default meaning of "projection" is "parallel projection."

Example 4.13. Calculate the projection and component of $\mathbf{u} = (1, -1, 1, 1)$ along $\mathbf{v} = (0, 1, -2, -1)$ and verify that $\mathbf{u} - \mathbf{p} \perp \mathbf{v}$.

Solution. We have that

$$\mathbf{v} \cdot \mathbf{u} = 0 \cdot 1 + 1(-1) + (-2)1 + (-1)1 = -4,$$
$$\mathbf{v} \cdot \mathbf{v} = 0^2 + 1^2 + (-2)^2 + (-1)^2 = 6,$$

so that

$$\mathbf{p} = \operatorname{proj}_{\mathbf{v}} \mathbf{u} = \frac{-4}{6}(0, 1, -2, -1) = \frac{1}{3}(0, -2, 4, 2).$$

It follows that

$$\mathbf{u} - \mathbf{p} = \frac{1}{3}(3, -1, -1, 1)$$

and

$$(\mathbf{u} - \mathbf{p}) \cdot \mathbf{v} = \frac{1}{3}(3 \cdot 0 + 1(-1) + (-1)(-2) + 1(-1)) = 0.$$

Also, the component of $\mathbf{u}$ along $\mathbf{v}$ is

$$\operatorname{comp}_{\mathbf{v}} \mathbf{u} = \frac{\mathbf{v} \cdot \mathbf{u}}{\|\mathbf{v}\|} = \frac{-4}{\sqrt{6}}.$$

□

A hyperplane is a basic geometrical object on which inner product tools can shed light. Here is the definition.

Definition 4.9. **Hyperplane in $\mathbb{R}^n$** A *hyperplane* in $\mathbb{R}^n$ is the set of all $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{a} \cdot \mathbf{x} = b$, where the nonzero vector $\mathbf{a} \in \mathbb{R}^n$ and scalar b are given.

These are familiar objects. For example, a hyperplane in $\mathbb{R}^3$ is the set of points (x, y, z) that satisfy an equation $ax + by + cz = d$, which is simply a

plane in three dimensions. A hyperplane in $\mathbb{R}^2$ is the set of points (x, y) that satisfy an equation $ax + by = c$, which is just a line in two dimensions. (Notice that in the absence of homogeneous space, a tuple like (x, y, z) has a dual interpretation as point or vector.) Here is a general geometrical interpretation of hyperplanes.

Theorem 4.4. Geometry of Hyperplanes Let H be the hyperplane in $\mathbb{R}^n$ defined by the equation $\mathbf{a} \cdot \mathbf{x} = b$ and let $\mathbf{x}_* \in H$. Then

(1) $\mathbf{a}^\perp = \{\mathbf{y} \in \mathbb{R}^n \,|\, \mathbf{a} \cdot \mathbf{y} = 0\}$ is a subspace of $\mathbb{R}^n$ of dimension $n - 1$.
(2) $H = \mathbf{x}_* + \mathbf{a}^\perp = \{\mathbf{x}_* + \mathbf{y} \,|\, \mathbf{y} \in \mathbf{a}^\perp\}$.

Proof. For (1), observe that $\mathbf{a}^\perp = \mathcal{N}(\mathbf{a}^T)$, which is a subspace of $\mathbb{R}^n$. According to the projection formula for vectors, any element of $\mathbb{R}^n$ can be expressed as a sum of a multiple of $\mathbf{a}$ and a vector orthogonal to $\mathbf{a}$. Therefore, $\mathbb{R}^n$ is spanned by a basis of $\mathbf{a}^\perp$ and $\mathbf{a}$. Since $\dim \mathbb{R}^n = n$, a basis of $\mathbf{a}^\perp$ must have at least $n - 1$ elements. If it had n elements, then we would have $\mathbf{a}^\perp = \mathbb{R}^n$, which would imply that $\mathbf{a} \cdot \mathbf{a} = 0$ and therefore $\mathbf{a} = \mathbf{0}$, which is false. Therefore, $\dim \mathbf{a}^\perp = n - 1$. Part (2) follows from Theorem 3.16 since $\mathbf{x}_*$ is a particular solution to the linear system $\mathbf{a}^T\mathbf{x} = 0$. □

Notice that the vector $\mathbf{a}$ can be read off immediately from the defining equation. For example, we see by inspection that a vector orthogonal to the plane given by $2x - 3y + z = 4$ is $\mathbf{a} = (2, -3, 1)$. Finding the defining equation is a bit more work.

Example 4.14. Find an equation that defines the plane containing the three (noncollinear) points P, Q, and R with coordinates $(1, 0, 2)$, $(2, 1, 0)$, and $(3, 1, 1)$, respectively.

Solution. First calculate displacement vectors

$$\overrightarrow{PQ} = (2, 1, 0) - (1, 0, 2) = (1, 1, -2)$$
$$\overrightarrow{PR} = (3, 1, 1) - (1, 0, 2) = (2, 1, -1).$$

These vectors are parallel to the plane. Therefore, their cross product, which is orthogonal to each vector, will be orthogonal to the plane. We calculate

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 1 & -2 \\ 2 & 1 & -1 \end{vmatrix} = \mathbf{i} - 3\mathbf{j} - \mathbf{k}.$$

Hence, the equation of the plane is $x - 3y - z = b$. To determine b, plug in the coordinates of P and obtain that $1 \cdot 1 - 3 \cdot 0 - 2 \cdot 1 = -1 = b$. Hence, an equation of the plane is $x - 3y - z = -1$. □

Least Squares

Example 4.15. You are using a pound scale to measure weights for produce sales when you notice that your scale is broken. The vendor at the next stall is leaving and lends you another scale as she departs. You then realize that the new scale is in units you don't recognize. You happen to have some known weights that are approximately 2, 5, and 7 pounds respectively. When you weigh these items on the new scale you get the numbers 0.7, 2.4, and 3.2. You get your calculator out and hypothesize that the unit of weight should be some constant multiple of pounds. Model this information as a system of equations. Is it clear from this system what the units of the scale are?

Solution. Express the relationship between the weight p in pounds and the weight w in unknown units as $w \cdot c = p$, where c is an unknown constant of proportionality. Your data show that we have

$$\begin{aligned} 0.7c &= 2 \\ 2.4c &= 5 \\ 3.4c &= 7. \end{aligned}$$

As a system of three equations in one unknown you see immediately that this overdetermined system(too many equations) is inconsistent. After all, the pound weights were only approximate and there is always some error in measurement. What to do? You could just average the three inconsistent values of c, thereby obtaining

$$c = (2/0.7 + 5/2.4 + 7/3.4)/3 = 2.3331.$$

You get a number, but it isn't at all clear that this is a good strategy. □

There really is a better way, and it will lead to a slightly different estimate of the number c. This method, called the *method of least squares*, was invented by C. F. Gauss to handle uncertainties in orbital calculations in astronomy.

Method of Least Squares

Here is the basic problem: suppose we have data that leads to a system of equations for unknowns that we want to solve for, but the data has errors in it and consequently leads to an inconsistent linear system

$$A\mathbf{x} = \mathbf{b}.$$

How do we find the "best" approximate solution? One could answer this in many ways. One of the most commonly accepted ideas is one that Gauss proposed: the *residual* $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ should be $\mathbf{0}$, so its departure from $\mathbf{0}$ is a measure of our error. Thus, we should try to find a value of the unknown $\mathbf{x}$ that minimizes the norm of the residual squared, i.e., a "solution" $\mathbf{x}$ such that

Residual Vector

$$\|\mathbf{b} - A\mathbf{x}\|^2$$

is minimized. Such a solution is called a "least squares" solution to the system. This technique is termed "linear regression" by statisticians,who use it in situations in which one has many estimates for unknown parameters that taken together are not perfectly consistent. It can be shown that if A is known, errors in $\mathbf{b}$ are normally distributed and the least squares solution unique, then it is an unbiased estimator of the true solution in the statistical sense.

Let's try to get a fix on this problem. Even the one-variable case is instructive, so let's use the preceding example. In this case the coefficient matrix A is the column vector $\mathbf{a} = [0.7, 2.4, 3.4]^T$, and the right-hand-side vector is $\mathbf{b} = [2, 5, 7]^T$. What we are really trying to find is a value of the scalar $x = c$ such that $\mathbf{b} - A x = \mathbf{b} - x\mathbf{a}$ is a minimum. Here is a geometrical interpretation: we want to find the multiple of the vector $\mathbf{a}$ that is closest to $\mathbf{b}$. Geometry suggests that this minimum occurs when $\mathbf{b} - x\mathbf{a}$ is orthogonal to $\mathbf{a}$, in other words, when $x\mathbf{a}$ is the projection of $\mathbf{b}$ along $\mathbf{a}$. Inspection of the projection formula shows us that we must have

$$x = \frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{a} \cdot \mathbf{a}} = \frac{0.7 \cdot 2 + 2.4 \cdot 5 + 3.4 \cdot 7}{0.7 \cdot 0.7 + 2.4 \cdot 2.4 + 3.4 \cdot 3.4} \approx 2.0887.$$

Notice that this value doesn't solve any of the original equations exactly, but it is, in a certain sense, the best approximate solution to all three equations taken together. Also, this solution is *not* the same as the average of the solutions to the three equations, which we computed to be approximately 2.3331.

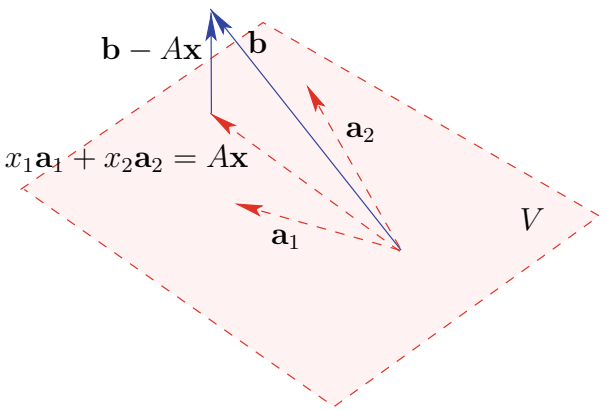


Fig. 4.3: The vector in the subspace $\mathcal{C}(A)$ nearest to $\mathbf{b}$.

Now how do we tackle the more general system $A\mathbf{x} = \mathbf{b}$? Since $A\mathbf{x}$ is just a linear combination of the columns, what we should find is the vector of this form that is closest to the vector $\mathbf{b}$. See Figure 4.3 for a picture of the situation with $n = 2$. Our experience with the 1-dimensional case suggests that we should require that the residual be orthogonal to each column of A, that is, $\mathbf{a}_i \cdot (\mathbf{b} - A\mathbf{x}) = \mathbf{a}_i^T(\mathbf{b} - A\mathbf{x}) = 0$, for all columns $\mathbf{a}_i$ of A. Each column gives rise to one equation. We can write all these equations at once in the form of

the *normal equations*:

Normal Equations

$$A^T A\mathbf{x} = A^T\mathbf{b}.$$

In fact, this is the same set of equations we get if we apply calculus to the scalar function of variables $x_1, x_2, \ldots, x_n$ given as $f(x) = \|\mathbf{b} - A\mathbf{x}\|^2$ and search for a local minimum by setting all partials equal to 0. Any solution to this system will minimize the norm of $\mathbf{b} - A\mathbf{x}$ as $\mathbf{x}$ ranges over all elements of $\mathbb{R}^n$.

Positive Semidefinite or Definite Matrix

The coefficient matrix $B = A^T A$ of the normal system has some pleasant properties. For one, it is a symmetric matrix. For another, it is a *positive semidefinite matrix*, by which we mean that B is a square $n \times n$ matrix such that $\mathbf{x}^T B\mathbf{x} \geq 0$ for all vectors $\mathbf{x} \in \mathbb{R}^n$. In fact, in some cases B is even better behaved because it is a *positive definite matrix*, by which we mean that B is a square $n \times n$ matrix such that $\mathbf{x}^T B\mathbf{x} > 0$ for all nonzero vectors $\mathbf{x} \in \mathbb{R}^n$. (For complex matrices, the condition is $\mathbf{x}^* B\mathbf{x} > 0$ for all nonzero vectors $\mathbf{x} \in \mathbb{C}^n$.)

Least Squares Solution and Genuine Solution

Does there exist a solution to the normal equations? The answer is yes. In general, any solution to the normal equations minimizes the residual norm and is called a *least squares solution* to the problem $A\mathbf{x} = \mathbf{b}$. Since we now have two versions of "solution" for the system $A\mathbf{x} = \mathbf{b}$, we should distinguish between them in situations that may refer to either. If the vector $\mathbf{x}$ actually satisfies the equation $A\mathbf{x} = \mathbf{b}$, we call $\mathbf{x}$ a *genuine solution* to the system to contrast it with a least squares solution. Certainly, every genuine solution is a least squares solution, but the converse will not be true if the original system is inconsistent. We leave the verifications as exercises.

The normal equations are guaranteed to be consistent—a nontrivial fact—and will have infinitely many solutions if $A^T A$ is a singular matrix. Consider the most common case, namely that in which A is a rank-n matrix. Recall that in this case we say that A has *full column rank*. We can show that the $n \times n$ matrix $A^T A$ is also of rank n. This means that it is an invertible matrix and therefore the solution to the normal equations is *unique*. Here is the necessary fact.

Theorem 4.5. Suppose that the real $m \times n$ matrix A has full column rank n. Then the $n \times n$ matrix $A^T A$ also has rank n and is invertible.

Proof. Assume that A has rank n. Now suppose that for some vector $\mathbf{x}$ we have

$$\mathbf{0} = A^T A\mathbf{x}.$$

Multiply on the left by $\mathbf{x}^T$ to obtain that

$$0 = \mathbf{x}^T\mathbf{0} = \mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T(A\mathbf{x}) = \|A\mathbf{x}\|^2,$$

so that $A\mathbf{x} = \mathbf{0}$. However, we know by Theorem 1.5 that the homogeneous system with A as its coefficient matrix must have a unique solution. Of course, this solution is the zero vector. Therefore, $\mathbf{x} = \mathbf{0}$. It follows that the square matrix A^TA has rank n and is also invertible by Theorem 2.6. □

Example 4.16. Two parameters, x_1 and x_2, are linearly related. Three samples are taken that lead to the system of equations

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

Show that this system is inconsistent, and find the least squares solution for $\mathbf{x} = (x_1, x_2)$. What is the minimum norm of the residual $\mathbf{b} - A\mathbf{x}$ in this case?

Solution. In this case it is obvious that the system is inconsistent: the first and third equations have the same quantity, $2x_1 + x_2$, equal to different values 0 and 2. Of course, we could have set up the augmented matrix of the system and found a pivot in the right-hand-side column as well. We see that the (rank 2) coefficient matrix A and right-hand side $\mathbf{b}$ are

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}, \text{ and } \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}.$$

Thus,

$$A^TA = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 5 \\ 5 & 3 \end{bmatrix}$$

and

$$A^T\mathbf{b} = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}.$$

As predicted by the preceding theorem, A^TA is invertible, and we use the 2×2 formula for the inverse:

$$(A^TA)^{-1} = \begin{bmatrix} 9 & 5 \\ 5 & 3 \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} 3 & -5 \\ -5 & 9 \end{bmatrix},$$

so that the unique least squares solution is

$$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b} = \frac{1}{2} \begin{bmatrix} 3 & -5 \\ -5 & 9 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The minimum value for the residual $\mathbf{b} - A\mathbf{x}$ occurs when $\mathbf{x}$ is a least squares solution, so we get

$$\mathbf{b} - A\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix},$$

and therefore

$$\|\mathbf{b} - A\mathbf{x}\| = \sqrt{2} \approx 1.414.$$

This isn't terribly small, but it's the best we can do with this system. This number tells us that the system is badly inconsistent. □

Computer Graphics

Cross products have important applications in computer graphics. Consider, e.g., the problem of rendering the reflective properties of a surface: such objects are typically stored in a computer as a mesh of adjacent triangles with the vertices of each triangle oriented in counterclockwise direction relative to a viewer on the outside of the mesh. (Some graphics systems such as Microsoft's Direct3D use a left-handed coordinate system and orient triangles in the clockwise direction.)

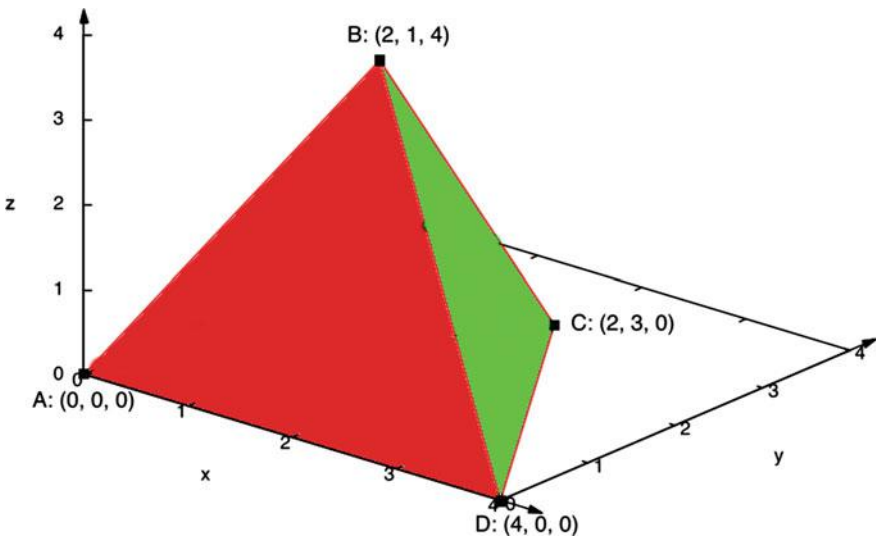


Fig. 4.4: Tetrahedron with labeled vertices A, B, C, D.

Example 4.17. Compute unit outward normals for two of the faces of the tetrahedron in Figure 4.4.

Solution. Consider the face of (correctly oriented) triangle DAC. Use the edges of this triangle in the correct order to obtain the vectors the vectors $\mathbf{v} = \overrightarrow{DA}$ and $\mathbf{w} = \overrightarrow{AC}$. Next compute the cross product of these vectors to obtain the outward orthogonal vector

$$\mathbf{v}\times\mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -4 & 0 & 0 \\ 2 & 3 & 0 \end{vmatrix} = (0\cdot 0 - 0\cdot 3)\,\mathbf{i} - (-4\cdot 0 - 0\cdot 2)\,\mathbf{j} + (-4\cdot 3 - 2\cdot 0)\,\mathbf{k} = -12\mathbf{k}.$$

Thus, a unit vector in the same direction as $\mathbf{v}\times\mathbf{w}$ is $\mathbf{u} = -\mathbf{k}$. This is what inspection of the tetrahedron tells us since triangle DAC lies in the xy-plane and the outward direction points in the negative direction of the z-axis.

Next consider the face of triangle ABC. The edges yield vectors $\mathbf{v} = \overrightarrow{AB}$ and $\mathbf{w} = \overrightarrow{BC}$. Compute the cross product of these vectors to obtain the outward orthogonal vector

$$\mathbf{v}\times\mathbf{w} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 2 & 1 & 4 \\ 0 & 2 & -4 \end{vmatrix} = (-4\cdot 1 - 2\cdot 4)\,\mathbf{i} - (-4\cdot 2 - 4\cdot 0)\,\mathbf{j} + (2\cdot 2 - 0\cdot 1)\,\mathbf{k}$$
$$= -12\mathbf{i} + 8\mathbf{j} + 4\mathbf{k}. \qquad \square$$

4.2 Exercises and Problems

In the following exercises, all vectors are real unless otherwise indicated.

Exercise 1. Find the angle θ in radians between the following pairs of vectors.
(a) $(2,-5)$, $(3,4)$ (b) $(4,5,-3,4)$, $(2,-4,1,3)$ (c) $(1,-2,3,4,1)$, $(2,3,1,5,5)$

Exercise 2. Find the angle θ between the following pairs of vectors.
(a) $(1,0,1,0,2)$, $(2,1,-3,2,4)$ (b) $(7,-3,1,1,2,-2)$, $(2,3,-4,-3,2,2)$

Exercise 3. Find the projection and component of $\mathbf{u}$ along $\mathbf{v}$, where the pair $\mathbf{u},\mathbf{v}$ are
(a) $(-4,3)$, $(2,1)$ (b) $(3,0,4)$, $(2,2,1)$ (c) $(1,0,-5,2)$, $(1,1,1,1)$

Exercise 4. Find the orthogonal projection of $\mathbf{u}$ to $\mathbf{v}$, where the pair $\mathbf{u},\mathbf{v}$ are
(a) $(1,-\sqrt{3})$, $(2,1)$ (b) $(2,1,3)$, $(8,2,-4)$ (c) $(3,2,1,1,1)$, $(1,1,1,0,1)$

Exercise 5. Verify the CBS inequality for the vectors $\mathbf{u}$ and $\mathbf{v}$, where the pair $\mathbf{u},\mathbf{v}$ are
(a) $\mathbf{i}-2\mathbf{j}$, $\mathbf{i}+\mathbf{j}-\mathbf{k}$ (b) $(3,-2,3)$, $(1,-5,2)$ (c) $(3,-2)$, $(-6,4)$

Exercise 6. Determine whether the following pairs of vectors $\mathbf{u}$, $\mathbf{v}$ are orthogonal, and if so, verify that the Pythagorean theorem holds for the pair.
(a) $(-2,1,3)$, $(1,2,0)$ (b) $(1,1,0,-1)$, $(1,-1,3,0)$ (c) $(\mathrm{i},2)$, $(2,\mathrm{i})$

Exercise 7. For the following orthogonal pairs $\mathbf{u},\mathbf{v}$ and matrix $M = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, determine whether $M\mathbf{u}$ and $M\mathbf{v}$ are orthogonal.
(a) $(2,1,1)$, $(1,0,-2)$ (b) $(0,1,1)$, $(1,-1,1)$ (c) $(3,1,-2)$, $(1,3,3)$

Exercise 8. For each of the pairs of Exercise 7, determine whether $M\mathbf{u}$ and $\left(M^{-1}\right)^T \mathbf{v}$ are orthogonal.

Exercise 9. Find equations for the following planes in $\mathbb{R}^3$.
(a) The plane containing the points $(1,1,2)$, $(-1,3,2)$, $(2,4,3)$.
(b) The plane containing the points $(-2,1,1)$ and $(0,1,2)$ and orthogonal to the plane $2x - y + z = 3$.

Exercise 10. Find equations for the following hyperplanes in $\mathbb{R}^4$.
(a) The plane parallel to the plane $2x_1 + x_2 - 3x_3 + x_4 = 2$ and containing the point $(2,1,1,3)$.
(b) The plane through the origin and orthogonal to the vector $(1,0,2,1)$.

Exercise 11. For each pair A, $\mathbf{b}$, solve the normal equations for the system $A\mathbf{x} = \mathbf{b}$ and find the residual vector and its norm. Are there any genuine solutions to the system?

(a) $\begin{bmatrix} 1 & 3 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & -2 \\ 1 & 1 \\ 3 & 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 2 & 2 \\ 1 & 1 & 0 \\ -1 & 1 & 2 \\ 1 & -2 & -3 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \\ 0 \\ 0 \end{bmatrix}$

Exercise 12. For each pair A, $\mathbf{b}$, solve the normal equations for the system $A\mathbf{x} = \mathbf{b}$ and find the residual vector and its norm. (Note: normal equations may not have unique solutions.)

(a) $\begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 2 & 0 \\ 1 & 0 & 2 \\ 1 & 2 & 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$

Exercise 13. (Linear regression) You have collected data points (x_k, y_k) that are theoretically linearly related by a line of the form $y = ax + b$. Each data point gives an equation for a and b. The collected data points are $(0,.3)$, $(1,1.1)$, $(2,2)$, $(3,3.5)$, and $(3.5,3.6)$. Write out the resulting system of 5 equations, solve the normal equations to find the line that best fits this data, and calculate the residual norm. A technology tool might be helpful.

Exercise 14. (Text retrieval) You are given the following *term-by-document* matrix, that is, a matrix whose (i,j)th entry is the number of times term t_i occurs in document D_j. Columns of this matrix are document vectors, as are queries. We measure the quality of a match between query and document by the cosine of the angle θ between the two vectors, larger cosine being better. Which of the following nine documents D_i matches the query $(0,1,0,1,1)$ above the threshold value $\cos\theta \geq 0.5$? Which is the best match to the query?

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
t_1	1	1	2	0	1	0	1	0	1
t_2	0	1	0	1	0	1	1	0	0
t_3	0	2	0	2	0	1	0	1	1
t_4	1	0	1	0	1	0	2	1	0
t_5	1	2	1	0	0	1	0	0	1

Exercise 15. Compute unit outward normals for all the faces of the tetrahedron of Example 4.17.

Exercise 16. Compute outward normals for all the faces of a tetrahedron ABCD, where coordinates of A, B, C, D are $(0,0,0)$, $(1,0,3)$, $(3,3,-2)$ and $(4,0,2)$, respectively.

*Problem 17. Show that if two vectors $\mathbf{u}$ and $\mathbf{v}$ satisfy the equation $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$, then $\mathbf{u}$ and $\mathbf{v}$ must be orthogonal.

Problem 18. Show that the CBS inequality is valid for complex vectors $\mathbf{u}$ and $\mathbf{v}$ by evaluating the nonnegative expression $\|\mathbf{u}+c\mathbf{v}\|^2$ with the complex dot product and evaluating it at $c = \|\mathbf{u}\|^2 / (\mathbf{u}\cdot\mathbf{v})$ in the case $\mathbf{u}\cdot\mathbf{v} \neq 0$.

Problem 19. Let A be an $m \times n$ real matrix and $B = A^T A$. Show the following:
(a) The matrix B is symmetric and positive semidefinite.
(b) If A has full column rank, then B is positive definite.

Problem 20. Show that if A is a real matrix and $A^T A$ is positive definite then A has full column rank.

Problem 21. In Example 4.15 two values of c are calculated: The average value and the least squares value. Calculate each resulting residual and its norm.

Problem 22. Let $\mathbf{u}$ and $\mathbf{v}$ be vectors of the same length. Show that $\mathbf{u}-\mathbf{v}$ is orthogonal to $\mathbf{u}+\mathbf{v}$. Sketch a picture in the plane and interpret it geometrically.

*Problem 23. Show that if A is a rank-one real matrix, then the normal equations with coefficient matrix A are consistent.

Problem 24. Show that if A is a complex matrix, then $A^* A$ is Hermitian and positive semidefinite.

*Problem 25. Show that Theorem 4.3 is valid for complex vectors.

Problem 26. It is hypothesized that sales of a certain product are linearly related to three factors. The sales output is quantified as z and the three factors as x_1, x_2, and x_3. Six samples are taken of the sales and the factor data. Results are contained in the following table. Does the hypothesis of a linear relationship seem reasonable? Explain your answer.

z	x_1	x_2	x_3
527	13	5	6
711	6	17	7
1291	12	16	23
625	11	13	4
1301	12	27	14
1350	5	14	31

Problem 27. Show that the volume of the parallelepiped with adjacent edges represented by vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in $\mathbb{R}^3$ is $|\mathbf{u}\times\mathbf{v}\cdot\mathbf{w}|$, the absolute value of the scalar triple product.

4.3 Orthogonal and Unitary Matrices

Orthogonal Sets of Vectors

In our discussion of bases in Section 3.3, we saw that linear independence of a set of vectors was a key idea for understanding the nature of vector spaces. One of our examples of a linearly independent set was the standard basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ of $\mathbb{R}^n$. Here $\mathbf{e}_i$ is the vector with a 1 in the ith coordinate and zeros elsewhere. In the case of geometrical vectors and $n = 3$, these are just the familiar vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$. These vectors have some particularly nice properties that go beyond linear independence. For one, each is a unit vector with respect to the standard norm. Furthermore, these vectors are mutually orthogonal to each other. These properties are so desirable that we elevate them to the status of a definition.

Definition 4.10. **Orthogonal and Orthonormal Set of Vectors** The set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in a standard vector space is said to be an *orthogonal set* if $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ whenever $i \neq j$. If, in addition, each vector has unit length, i.e., $\mathbf{v}_i \cdot \mathbf{v}_i = 1$, then the set of vectors is said to be an *orthonormal set* of vectors.

Example 4.18. Which of the following sets of vectors are orthogonal? Orthonormal? Use the standard inner product in each case.

(a)$\{(3/5, 4/5), (-4/5, 3/5)\}$ (b) $\{(1, -1, 0), (1, 1, 0), (0, 0, 1)\}$ (c) $\{(1, \mathrm{i}), (\mathrm{i}, 1)\}$

Solution. For (a) let $\mathbf{v}_1 = (3/5, 4/5)$, $\mathbf{v}_2 = (-4/5, 3/5)$ to obtain that

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \frac{-12}{25} + \frac{12}{25} = 0 \text{ and } \mathbf{v}_1 \cdot \mathbf{v}_1 = \frac{9}{25} + \frac{16}{25} = 1 = \mathbf{v}_2 \cdot \mathbf{v}_2.$$

It follows that the first set of vectors is an orthonormal set.

For (b) let $\mathbf{v}_1 = (1, -1, 0), \mathbf{v}_2 = (1, 1, 0), \mathbf{v}_3 = (0, 0, 1)$ and check that

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = 1 \cdot 1 - 1 \cdot 1 + 0 \cdot 0 = 0 \text{ and } \mathbf{v}_1 \cdot \mathbf{v}_3 = 1 \cdot 0 - 1 \cdot 0 + 0 \cdot 1 = 0 = \mathbf{v}_2 \cdot \mathbf{v}_3.$$

Hence, this set of vectors is orthogonal, but $\mathbf{v}_1 \cdot \mathbf{v}_1 = 1 \cdot 1 + (-1) \cdot (-1) + 0 = 2$, which is sufficient to show that the vectors do not form an orthonormal set.

For (c) let $\mathbf{v}_1 = (1, \mathrm{i})$, $\mathbf{v}_2 = (\mathrm{i}, 1)$ to obtain that

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \bar{1}\mathrm{i} + \bar{\mathrm{i}}1 = \mathrm{i} - \mathrm{i} = 0 \text{ and } \mathbf{v}_1 \cdot \mathbf{v}_1 = 1 + 1 = 2 = \mathbf{v}_2 \cdot \mathbf{v}_2.$$

It follows that this set is orthogonal, but not orthonormal. □

One of the principal reasons that orthogonal sets are so desirable is the following key fact, which we call the *orthogonal coordinates theorem.*

Theorem 4.6. Orthogonal Coordinates Theorem Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be an orthogonal set of nonzero vectors and suppose that $\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Then $\mathbf{v}$ can be expressed uniquely (up to order) as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, namely

$$\mathbf{v} = \frac{\mathbf{v}_1 \cdot \mathbf{v}}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 + \frac{\mathbf{v}_2 \cdot \mathbf{v}}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 + \cdots + \frac{\mathbf{v}_n \cdot \mathbf{v}}{\mathbf{v}_n \cdot \mathbf{v}_n}\mathbf{v}_n.$$

Proof. Since $\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, we know that $\mathbf{v}$ is expressible as *some* linear combination of the $\mathbf{v}_i$'s, say

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n.$$

Now we carry out a simple but wonderful trick that is used frequently with orthogonal sets, namely, take the inner product of both sides with the vector $\mathbf{v}_k$. Since $\mathbf{v}_k \cdot \mathbf{v}_j = 0$ if $j \neq k$, we obtain

$$\begin{aligned}\mathbf{v}_k \cdot \mathbf{v} &= \mathbf{v}_k \cdot (c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots \cdots + c_n\mathbf{v}_n) \\ &= c_1\mathbf{v}_k \cdot \mathbf{v}_1 + c_2\mathbf{v}_k \cdot \mathbf{v}_2 + \cdots + c_n\mathbf{v}_k \cdot \mathbf{v}_n = c_k\mathbf{v}_k \cdot \mathbf{v}_k.\end{aligned}$$

Since $\mathbf{v}_k \neq 0$, we have $\|\mathbf{v}_k\|^2 = \mathbf{v}_k \cdot \mathbf{v}_k \neq 0$, so solve for c_k to obtain that

$$c_k = \frac{\mathbf{v}_k \cdot \mathbf{v}}{\mathbf{v}_k \cdot \mathbf{v}_k}.$$

This proves that the coefficients c_k are unique and establishes the formula of the theorem. □

The vector $\dfrac{\mathbf{v}_k \cdot \mathbf{v}}{\mathbf{v}_k \cdot \mathbf{v}_k}\mathbf{v}_k$ should look familiar. In fact, it is the projection of the vector $\mathbf{v}$ along the vector $\mathbf{v}_k$. Thus, Theorem 4.6 says that any linear combination of an orthogonal set of nonzero vectors is the sum of its projections in the direction of each vector in the set.

The coefficients c_k of Theorem 4.6 are also familiar: they are the *coordinates* of $\mathbf{v}$ relative to the basis $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, so that $[\mathbf{v}]_B = (c_1, c_2, \ldots, c_n)$. This terminology was introduced in Section 3.3. Theorem 4.6 shows us that coordinates are rather easy to calculate with respect to an orthogonal basis. Contrast this with Example 3.25.

Corollary 4.1. Orthogonal Implies Linearly Independent Every orthogonal set of nonzero vectors is linearly independent.

Proof. Consider a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. If some linear combination were to have value zero, say

$$\mathbf{0} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n,$$

it would follow from the preceding theorem that

$$c_k = \frac{\mathbf{v}_k \cdot 0}{\mathbf{v}_k \cdot \mathbf{v}_k} = 0.$$

It follows from the definition of linear independence that vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent. □

Caution: The converse of the corollary is false, that is, not every linearly independent set of vectors is orthogonal.

For an example, consider the linearly independent vectors $\mathbf{v}_1 = (1, 0)$, $\mathbf{v}_2 = (1, 1)$ in $V = \mathbb{R}^2$.

Given an orthogonal set of nonzero vectors, it is easy to manufacture an orthonormal set of vectors from them. Simply replace every vector in the original set by the vector divided by its length. The formula of Theorem 4.6 simplifies very nicely if the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ form an orthonormal set (which automatically consists of nonzero vectors!), namely

$$\mathbf{v} = (\mathbf{v}_1 \cdot \mathbf{v})\ \mathbf{v}_1 + (\mathbf{v}_2 \cdot \mathbf{v})\ \mathbf{v}_2 + \cdots + (\mathbf{v}_n \cdot \mathbf{v})\ \mathbf{v}_n.$$

The following theorem gives us a nice analogue to the fact that every linearly independent set of vectors can be expanded to a basis.

Theorem 4.7. Every orthogonal set of nonzero vectors in a standard vector space can be expanded to an orthogonal basis of the space.

Proof. Suppose that we have expanded our original orthogonal set in $\mathbb{R}^n$ to the orthogonal set of nonzero vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$, where $k < n$. We show how to add one more element. This is sufficient, because by repeating this step we eventually fill up $\mathbb{R}^n$. Let $A = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]^T$ and let $\mathbf{v}_{k+1}$ be any nonzero solution to $A\mathbf{x} = \mathbf{0}$, which exists since $k < n$. This vector is orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$. □

Orthogonal and Unitary Matrices

In general, if we want to determine the coordinates of a vector $\mathbf{b}$ with respect to a certain basis of vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$, we stack the basis vectors together to form a matrix A, then solve the system $A\mathbf{x} = \mathbf{b}$ for the vector of coordinates $\mathbf{x}$ of $\mathbf{b}$ with respect to this basis. In fact, $\mathbf{x} = A^{-1}\mathbf{b}$. Now we have seen that if the basis vectors happen to form an orthonormal set, the situation is much simpler and we definitely don't have to find A^{-1}. Is this simplicity reflected in properties of the matrix A? The answer is yes and we can see this as follows: suppose that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ is an orthonormal basis of $\mathbb{R}^n$ and let $A = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$. Orthonormality says that $\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta. This means that the matrix $A^T A$, whose (i, j)th entry is $\mathbf{u}_n^T \mathbf{u}_n$, is simply $[\delta_{ij}] = I$, that is, $A^T A = I$. Now recall that Theorem 2.4 shows that a square one-sided inverse of a square matrix is really the two-sided inverse. Hence, $A^{-1} = A^T$. A similar argument works if $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ is an

orthonormal basis of $\mathbb{C}^n$ except that we use conjugate transpose instead of transpose. Matrices with these properties are important enough to be named.

Definition 4.11. Orthogonal and Unitary Matrix A square real matrix Q is called *orthogonal if* $Q^T = Q^{-1}$. A square matrix U is called *unitary* if $U^* = U^{-1}$.

One could allow orthogonal matrices to be complex as well, but these are not particularly useful for us, so in this text we will always assume that orthogonal matrices have real entries. For real matrices Q, we have $Q^* = Q^T$. Hence, we see from the definition that orthogonal matrices are exactly the real unitary matrices. The naming of orthogonal matrices is traditional in matrix theory, but a bit unfortunate because it can be a source of confusion.

Caution: Do not confuse "orthogonal vectors" and "orthogonal matrix." The objects and meanings are different.

By orthogonal *vectors* we mean a set of vectors with a certain relationship to each other, while an orthogonal *matrix* is a real matrix whose inverse is its transpose. To make matters more confusing, there actually is a close connection between the two terms, because a square matrix is orthogonal exactly when its columns form an orthonormal set.

Example 4.19. Show that the matrix $U = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{bmatrix}$ is unitary and that for any angle θ, the matrix $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ is orthogonal.

Solution. It is sufficient to check that $U^*U = I$ and $R(\theta)^T R(\theta) = I$. So we calculate

$$\begin{aligned} U^*U &= \left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{bmatrix}\right)^* \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -\mathrm{i} \\ -\mathrm{i} & 1 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1 \end{bmatrix} \\ &= \frac{1}{2}\begin{bmatrix} 1-\mathrm{i}^2 & \mathrm{i}-\mathrm{i} \\ -\mathrm{i}+\mathrm{i} & 1-\mathrm{i}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned}$$

which shows that U is unitary. For the real matrix $R(\theta)$ we have

$$\begin{aligned} R(\theta)^T R(\theta) &= \left(\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\right)^T \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos^2\theta+\sin^2\theta & \cos\theta\sin\theta-\sin\theta\cos\theta \\ -\cos\theta\sin\theta+\sin\theta\cos\theta & \sin^2\theta+\cos^2\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned}$$

which shows that $R(\theta)$ is orthogonal. □

Orthogonal and unitary matrices have a certain "rigidity" quality about them that is nicely illustrated by the rotation matrix $R(\theta)$. We first saw this matrix in Example 2.18 of Chapter 2. The effect of multiplying a vector $\mathbf{x} \in \mathbb{R}^2$ by $R(\theta)$ is to rotate the vector counterclockwise through an angle of θ. This is illustrated in Figure 2.3 of Chapter 2. In particular, angles between vectors and lengths of vectors are preserved by such a multiplication. This is no accident of $R(\theta)$, but rather a property of orthogonal and unitary matrices in general. Here is a statement of these properties for orthogonal matrices. An analogous fact holds for complex unitary matrices with vectors in $\mathbb{C}^n$.

Theorem 4.8. Let Q be an orthogonal $n \times n$ matrix and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with the standard inner (dot) product. Then

$$\|Q\mathbf{x}\| = \|\mathbf{x}\| \qquad \text{and} \qquad Q\mathbf{x} \cdot Q\mathbf{y} = \mathbf{x} \cdot \mathbf{y}.$$

Proof. We calculate the norm of $Q\mathbf{x}$:

$$\|Q\mathbf{x}\|^2 = Q\mathbf{x} \cdot Q\mathbf{x} = (Q\mathbf{x})^T Q\mathbf{x} = \mathbf{x}^T Q^T Q\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2,$$

which proves the first assertion, while similarly

$$Q\mathbf{x} \cdot Q\mathbf{y} = (Q\mathbf{x})^T Q\mathbf{y} = \mathbf{x}^T Q^T Q\mathbf{y} = \mathbf{x}^T\mathbf{y} = \mathbf{x} \cdot \mathbf{y}. \qquad \square$$

Here is another kind of orthogonal matrix that has turned out to be very useful in numerical calculations and has a very nice geometrical interpretation as well. As with rotation matrices, it gives us a simple way of forming orthogonal matrices directly without explicitly constructing an orthonormal basis. The proof that $H_{\mathbf{v}}$ is orthogonal and symmetric is left as an exercise.

Definition 4.12. Householder Matrix A matrix of the form

$$H_{\mathbf{v}} = I - 2\left(\mathbf{v}\mathbf{v}^T\right) / \left(\mathbf{v}^T\mathbf{v}\right),$$

where $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$, is called a *Householder* matrix.

Example 4.20. Let $\mathbf{v} = (3, 0, 4)$ and compute the Householder matrix $H_{\mathbf{v}}$. What is the effect of multiplying it by the vector $\mathbf{v}$?

Solution. We calculate $H_{\mathbf{v}}$ to be

$$\begin{aligned} I - \frac{2}{\mathbf{v}^T\mathbf{v}}\mathbf{v}\mathbf{v}^T &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{2}{3^2 + 4^2} \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} \begin{bmatrix} 3 & 0 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{2}{25} \begin{bmatrix} 9 & 0 & 12 \\ 0 & 0 & 0 \\ 12 & 0 & 16 \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 7 & 0 & -24 \\ 0 & 25 & 0 \\ -24 & 0 & -7 \end{bmatrix}. \end{aligned}$$

Therefore, multiplying $H_{\mathbf{v}}$ by $\mathbf{v}$ gives

$$H_{\mathbf{v}}\mathbf{v} = \frac{1}{25}\begin{bmatrix} 7 & 0 & -24 \\ 0 & 25 & 0 \\ -24 & 0 & -7 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} = \frac{1}{25}\begin{bmatrix} -75 \\ 0 \\ -100 \end{bmatrix} = -\begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} = -\mathbf{v}. \qquad \square$$

Multiplication by a Householder matrix can be thought of as a geometrical reflection that reflects the vector $\mathbf{v}$ to $-\mathbf{v}$ and leaves any vector orthogonal to $\mathbf{v}$ unchanged. This is implied by the following theorem. For a picture of this geometrical interpretation, see Figure 4.5. Notice that in this figure V is the plane perpendicular to $\mathbf{v}$ and the reflections are across this plane.

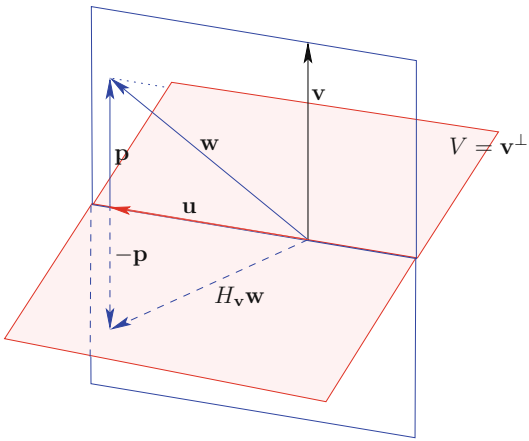


Fig. 4.5: Action of $\mathbf{H_v}$ on $\mathbf{w}$ as a reflection across the plane V perpendicular to $\mathbf{v}$.

Theorem 4.9. Let $H_{\mathbf{v}}$ be the Householder matrix defined by $\mathbf{v} \in \mathbb{R}^n$ and let $\mathbf{w} \in \mathbb{R}^n$ be written as $\mathbf{w} = \mathbf{u} + \mathbf{p}$, where $\mathbf{p}$ is the projection of $\mathbf{w}$ along $\mathbf{v}$ and $\mathbf{u} = \mathbf{w} - \mathbf{p}$. Then

$$H_{\mathbf{v}}\mathbf{w} = \mathbf{u} - \mathbf{p}.$$

Proof. With notation as in the statement of the theorem, we have $\mathbf{p} = \dfrac{\mathbf{v}^T\mathbf{w}}{\mathbf{v}^T\mathbf{v}}\mathbf{v}$, $\mathbf{w} = \mathbf{p} + \mathbf{u}$ and $\mathbf{v} \perp \mathbf{u}$ by Theorem 4.3. So we calculate that

$$H_{\mathbf{v}}\mathbf{w} = \left(I - \frac{2}{\mathbf{v}^T\mathbf{v}}\mathbf{v}\mathbf{v}^T\right)(\mathbf{p} + \mathbf{u}) = \mathbf{p} + \mathbf{u} - 2\frac{\mathbf{v}^T\mathbf{w}}{(\mathbf{v}^T\mathbf{v})^2}\mathbf{v}\mathbf{v}^T\mathbf{v} - 2\frac{\mathbf{v}^T\mathbf{w}}{\mathbf{v}^T\mathbf{v}}\mathbf{v}\mathbf{v}^T\mathbf{u}$$

$$= \mathbf{p} + \mathbf{u} - 2\frac{\mathbf{v}^T\mathbf{w}}{\mathbf{v}^T\mathbf{v}}\mathbf{v} - \mathbf{0} = \mathbf{p} + \mathbf{u} - 2\mathbf{p} = \mathbf{u} - \mathbf{p}. \qquad \square$$

Corollary 4.2. Let $\mathbf{w}, \mathbf{q} \in \mathbb{R}^n$ with $\|\mathbf{w}\| = \|\mathbf{q}\|$ and $\mathbf{v} = \mathbf{q} - \mathbf{w}$. Then $\mathbf{H_v w} = \mathbf{q}$.

Proof. We have that

$$(\mathbf{q}+\mathbf{w})\cdot\mathbf{v} = (\mathbf{q}+\mathbf{w})\cdot(\mathbf{q}-\mathbf{w}) = \mathbf{q}^T\mathbf{q}-\mathbf{q}^T\mathbf{w}+\mathbf{w}^T\mathbf{q}-\mathbf{w}^T\mathbf{w} = \|\mathbf{w}\|^2-\|\mathbf{q}\|^2 = 0$$

and

$$\mathbf{w} = \frac{1}{2}(\mathbf{q}+\mathbf{w}) - \frac{1}{2}(\mathbf{q}-\mathbf{w}).$$

It follows therefore that $(\mathbf{q}+\mathbf{w}) \perp \mathbf{v}$ and that $\mathbf{p} = -\frac{1}{2}(\mathbf{q}-\mathbf{w}) = \frac{1}{2}\mathbf{v}$ is the projection of $\mathbf{w}$ along $\mathbf{v}$. With notation as in Theorem 4.9, we set $\mathbf{u} = \mathbf{w}-\mathbf{p} = \mathbf{w}+\frac{1}{2}(\mathbf{q}-\mathbf{w}) = \frac{1}{2}(\mathbf{q}+\mathbf{w})$ and obtain $\mathbf{w} = \mathbf{p}+\mathbf{u}$. Thus, the theorem yields

$$H_{\mathbf{v}}\mathbf{w} = -\mathbf{p}+\mathbf{u} = \frac{1}{2}(\mathbf{q}-\mathbf{w}) + \frac{1}{2}(\mathbf{q}+\mathbf{w}) = \mathbf{q}. \qquad \square$$

Corollary 4.2 enables us to map a nonzero vector to any other vector of the same length by way of an orthogonal transformation, as in the following example.

Example 4.21. Find an orthogonal matrix that maps the vector $\mathbf{w} = (3,0,4)$ to a multiple of $(0,1,0)$ and confirm that it works.

Solution. We have $\|\mathbf{w}\|^2 = 3^2+4^2 = 25 = 5^2$, so set $\mathbf{q} = (0,5,0)$ and apply Corollary 4.2 with $\mathbf{v} = \mathbf{q}-\mathbf{w} = (-3,5,-4)$. We calculate $H_{\mathbf{v}}$ to be

$$I - \frac{2}{\mathbf{v}^T\mathbf{v}}\mathbf{v}\mathbf{v}^T = \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \frac{2}{(-3)^2+5^2+(-4)^2}\begin{bmatrix}-3\\5\\-4\end{bmatrix}\begin{bmatrix}-3&5&-4\end{bmatrix}$$
$$= \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} - \frac{1}{25}\begin{bmatrix}9&-15&12\\-15&25&-20\\12&-20&16\end{bmatrix} = \frac{1}{25}\begin{bmatrix}16&15&-12\\15&0&20\\-12&20&9\end{bmatrix}.$$

Therefore, multiplying $H_{\mathbf{v}}$ by $\mathbf{w}$ gives

$$H_{\mathbf{v}}\mathbf{w} = \frac{1}{25}\begin{bmatrix}16&15&-12\\15&0&20\\-12&20&9\end{bmatrix}\begin{bmatrix}3\\0\\4\end{bmatrix} = \frac{1}{25}\begin{bmatrix}0\\125\\0\end{bmatrix} = \begin{bmatrix}0\\5\\0\end{bmatrix}. \qquad \square$$

Example 4.22. Let $\mathbf{v} = (3,0,4)$ and $H_{\mathbf{v}}$ the corresponding Householder matrix (as in Example 4.20). The columns of this matrix form an orthonormal basis for the space $\mathbb{R}^3$. Find the coordinates of the vector $\mathbf{w} = (2,1,-4)$ relative to this basis.

Solution. We have already calculated $H_{\mathbf{v}} = [\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3]$ in Example 4.20. The vector $\mathbf{c} = (c_1,c_2,c_3)$ of coordinates of $\mathbf{w}$ must satisfy the equations

$$\mathbf{w} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3 = H_{\mathbf{v}}\mathbf{c}.$$

Since $H_{\mathbf{v}}$ is orthogonal, it follows that

$$\mathbf{c} = H_{\mathbf{v}}^{-1}\mathbf{w} = H_{\mathbf{v}}^{T}\mathbf{w} = H_{\mathbf{v}}\mathbf{w} = \frac{1}{25}\begin{bmatrix} 7 & 0 & -24 \\ 0 & 25 & 0 \\ -24 & 0 & -7 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \\ -4 \end{bmatrix} = \begin{bmatrix} 4.4 \\ 1.0 \\ -0.8 \end{bmatrix}. \quad \square$$

Usually we work with real Householder matrices. Occasionally, complex numbers are a necessary part of the scenery. In such situations we can define the *complex* Householder matrix by the formula $H_{\mathbf{v}} = I - 2(\mathbf{v}\mathbf{v}^*)/(\mathbf{v}^*\mathbf{v})$. The projection formula (Theorem 4.3) remains valid for complex vectors, so that the proof of Theorem 4.9 carries over to complex vectors provided that we replace all transposes by conjugate transposes. Also, $H_{\mathbf{v}} = H_{\mathbf{v}}^* = H_{\mathbf{v}}^{-1}$ is unitary.

Our next example is to generate orthogonal matrices with specified columns.

Example 4.23. Find orthogonal matrices with these orthonormal vectors as columns: (a) $\frac{1}{\sqrt{3}}(1,1,1)$ (b) $\frac{1}{3}(1,2,2,0)$, $\frac{1}{3}(-2,1,0,2)$

Solution. For (a), set $\mathbf{u}_1 = \frac{1}{\sqrt{3}}(1,1,1)$, and we see by inspection that a second orthonormal vector is $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1,-1,0)$. To obtain a third, take the cross product $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2 = \frac{1}{\sqrt{6}}(1,1,-2)$. This vector is orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_2$ and has unit length. Hence, the desired matrix is

$$P = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3] = \frac{1}{\sqrt{6}}\begin{bmatrix} \sqrt{2} & \sqrt{3} & 1 \\ \sqrt{2} & -\sqrt{3} & 1 \\ \sqrt{2} & 0 & -2 \end{bmatrix}.$$

To keep the arithmetic simple in (b), form the system $A\mathbf{x} = \mathbf{0}$ where the rows of A are $(1,2,2,0)$ and $(-2,1,0,2)$. These are nonzero orthogonal vectors. Solve the system to get a general solution (the reader should check this) $\mathbf{x} = \left(-\frac{2}{5}x_3 + \frac{4}{5}x_4, -\frac{4}{5}x_3 - \frac{2}{5}x_4, x_3, x_4\right)$. So take $x_3 = 5$, $x_4 = 0$ and get a particular solution $(-2,-4,5,0)$. Take $x_3 = 0$, $x_4 = 5$ and get a particular solution $(4,-2,0,5)$. Normalize all four vectors to obtain the desired orthogonal matrix

$$P = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4] = \frac{1}{3\sqrt{5}}\begin{bmatrix} \sqrt{5} & -2\sqrt{5} & -2 & 4 \\ 2\sqrt{5} & \sqrt{5} & -4 & -2 \\ 2\sqrt{5} & 0 & 5 & 0 \\ 0 & 2\sqrt{5} & 0 & 5 \end{bmatrix}. \quad \square$$

Orthogonal bases have some very pleasant properties, such as easy coordinate calculations. So given a linearly independent set $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ of vectors, we would like a straightforward algorithm to turn this into an orthogonal basis. The tool we need is the Gram–Schmidt algorithm.

Theorem 4.10. Gram–Schmidt Algorithm Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ be linearly independent vectors in a standard space. Define vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ recursively by the formula

$$\mathbf{v}_k = \mathbf{w}_k - \frac{\mathbf{v}_1 \cdot \mathbf{w}_k}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 - \frac{\mathbf{v}_2 \cdot \mathbf{w}_k}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 - \cdots - \frac{\mathbf{v}_{k-1} \cdot \mathbf{w}_k}{\mathbf{v}_{k-1} \cdot \mathbf{v}_{k-1}}\mathbf{v}_{k-1}, \quad k = 1, \ldots, n.$$

Then

(1) The vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ form an orthogonal set.
(2) For each index $k = 1, \ldots, n$,

$$\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\} = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}.$$

Proof. In the case $k = 1$, we have that the single vector $\mathbf{v}_1 = \mathbf{w}_1$ is an orthogonal set and so $\operatorname{span}\{\mathbf{w}_1\} = \operatorname{span}\{\mathbf{v}_1\}$. Now suppose that for some index $k > 1$ we have shown that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}$ is an orthogonal set such that $\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k-1}\} = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}\}$. Then it is true that $\mathbf{v}_r \cdot \mathbf{v}_s = 0$ for any distinct indices r, s both less than k. Take the inner product of $\mathbf{v}_k$, as given by the formula above, with the vector $\mathbf{v}_j$, where $j < k$, and we obtain

$$\begin{aligned}
\mathbf{v}_j \cdot \mathbf{v}_k &= \mathbf{v}_j \cdot \left(\mathbf{w}_k - \frac{\mathbf{v}_1 \cdot \mathbf{w}_k}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 - \frac{\mathbf{v}_2 \cdot \mathbf{w}_k}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 - \cdots - \frac{\mathbf{v}_{k-1} \cdot \mathbf{w}_k}{\mathbf{v}_{k-1} \cdot \mathbf{v}_{k-1}}\mathbf{v}_{k-1} \right) \\
&= \mathbf{v}_j \cdot \mathbf{w}_k - \mathbf{v}_j \cdot \mathbf{v}_1 \frac{\mathbf{v}_j \cdot \mathbf{w}_k}{\mathbf{v}_1 \cdot \mathbf{v}_1} - \cdots - \mathbf{v}_j \cdot \mathbf{v}_{k-1} \frac{\mathbf{v}_{k-1} \cdot \mathbf{w}_k}{\mathbf{v}_{k-1} \cdot \mathbf{v}_{k-1}} \\
&= \mathbf{v}_j \cdot \mathbf{w}_k - \mathbf{v}_j \cdot \mathbf{v}_j \frac{\mathbf{v}_j \cdot \mathbf{w}_k}{\mathbf{v}_j, \cdot \mathbf{v}_j} = 0.
\end{aligned}$$

It follows that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ is an orthogonal set. The Gram–Schmidt formula show us that one of $\mathbf{v}_k$ or $\mathbf{w}_k$ can be expressed as a linear combination of the other and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}$. Therefore,

$$\begin{aligned}
\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k-1}, \mathbf{w}_k\} &= \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}, \mathbf{w}_k\} \\
&= \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}, \mathbf{v}_k\},
\end{aligned}$$

which is the second part of the theorem. Repeat this argument for each index $k = 2, \ldots, n$ to complete the proof of the theorem. □

The Gram–Schmidt formula is easy to remember: subtract from the vector $\mathbf{w}_k$ all of the projections of $\mathbf{w}_k$ along the directions $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}$ to obtain the vector $\mathbf{v}_k$.

Example 4.24. Let $V = \mathcal{C}(A)$ with the standard inner product and compute an orthonormal basis of V, where

$$A = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 1 & -1 & 3 & 2 \\ 1 & -1 & 3 & 2 \\ -1 & 1 & -3 & 1 \end{bmatrix}.$$

Solution. We know that V is spanned by the four columns of A. However, the Gram–Schmidt algorithm requests a basis of V and we don't know that the columns are linearly independent. We leave it to the reader to check that the reduced row echelon form of A is the matrix

$$R = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

It follows from the column space algorithm that columns 1, 2, and 4 of the matrix A yield a basis of V. So let $\mathbf{w}_1 = (1, 1, 1, -1)$, $\mathbf{w}_2 = (2, -1, -1, 1)$, $\mathbf{w}_3 = (-1, 2, 2, 1)$, and apply the Gram–Schmidt algorithm to obtain

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{w}_1 = (1, 1, 1, -1), \\
\mathbf{v}_2 &= \mathbf{w}_2 - \frac{\mathbf{v}_1 \cdot \mathbf{w}_2}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 \\
&= (2, -1, -1, 1) - \frac{-1}{4}(1, 1, 1, -1) = \frac{1}{4}(9, -3, -3, 3), \\
\mathbf{v}_3 &= \mathbf{w}_3 - \frac{\mathbf{v}_1 \cdot \mathbf{w}_3}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 - \frac{\mathbf{v}_2 \cdot \mathbf{w}_3}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 \\
&= (-1, 2, 2, 1) - \frac{2}{4}(1, 1, 1, -1) - \frac{-18}{108}(9, -3, -3, 3) \\
&= \frac{1}{4}(-4, 8, 8, 4) - \frac{1}{4}(2, 2, 2, -2) + \frac{1}{4}(6, -2, -2, 2) = (0, 1, 1, 2).
\end{aligned}$$

Finally, normalize each vector to obtain the orthonormal basis

$$\begin{aligned}
\mathbf{u}_1 &= \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} = \frac{1}{2}(1, 1, 1, -1), \\
\mathbf{u}_2 &= \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} = \frac{1}{\sqrt{108}}(9, -3, -3, 3) = \frac{1}{2\sqrt{3}}(3, -1, -1, 1), \\
\mathbf{u}_3 &= \frac{\mathbf{v}_3}{\|\mathbf{v}_3\|} = \frac{1}{\sqrt{6}}(0, 1, 1, 2).
\end{aligned}$$

□

For hand calculations these observations are useful:

- If one encounters an inconvenient fraction, such as the $\frac{1}{4}$ in $\mathbf{v}_2$, replace the calculated $\mathbf{v}_2$ by $4\mathbf{v}_2$, thereby eliminating the fraction, and yet achieving the same results in subsequent calculations. The idea here is that for any nonzero scalar c,
$$\frac{\mathbf{v}_2 \cdot \mathbf{w}}{\mathbf{v}_2, \cdot \mathbf{v}}\mathbf{v}_2 = \frac{c\mathbf{v}_2 \cdot \mathbf{w}}{c\mathbf{v}_2 \cdot c\mathbf{v}_2}c\mathbf{v}_2.$$
So we could have replaced $\frac{1}{4}(9, -3, -3, 3)$ by $(3, -1, -1, 1)$ and achieved the same results.
- The same remark applies to the normalizing process, since in general,
$$\frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} = \frac{c\mathbf{v}_2}{\|c\mathbf{v}_2\|}.$$

The Gram–Schmidt algorithm is capable of handling linearly dependent spanning sets gracefully, provided that all zero vectors are discarded. We illustrate this fact with the following example:

Example 4.25. Suppose we had used all the columns of A in Example 4.24 instead of linearly independent ones, labeling them $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4$. How would the Gram–Schmidt calculation work out?

Solution. Everything would have proceeded as above until we reached the calculation of $\mathbf{v}_3$, which would then yield

$$\begin{aligned}\mathbf{v}_3 &= \mathbf{w}_3 - \frac{\mathbf{v}_1 \cdot \mathbf{w}_3}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 - \frac{\mathbf{v}_2 \cdot \mathbf{w}_3}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 \\ &= (0,3,3,-3) - \frac{9}{4}(1,1,1,-1) + \frac{1}{4}(9,-3,-3,3) = (0,0,0,0).\end{aligned}$$

This tells us that $\mathbf{w}_3$ is a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$, which mirrors the fact that $\mathbf{w}_3$ is a linear combination of $\mathbf{w}_1$ and $\mathbf{w}_2$. Now discard $\mathbf{v}_3$ and continue the calculations to get that

$$\begin{aligned}\mathbf{v}_4 &= \mathbf{w}_4 - \frac{\mathbf{v}_1 \cdot \mathbf{w}_4}{\mathbf{v}_1, \cdot \mathbf{v}}\mathbf{v}_1 - \frac{\mathbf{v}_2 \cdot \mathbf{w}_4}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 \\ &= (-1,2,2,1) - \frac{2}{4}(1,1,1,-1) - \frac{-18}{108}(9,-3,-3,3) = (0,1,1,2)\,. \qquad \square\end{aligned}$$

Interestingly enough, this calculation yields the same third vector that we obtained in Example 4.24. The upshot is that Gram–Schmidt can be applied to any spanning set, provided that any zero vectors that result from this calculation are discarded. The net result is still an orthogonal basis.

4.3 Exercises and Problems

Exercise 1. Determine whether the following sets of vectors are orthogonal, orthonormal, or linearly independent.
(a) $(1,-1,2)$, $(2,2,0)$ (b) $(3,-1,1)$, $(1,2,-1)$, $(2,-1,0)$ (c) $\frac{1}{5}(3,4)$, $\frac{1}{5}(4,-3)$

Exercise 2. Determine whether these sets are orthogonal or orthonormal. If orthogonal but not orthonormal, normalize the set to form an orthonormal set.
(a) $(2,-3,2,1)$, $(2,1,-1,1)$ (b) $\frac{1}{3}(2,2,1)$, $\frac{1}{\sqrt{5}}(1,0,-2)$ (c) $(1+\mathrm{i},-1)$, $(1,1-\mathrm{i})$

Exercise 3. Let $\mathbf{v}_1 = (1,1,0)$, $\mathbf{v}_2 = (-1,1,1)$, and $\mathbf{v}_3 = \frac{1}{2}(1,-1,2)$. Show that this set is an orthogonal basis of $\mathbb{R}^3$ and find the coordinates of the following vectors $\mathbf{v}$ with respect to this basis by using the orthogonal coordinates theorem.
(a) $(1,2,-2)$ (b) $(1,0,0)$ (c) $(4,-3,2)$

Exercise 4. Let $\mathbf{v}_1 = (-1, 1, 1)$ and $\mathbf{v}_2 = (1, -1, 2)$. Determine whether each of the following vectors $\mathbf{v}$ is in $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ by testing the orthogonal coordinates theorem (if $\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ then Theorem 4.6 should yield an equality).

(a) $(1, -1, 8)$ (b) $(-2, 1, 3)$ (c) $(-4, 4, 1)$

Exercise 5. Determine whether the following matrices are orthogonal or unitary and if so, find their inverse.

(a) $\frac{1}{5}\begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}$ (b) $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 & -1 \\ 0 & \sqrt{2} & 0 \\ -1 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{bmatrix}$

(d) $\frac{1}{2}\begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}$ (e) $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2}\mathrm{i} & 0 \\ \mathrm{i} & 0 & -\mathrm{i} \end{bmatrix}$ (f) $\frac{1}{\sqrt{3}}\begin{bmatrix} 1+\mathrm{i} & \mathrm{i} \\ \mathrm{i} & 1-\mathrm{i} \end{bmatrix}$

Exercise 6. Find the coordinates of the following vectors with respect to the basis of column vectors of the corresponding matrices of Exercise 5.

(a) $(2, 4)$ (b) $(3, 1, 1)$ (c) $(4, -3, 1)$
(d) $(3, -2, 4, 1)$ (e) $(\mathrm{i}, -2, 1)$ (f) $(1, 2)$

Exercise 7. Let $\mathbf{u} = (1, 2, -2)$, $\mathbf{w} = (3, 0, 0)$, and $\mathbf{v} = \mathbf{u} - \mathbf{w}$. Construct the Householder matrix $H_{\mathbf{v}}$ and calculate $H_{\mathbf{v}}\mathbf{u}$ and $H_{\mathbf{v}}\mathbf{w}$.

Exercise 8. Find a matrix reflecting vectors in $\mathbb{R}^3$ across the plane $x+y+z = 0$.

Exercise 9. Find orthogonal or unitary matrices that include the following orthonormal vectors in their columns.

(a) $\frac{1}{\sqrt{6}}(1, 2, -1)$, $\frac{1}{\sqrt{3}}(-1, 1, 1)$ (b) $\frac{1}{5}(-4, 3)$ (c) $(0, \mathrm{i})$

Exercise 10. Repeat Exercise 9 for these vectors.

(a) $\frac{1}{3}(1, 2, -2)$ (b) $\frac{1}{2}(1, 1, -1, -1)$, $\frac{1}{2}(1, -1, 1, -1)$ (c) $\frac{1}{2}(1 + \mathrm{i}, 1 - \mathrm{i})$

Exercise 11. Let $P = \frac{1}{2}\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$. Verify that P is a *projection matrix*, that is, $P^T = P$ and $P^2 = P$. Also verify that that $R = I - 2P$ is a *reflection matrix*, that is, R is a symmetric orthogonal matrix.

Exercise 12. Let $R = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ and $P = \frac{1}{2}(I - R)$. Verify that R is a reflection matrix and P is a projection matrix.

Exercise 13. Apply the Gram–Schmidt algorithm to the columns of the following matrices in left-to-right order.

(a) $\begin{bmatrix} 1 & -1 & 1 \\ 1 & 2 & 3 \\ -1 & 2 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 4 \\ 1 & 2 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

Exercise 14. Apply the Gram–Schmidt algorithm to the following vectors:
(a) $(1,-2,0)$, $(0,1,1)$, $(1,0,2)$.
(b) $(1,0,0)$, $(1,1,0)$, $(1,1,1)$.
(c) $(1,1,-1,-1)$, $(0,1,1,0)$, $(2,2,1,0)$, $(1,2,3,1)$.

Problem 15. Show that if the real $n \times n$ matrix M is invertible and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are orthogonal, then so are $M\mathbf{u}$ and $\left(M^T\right)^{-1}\mathbf{v}$. What does this imply for orthogonal matrices?

*Problem 16. Show that if P is an orthogonal matrix, then $e^{i\theta}P$ is a unitary matrix for any real θ.

Problem 17. Let P be a real projection matrix and $R = I - 2P$. Prove that R is a reflection matrix. (See Exercise 11 for definitions.)

Problem 18. Let R be a reflection matrix. Prove that $P = \frac{1}{2}(I - R)$ is a projection matrix.

Problem 19. Prove that every Householder matrix is a reflection matrix.

Problem 20. Show that the product of orthogonal matrices is orthogonal, and by example that the sum need not be orthogonal.

Problem 21. Let the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ be defined by the formula $y = f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, where A is a real matrix. Suppose that an orthogonal change of variablesis made in the domain, say $\mathbf{x} = Q\mathbf{x}'$, where Q is orthogonal. Show that in the new coordinates $y = \mathbf{x}'^T(Q^T AQ)\mathbf{x}'$.

4.4 *Applications and Computational Notes

The QR Factorization

We are going to use orthogonality ideas to develop one more way of solving the linear system $A\mathbf{x} = \mathbf{b}$, where the $m \times n$ real matrix A has full column rank. In fact, if the system is inconsistent, then this method will find the unique least squares solution to the system. Here is the basic idea: express the matrix A in the form $A = QR$, where the columns of the $m \times n$ matrix Q are orthonormal vectors and the $n \times n$ matrix R is upper triangular with nonzero

diagonal entries. Such a factorization of A is called a *QR factorization* of A. It follows that the product Q^TQ is equal to I_n. Now multiply both sides of the linear system on the left by Q^T to obtain that

$$Q^TA\mathbf{x} = Q^TQR\mathbf{x} = IR\mathbf{x} = R\mathbf{x} = Q^Tb.$$

The net result is a simple square system with a triangular matrix, which we can solve by back solving. That is, we use the last equation to solve for x_n, then the next to the last to solve for x_{n-1}, and so forth. This is the back solving phase of Gaussian elimination as we first learned it in Chapter 1, before we were introduced to Gauss–Jordan elimination.

One has to wonder why we have any interest in such a factorization, since we already have Gauss–Jordan elimination for system solving. Furthermore, it can be shown that finding a QR factorization is harder by a factor of about 2, that is, requires about twice as many floating-point operations to accomplish. So why bother? There are many answers. For one, it can be shown that using the QR factorization has an advantage of higher accuracy than Gauss–Jordan elimination in certain situations. For another, QR factorization gives us a method for solving least squares problems. We'll see an example of this method at the end of this section.

Where can we find such a factorization? As a matter of fact, we already have the necessary tools, compliments of the Gram–Schmidt algorithm. To explain matters, let's suppose that we have a matrix $A = [\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3]$ with linearly independent columns. Application of the Gram–Schmidt algorithm leads to orthogonal vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ by the following formulas:

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{w}_1 \\
\mathbf{v}_2 &= \mathbf{w}_2 - \frac{\mathbf{v}_1 \cdot \mathbf{w}_2}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 \\
\mathbf{v}_3 &= \mathbf{w}_3 - \frac{\mathbf{v}_1 \cdot \mathbf{w}_3}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 - \frac{\mathbf{v}_2 \cdot \mathbf{w}_3}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2.
\end{aligned}$$

Next, solve for $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ in the above equations to obtain

$$\begin{aligned}
\mathbf{w}_1 &= \mathbf{v}_1 \\
\mathbf{w}_2 &= \frac{\mathbf{v}_1 \cdot \mathbf{w}_2}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 + \mathbf{v}_2 \\
\mathbf{w}_3 &= \frac{\mathbf{v}_1 \cdot \mathbf{w}_3}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 + \frac{\mathbf{v}_2 \cdot \mathbf{w}_3}{\mathbf{v}_2 \cdot \mathbf{v}_2}\mathbf{v}_2 + \mathbf{v}_3.
\end{aligned}$$

In matrix form, these equations become

$$A = [\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3] = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \begin{bmatrix} 1 & \frac{\mathbf{v}_1 \cdot \mathbf{w}_2}{\mathbf{v}_1 \cdot \mathbf{v}_1} & \frac{\mathbf{v}_1 \cdot \mathbf{w}_3}{\mathbf{v}_1 \cdot \mathbf{v}_1} \\ 0 & 1 & \frac{\mathbf{v}_2 \cdot \mathbf{w}_3}{\mathbf{v}_2 \cdot \mathbf{v}_2} \\ 0 & 0 & 1 \end{bmatrix}.$$

Now normalize the $\mathbf{v}_j$'s by setting $\mathbf{q}_j = \mathbf{v}_j / \|\mathbf{v}_j\|$ and observe that

$$A = [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3] \begin{bmatrix} \|\mathbf{v}_1\| & 0 & 0 \\ 0 & \|\mathbf{v}_2\| & 0 \\ 0 & 0 & \|\mathbf{v}_3\| \end{bmatrix} \begin{bmatrix} 1 & \frac{\mathbf{v}_1 \cdot \mathbf{w}_2}{\mathbf{v}_1 \cdot \mathbf{v}_1} & \frac{\mathbf{v}_1 \cdot \mathbf{w}_3}{\mathbf{v}_1 \cdot \mathbf{v}_1} \\ 0 & 1 & \frac{\mathbf{v}_2 \cdot \mathbf{w}_3}{\mathbf{v}_2, \mathbf{v}_2} \\ 0 & 0 & 1 \end{bmatrix}$$
$$= [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3] \begin{bmatrix} \|\mathbf{v}_1\| & \frac{\mathbf{v}_1 \cdot \mathbf{w}_2}{\|\mathbf{v}_1\|} & \frac{\mathbf{v}_1 \cdot \mathbf{w}_3}{\|\mathbf{v}_1\|} \\ 0 & \|\mathbf{v}_2\| & \frac{\mathbf{v}_2 \cdot \mathbf{w}_3}{\|\mathbf{v}_2\|} \\ 0 & 0 & \|\mathbf{v}_3\| \end{bmatrix}.$$

This gives our QR factorization, which can be alternatively written as

$$A = [\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3] = [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3] \begin{bmatrix} \|\mathbf{v}_1\| & \mathbf{q}_1 \cdot \mathbf{w}_2 & \mathbf{q}_1 \cdot \mathbf{w}_3 \\ 0 & \|\mathbf{v}_2\| & \mathbf{q}_2 \cdot \mathbf{w}_3 \\ 0 & 0 & \|\mathbf{v}_3\| \end{bmatrix} = QR.$$

In general, the columns of A are linearly independent exactly when A has full column rank. It is easy to see that the argument we have given extends to any such matrix, so we have the following theorem.

Theorem 4.11. QR Factorization If A is an $m \times n$ full-column-rank matrix, then $A = QR$, where the columns of the $m \times n$ matrix Q are orthonormal vectors and the $n \times n$ matrix R is upper triangular with nonzero diagonal entries.

Example 4.26. Let the full-column-rank matrix A be given as

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 2 \\ 1 & -1 & 2 \\ -1 & 1 & 1 \end{bmatrix}.$$

Find a QR factorization of A and use this to find the least squares solution to the problem $A\mathbf{x} = \mathbf{b}$, where $\mathbf{b} = (1, 1, 1, 1)$. What is the norm of the residual $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ in this problem?

Solution. Notice that the columns of A are just the vectors $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ of Example 4.24. Furthermore, the vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ calculated in that example are just the $\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3$ that we require. Thus, we have from those calculations that

$$\|\mathbf{v}_1\| = \|(1, 1, 1, -1)\| = 2 \text{ and } \mathbf{q}_1 = \frac{1}{2}(1, 1, 1, -1),$$
$$\|\mathbf{v}_2\| = \left\|\frac{1}{4}(9, -3, -3, 3)\right\| = \frac{3}{2}\sqrt{3} \text{ and } \mathbf{q}_2 = \frac{1}{2\sqrt{3}}(3, -1, -1, 1),$$
$$\|\mathbf{v}_3\| = \|(0, 1, 1, 2)\| = \sqrt{6} \text{ and } \mathbf{q}_3 = \frac{1}{\sqrt{6}}(0, 1, 1, 2).$$

Now we calculate

$$\langle \mathbf{q}_1, \mathbf{w}_2 \rangle = \frac{1}{2}(1,1,1,-1) \cdot (2,-1,-1,1) = -\frac{1}{2}$$
$$\langle \mathbf{q}_1, \mathbf{w}_3 \rangle = \frac{1}{2}(1,1,1,-1) \cdot (-1,2,2,1) = 1$$
$$\langle \mathbf{q}_2, \mathbf{w}_3 \rangle = \frac{1}{2\sqrt{3}}(3,-1,-1,1) \cdot (-1,2,2,1) = -\sqrt{3}.$$

It follows that

$$A = \begin{bmatrix} 1/2 & 3/(2\sqrt{3}) & 0 \\ 1/2 & -1/(2\sqrt{3}) & 1/\sqrt{6} \\ 1/2 & -1/(2\sqrt{3}) & 1/\sqrt{6} \\ -1/2 & 1/(2\sqrt{3}) & 2/\sqrt{6} \end{bmatrix} \begin{bmatrix} 2 & -1/2 & 1 \\ 0 & \frac{3}{2}\sqrt{3} & -\sqrt{3} \\ 0 & 0 & \sqrt{6} \end{bmatrix} = QR.$$

Solving the system $R\mathbf{x} = Q^T\mathbf{b}$, where $\mathbf{b} = (1,1,1,1)$, by hand is rather tedious even though the system is a simple triangular one. We leave the detailed calculations to the reader. Better yet, use a technology tool to obtain the solution $\mathbf{x} = \left(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}\right)$. Thus,

$$\mathbf{r} = \mathbf{b} - A\mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 2 \\ 1 & -1 & 2 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

It follows that the system $A\mathbf{x} = \mathbf{b}$ is actually consistent, since the least squares solution turns out to be a genuine solution to the problem. □

Does this method really solve least squares problems? It does, and to see why, observe that with the above notation we have $A^T = (QR)^T = R^TQ^T$, so that the normal equations for the system $A\mathbf{x} = \mathbf{b}$, given by $A^TA\mathbf{x} = A^T\mathbf{b}$, become

QR as Least Squares Solver

$$A^TA\mathbf{x} = R^TQ^TQR\mathbf{x} = R^TIR\mathbf{x} = R^TR\mathbf{x} = A^T\mathbf{b} = R^TQ^T\mathbf{b}.$$

But the triangular matrix R is invertible because its diagonal entries are nonzero; cancel it and obtain that the normal equations are equivalent to $R\mathbf{x} = Q^T\mathbf{b}$, which is exactly what the method we have described solves.

A Practical Algorithm for the QR Factorization

In the preceding section we saw that the QR factorization can be used to solve systems including least squares. We also saw the factorization as a consequence of the Gram–Schmidt algorithm. However, the *classical* Gram–Schmidt algorithm that we have presented has certain numerical stability problems when used in practice. There is another approach to QR factorization that uses the Householder matrices we introduced in Section 4.3. It is more efficient and stable than Gram–Schmidt. If you use a technology tool to findthe QR factorization of a matrix, it is likely that this is the method used by the system.

The basic idea behind this Householder QR is to use a succession of Householder matrices to zero out the lower triangle of a matrix, one column at a time. The key fact about Householder matrices is Corollary 4.2, which says that any two vectors of the same length can be reflected to each other by way of a Householder matrix. Thus, we have a tool for massively zeroing out entries in a vector of the form $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Set $\mathbf{y} = (\pm \|\mathbf{x}\|, 0, \ldots, 0)$ and $\mathbf{v} = \mathbf{y} - \mathbf{x}$ to construct Householder H such that $H_{\mathbf{v}}\mathbf{x} = \mathbf{y}$. It is standard to choose the $\pm$ to be the negative of the sign of x_1. In this way, the first term will not cause any loss of accuracy to subtractive cancellation. We can extend this idea to zeroing out lower parts of $\mathbf{x}$ only, say

$$\mathbf{x} = \begin{bmatrix} \mathbf{z} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{z} \\ \times \\ \times \\ \times \end{bmatrix} \text{ by using } \mathbf{y} = \begin{bmatrix} \mathbf{z} \\ \pm\|\mathbf{w}\| \\ 0 \\ 0 \end{bmatrix} \text{ so } \mathbf{v} = \begin{bmatrix} \mathbf{0} \\ \times \\ \times \\ \times \end{bmatrix} \text{ and } H_{\mathbf{v}}\mathbf{x} = \begin{bmatrix} \mathbf{z} \\ \times \\ 0 \\ 0 \end{bmatrix}.$$

Note that the work of computing $H_{\mathbf{v}}\mathbf{x}$ is reduced as the size of $\mathbf{z}$ increases. The details are left as an exercise. We can apply this idea to systematically zero out subdiagonal entries by successive multiplication by Householder (hence orthogonal) matrices; schematically we have this representation of a full-rank $m \times n$ matrix A:

$$A = \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix} \xrightarrow{H_1} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{H_2} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{H_3} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{bmatrix} = R,$$

so that $H_3H_2H_1A = R$. Now we can check easily from the definition of a Householder matrix H that $H^T = H = H^{-1}$. Thus, if we set $Q = H_1^{-1}H_2^{-1}H_3^{-1} = H_1H_2H_3$, it follows that $A = QR$. Notice that we don't actually have to carry out the multiplications to compute Q unless they are needed, and the vectors needed to define these Householder matrices are themselves easily stored in a single matrix. What we have here is just a bit different from the QR factorization discussed in the last section. Here the matrix Q is a full $m \times m$ matrix and R is the same size as A. Even if A is not of full column rank, this procedure will work, provided we simply skip construction of H in the case that there are no nonzero elements to zero out in some column. Consequently, we have essentially proved the following theorem, which is sometimes called a *full* QR factorization, in contrast to the *reduced* QR factorization of Theorem 4.11.

Theorem 4.12. Full QR Factorization Let A be a real $m \times n$ matrix. Then there exist an $m \times m$ orthogonal matrix Q and $m \times n$ upper triangular matrix R such that $A = QR$.

All of the results we have discussed regarding QR factorization work for complex matrices, provided we use unitary matrices and conjugate transpose.

Data Compression and the Haar Wavelet Transform

Our goal here is to illustrate a compression method for graphics images. Data compression is a fundamental tool in the efficient transmission of information. For example, such methodologies lie behind graphics formats such as the widely used JPEG standard. Moreover, it turn out that these tools can be used for other problems such as edge detection in images as well. This is a very large area of current activity and a full treatment requires delving into topics such as Fourier series and transforms. We shall only consider a fairly simple version of these tools, and our treatment of them is an compacted variation on the very full development of this subject in the textbook Discrete Wavelet Transformations [10] by Patrick van Fleet.

Let's return to the ideas of DSP introduced in Section 2.8 of Chapter 2. We begin with what could be considered the simplest possible smoothing digital filter of the bi-infinite data sequence $\{x_k\}_{k=-\infty}^{\infty}$, namely, averaging terms:

$$y_k = \frac{1}{2}\left(x_k + x_{k-1}\right), \quad k \in \mathbb{Z}. \tag{4.4}$$

Suppose we restrict ourselves to a finite sequence of N data points $\{x_k\}_{k=1}^{N}$ and apply the filter to it. If we pad the sequence with $x_0 = x_1$ for equation (4.4), the output consists of as many points as we input. Yet we have actually lost some information in the sense that we cannot recover the original signal from the smoothed output $\{y_k\}_{k=1}^{N}$. There is a simple solution to this problem: In parallel, apply the unsmoothing filter

$$z_k = \frac{1}{2}\left(x_k - x_{k-1}\right), \quad k = 1, \ldots, N. \tag{4.5}$$

Now we can fully recover the data because $y_k + z_k = x_k$, $k = 1, ..., N$. To visualize the action of these filters, consider the following example.

Example 4.27. Sample the curve $g(t) = \frac{3}{2} + \cos\left(\frac{\pi}{4}t\right) - \frac{1}{4}\cos\left(\frac{7\pi}{4}t\right)$, $0 \le t \le 8$ at times $t_k = k/5$, $k = 0, 1, \ldots, 40$, to obtain data points $x_k = g(t_k)$ and plot (in dot-to-dot format) the data x_k along with the y_k and z_k obtained by the filters of equations (4.4) and (4.5).

Solution. See Figure 4.6 for the results of these calculations. Notice that the graph of the z_k's is positive where the y_k's are below the x_k's and negative where the y_k's are above the x_k's, which confirms that $x_k = y_k + z_k$. □

There is a curious inefficiency about the actions of the two filters in equations (4.4) and (4.5): In order to keep full information on the filtered data x_k, $k = 1, 2, \ldots, N$, we must keep $2N$ bits of data y_k, z_k, $k = 1, 2, \ldots, N$, which doubles the amount of storage the original data required. However, there is a clever way around this issue. Notice that we only used the fact that $x_k = y_k + z_k$, but there is another bit of arithmetic that we have neglected to mention: $x_{k-1} = y_k - z_k$. Thus, complete information about two

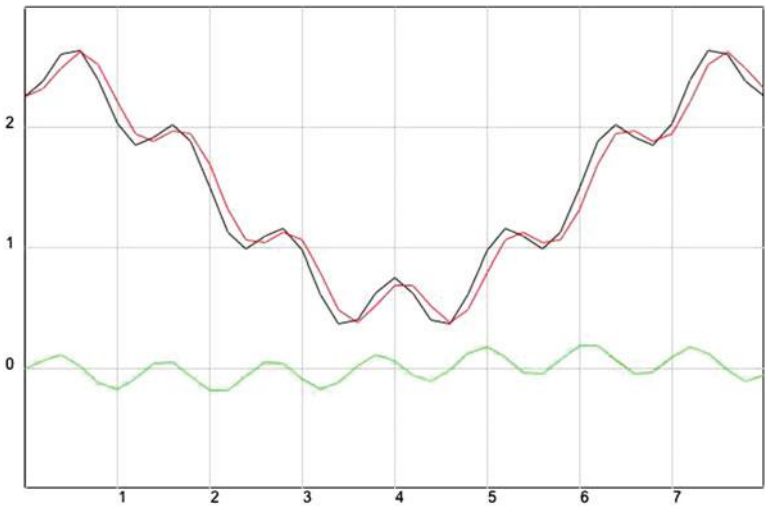


Fig. 4.6: Graph of data from Example 4.27: x_k (——), y_k (——) and z_k (——), $k = 0, 1, \ldots, 40$.

successive data bits x_{k-1}, x_k is contained in a single pass of the two filters. So why not downsample our collection of data to half as many repetitions of the filters in (4.4) and (4.5)? Of course, this requires the number of input data points $N = 2M$ to be even and that we organize it in groups of two: x_k, x_{k-1}, $k = 2, 4, \ldots, N$. The entire procedure can be expressed in matrix form, as the following example illustrates.

Example 4.28. Apply filters of equations (4.4) and (4.5) to a sample of $N = 6$ samples with downsampling and display this process as a single matrix multiplication.

Solution. The equations in question in this case are (in increasing data index):

$$\begin{array}{lll} \frac{1}{2}(x_1 + x_2) = y_2 & \frac{1}{2}(x_3 + x_4) = y_4 & \frac{1}{2}(x_5 + x_6) = y_6 \\ \frac{1}{2}(-x_1 + x_2) = z_2 & \frac{1}{2}(-x_3 + x_4) = z_4 & \frac{1}{2}(-x_5 + x_6) = z_6 \end{array}$$

All of this can be expressed as the single matrix product

$$A_6\mathbf{x} \equiv \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} y_2 \\ y_4 \\ y_6 \\ z_2 \\ z_4 \\ z_6 \end{bmatrix}. \qquad \square$$

A note of caution: No claim of efficiency is made for this matrix form of the transformation; however it allows for an important conceptualization of the process, which is made clear by this simple calculation:

$$A_6 A_6^T = \frac{1}{4}\begin{bmatrix} 1\,1 & 0\,0 & 0 & 0 \\ 0\,0 & 1\,1 & 0 & 0 \\ 0\,0 & 0\,0 & 1 & 1 \\ -1\,1 & 0\,0 & 0 & 0 \\ 0\,0 & -1\,1 & 0 & 0 \\ 0\,0 & 0\,0 & -1 & 1 \end{bmatrix}\begin{bmatrix} 1\,0\,0 & -1 & 0 & 0 \\ 1\,0\,0 & 1 & 0 & 0 \\ 0\,1\,0 & 0 & -1 & 0 \\ 0\,1\,0 & 0 & 1 & 0 \\ 0\,0\,1 & 0 & 0 & -1 \\ 0\,0\,1 & 0 & 0 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix} = \frac{1}{2} I_6.$$

In general, given an even number $N = 2M$, the $N \times N$ matrix A_N is defined as

$$A_N = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \ldots & 0 & 0 \\ 0 & 0 & 1 & 1 \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \ldots & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -1 & 1 & 0 & 0 \ldots & 0 & 0 \\ 0 & 0 & -1 & 1 \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \ldots & -1 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} U \\ \cdots \\ L \end{bmatrix} \tag{4.6}$$

where U and L are $M \times N$ matrices representing upper and lower halves of $2A_N$. A calculation similar to the above for A_6 shows that $A_N A_N^T = \frac{1}{2} I_N$. Thus, the matrix A_N is *nearly* orthogonal.

A slight adjustment in the filters that led to the matrix A_N yields formulas that lead to an orthogonal matrix:

Definition 4.13. Haar Filters The Haar filter and Haar wavelet filter for the data sequence $\{x_k\}_{k=-\infty}^{\infty}$ are defined respectively as

$$y_k = \frac{\sqrt{2}}{2}\left(x_k + x_{k-1}\right), \quad k \in \mathbb{Z}. \tag{4.7}$$

and

$$z_k = \frac{\sqrt{2}}{2}\left(x_k - x_{k-1}\right), \quad k \in \mathbb{Z}. \tag{4.8}$$

If we repeat the construction of matrices like A_N with the above filters, we obtain the following:

Definition 4.14. Haar Wavelet Transform Matrix Given an even number N, the $N \times N$ Haar wavelet transform (HWT) matrix is defined as

$$W_N = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \dots & 0 & 0 \\ 0 & 0 & 1 & 1 \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \dots & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -1 & 1 & 0 & 0 \dots & 0 & 0 \\ 0 & 0 & -1 & 1 \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \dots & -1 & 1 \end{bmatrix} = \frac{\sqrt{2}}{2}\begin{bmatrix} U \\ \cdots \\ L \end{bmatrix}. \tag{4.9}$$

A calculation like that of Example 4.28, which is left as an exercise, shows the following key fact:

Theorem 4.13. The Haar wavelet transform matrix W_N is orthogonal.

What is preserved by using the Haar matrix on a signal vector $\mathbf{x}$ is its magnitude, i.e., $\|W_N\mathbf{x}\| = \|\mathbf{x}\|$. If we view that magnitude as a measure of the "energy" of the signal, then rather than reduce that energy by transforming it via A_N, the Haar transform W_N redistributes it among coordinates of $\mathbf{x}$.

An intriguing possibility for the filter matrices that we have introduced is the application of these filters to an entire image which is itself written out as a matrix with each entry representing a pixel of the image. We consider only the simplest kind of image: A grayscale image in which each pixel is determined by a number representing a shade between black and white. The simplest values for pixels are integer values between 0 (black) and 255 (white), since such a value can be stored in a computer as a single byte.

Let such a graphic be represented by an $m \times n$ matrix A of integers between 0 and 255, where m and n are even. Now apply the Haar wavelet transform to each column of A resulting in the new $m \times n$ matrix $W_m A$. This results in a somewhat flattened new image (low frequency data) in the top half of $W_m A$ created by applying the Haar filter (equation 4.7) to the columns of A, and a crude picture of edges (high frequency data) in the lower half of $W_m A$ created by applying the Haar wavelet filter (equation 4.8) to the columns of A. A similar argument could be applied to the rows of A which are simply the columns of A^T. Notice that $\left(W_n A^T\right)^T = A W_n^T$. Thus, we can apply both operations simultaneously and obtain the equation

$$W_m A W_n^T = 2\begin{bmatrix} B & V \\ H & D \end{bmatrix}. \tag{4.10}$$

(a) Digital image of 512×344 matrix A (b) Digital image of matrix $\frac{1}{2}W_{512}AW_{344}^T$

Fig. 4.7: Image and resulting image via an application of the HWT. (Master of the Castello Nativity, Portrait of a Woman, ca. 1450s. The Metropolitan Museum of Art; The Jules Bache Collection, 1949.)

The matrix $W_mAW_n^T$ is called the *Haar wavelet transform* of A. Each of the blocks in this matrix are $\frac{m}{2} \times \frac{n}{2}$ and we interpret them as follows: B represents the blurred image of A, while V, H and D represent edges of the image A along vertical, horizontal and diagonal directions, respectively. See Figure 4.7 for images defined by a 512×344 matrix A and transformed matrix $\frac{1}{2}W_{512}AW_{344}^T$. But there is an obvious question here: Why the factor of 2 in equation (4.10)? This is best explained by the following calculation:

Haar Wavelet Transform

Example 4.29. Use block arithmetic to determine the range of possible values of a matrix obtained by the Haar wavelet transform applied to the $m \times n$ matrix A (m and n even), given that the range of possible values of the entries of A are in the interval $[0, M]$.

Solution. Use the notation U_N, L_N for the $\frac{N}{2} \times N$ upper and lower blocks of equation (4.9). From this equation and block arithmetic we see that

$$\begin{aligned} W_mAW_n^T &= \frac{\sqrt{2}}{2}\begin{bmatrix} U_m \\ L_m \end{bmatrix} A \frac{\sqrt{2}}{2}\begin{bmatrix} U_n \\ L_n \end{bmatrix}^T \\ &= \frac{2}{4}\begin{bmatrix} U_mA \\ L_mA \end{bmatrix}\begin{bmatrix} U_n^T & L_n^T \end{bmatrix} = \frac{1}{2}\begin{bmatrix} U_mAU_n^T & U_mAL_n^T \\ L_mAU_n^T & L_mAL_n^T \end{bmatrix}. \end{aligned} \tag{4.11}$$

(a) Enlarged blur (b) Edge image

Fig. 4.8: Enlarged blur and edge image obtained from Figure 4.7(b). (Master of the Castello Nativity, Portrait of a Woman, ca. 1450s. The Metropolitan Museum of Art; The Jules Bache Collection, 1949.)

Observe that multiplication on either side of a matrix by a U_N or U_N^T results in adding pairs of elements of the matrix. Thus, elements of $U_m A$ and AU_n^T are in the range of $[0, 2M]$ and elements of $U_m A U_n^T$ are in the range of $[0, 4M]$. On the other hand multiplication on either side by an L_N or L_N^T results in subtraction of elements of the matrix. Thus, elements of $L_m A$ are in the range of $[-M, M]$, elements of $L_m A L_n^T$ are in the range of $[-2M, 2M]$ and elements of $L_m A U_n^T$ and $U_m A L_n^T$ are in the range of $[-2M, 2M]$. □

This example indicates that the correct scaling of a transformed matrix should be $\frac{1}{2}$ since the block matrix on the right-hand side of equation (4.11) already has a factor of $\frac{1}{2}$ in front of it. This brings the positive elements into the range of $[0, M]$ and the negative elements into the range of $\left[-\frac{M}{2}, 0\right]$. No positive scaling by itself can bring the negative numbers into a positive range, so these have to be dealt with separately. One option is to zero out the negative numbers. Another is to replace them by their absolute values, which is what we did in Figures 4.7(b) and 4.8(b).

If we did not want to sacrifice any information, we could use the transformed matrix $C = \frac{1}{2} W_m A W_n^T$ instead of A for storage. Because W_N is orthogonal, A is easily recovered from this matrix since $A = 2W_m^T C W_n$. Moreover, because the number of different bytes required to represent the data of V, H and D is much lower than those required for B, the transformed matrix is

more amenable to data compression techniques than B. This type of compression is termed *lossless compression* since no information is lost by using it.

Lossless Compression

If, on the other hand, we are willing to sacrifice image accuracy for improved storage, we could simply store the blur B in place of the full matrix A and reduce our storage requirements from mn to $mn/4$. We could store the information of B with no modification or compress it by merging the least significant (in terms of frequency of occurrence) bits of data to zero. In either case we lose information contained in the original image, so this is termed *lossy compression*. And what is lost? Consider the blur B of Figure 4.7(a) shown in Figure 4.7(b): It requires one fourth the storage requirements of the original and is enlarged in Figure 4.8(a) for comparison to Figure 4.7(a). Note the differences are visually slight. This idea can be pushed even further by a repeat application of the HWT to the smoothed image B resulting in a new smoothed image that reduces storage requirements from $mn/4$ to $mn/16$.

Lossy Compression

Finally, there is another important graphics application that comes from application of the HWT, namely *edge detection*. We can obtain an image of edges exclusively if we eliminate the blur B from consideration. To achieve this, we first form the transformed matrix $W_m A W_n^T$ as in Equation (4.10). Next we zero out the B portion of the matrix and perform the inverse transformation to obtain a new image with the same dimensions as the original A:

Edge Detection

$$A_e = 2W_m^T \begin{bmatrix} 0 & V \\ H & D \end{bmatrix} W_n.$$

This image reassembles the edge data from matrices V, H and D. For an example of what this procedure produces, see Figure 4.8(b).

Once again we have just scratched the surface of an important application of linear algebra. For a detailed in-depth discussion of topics introduced here the reader should consult Chapter 6 of the text [10].

4.4 Exercises and Problems

Exercise 1. Use Gram–Schmidt to find QR factorizations for these matrices and use them to compute the least squares solutions of $A\mathbf{x} = \mathbf{b}$ with these pairs $A, \mathbf{b}$.

(a) $\begin{bmatrix} 3 & 2 \\ 0 & 1 \\ 4 & 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -2 \\ 5 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 & 2 \\ 0 & 1 & 2 \\ -2 & 1 & 6 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 8 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & 2 \\ -1 & 1 & 1 \\ -1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 1 \\ 3 \\ 1 \end{bmatrix}$

Exercise 2. Let $A = \begin{bmatrix} 3 & 2 \\ 0 & 1 \\ 4 & 1 \end{bmatrix}$ and use Householder matrices to find a full QR factorization of A. Use this result to find the least squares solution to the system $A\mathbf{x} = \mathbf{b}$, where $\mathbf{b} = (1, 2, 3)$, and resulting residual.

Exercise 3. Calculate the B, V, H, D blocks for a grayscale image with this matrix form:

$$A = \begin{bmatrix} 12 & 0 & 100 & 0 & 0 & 0 \\ 0 & 120 & 0 & 120 & 16 & 0 \\ 140 & 20 & 0 & 12 & 120 & 0 \\ 0 & 80 & 100 & 140 & 80 & 100 \end{bmatrix}$$

Exercise 4. Calculate the B, V, H, D blocks for a grayscale image with this matrix form and repeat the calculation on the blur B:

$$A = \begin{bmatrix} 0 & 0 & 0 & 100 & 20 & 0 & 0 & 100 \\ 20 & 12 & 0 & 120 & 0 & 0 & 140 & 0 \\ 0 & 0 & 0 & 0 & 80 & 100 & 0 & 0 \\ 40 & 0 & 120 & 0 & 0 & 0 & 0 & 20 \\ 0 & 100 & 0 & 140 & 0 & 0 & 0 & 0 \\ 0 & 80 & 0 & 0 & 100 & 0 & 20 & 12 \\ 120 & 0 & 24 & 0 & 80 & 0 & 24 & 0 \\ 100 & 0 & 0 & 0 & 60 & 0 & 0 & 0 \end{bmatrix}$$

Problem 5. Calculate a full QR factorization of $\begin{bmatrix} 1 & 10 & 20 \\ 10 & 100 & 201 \\ 1000 & 10001 & 20001 \end{bmatrix}$ with a technology tool. Inspect the R matrix and estimate the rank of A. Use QR to find the least squares solution to $A\mathbf{x} = \mathbf{b}$, where $\mathbf{b} = (1, 2, 3)$, and resulting residual.

Problem 6. The following is a simplified description of the *QR algorithm* (which is separate from the QR factorization, but involves it) for a real $n \times n$ matrix A :

$T_0 = A$, $Q_0 = I_n$
for $k = 0, 1, \ldots$
$T_k = Q_{k+1} R_{k+1}$ (QR factorization of T_k)
$T_{k+1} = R_{k+1} Q_{k+1}$
end

Apply this algorithm to the following two matrices and, based on your results, speculate about what it is supposed to compute. You will need a technology tool for this exercise and you will stop in a finite number of steps, but expect to take more than a few.

(a) $A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & -2 \\ 0 & -2 & 1 \end{bmatrix}$ (b) $A = \begin{bmatrix} -8 & -5 & 8 \\ 6 & 3 & -8 \\ -3 & 1 & 9 \end{bmatrix}$

*Problem 7. The flop count for a matrix-vector multiplication $A\mathbf{x}$, where A is $n \times n$ is of order n^2. Prove this. However, if $A = H_{\mathbf{v}}$, a Householder matrix, the calculation is of order n. Exhibit an algorithm that proves this and determine its flop count.

Problem 8. Prove that the Haar wavelet transform matrix W_N is orthogonal.

Problem 9. Confirm that the extremes of the interval ranges of Example 4.29 do occur by applying the Haar wavelet transform to the matrix $A = \begin{bmatrix} x & y \\ z & w \end{bmatrix}$ and making suitable choices of $x, y, z, w \in [0, M]$.

Problem 10. Apply the Haar filter and wavelet filter to the data of Example 4.27 and plot a dot-to-dot of the the data and the output as in Figure 4.6.

4.5 *Projects and Reports

Project: Rotations in Computer Graphics I
Problem Description: the objective of this project is to implement a counterclockwise rotation of θ radians about an axis specified by the nonzero three-dimensional vector $\mathbf{v}$ using matrix multiplication. Assume that you are given this vector. Show how to calculate the appropriate matrix $R_{\mathbf{v}}$ and offer some justification (proofs aren't required). Illustrate the method with examples.

Implementation Notes: in principle, the desired matrix can be constructed in three steps: (1) Construct an orthonormal set of vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ such that $\mathbf{v}_1 \times \mathbf{v}_2 = \mathbf{v}_3 = \mathbf{v}/\|\mathbf{v}\|$. (2) Construct the orthogonal matrix P that maps $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ to $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. (3) To construct $R_{\mathbf{v}}$, apply P, do a rotation θ of the xy-plane about the z-axis via $R(\theta)$, then apply P^{-1}. Your job is to elaborate on the details of these steps and illustrate the result with examples.

Project: Rotations in Computer Graphics II
Problem Description: the objective of this project is to implement a counterclockwise rotation of θ radians about an axis specified by the nonzero three-dimensional vector $\mathbf{v}$ using quaternions. Assume that you are given this vector. Show how to calculate the appropriate quaternion $\mathbf{q}_{\mathbf{v}}$. Illustrate the method (and your mastery of quaternion arithmetic) with examples.

Background: Quaternions have a long and storied history in mathematics dating back to 1843, when they were discovered by Sir William Rowan Hamilton as a generalization of complex numbers. Three-dimensional vector dot and cross products originated as aids to quaternion arithmetic. In 1985 Ken Shoemake [22] showed that quaternions were well suited for certain transforms in computer graphics, namely rotations about an axis in three-dimensional space. A quaternion that does the job requires only four numbers, in contrast to the nine needed for an orthogonal transform.

Quaternions

Implementation Notes: A quick google of "quaternions" will give you more than enough information. A brief précis: quaternion objects are simply elements of $\mathbb{H} = \mathbb{R}^4$,homogeneous space (see Section 3.1.) As such, $\mathbb{H}$ immediately has a vector space structure, standard inner product, and norm. Standard basis elements are denoted by $\mathbf{i} = \mathbf{e}_1$, $\mathbf{j} = \mathbf{e}_2$, $\mathbf{k} = \mathbf{e}_3$, and $\mathbf{h} = \mathbf{e}_4$. Hence, quaternions can be written as $\mathbf{q} = q_x\mathbf{i} + q_y\mathbf{j} + q_z\mathbf{k} + q_w\mathbf{h} = \mathbf{q}_v + q_w\mathbf{h}$. The vector $\mathbf{q}_v$ is called the "imaginary" part of $\mathbf{q}$, and $q_w\mathbf{h}$ the "real" part. Inspired by complex numbers, we define the *conjugate* quaternion $q^* = q_w - \mathbf{q}_v$. Unlike homogeneous space, $\mathbb{H}$ carries a multiplicative structure. Multiplication is indicated by juxtaposition. We only need to know how to multiply basis elements, since the rest follows from using distributive and associative laws, which we assume to hold for quaternions (of course, everything can be proved formally). Here are the fundamental rules:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{h} = -\mathbf{h}^2.$$

It is a customary abuse of language to identify $\mathbf{h}$ with 1 and write $\mathbf{q} = \mathbf{q_v} + q_w$. From these laws we can deduce that $\mathbf{ij} = \mathbf{k}$, $\mathbf{jk} = \mathbf{i}$, $\mathbf{ki} = \mathbf{j}$, $\mathbf{ik} = -\mathbf{j}$, $\mathbf{kj} = -\mathbf{i}$ and $\mathbf{ji} = -\mathbf{k}$, which is everything we need to know to do arithmetic. A remarkable property of quaternions is that every nonzero element has a multiplicative inverse, namely

$$\mathbf{q}^{-1} = \frac{1}{\|\mathbf{q}\|^2}\mathbf{q}^*.$$

Finally, the connection to rotations can be spelled out as follows: let $\mathbf{p}, \mathbf{q} \in \mathbb{H}$, with $\mathbf{q}$ a unit quaternion, i.e., $\|\mathbf{q}\| = 1$, and $\mathbf{p}$ a quaternion that represents a geometrical point or vector in homogeneous space. Then (1) we can write $\mathbf{q} = \cos\phi + \sin\phi\mathbf{q}_v$ for some angle ϕ and unit vector $\mathbf{q}_v$ and (2) $\mathbf{qpq}^{-1}$ is the result of rotating $\mathbf{p}$ counterclockwise about the axis $\mathbf{q}_v$ through an angle of 2ϕ. Your job is elaborate on the details of this calculation and illustrate the result with examples. As an exercise in manipulation, prove item (1) (it isn't hard), but assume everything else.

Report: Image Compression and Edge Detection

Problem Description: In this report you will test some limits of data compression by experimenting with an interesting image of your own choosing. It could be a photograph you have taken or some reasonably complex image from the internet that piqued your interest. You are to transform the image into suitable format and then see how much you can can compress that data storage requirements for that image while losing an acceptable amount of detail.

Implementation Notes: First, you must convert the image to a grayscale format without layers with pixels stored as unsigned eight bit integers. (We are not going to deal with the additional details of color images.) For this you will need an image manipulation program such as the GNU program Gimp or

commercial software such as Adobe Photoshop. Figure 4.7(a) is an example of the sort of file you with which you will start experimenting. Secondly, you will need a technology tool capable of importing standard flattened image grayscale images (such as .png, etc.) into matrices and vice versa. The freely available R programming language and Octave, as well as commercial Matlab and others are perfectly capable of these tasks. For the record, Figure 4.7 and its relatives in this chapter were converted from .pdf into grayscale .png images via Gimp and read, manipulated and written as matrices via Octave.

Next, you must apply the Haar wavelet transform repeatedly to your initial image until you reach a blur that is unacceptably far from the initial image. These computations will require a mild bit of programming on your part with the technology tool of your choice. Each application of the transform will reduce storage requirements by a factor of four. How much are you able to save? Of course, "unacceptably far" is in the eye of the beholder, but here's a pretty reasonable case: Start with a some text and compress it until you can no longer read the text.

A picture is worth a thousand words, so be lavish with them in your write-up. Consider the amount of savings if, in addition to saving the blurs in all their detail, you were to to save a very good approximation to the edges portion of the transformed picture. For example, consider what you might achieve by first applying some thresholding condition to edge portions of the picture that sets all pixels below a certain level to zero, then accounting for the large number of resulting zeros by some compression technique. You might even suggest a format for such a compression format.

Report: Least Squares

Note to the Instructor: the data below comes from a hypothetical conference. This project works better when adapted to your local environment. Pick a sport in season at your institution or locale. Have students collect the data themselves, make out a data table as below, and predict the spread for some (as yet) unplayed games of local interest. It can be very interesting to make it an ongoing project, where for a number of weeks the students collect the previous week's data and make predictions for the following week based on all data collected to date.

The Big Eight needs your help! Below is a table of scores from the games played thus far: The (i, j)th entry is team i's score in the game with team j. Your assignment is twofold. First, write a notebook (or script) in a technology tool available to you that obtains team ratings and predicted point spreads based on the least squares and graph theory ideas you have seen. Include instructions for the illiterate on how to plug in data. Second, you are to write a brief report on your project that describes the problem, your solution to it, its limitations, and the ideas behind it.

	CU	IS	KS	KU	MU	NU	OS	OU
CU		24		21	45		21	14
IS	12			42	21	16		7
KS				12	21	3	27	24
KU	9	14	30			10		14
MU	8	3	52			18	21	
NU		51	48	63	26		63	
OS	41		45		49	42		28
OU	17	35	70	63			31	

Implementation Notes: You will need to set up a suitable system of equations, form the normal equations, and have a technology tool solve the problem. The equations in question are formed by letting the variables be a vector $\mathbf{x}$ of "potentials" $x(i)$, one for each team i, so that the "potential differences" best approximate the actual score differences (i.e., point spreads) of the games. To find the vector $\mathbf{x}$ of potentials you solve the system $A\mathbf{x} = \mathbf{b}$, where $\mathbf{b}$ is the vector of observed potential differences. N.B: the matrix A is *not* the table given above. You will get one equation for each game played. For example, by checking the $(1,2)$th and $(2,1)$th entries, we see that CU beat IS by a score of 24 to 12. So the resulting equation for this game is $x(1) - x(2) = 24 - 12 = 12$. Ideally, the resulting potentials would give numbers that would enable you to predict the point spread of an as yet unplayed game: all you would have to do to determine the spread for team i versus team j is calculate the difference $x(j) - x(i)$. Of course, it doesn't really work out this way, but this is a reasonable use of the known data. When you set up this system, you obtain an inconsistent system. This is where least squares enter the picture. You will need to set up and solve the normal equations, one way or another. You might notice that the null space of the resulting coefficient matrix is nontrivial, so this matrix does not have full column rank. This makes sense: potentials are unique only up to a constant. To fix this, you could arbitrarily fix the value of one team's potential, that is, set the weakest team's potential value to zero by adding one additional equation to the system of the form $x(i) = 0$.

5

THE EIGENVALUE PROBLEM

The first major problem of linear algebra is to understand how to solve the basis linear system $A\mathbf{x} = \mathbf{b}$ and what the solution means. We have explored this system from three points of view: In Chapter 1 we approached the problem from an operational point of view and learned the mechanics of computing solutions. In Chapter 2, we took a more sophisticated look at the system from the perspective of matrix theory. Finally, in Chapter 3, we viewed the problem from the vantage of vector space theory.

Now we begin a study of the second major problem of linear algebra, namely the eigenvalue problem. We had to tackle linear systems first because the eigenvalue problem is more sophisticated and will require most of the tools that we have thus far developed. This subject has many important applications, such as the analysis of discrete dynamical systems that we have seen in earlier chapters.

5.1 Definitions and Basic Properties

What Are They?

Good question. Let's get right to the point.

Definition 5.1. **Eigenvector, Eigenvalue, Eigenpair** Let A be a square $n \times n$ matrix. An *eigenvector* of A is a nonzero vector $\mathbf{x}$ in $\mathbb{R}^n$ (or $\mathbb{C}^n$, if we are working over complex numbers) such that for some scalar λ, we have

$$A\mathbf{x} = \lambda\mathbf{x}.$$

The scalar λ is called an *eigenvalue* of the matrix A, and we say that the vector $\mathbf{x}$ is an *eigenvector belonging to the eigenvalue* λ. The pair $\{\lambda, \mathbf{x}\}$ is called an *eigenpair* for the matrix A.

Eigenvalues and eigenvectors are also known as characteristic values and characteristic vectors. In fact, the word "eigen" means (among other things) "characteristic" in German.

Eigenvectors of A, as defined above, are also called *right eigenvectors* of A.

Right and Left Eigenvectors

Notice that if $A^T\mathbf{x} = \lambda\mathbf{x}$, then

$$\lambda\mathbf{x}^T = (\lambda\mathbf{x})^T = \left(A^T\mathbf{x}\right)^T = \mathbf{x}^T A.$$

For this reason, eigenvectors of A^T are called *left eigenvectors* of A.

The only kinds of matrices for which these objects are defined are square matrices, so unless otherwise stated, we'll assume throughout this chapter that we are dealing with such matrices.

Zero Not Eigenvector

Caution: Be aware that the eigenvalue λ is allowed to be the 0 scalar, but an eigenvector $\mathbf{x}$ is, by definition, *never the* $\mathbf{0}$ *vector.*

As a matter of fact, it is quite informative to have an eigenvalue 0. This says that the system $A\mathbf{x} = 0\mathbf{x} = \mathbf{0}$ has a nontrivial solution $\mathbf{x}$. Therefore, A is not invertible by Theorem 2.6. There are other reasons for the usefulness of the eigenvector/value concept that we will develop later, but we already see that knowledge of eigenvalues tells us about invertibility of a matrix.

Here are a few simple examples of eigenvalues and eigenvectors. Let $A = \begin{bmatrix} 7 & 4 \\ 3 & 6 \end{bmatrix}$, $\mathbf{x} = (-1, 1)$, and $\mathbf{y} = (4, 3)$. One checks that $A\mathbf{x} = (-3, 3) = 3\mathbf{x}$ and $A\mathbf{y} = (40, 30) = 10\mathbf{y}$. It follows that $\mathbf{x}$ and $\mathbf{y}$ are eigenvectors corresponding to eigenvalues 3 and 10, respectively.

Why should we have any interest in these quantities? A general answer goes something like this: knowledge of eigenvectors and eigenvalues gives us deep insights into the structure of the matrix A. Here is just one example: suppose that we would like to have a better understanding of the effect of multiplication of a vector $\mathbf{x}$ by powers of the matrix A, that is, of $A^k\mathbf{x}$. Let's start with the first power, $A\mathbf{x}$. If we knew that $\mathbf{x}$ were an eigenvector of A, then we would have that for some scalar λ,

$$\begin{aligned} A\mathbf{x} &= \lambda\mathbf{x} \\ A^2\mathbf{x} &= A(A\mathbf{x}) = A\lambda\mathbf{x} = \lambda A\mathbf{x} = \lambda^2\mathbf{x} \\ &\vdots \\ A^k\mathbf{x} &= A(A^{k-1}\mathbf{x}) = \cdots = \lambda^k\mathbf{x}. \end{aligned}$$

This is very nice, because it reduces the more complicated matrix–vector multiplication to a simpler scalar–vector multiplication.

We need some handles on these quantities. Let's ask how we could figure out what they are for specific matrices. Here are some of the basic points about eigenvalues and eigenvectors.

Theorem 5.1. Let A be a square $n \times n$ matrix. Then

(1) The eigenvalues of A are all the scalars λ that are solutions to the nth-degree polynomial equation

$$\det(\lambda I - A) = 0.$$

(2) For eigenvalue λ, the eigenvectors of the matrix A belonging to that eigenvalue are all the nonzero elements of $\mathcal{N}(\lambda I - A)$.

Proof. Note that $\lambda\mathbf{x} = \lambda I\mathbf{x}$. Thus, we have the following chain of thought: A has eigenvalue λ if and only if $A\mathbf{x} = \lambda\mathbf{x}$, for some nonzero vector $\mathbf{x}$, which is true if and only if

$$\mathbf{0} = \lambda\mathbf{x} - A\mathbf{x} = \lambda I\mathbf{x} - A\mathbf{x} = (\lambda I - A)\mathbf{x}$$

for some nonzero vector $\mathbf{x}$. This last statement is equivalent to the assertion that $\mathbf{0} \neq \mathbf{x} \in \mathcal{N}(\lambda I - A)$. The matrix $\lambda I - A$ is square, so it has a nontrivial null space precisely when it is singular (Theorem 2.6). This occurs only when $\det(\lambda I - A) = 0$. If we expand this determinant down the first column, we see that the highest-order term involving λ that occurs is the product of the diagonal terms $(\lambda - a_{ii})$, so that the degree of the expression $\det(\lambda I - A)$ as a polynomial in λ is n. This proves (1).

We have seen that if λ is an eigenvalue of A, then the eigenvectors belonging to that eigenvalue are precisely the nonzero vectors $\mathbf{x}$ such that $(\lambda I - A)\mathbf{x} = \mathbf{0}$, that is, the nonzero elements of $\mathcal{N}(\lambda I - A)$, which is what (2) asserts. □

Monic Polynomial

Here is some terminology that we will use throughout this chapter. We call a polynomial *monic* if the leading coefficient is 1. For example, $\lambda^2 + 2\lambda + 3$ is a monic polynomial in λ while $2\lambda^2 + \lambda + 1$ is not.

Definition 5.2. Characteristic Equation and Polynomial If A is a square $n \times n$ matrix, the equation $\det(\lambda I - A) = 0$ is called the *characteristic equation* of A, and the nth-degree monic polynomial $p(\lambda) = \det(\lambda I - A)$ is called the *characteristic polynomial* of A.

Suppose we already know the eigenvalues of A and want to find the eigenvalues of something like $3A + 4I$. Do we have to start over to find them? The next calculation is really a useful tool for answering such questions.

Theorem 5.2. If $B = cA + dI$ for scalars d and $c \neq 0$, then the eigenvalues of B are of the form $\mu = c\lambda + d$, where λ runs over the eigenvalues of A, and the eigenvectors of A and B are identical.

Proof. Let $\mathbf{x}$ be an eigenvector of A corresponding to the eigenvalue λ. Then by definition, $\mathbf{x} \neq \mathbf{0}$ and $A\mathbf{x} = \lambda\mathbf{x}$. Also, we have that $dI\mathbf{x} = d\mathbf{x}$. Now multiply the first equation by the scalar c and add these two equations to obtain

$$(cA + dI)\mathbf{x} = B\mathbf{x} = (c\lambda + d)\mathbf{x}.$$

It follows that every eigenvector of A belonging to λ is also an eigenvector of B belonging to the eigenvalue $c\lambda + d$. Conversely, if $\mathbf{y}$ is an eigenvector of B belonging to μ, then

$$B\mathbf{y} = \mu\mathbf{y} = (cA + dI)\mathbf{y}.$$

Now solve for $A\mathbf{y}$ to obtain that

$$A\mathbf{y} = \frac{1}{c}(\mu - d)\mathbf{y},$$

so that $\lambda = (\mu - d)/c$ is an eigenvalue of A with corresponding eigenvector $\mathbf{y}$. It follows that A and B have the same eigenvectors, and their eigenvalues are related by the formula $\mu = c\lambda + d$. □

Example 5.1. Let $A = \begin{bmatrix} 7 & 4 \\ 3 & 6 \end{bmatrix}$, $\mathbf{x} = (-1, 1)$, and $\mathbf{y} = (4, 3)$, so that $A\mathbf{x} = (-3, 3) = 3\mathbf{x}$ and $A\mathbf{y} = (40, 30) = 10\mathbf{y}$. Find the eigenvalues and corresponding eigenvectors for the matrix $B = 3A + 4I$.

Solution. From the calculations given to us, we observe that $\mathbf{x}$ and $\mathbf{y}$ are eigenvectors corresponding to the eigenvalues 3 and 10, respectively, for A. These are all the eigenvalues of A, since the characteristic polynomial of A is of degree 2, so has only two roots. According to Theorem 5.2, the eigenvalues of $3A+4I$ must be $\mu_1 = 3\cdot 3+4 = 13$ with corresponding eigenvector $\mathbf{x} = (-1, 1)$, and $\mu_2 = 3 \cdot 10 + 4 = 34$ with corresponding eigenvector $\mathbf{y} = (4, 3)$. □

Definition 5.3. **Eigenspace** The *eigenspace* corresponding to eigenvalue λ is the subspace $\mathcal{N}(\lambda I - A)$ of $\mathbb{R}^n$ (or $\mathbb{C}^n$). We write $\mathcal{E}_\lambda(A) = \mathcal{N}(\lambda I - A)$.

Definition 5.4. **Eigensystem** By an *eigensystem* of the matrix A, we mean a list of all the eigenvalues of A and, for each eigenvalue λ, a complete description of the eigenspace corresponding to λ.

The usual way to give a complete description of an eigenspace is to list a basis for the space. Remember that there is one vector in the eigenspace $\mathcal{N}(\lambda I - A)$ that is *not* an eigenvector, namely $\mathbf{0}$. In any case, the computational route is now clear. To call the following formula an algorithm is bit of an exaggeration, since we don't specify a complete computational description of the eigenvalue phase (1), but here it is:

Eigensystem Algorithm

Given $n \times n$ matrix A, to find an eigensystem for A:

(1) Find the eigenvalues of A.
(2) For each scalar λ in (1), use the null space algorithm to find a basis of the eigenspace $\mathcal{N}(\lambda I - A)$.

For a deeper look at numerical methods for finding eigenvalues, the reader is encouraged to consult Section 5.7. Here we confine ourselves to relatively simple cases where we can solve the characteristic equation by explicit means. As a matter of convenience, it is sometimes a little easier to work with $A - \lambda I$ when calculating eigenspaces (because there are fewer extra minus signs to worry about). This is perfectly OK, since $\mathcal{N}(A - \lambda I) = \mathcal{N}(\lambda I - A)$. It doesn't affect the eigenvalues either, since $\det(\lambda I - A) = \pm \det(A - \lambda I)$. Here is our first eigensystem calculation.

Example 5.2. Find an eigensystem for the matrix $A = \begin{bmatrix} 7 & 4 \\ 3 & 6 \end{bmatrix}$.

Solution. First solve the characteristic equation

$$\begin{aligned} 0 = \det(\lambda I - A) &= \det \begin{bmatrix} \lambda - 7 & -4 \\ -3 & \lambda - 6 \end{bmatrix} \\ &= (\lambda - 7)(\lambda - 6) - (-3)(-4) \\ &= \lambda^2 - 13\lambda + 42 - 12 \\ &= \lambda^2 - 13\lambda + 30 \\ &= (\lambda - 3)(\lambda - 10). \end{aligned}$$

Hence, the eigenvalues are $\lambda = 3, 10$. Next, for each eigenvector calculate the corresponding eigenspace.

$\lambda = 3$: Then $A - 3I = \begin{bmatrix} 7-3 & 4 \\ 3 & 6-3 \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 3 & 3 \end{bmatrix}$ and row reduction gives

$$\begin{bmatrix} 4 & 4 \\ 3 & 3 \end{bmatrix} \xrightarrow[E_1(1/4)]{E_{21}(-3/4)} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

so the general solution is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -x_2 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Therefore, a basis of $\mathcal{E}_3(A)$ is $\{(-1, 1)\}$.

$\lambda = 10$: Then $A - 10I = \begin{bmatrix} 7-10 & 4 \\ 3 & 6-10 \end{bmatrix} = \begin{bmatrix} -3 & 4 \\ 3 & -4 \end{bmatrix}$ and row reduction gives

$$\begin{bmatrix} -3 & 4 \\ 3 & -4 \end{bmatrix} \xrightarrow[E_1(-1/3)]{E_{21}(1)} \begin{bmatrix} 1 & -4/3 \\ 0 & 0 \end{bmatrix},$$

so the general solution is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} (4/3)x_2 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} 4/3 \\ 1 \end{bmatrix}.$$

Therefore, a basis of $\mathcal{E}_{10}(A)$ is $\{(4/3, 1)\}$. □

Concerning this example, there are several observations worth noting:

- Since the 2×2 matrix $A - \lambda I$ is singular for the eigenvalue λ, one row should always be a multiple of the other. Knowing this, we didn't have to do even the little row reduction we did above. However, it's a good idea to check; it helps you avoid mistakes. Remember: any time that row reduction of $A - \lambda I$ leads to full rank (only trivial solutions), you have either made an arithmetic error or you do not have an eigenvalue.
- This matrix is familiar. In fact, $B = \frac{1}{10}A$ is the Markov chain transition matrix from Example 2.19. Therefore, the eigenvalues of B are 0.3 and 1, by Example 5.2 and Theorem 5.2 with $c = 0.1$ and $d = 0$. The eigenvector belonging to $\lambda = 1$ is just a solution to the equation $B\mathbf{x} = \mathbf{x}$, which was discussed in Example 3.31.
- The vector
$$\mathbf{x} = \begin{bmatrix} 4/7 \\ 3/7 \end{bmatrix} = \frac{3}{7} \begin{bmatrix} 4/3 \\ 1 \end{bmatrix}$$
is an eigenvector of A belonging to the eigenvalue $\lambda = 10$ of A, so that from $A\mathbf{x} = 10\mathbf{x}$ we see that $B\mathbf{x} = 1\mathbf{x}$. Hence, $\mathbf{x} \in \mathcal{E}_1(B)$.

Example 5.3. How do we find eigenvalues of a triangular matrix? Illustrate the method with $A = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$.

Solution. Eigenvalues are just the roots of the characteristic equation $\det(\lambda I - A) = 0$. Notice that $-A$ is triangular if A is. Also, the only entries in $\lambda I - A$ that are any different from the entries of $-A$ are the diagonal entries, which change from $-a_{ii}$ to $\lambda - a_{ii}$. Therefore, $\lambda I - A$ is triangular if A is. We already know that the determinant of a triangular matrix is easy to compute: just form the product of the diagonal entries. Therefore, the roots of the characteristic equation are the solutions to

$$0 = \det(\lambda I - A) = (\lambda - a_{11})(\lambda - a_{22}) \cdots (\lambda - a_{nn}),$$

that is, $\lambda = a_{11}, a_{22}, \ldots, a_{nn}$. In other words, for a triangular matrix the eigenvalues are simply the diagonal elements! Thus, for the example A given above, we see with no calculations that the eigenvalues are $\lambda = 2, 1, -1$. □

Notice, by the way, that we don't quite get off the hook in the preceding example if we are required to find the eigenvectors. It will still be some work

to compute each of the relevant null spaces, but much less than for a general matrix.

Example 5.3 can be used to illustrate another very important point. The reduced row echelon form of the matrix of that example is clearly the identity matrix I_3. This matrix has eigenvalues $1, 1, 1$, which are *not* the same as the eigenvalues of A (would that eigenvalue calculations were so easy!). In fact, a single elementary row operation on a matrix can change the eigenvalues. For example, simply multiply the first row of A above by $\frac{1}{2}$. This point warrants a warning, since it is the source of a fairly common mistake.

Caution: The eigenvalues of a matrix A and the matrix EA, where E is an elementary matrix, need not be the same.

Example 5.4. Find an eigensystem for the matrix $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$.

Solution. For eigenvalues, compute the roots of the equation

$$\begin{aligned} 0 = \det(A - \lambda I) &= \det \begin{bmatrix} 1-\lambda & -1 \\ 1 & 1-\lambda \end{bmatrix} \\ &= (1-\lambda)^2 - (-1) = \lambda^2 - 2\lambda + 2. \end{aligned}$$

Now we have a little problem. Do we allow complex numbers? If not, we are stuck because the roots of this equation are

$$\lambda = \frac{-(-2) \pm \sqrt{(-2)^2 - 4 \cdot 2}}{2} = 1 \pm \mathrm{i}.$$

In other words, if we did not enlarge our field of scalars to the complex numbers, we would have to conclude that there are *no* eigenvalues or eigenvectors! Somehow, this doesn't seem like a good idea. It is throwing information away. Perhaps it comes as no surprise that complex numbers would eventually figure into the eigenvalue story. After all, finding eigenvalues is all about solving polynomial equations, and complex numbers were invented to overcome the inability of real numbers to provide solutions to all polynomial equations. Let's allow complex numbers as the scalars. Now our eigenspace calculations are really going on in the complex space $\mathbb{C}^2$ instead of $\mathbb{R}^2$.

$\lambda = 1 + \mathrm{i}$: Then $A - (1+\mathrm{i})I = \begin{bmatrix} 1-(1+\mathrm{i}) & -1 \\ 1 & 1-(1+\mathrm{i}) \end{bmatrix} = \begin{bmatrix} -\mathrm{i} & -1 \\ 1 & -\mathrm{i} \end{bmatrix}$ and row reduction gives (recall that $1/\mathrm{i} = -\mathrm{i}$)

$$\begin{bmatrix} -\mathrm{i} & -1 \\ 1 & -\mathrm{i} \end{bmatrix} \xrightarrow[E_1(1/(-\mathrm{i}))]{E_{21}(-\mathrm{i})} \begin{bmatrix} 1 & -\mathrm{i} \\ 0 & 0 \end{bmatrix},$$

so the general solution is

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \mathrm{i}z_2 \\ z_2 \end{bmatrix} = z_2 \begin{bmatrix} \mathrm{i} \\ 1 \end{bmatrix}.$$

Therefore, a basis of $\mathcal{E}_{1+\mathrm{i}}(A)$ is $\{(\mathrm{i}, 1)\}$.

$\lambda = 1 - \mathrm{i}$: Then $A - (1-\mathrm{i})I = \begin{bmatrix} 1-(1-\mathrm{i}) & -1 \\ 1 & 1-(1-\mathrm{i}) \end{bmatrix} = \begin{bmatrix} \mathrm{i} & -1 \\ 1 & \mathrm{i} \end{bmatrix}$ and row reduction gives

$$\begin{bmatrix} \mathrm{i} & -1 \\ 1 & \mathrm{i} \end{bmatrix} \overrightarrow{\begin{matrix} E_{21}(\mathrm{i}) \\ E_1(1/\mathrm{i}) \end{matrix}} \begin{bmatrix} 1 & \mathrm{i} \\ 0 & 0 \end{bmatrix},$$

so the general solution is

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} -\mathrm{i}z_2 \\ z_2 \end{bmatrix} = z_2 \begin{bmatrix} -\mathrm{i} \\ 1 \end{bmatrix}.$$

Therefore, a basis of $\mathcal{E}_{1-\mathrm{i}}(A)$ is $\{(-\mathrm{i}, 1)\}$. □

In view of the previous example, we are going to adopt the following practice: if the eigenvalue calculation leads us to complex numbers, we take the point of view that the field of scalars should be enlarged to include the complex numbers and the eigenvalues in question. One small consolation for having to deal with complex eigenvalues is that in some cases our work may be cut in half.

Example 5.5. Show that if $\{\lambda, \mathbf{x}\}$ is an eigenpair for real matrix A, then so is $\{\overline{\lambda}, \overline{\mathbf{x}}\}$.

Solution. By hypothesis, $A\mathbf{x} = \lambda\mathbf{x}$. Apply complex conjugation to both sides and use the fact that A is real to obtain

$$\overline{A\mathbf{x}} = \overline{A}\overline{\mathbf{x}} = A\overline{\mathbf{x}} = \overline{\lambda\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}}.$$

Thus, $\{\overline{\lambda}, \overline{\mathbf{x}}\}$ is also an eigenpair for A. □

In view of this fact, we could have stopped with the calculation of eigenpair $\{1+\mathrm{i}, (\mathrm{i}, 1)\}$ in Example 5.4, since we automatically have that $\{1-\mathrm{i}, (-\mathrm{i}, 1)\}$ is also an eigenpair.

Multiplicity of Eigenvalues

The following example presents yet another curiosity about eigenvalues and eigenvectors.

Example 5.6. Find an eigensystem for the matrix $A = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$.

Solution. Here the eigenvalues are easy. This matrix is triangular, so they are $\lambda = 2, 2$. Next we calculate eigenvectors.

$\lambda = 2$: Then $A - 2I = \begin{bmatrix} 2-2 & 1 \\ 0 & 2-2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and row reduction is not necessary here. Notice that the variable x_1 is free here, while x_2 is bound. The general solution is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Therefore, a basis of $\mathcal{E}_2(A)$ is $\{(1,0)\}$. □

The manner in which we list the eigenvalues in this example is intentional. The number 2 occurs twice on the diagonal, suggesting that it should be counted twice. As a matter of fact, $\lambda = 2$ is a root of the characteristic equation $(\lambda - 2)^2 = 0$ of multiplicity 2. Yet there is a curious mismatch here. In all of our examples to this point, we have been able to come up with as many eigenvectors as eigenvalues, namely the size of the matrix if we allow complex numbers. In this case there is a deficiency in the number of eigenvectors, since there is only one eigenspace and it is one-dimensional. Does this failing always occur with multiple eigenvalues? The answer is no. The situation is a bit more complicated, as the following example shows.

Example 5.7. Discuss the eigenspace corresponding to the eigenvalue $\lambda = 2$ for these two matrices:

$$\text{(a)} \begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 2 \end{bmatrix} \qquad \text{(b)} \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

Solution. Notice that each of these matrices has eigenvalues $\lambda = 1, 2, 2$. Now for the eigenspace $\mathcal{E}_2(A)$.

(a) For this eigenspace calculation we have

$$A - 2I = \begin{bmatrix} 2-2 & 1 & 2 \\ 0 & 1-2 & -2 \\ 0 & 0 & 2-2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{bmatrix},$$

and row reduction gives

$$\begin{bmatrix} 0 & 1 & 2 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow[E_{21}(1)]{} \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

so that free variables are x_1, x_3 and the general solution is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ -2x_3 \\ x_3 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ -2 \\ 1 \end{bmatrix}.$$

Thus, a basis for $\mathcal{E}_2(A)$ is $\{(1,0,0), (0,-2,1)\}$. Notice that in this case we get as many independent eigenvectors as the number of times that the eigenvalue $\lambda = 2$ occurs.

(b) For this eigenspace calculation we have

$$A - 2I = \begin{bmatrix} 2-2 & 1 & 1 \\ 0 & 1-2 & 1 \\ 0 & 0 & 2-2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

and row reduction gives

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow[E_{21}(1)]{} \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow[E_{12}(-1)]{E_2(1/2)} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

so that the only free variable is x_1 and the general solution is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \\ 0 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Thus, a basis for $\mathcal{E}_2(A)$ is $\{(1,0,0)\}$. Notice that in this case we don't get as many independent eigenvectors as the number of times that the eigenvalue $\lambda = 2$ occurs. □

This example shows that there are two kinds of "multiplicities" of an eigenvector. On the one hand there is the number of times that the eigenvalue occurs as a root of the characteristic equation. On the other hand there is the dimension of the corresponding eigenspace. One of these is algebraic in nature, the other is geometric. Here are the appropriate definitions.

Definition 5.5. Algebraic and Geometric Multiplicity Let λ be a root of the characteristic equation $\det(\lambda I - A) = 0$. The *algebraic* multiplicity of λ is the multiplicity of λ as a root of the characteristic equation. The *geometric* multiplicity of λ is the dimension of the space $\mathcal{E}_\lambda(A) = \mathcal{N}(\lambda I - A)$.

We categorize eigenvalues as simple or repeated, according to the following definition.

Definition 5.6. Simple Eigenvalue The eigenvalue λ of A is said to be *simple* if its algebraic multiplicity is 1, that is, the number of times it occurs as a root of the characteristic equation is 1. Otherwise, the eigenvalue is said to be *repeated*.

In Example 5.7 we saw that the repeated eigenvalue $\lambda = 2$ has algebraic multiplicity 2 in both (a) and (b), but geometric multiplicity 2 in (a) and 1 in (b). What can be said in general? The following theorem summarizes the facts. In particular, (2) says that *algebraic multiplicity is always greater than or equal to geometric multiplicity*. Item (1) is immediate since a polynomial of degree n has n roots, counting complex roots and multiplicities. We defer the proof of (2) to Section 5.3.

Theorem 5.3. Let A be an $n \times n$ matrix with characteristic polynomial $p(\lambda) = \det(\lambda I - A)$. Then:

(1) The number of eigenvalues of A, counting algebraic multiplicities and complex numbers, is n.
(2) For each eigenvalue λ of A, if $m(\lambda)$ is the algebraic multiplicity of λ, then

$$1 \leq \dim \mathcal{E}_\lambda(A) \leq m(\lambda).$$

Now when we wrote that each of the matrices of Example 5.7 has eigenvalues $\lambda = 1, 2, 2$, what we intended to indicate was a complete listing of the eigenvalues of the matrix, counting algebraic multiplicities. In particular, $\lambda = 1$ is a simple eigenvalue of the matrices, while $\lambda = 2$ is not. The geometric multiplicities of (a) are identical to the algebraic multiplicities in (a) but not those in (b). The latter kind of matrix is harder to deal with than the former. Following a time-honored custom of mathematicians, we call the more difficult matrix by a less than flattering name, namely, "defective."

Definition 5.7. Defective Matrix A matrix is *defective* if one of its eigenvalues has geometric multiplicity less than its algebraic multiplicity.

Notice that the sum of the algebraic multiplicities of an $n \times n$ matrix is the size n of the matrix. This is due to the fact that the characteristic polynomial of the matrix has degree n, hence exactly n roots, counting multiplicities. Therefore, the sum of the geometric multiplicities of a defective matrix will be less than n.

5.1 Exercises and Problems

Exercise 1. Exhibit all eigenvalues of these matrices.

(a) $\begin{bmatrix} 7 & -10 \\ 5 & -8 \end{bmatrix}$ (b) $\begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 1 & 1 \\ 0 & 3 & 1 \\ 0 & 0 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}$

Exercise 2. Compute the eigenvalues of these matrices.

(a) $\begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 6 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 1+\mathrm{i} & 3 \\ 0 & \mathrm{i} \end{bmatrix}$ (d) $\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 0 & 0 & 1 \end{bmatrix}$ (e) $\begin{bmatrix} 2 & 1 & -1 & -2 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

Exercise 3. Find eigensystems for the matrices of Exercise 1. Specify the algebraic and geometric multiplicity of each eigenvalue.

Exercise 4. Find eigensystems for the matrices of Exercise 2 and identify any defective matrices.

Exercise 5. You are given that the matrix $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ has eigenvalues $1, -1$ and respective eigenvectors $(1, 1)$, $(1, -1)$. Use Theorem 5.2 to determine an eigensystem for $B = \begin{bmatrix} 3 & -5 \\ -5 & 3 \end{bmatrix}$ without further eigensystem calculations.

Exercise 6. You are given that $A = \begin{bmatrix} 2 & -2 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 2 \end{bmatrix}$ and that $\{2, (-1, 0, 1)\}$ and $\{1 + \mathrm{i}, (2, 1 - \mathrm{i}, 0)\}$ are eigenpairs of A. Determine an eigensystem of A without further eigensystem calculations.

Exercise 7. The *trace* of a matrix A is the sum of all the diagonal entries of the matrix and is denoted by $\operatorname{tr} A$. Find the trace of each matrix in Exercise 1 and verify that it is the sum of the eigenvalues of the matrix.

Exercise 8. For each of the matrices in Exercise 2 show that the product of all eigenvalues is the determinant of the matrix.

Exercise 9. Show that for each matrix A of Exercise 1, A and A^T have the same eigenvalues.

Exercise 10. Find all left eigenvectors of each matrix in Exercise 1. Are right and left eigenspaces for each eigenvalue the same?

Exercise 11. For each matrix A of Exercise 1 determine whether $A^T A$ and A^2 have the same eigenvalues. (*Hint:* test eigenvalues of one matrix on the other.)

Exercise 12. For each matrix A of Exercise 2 show that the matrix $B = A^* A$ has nonnegative eigenvalues.

Exercise 13. Let $A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, and let α be an eigenvalue of A, β an eigenvalue of B. Confirm or deny the hypotheses that (a) $\alpha + \beta$ is an eigenvalue of $A + B$, and (b) $\alpha\beta$ is an eigenvalue of AB.

Exercise 14. Let $A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$. Confirm or deny the hypothesis that eigenvalues of AB and BA are the same.

Problem 15. Show that if A is Hermitian, then right and left eigenvalues and eigenvectors coincide.

Problem 16. Show from the definition of eigenvector that if $\mathbf{x}$ is an eigenvector for the matrix A belonging to the eigenvalue λ, then so is $c\mathbf{x}$ for any scalar $c \neq 0$.

*Problem 17. Prove that if A is invertible and λ is an eigenvalue of A, then $1/\lambda$ is an eigenvalue of A^{-1}.

Problem 18. Show that if λ is an eigenvalue of an orthogonal matrix P, then $|\lambda| = 1$.

*Problem 19. Let A be a matrix whose eigenvalues are all less than 1 in absolute value. Show that every eigenvalue of $I - A$ is nonzero and deduce that $I - A$ is invertible.

*Problem 20. Show that A and A^T have the same eigenvalues.

Problem 21. Let A be a real matrix and $\{\lambda, \mathbf{x}\}$ an eigenpair for A. Show that $\{\overline{\lambda}, \overline{\mathbf{x}}\}$ is also an eigenpair for A.

Problem 22. Let A be a square matrix and $f(x)$ an arbitrary polynomial. Show that if λ is an eigenvalue A, then $f(\lambda)$ is an eigenvalue of $f(A)$.

*Problem 23. Show that if A and B are the same size, then AB and BA have the same eigenvalues.

Problem 24. Let T_k be the $k \times k$ tridiagonal matrix whose diagonal entries are 2 and off-diagonal nonzero entries are -1. Use a technology tool to build an array y of length 30 whose kth entry is the minimum of the absolute value of the eigenvalues of T_{k+1}. Plot this array. Use the graph as a guide and try to approximate $y(k)$ as a simple function of k.

5.2 Similarity and Diagonalization

Diagonalization and Matrix Powers

Eigenvalues: Why are they important? This is a good question and has many answers. We will try to demonstrate their importance by focusing on one special class of problems, namely, *discrete linear dynamical systems,* which were defined in Section 2.3. We have seen examples of this kind of system before, namely in Markov chains and difference equations. Here is the sort of question that we would like to answer: when is it the case that there is a limiting vector $\mathbf{x}$ for this sequence of vectors, and if so, how does one compute this vector? The answer to this question will explain the behavior of the Markov chain that was introduced in Example 2.19.

Discrete Linear Dynamical System

If there is such a limiting vector $\mathbf{x}$ for a Markov chain, we saw in Example 3.31 how to proceed: find the null space of the matrix $I - A$, that is, the set of all solutions to the system $(I - A)\mathbf{x} = 0$. However, the question whether

all initial states $\mathbf{x}^{(0)}$ lead to this limiting vector is a more subtle issue, which requires the insights of the next section. We've already done some work on this problem. We saw in Section 2.3 that the entire sequence of vectors is uniquely determined by the initial vector and the transition matrix A in the explicit formula

$$\mathbf{x}^{(k)} = A^k \mathbf{x}^{(0)}.$$

Before proceeding further, let's consider another example that will indicate why we would be interested in limiting vectors.

Example 5.8. By some unfortunate accident a new species of frog has been introduced into an area where it has too few natural predators. In an attempt to restore the ecological balance, a team of scientists is considering introducing a species of bird that feeds on this frog. Experimental data suggests that the population of frogs and birds from one year to the next can be modeled by linear relationships. Specifically, it has been found that if the quantities F_k and B_k represent the populations of the frogs and birds in the kth year, then

$$\begin{aligned} B_{k+1} &= 0.6B_k + 0.4F_k, \\ F_{k+1} &= -rB_k + 1.4F_k, \end{aligned}$$

is a system that models their joint behavior reasonably well. Here the positive number r is a kill rate that measures the consumption of frogs by birds. It varies with the environment, depending on factors such as the availability of other food for the birds. Experimental data suggests that in the environment where the birds are to be introduced, $r = 0.35$. The question is this: in the long run, will the introduction of the birds reduce or eliminate growth of the frog population?

Solution. The discrete dynamical system concept introduced in the preceding discussion fits this situation very nicely. Let the population vector in the kth year be $\mathbf{x}^{(k)} = (B_k, F_k)$. Then the linear relationship above becomes

$$\begin{bmatrix} B_{k+1} \\ F_{k+1} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 \\ -0.35 & 1.4 \end{bmatrix} \begin{bmatrix} B_k \\ F_k \end{bmatrix},$$

which is a discrete linear dynamical system. Notice that this is different from the Markov chains we studied earlier, since one of the entries of the coefficient matrix is negative. Before we can finish solving this example we need to have a better understanding of discrete dynamical systems and the relevance of eigenvalues. □

Let's try to understand how state vectors change in the general discrete dynamical system. We have $\mathbf{x}^{(k)} = A^k \mathbf{x}^{(0)}$. So what we really need to know is how the powers of the transition matrix A behave. In general, this is very hard!

Here is an easy case we can handle: what if $A = [a_{ij}]$ is diagonal? Since we'll make extensive use of diagonal matrices, let's recall a notation that was

introduced in Chapter 2. The matrix $\operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the $n \times n$ diagonal matrix with entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ down the diagonal. For example,

$$\operatorname{diag}\{\lambda_1, \lambda_2, \lambda_3\} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}.$$

By matching up the ith row and jth column of A we see that the only time we could have a nonzero entry in A^2 is when $i = j$, and in that case the entry is a_{ii}^2. A similar argument applies to any power of A. In summary, we have this handy fact.

Theorem 5.4. If $D = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, then for all positive integers k, $D^k = \operatorname{diag}\{\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k\}$.

Just as an aside, this theorem has a very interesting consequence. We have seen in some exercises that if $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ is a polynomial, we can evaluate $f(x)$ at the square matrix A as long as we understand that the constant term a_0 is evaluated as $a_0 I$. In the case of a diagonal A, the following fact reduces evaluation of $f(A)$ to scalar calculations.

Corollary 5.1. If $D = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $f(x)$ is a polynomial, then

$$f(D) = \operatorname{diag}\{f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)\}.$$

Proof. Observe that if $f(x) = a_0 + a_1 x + \cdots + a_n x^n$, then $f(D) = a_0 I + a_1 D + \cdots + a_n D^n$. Now apply the preceding theorem to each monomial D^k and add up the resulting terms in $f(D)$. □

Now for the powers of a more general A. For ease of notation, let's consider a 3×3 matrix A. What if we could find three linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$? We would have $A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$, $A\mathbf{v}_2 = \lambda_2 \mathbf{v}_2$, and $A\mathbf{v}_3 = \lambda_3 \mathbf{v}_3$. In matrix form,

$$A[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \operatorname{diag}\{\lambda_1, \lambda_2, \lambda_3\}.$$

Now set $P = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$ and $D = \operatorname{diag}\{\lambda_1, \lambda_2, \lambda_3\}$. Then P is invertible since the columns of P are linearly independent. (Remember that any nonzero solution to $A\mathbf{x} = \mathbf{0}$ would give rise to a nontrivial linear combination of the column of A that sums to $\mathbf{0}$.) So the equation $AP = PD$, if multiplied on the left by P^{-1}, gives the equation

$$P^{-1}AP = D.$$

This is a beautiful equation, because it makes the powers of A simple to understand. The procedure we just went through is reversible as well. In other

words, if P is an invertible matrix such that $P^{-1}AP = D$, then we deduce that $AP = PD$, identify the columns of P by the equation $P = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$, and conclude that the columns of P are linearly independent eigenvectors of A. We make the following definition and follow it with a simple but key theorem relating similar matrices.

Definition 5.8. **Similar Matrices** A matrix A is said to be *similar* to matrix B if there exists an invertible matrix P such that

$$P^{-1}AP = B.$$

The matrix P is called a *similarity transformation* matrix.

Similarity has an interesting interpretation from the perspective of matrix operators. Recall some facts about coordinates from Section 3.3: If $B = \{\mathbf{e}_1, \ldots \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$ or $\mathbb{C}^n$ and $C = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is any other basis, then the change of basis matrix is $P = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$. The value of this matrix is that if a vector $\mathbf{v}$ is expressed in standard coordinates $\mathbf{v} = (v_1, \ldots, v_n) = [\mathbf{v}]_B$ or coordinates with respect to C, $[\mathbf{v}]_C = (c_1, \ldots, c_n)$ meaning $\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, then $[\mathbf{v}]_B = P[\mathbf{v}]_C$ (Theorem 3.8). Now suppose that the linear operator $T = T_A$ maps $\mathbb{R}^n$ or $\mathbb{C}^n$ into itself and is expressed in standard coordinates: $\mathbf{w} = T(\mathbf{v})$ means that $[\mathbf{w}]_B = A[\mathbf{v}]_B$. To see how do we express this in terms of C coordinates, simply plug in the change of coordinates formula to obtain $P[\mathbf{w}]_C = AP[\mathbf{v}]_C$, from which it follows that $[\mathbf{w}]_C = P^{-1}AP[\mathbf{v}]_C$. Consequently the operator T acts like matrix multiplication by the matrix $P^{-1}AP$ on the C coordinates of a vector.

A simple size check shows that similar matrices have to be square and of the same size. Furthermore, if A is similar to B, then B is similar to A. To see this, suppose that $P^{-1}AP = B$ and multiply by P on the left and P^{-1} on the right to obtain that

$$A = PP^{-1}APP^{-1} = PBP^{-1} = (P^{-1})^{-1}BP^{-1}.$$

Similar matrices have much in common. For example, suppose that $B = P^{-1}AP$ and λ is an eigenvalue of A, say $A\mathbf{x} = \lambda\mathbf{x}$. One calculates

$$\lambda P^{-1}\mathbf{x} = P^{-1}A\mathbf{x} = P^{-1}AP\left(P^{-1}\mathbf{x}\right),$$

from which it follows that λ is an eigenvalue of B. Here is a slightly stronger statement.

Theorem 5.5. Suppose that A is similar to M, say $P^{-1}AP = M$. Then:

(1) For every polynomial $q(x)$,

$$q(M) = P^{-1}q(A)P.$$

(2) The matrices A and M have the same characteristic polynomial, hence the same eigenvalues.
(3) If P is a change of basis matrix from basis C to the standard basis B, then the matrix of the linear operator T_A with respect to the basis C is $P^{-1}AP$.

Proof. We see that successive terms $P^{-1}P$ cancel out in the k-fold product

$$M^k = (P^{-1}AP)(P^{-1}AP)\cdots(P^{-1}AP)$$

to give that

$$M^k = P^{-1}A^kP.$$

It follows easily that

$$a_0I + a_1M + \cdots + a_mM^m = P^{-1}(a_0I + a_1A + \cdots + a_mA^m)P,$$

which proves (1). For (2), remember that the determinant distributes over products, so that we can pull this clever little trick:

$$\begin{aligned}\det(\lambda I - M) &= \det(\lambda P^{-1}IP - P^{-1}AP) = \det(P^{-1}(\lambda I - A)P)\\ &= \det(P^{-1})\det(\lambda I - A)\det(P)\\ &= \det(\lambda I - A)\det(P^{-1}P) = \det(\lambda I - A).\end{aligned}$$

This proves (2). Item (3) was proved in the discussion preceding this theorem. □

Now we can see the significance of the equation $P^{-1}AP = D$, where D is diagonal: For any positive integer k, we have $P^{-1}A^kP = D^k$, so multiplying on the left by P and on the right by P^{-1} yields

$$A^k = PD^kP^{-1}. \tag{5.1}$$

As we have seen, the term PD^kP^{-1} is easily computed. This gives us a way of constructing a formula for A^k.

We can also use this identity to extend part (1) to transcendental functions like $\sin x$, $\cos x$, and e^x, which can be defined in terms of an infinite series (a limit of polynomials functions). One can show that for such functions $f(x)$, if $f(D)$ is defined, then $f(A) = Pf(D)P^{-1}$ uniquely defines $f(A)$. In particular, if $D = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, then we have $f(D) = \operatorname{diag}\{f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)\}$. Thus, we can define $f(A)$ for any matrix A similar to a diagonal matrix provided that $f(x)$ is defined for all scalars x.

Functions of Matrices

Example 5.9. Illustrate the preceding discussion with the matrix in part (a) of Example 5.7 and $f(x) = \sin\left(\frac{\pi}{2}x\right)$.

Solution. The eigenvalues of this problem are $\lambda = 1, 2, 2$. We already found the eigenspace for $\lambda = 2$. Denote the two basis vectors by $\mathbf{v}_1 = (1, 0, 0)$ and $\mathbf{v}_2 = (0, -2, 1)$. For $\lambda = 1$, apply Gauss–Jordan elimination to the matrix

$$A - 1I = \begin{bmatrix} 2-1 & 1 & 2 \\ 0 & 1-1 & -2 \\ 0 & 0 & 2-1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & -2 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow[E_{13}(-2)]{E_{23}(2)} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

which gives a general eigenvector of the form

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -x_2 \\ x_2 \\ 0 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}.$$

Hence, the eigenspace $\mathcal{E}_1(A)$ has basis $\{(-1, 1, 0)\}$. Now set $\mathbf{v}_3 = (-1, 1, 0)$. Form the matrix

$$P = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -2 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

This matrix is nonsingular since $\det P = -1$, and a calculation, which we leave to the reader, shows that

$$P^{-1} = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}.$$

The discussion of the first part of this section shows us that P is a similarity transformation matrix that diagonalizes A, that is,

$$P^{-1}AP = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = D.$$

As we have seen, this means that for any positive integer k, we have

$$A^k = PD^kP^{-1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -2 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2^k & 0 & 0 \\ 0 & 2^k & 0 \\ 0 & 0 & 1^k \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$
$$= \begin{bmatrix} 2^k & 2^k - 1 & 2^{k+1} - 2 \\ 0 & 1 & -2^{k+1} + 2 \\ 0 & 0 & 2^k \end{bmatrix}.$$

This is the formula we were looking for. It's *much* easier than calculating A^k directly!

For $\sin\left(\frac{\pi}{2}A\right)$, we have $\frac{\pi}{2}A = P\frac{\pi}{2}DP^{-1}$, so that $f\left(\frac{\pi}{2}A\right) = Pf\left(\frac{\pi}{2}D\right)P^{-1}$. Thus,

$$\sin\left(\frac{\pi}{2}A\right) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -2 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \sin\left(2\frac{\pi}{2}\right) & 0 & 0 \\ 0 & \sin\left(2\frac{\pi}{2}\right) & 0 \\ 0 & 0 & \sin\left(1\frac{\pi}{2}\right) \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & -1 & -2 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}.$$

Similarly, we could evaluate this matrix A at any other transcendental function that is defined at the eigenvalues of A. □

This example showcases some very nice calculations. When can we pull off the same sort of calculation for a general matrix A? First, let's give the favorable case a name.

Definition 5.9. Diagonalizable Matrix The matrix A is *diagonalizable* if it is similar to a diagonal matrix, that is, there is an invertible matrix P and diagonal matrix D such that $P^{-1}AP = D$. In this case we say that P is a *diagonalizing matrix* for A or that P *diagonalizes* A.

Can we be more specific about when a matrix is diagonalizable? We can. As a first step, notice that the calculations that we began the section with can easily be written in terms of an $n \times n$ matrix instead of a 3×3 matrix. What these calculations prove is the following basic fact.

Theorem 5.6. Diagonalization Theorem The $n \times n$ matrix A is diagonalizable if and only if there exists a linearly independent set of eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of A, in which case $P = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ is a diagonalizing matrix for A.

Can we be more specific about when a linearly independent set of eigenvectors exists? Actually, we can. Clues about what is really going on can be gleaned from a reexamination of Example 5.7.

Example 5.10. Apply the results of the preceding discussion to the matrix in part (b) of Example 5.7 or explain why they fail to apply.

Solution. The eigenvalues of this problem are $\lambda = 1, 2, 2$. We already found the eigenspace for $\lambda = 2$. Denote the single basis vector of $\mathcal{E}_2(A)$ by $\mathbf{v}_1 = (1, 0, 0)$. For $\lambda = 1$, apply Gauss–Jordan elimination to the matrix

$$A - 1I = \begin{bmatrix} 2-1 & 1 & 1 \\ 0 & 1-1 & 1 \\ 0 & 0 & 2-1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow[E_{21}(-1)]{E_{32}(-1)} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

which gives a general eigenvector of the form

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -x_2 \\ x_2 \\ 0 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}.$$

Hence, the eigenspace $\mathcal{E}_1(A)$ has basis $\{(-1,1,0)\}$. All we could come up with here is two eigenvectors. As a matter of fact, they are linearly independent since one is not a multiple of the other. But they aren't enough and there is no way to find a third eigenvector, since we have found them all! Therefore, we have no hope of diagonalizing this matrix according to the diagonalization theorem. The problem is that A is defective, since the algebraic multiplicity of $\lambda = 2$ exceeds the geometric multiplicity of this eigenvalue. □

It would be very handy to have some working criterion for when we can manufacture linearly independent sets of eigenvectors. The next theorem gives us such a criterion.

Theorem 5.7. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ be a set of eigenvectors of the matrix A such that corresponding eigenvalues are all distinct. Then the set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ is linearly independent.

Proof. Suppose the set is linearly dependent. Then there is some nontrivial linear combination with the fewest terms of the form

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_m\mathbf{v}_m = \mathbf{0} \tag{5.2}$$

with each $c_j \neq 0$ and $\mathbf{v}_j$ belonging to the eigenvalue λ_j, where the $\mathbf{v}_i$ are relabelled, if necessary. Multiply (5.2) by λ_1 to obtain the equation

$$c_1\lambda_1\mathbf{v}_1 + c_2\lambda_1\mathbf{v}_2 + \cdots + c_m\lambda_1\mathbf{v}_m = \mathbf{0}. \tag{5.3}$$

Next multiply (5.2) on the left by A to obtain

$$\mathbf{0} = A(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_m\mathbf{v}_m) = c_1A\mathbf{v}_1 + c_2A\mathbf{v}_2 + \cdots + c_mA\mathbf{v}_m,$$

that is,

$$c_1\lambda_1\mathbf{v}_1 + c_2\lambda_2\mathbf{v}_2 + \cdots + c_k\lambda_m\mathbf{v}_m = \mathbf{0}. \tag{5.4}$$

Now subtract (5.4) from (5.3) to obtain

$$0\mathbf{v}_1 + c_2(\lambda_1 - \lambda_2)\mathbf{v}_2 + \cdots + c_k(\lambda_1 - \lambda_m)\mathbf{v}_m = \mathbf{0}.$$

This is a new nontrivial linear combination (since $c_2(\lambda_1 - \lambda_2) \neq 0$) of fewer terms, that contradicts our choice of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$. It follows that the original set of vectors must be linearly independent. □

Actually, a little bit more is true: if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ is such that for any eigenvalue λ of A, the subset of all these vectors belonging to λ is linearly independent, then the conclusion of the theorem is valid. We leave this as an exercise. Here's an application of the theorem that is useful for many problems.

Corollary 5.2. If the $n \times n$ matrix A has n distinct eigenvalues, then A is diagonalizable.

Proof. We can always find one nonzero eigenvector $\mathbf{v}_i$ for each eigenvalue λ_i of A. By the preceding theorem, the set $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is linearly independent. Thus, A is diagonalizable by the diagonalization theorem. □

Caution: Just because the $n \times n$ matrix A has fewer than n distinct eigenvalues, you may not conclude that it is not diagonalizable.

A simple example is the identity matrix, which is certainly diagonalizable (it's already diagonal!) but has only 1 as an eigenvalue.

5.2 Exercises and Problems

Exercise 1. Are the following matrices diagonalizable?

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

Exercise 2. Use eigensystems to determine whether the following matrices are diagonalizable.

(a) $\begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 0 \\ 3 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 1 & -1 & -1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}$

Exercise 3. Find a matrix P such that $P^{-1}AP$ is diagonal.

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

Exercise 4. For each matrix A in Exercise 3 use the matrix P to find a formula for A^k, k a positive integer.

Exercise 5. Given a matrix A, let $q(x)$ be the product of linear factors $x - \lambda$, where λ runs over each eigenvalue of A exactly once. For each of the following matrices, confirm or deny the hypothesis that $q(A) = 0$ if and only if A is diagonalizable.

(a) $\begin{bmatrix} 2 & 0 \\ 3 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 0 \\ 3 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$

Exercise 6. Given a matrix A, let $p(x)$ be the characteristic polynomial of A. For each of the matrices of Exercise 5, confirm or deny the hypothesis that if $p(A) = 0$, then A is diagonalizable.

Exercise 7. Show that the matrix $J_2(\lambda) = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$ is not diagonalizable for any scalar λ and calculate the second, third, and fourth powers of the matrix. What is a formula for $J_2(\lambda)^k$, k a positive integer, based on these calculations?

Exercise 8. Show that the matrix $J_3(\lambda) = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$ is not diagonalizable and calculate the third, fourth, and fifth powers of the matrix. What is a formula for $J_3(\lambda)^k$, $k > 2$, based on these calculations?

Exercise 9. Show that the matrices $A = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 6 \\ 0 & -2 \end{bmatrix}$ are similar as follows: find diagonalizing matrices P, Q for A, B, respectively, that yield identical diagonal matrices, set $S = PQ^{-1}$, and confirm that $S^{-1}AS = B$.

Exercise 10. Repeat Exercise 9 for the pair $A = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ and $B = \begin{bmatrix} 3 & 2 & 2 \\ 0 & 3 & 0 \\ 0 & -1 & 2 \end{bmatrix}$.

Exercise 11. Compute $\sin\left(\frac{\pi}{6}A\right)$ and $\cos\left(\frac{\pi}{6}A\right)$, where $A = \begin{bmatrix} 2 & 4 \\ 0 & -3 \end{bmatrix}$.

Exercise 12. Compute $\exp(A)$ and $\arctan(A)$, where $A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$.

*Problem 13. Show by example that a non-diagonal matrix A can be diagonalized by more than one matrix P.

*Problem 14. Show that any upper triangular matrix with identical diagonal entries is diagonalizable if and only if it is already diagonal.

Problem 15. Suppose that A is an invertible matrix that is diagonalized by the matrix P, that is, $P^{-1}AP = D$ is a diagonal matrix. Use this information to find a diagonalization for A^{-1}.

*Problem 16. Show that if A has no repeated eigenvalues, then the only matrices that commute with A are matrices with the same eigenvectors as A.

Problem 17. Show that if A is diagonalizable, then so is A^*.

*Problem 18. Prove the Cayley–Hamilton theorem for diagonalizable matrices: If $p(x)$ is the characteristic polynomial of the diagonalizable matrix A, then A satisfies its characteristic equation, that is, $p(A) = 0$.

Problem 19. Adapt the proof of Theorem 5.7 to prove that if eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are such that for any eigenvalue λ of A, the subset of all these vectors belonging to λ is linearly independent, then the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are linearly independent.

*Problem 20. The thirteenth-century mathematician Leonardo Fibonacci discovered the sequence of integers $1, 1, 2, 3, 5, 8, \ldots$ called the *Fibonacci sequence*. These numbers have a way of turning up in many applications. They can be specified by the formulas

$$\begin{aligned} f_0 &= 1 \\ f_1 &= 1 \\ f_{k+2} &= f_{k+1} + f_k, \qquad k = 0, 1, \ldots . \end{aligned}$$

(a) Let $\mathbf{x}^{(k)} = (f_{k+1}, f_k)$ and show that these equations are equivalent to the matrix equations $\mathbf{x}^{(0)} = (1, 1)$ and $\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}$, $n = 0, 1, \ldots$, where $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$.
(b) Use part (a) and the diagonalization theorem to find an explicit formula for the kth Fibonacci number.

Problem 21. Suppose that the kill rate r of Example 5.8 is viewed as a variable positive parameter. There is a value of the number r for which the eigenvalues of the corresponding matrix are equal.

(a) Find this value of r and the corresponding eigenvalues by examining the characteristic polynomial of the matrix.

(b) Use a technology tool to determine experimentally the long-term behavior of populations for the value of r found in (a). Your choices of initial states should include $(100, 1000)$.

*Problem 22. Let A and B be matrices of the same size and suppose that A has no repeated eigenvalues. Show that $AB = BA$ if and only if A and B are simultaneously diagonalizable, that is, a single matrix P diagonalizes both A and B.

Problem 23. Prove or disprove: If $n \times n$ matrices A, B are similar to C, D, respectively, then AB is similar to CD.

Problem 24. Show that if $S^{-1}AS = B$ and λ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}$, then λ is an eigenvalue of B with corresponding eigenvector $S^{-1}\mathbf{x}$.

Problem 25. Show that similar matrices have the same rank.

5.3 Applications to Discrete Dynamical Systems

Now we have enough machinery to come to a fairly complete understanding of the discrete dynamical system

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}.$$

Diagonalizable Transition Matrix

Let us first examine the case that A is diagonalizable. So we assume that the $n \times n$ matrix A is diagonalizable and that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a complete linearly independent set of eigenvectors of A belonging to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of A. Let us further suppose that these eigenvalues are ordered so that

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

The eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ form a basis of $\mathbb{R}^n$ or $\mathbb{C}^n$, whichever is appropriate. In particular, we may write $\mathbf{x}^{(0)}$ as a linear combination of these vectors by solving the system $[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]\,\mathbf{c} = \mathbf{x}^{(0)}$ to obtain the coefficients $c_1, c_2, \ldots, c_n$ of the equation

$$\mathbf{x}^{(0)} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n. \tag{5.5}$$

Now we can see what the effect of multiplication by A is:

$$\begin{aligned} A\mathbf{x}^{(0)} &= A(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n) \\ &= c_1(A\mathbf{v}_1) + c_2(A\mathbf{v}_2) + \cdots + c_n(A\mathbf{v}_n) \\ &= c_1\lambda_1\mathbf{v}_1 + c_2\lambda_2\mathbf{v}_2 + \cdots + c_n\lambda_n\mathbf{v}_n. \end{aligned}$$

Now apply A on the left repeatedly. Since $\mathbf{x}^{(k)} = A^k\mathbf{x}^{(0)}$, we see that

$$\mathbf{x}^{(k)} = c_1\lambda_1^k\mathbf{v}_1 + c_2\lambda_2^k\mathbf{v}_2 + \cdots + c_n\lambda_n^k\mathbf{v}_n. \tag{5.6}$$

Equation (5.6) is the key to understanding how the state vector changes in a discrete dynamical system. Now we can see clearly that it is the size of the eigenvalues that governs the growth of successive states. Because of this fact, a handy quantity that can be associated with a matrix A (whether it is diagonalizable or not) is the following.

Definition 5.10. **Spectral Radius and Dominant Eigenvalue** The *spectral radius* $\rho(A)$ of a matrix A with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ is defined to be the number

$$\rho(A) = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}.$$

If $|\lambda_k| = \rho(A)$ and λ_k is the only eigenvalue with this property, then λ_k is the *dominant eigenvalue* of A.

Thus, $\rho(A)$ is the largest absolute value of the eigenvalues of A. We summarize a few of the conclusions about a matrix that can be drawn from the spectral radius.

Theorem 5.8. Let the transition matrix for a discrete dynamical system be the $n \times n$ diagonalizable matrix A. Let $\mathbf{x}^{(0)}$ be an initial state vector given as in equation (5.5). Then the following are true:

(1) If $\rho(A) < 1$, then $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{0}$.
(2) If $\rho(A) = 1$, then the sequence of norms $\left\{\left\|\mathbf{x}^{(k)}\right\|\right\}_{k=0}^{\infty}$ is bounded.
(3) If $\rho(A) = 1$ and $\lambda = 1$ is the dominant eigenvalue of A, then $\lim_{k\to\infty} \mathbf{x}^{(k)}$ is an element of $\mathcal{E}_1(A)$, hence either an eigenvector or $\mathbf{0}$.
(4) If $\rho(A) > 1$, then for some choices of $\mathbf{x}^{(0)}$ we have $\lim_{k\to\infty} \|\mathbf{x}\| = \infty$.

Proof. We may assume that the eigenvalues and eigenvectors of A are organized as in the discussion preceding this theorem. Suppose that $\rho(A) < 1$. Then for all i, $\lambda_i^k \to 0$ as $k \to \infty$, so we see from equation (5.6) that $\mathbf{x}^{(k)} \to 0$ as $k \to \infty$, which is what (1) says. Next suppose that $\rho(A) = 1$. Then take norms of equation (5.6) to obtain that, since each $|\lambda_i| \leq 1$,

$$\begin{aligned}\left\|\mathbf{x}^{(k)}\right\| &= \left\|c_1\lambda_1^k\mathbf{v}_1 + c_2\lambda_2^k\mathbf{v}_2 + \cdots + c_n\lambda_n^k\mathbf{v}_n\right\| \\ &\leq |\lambda_1|^k \|c_1\mathbf{v}_1\| + |\lambda_2|^k \|c_2\mathbf{v}_2\| + \cdots + |\lambda_n|^k \|c_n\mathbf{v}_n\| \\ &\leq \|c_1\mathbf{v}_1\| + \|c_2\mathbf{v}_2\| + \cdots + \|c_n\mathbf{v}_n\|.\end{aligned}$$

Therefore, the sequence of norms $\left\|\mathbf{x}^{(k)}\right\|$ is bounded by a constant that depends only on $\left\|\mathbf{x}^{(0)}\right\|$, which proves (2). The proof of (3) follows from inspection of equation (5.6): observe that the eigenvalue powers λ_j^k are equal to 1 if $\lambda = 1$, and otherwise the powers tend to zero, since all other eigenvalues are less than 1 in absolute value. Hence, if any coefficient c_j of an eigenvector $\mathbf{v}_j$ corresponding to 1 is not zero, the limiting vector is an eigenvector corresponding to $\lambda = 1$. Otherwise, the coefficients all tend to 0 and the limiting vector is $\mathbf{0}$. Finally, if $\rho(A) > 1$, then for $\mathbf{x}^{(0)} = c_1\mathbf{v}_1$, we have that $\mathbf{x}^{(k)} = c_1\lambda_1^k\mathbf{v}_1$. However, $|\lambda_1| > 1$, so that $\left|\lambda_1^k\right| \to \infty$, as $k \to \infty$, from which (4) follows. □

We should note that the cases of the preceding theorem are not quite exhaustive. One possibility that is not covered is the case that $\rho(A) = 1$ and A has other eigenvalues of absolute value 1. In this case the sequence of vectors $\mathbf{x}^{(k)}$ is bounded in norm, i.e., $\left\|\mathbf{x}^{(k)}\right\| \leq K$ for some constant K and indices $k = 0, 1, \ldots$, but need not converge to anything. An example of this phenomenon is given in Example 5.13.

Example 5.11. Apply the preceding theory to the population of Example 5.8.

Solution. We saw in this example that the transition matrix is

$$A = \begin{bmatrix} 0.6 & 0.4 \\ -0.35 & 1.4 \end{bmatrix}.$$

The characteristic equation of this matrix is

$$\begin{aligned} \det \begin{bmatrix} 0.6-\lambda & 0.4 \\ -0.35 & 1.4-\lambda \end{bmatrix} &= (0.6-\lambda)(1.4-\lambda) + 0.35 \cdot 0.4 \\ &= \lambda^2 - 2\lambda + 0.84 + 0.14 \\ &= \lambda^2 - 2\lambda + 0.98, \end{aligned}$$

whence we see that the eigenvalues of A are

$$\lambda = 1.0 \pm \sqrt{4 - 3.92}/2 \approx 1.1414,\ 0.85858.$$

A calculation that we leave to the reader also shows that the eigenvectors of A corresponding to these eigenvalues are approximately $\mathbf{v}_1 = (1.684, 2.2794)$ and $\mathbf{v}_2 = (.8398, .54289)$, respectively. Since $\rho(A) \approx 1.1414 > 1$, it follows from (1) of Theorem 5.8 that for every initial state except a multiple of $\mathbf{v}_2$, the limiting state will grow without bound. Now if we imagine an initial state to be a random choice of values for the coefficients c_1 and c_2, we see that the probability of selecting $c_1 = 0$ is for all practical purposes 0. Therefore, with probability 1, we will make a selection with $c_1 \neq 0$, from which it follows that the subsequent states will tend to arbitrarily large multiples of the vector $\mathbf{v}_1 = (1.684, 2.2794)$.

Finally, we can offer some advice to the scientists who are thinking of introducing a predator bird to control the frog population of this example: Don't do it! Almost any initial distribution of birds and frogs will result in a population of birds and frogs that grows without bound. Therefore, we will be stuck with both non-indigenous frogs *and* birds. To drive the point home, start with a population of 10,000 frogs and 100 birds. In 20 years we will have a population state of

$$\begin{bmatrix} 0.6 & 0.4 \\ -0.35 & 1.4 \end{bmatrix}^{20} \begin{bmatrix} 100 \\ 10{,}000 \end{bmatrix} \approx \begin{bmatrix} 197{,}320 \\ 267{,}550 \end{bmatrix}.$$

In view of our eigensystem analysis, we know that these numbers are no fluke. Almost any initial population will grow similarly. The conclusion is that we should try another strategy or leave well enough alone in this ecology. □

Example 5.12. Apply the preceding theory to the Markov chain Example 2.19.

Solution. Recall that this example led to a Markov chain whose transition matrix is

$$A = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix}.$$

Conveniently, we have already computed the eigenvalues and vectors of $10A$ in Example 5.2. There we found eigenvalues $\lambda = 10, 3$, with corresponding

eigenvectors $\mathbf{v}_2 = (1, -1)$ and $\mathbf{v}_1 = (4/3, 1)$, respectively. It follows that the eigenvalues of A are $\lambda = 1, 0.3$, with the same eigenvectors. Therefore, 1 is the dominant eigenvalue. Any initial state will necessarily involve $\mathbf{v}_1$ nontrivially, since multiples of $\mathbf{v}_2$ are not probability distribution vectors (the entries are of opposite signs). Thus, we may apply part 3 of Theorem 5.8 to conclude that for any initial state, the only possible nonzero limiting state vector is some multiple of $\mathbf{v}_1$. Which multiple? Since the sum of the entries of each state vector $\mathbf{x}^{(k)}$ sum to 1, the same must be true of the initial vector. Since

$$\mathbf{x}^{(0)} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = c_1 \begin{bmatrix} 4/3 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} c_1\,(4/3) + c_2 1 \\ c_1 1 + c_2\,(-1) \end{bmatrix},$$

we see that

$$1 = c_1\,(4/3) + c_2 1 + c_1 1 + c_2\,(-1) = c_1(7/3),$$

so that $c_1 = 3/7$. Now use the facts that $\lambda_1 = 1$, $\lambda_2 = 0.3$, and equation (5.6) with $n = 2$ to see that the limiting state vector is

$$\lim_{k\to\infty} c_1 1^k \mathbf{v}_1 + c_2\,(0.3)^k\,\mathbf{v}_2 = c_1\mathbf{v}_1 = \begin{bmatrix} 4/7 \\ 3/7 \end{bmatrix} \approx \begin{bmatrix} .57143 \\ .42857 \end{bmatrix}.$$

Compare this with the calculations in Example 2.20. □

When do complex eigenvalues occur and what do they mean? In general, all we can say is that the characteristic polynomial of a matrix, even if it is real, may have complex roots. This is an unavoidable fact, but it can be instructive. To see how this is so, consider the following example.

Example 5.13. Suppose that a discrete dynamical system has transition matrix $A = \begin{bmatrix} 0 & a \\ -a & 0 \end{bmatrix}$, where a is a positive real number. What can be said about the states $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$, if the initial state $\mathbf{x}^{(0)}$ is an arbitrary nonzero vector?

Solution. The eigenvalues of A are $\pm a\mathrm{i}$. Since the eigenvalues of A are distinct, there is an invertible matrix P such that

$$P^{-1}AP = D = \begin{bmatrix} a\mathrm{i} & 0 \\ 0 & -a\mathrm{i} \end{bmatrix}.$$

So we see from equation (5.1) that

$$A^k = PD^kP^{-1} = P \begin{bmatrix} (a\mathrm{i})^{\mathrm{k}} & 0 \\ 0 & (-a\mathrm{i})^{\mathrm{k}} \end{bmatrix} P^{-1}.$$

The columns of P are eigenvectors of A, hence complex. We may take real parts of the matrix D^k to get a better idea of what the powers of A do. Now $\mathrm{i} = \mathrm{e}^{\mathrm{i}\frac{\pi}{2}}$, so we may use de Moivre's formula to get

$$\Re((ai)^k) = a^k \cos(k\frac{\pi}{2}) = (-1)^{k/2} a^k \quad \text{if } k \text{ is even.}$$

We know that $\mathbf{x}^{(k)} = A^k \mathbf{x}^{(0)}$. In view of the above equation, we see that the states $\mathbf{x}^{(k)}$ will oscillate around the origin. If $a < 1$, the values tend to 0. In the case that $a = 1$ we expect the states to remain bounded, but if $a > 1$, we expect the values to become unbounded. In all cases the values oscillate in sign. This oscillation is fairly typical of what happens when complex eigenvalues are present, though it need not be as rapid as in this example. □

Nondiagonalizable Transition Matrix

How can a matrix be nondiagonalizable? All the examples we have considered so far suggest that nondiagonalizability is the same as being defective. Put another way, diagonalizable equals nondefective. This is exactly right, as the following shows.

Theorem 5.9. The matrix A is diagonalizable if and only if the geometric multiplicity of every eigenvalue equals its algebraic multiplicity.

Proof. Suppose that the $n \times n$ matrix A is diagonalizable. According to the diagonalization theorem, there exists a complete linearly independent set of eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of the matrix A. The number of these vectors belonging to an eigenvalue λ of A is a number $d(\lambda)$ at most the geometric multiplicity of λ, since they form a basis of the eigenspace $\mathcal{E}_\lambda(A)$. Hence, their number is at most the algebraic multiplicity $m(\lambda)$ of λ by Theorem 5.3. Since the sum of all the numbers $d(\lambda)$ is n, as is the sum of all the algebraic multiplicities $m(\lambda)$, it follows that the sum of the geometric multiplicities must also be n. The only way for this to happen is that for each eigenvalue λ, we have that geometric multiplicity equals algebraic multiplicity. Thus, A is nondefective.

Conversely, if geometric multiplicity equals algebraic multiplicity, we can produce $m(\lambda)$ linearly independent eigenvectors belonging to each eigenvalue λ. Assemble all of these vectors and we have n eigenvectors such that for any eigenvalue λ of A, the subset of all these vectors belonging to λ is linearly independent. Therefore, the entire set of eigenvectors is linearly independent by the remark following Theorem 5.7. Now apply the diagonalization theorem to obtain that A is diagonalizable. □

The last item of business in our examination of diagonalization is to prove part 2 of Theorem 5.3, which asserts*: for each eigenvalue μ of A, if $m(\mu)$ is the algebraic multiplicity of μ, then*

$$1 \leq \dim \mathcal{E}_\mu(A) \leq m(\mu).$$

To see why this is true, suppose the eigenvalue μ has geometric multiplicity k and that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ is a basis for the eigenspace $\mathcal{E}_\mu(A)$. We know from

the Steinitz substitution theorem that this set can be expanded to a basis of the vector space $\mathbb{R}^n$ (or $\mathbb{C}^n$), say $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$. Form the nonsingular matrix $S = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ and let

$$B = [S^{-1}A\mathbf{v}_{k+1}, S^{-1}A\mathbf{v}_{k+2}, \ldots, S^{-1}A\mathbf{v}_n] = \begin{bmatrix} F \\ G \end{bmatrix},$$

where F consists of the first k rows of B and G the remaining rows. Thus, we obtain that

$$\begin{aligned} AS &= [A\mathbf{v}_1, A\mathbf{v}_2, \ldots, A\mathbf{v}_n] \\ &= [\mu\mathbf{v}_1, \mu\mathbf{v}_2, \ldots, \mu\mathbf{v}_k, A\mathbf{v}_{k+1}, \ldots, A\mathbf{v}_n] \\ &= S \begin{bmatrix} \mu I_k & F \\ 0 & G \end{bmatrix}. \end{aligned}$$

Now multiply both sides on the left by S^{-1}, and we have

$$C = S^{-1}AS = \begin{bmatrix} \mu I_k & F \\ 0 & G \end{bmatrix}.$$

We see that the block upper triangular matrix C is similar to A. By part 2 of Theorem 5.5 we see that A and C have the same characteristic polynomial. However, the characteristic polynomial of C is

$$\begin{aligned} p(\lambda) &= \det\left(\lambda I_n - \begin{bmatrix} \mu I_k & F \\ 0 & G \end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} (\lambda-\mu) I_k & F \\ 0 & G - \lambda I_{n-k} \end{bmatrix}\right) \\ &= \det(\lambda-\mu) I_k \cdot \det\left(G - \lambda I_{n-k}\right) \\ &= (\lambda-\mu)^k \det\left(G - \lambda I_{n-k}\right). \end{aligned}$$

The product term above results from Exercise 26 of Section 2.6. It follows that the algebraic multiplicity of μ as a root of $p(\lambda)$ is at least as large as k, which is what we wanted to prove. □

Our newfound insight into nondiagonalizable matrices is somewhat of a negative nature: they are defective. Unfortunately, this isn't much help in determining the behavior of discrete dynamical systems with a nondiagonalizable transition matrix. If matrices are not diagonalizable, what simple kind of matrix *are* they reducible to? There is a very nice answer to this question; this answer requires the notion of a Jordan block, which can be defined as a $d \times d$ matrix of the form

Jordan Block

$$J_d(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix},$$

where the entries off the main diagonal and first superdiagonal are understood to be zeros. This matrix is very close to being a diagonal matrix. Its true value comes from the following classical theorem, whose proof we defer to Section 5.7. We refer the reader to [17] and [19] of the bibliography for other applications of this important theorem. These texts are excellent references for higher-level linear algebra and matrix theory.

Theorem 5.10. Jordan Canonical Form Every matrix A is similar to a block diagonal matrix that consists of Jordan blocks down the diagonal. Moreover, these blocks are uniquely determined by A up to order.

In particular, if $J = S^{-1}AS$, where J consists of Jordan blocks down the diagonal, we call J "the" Jordan canonical form of the matrix A, which suggests there is only one. This is a slight abuse of language, since the order of occurrence of the Jordan blocks of J could vary. To fix ideas, let's consider an example.

Example 5.14. Find all possible Jordan canonical forms for a 3×3 matrix A whose eigenvalues are $-2, 3, 3$.

Solution. Notice that each Jordan block $J_d(\lambda)$ contributes d eigenvalues λ to the matrix. Therefore, there can be only one 1×1 Jordan block for the eigenvalue -2 and either two 1×1 Jordan blocks for the eigenvalue 3 or one 2×2 block for the eigenvalue 3. Thus, the possible Jordan canonical forms for A (up to order of blocks) are

$$\begin{bmatrix} -2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}. \qquad \square$$

Notice that if all Jordan blocks are 1×1, then the Jordan canonical form of a matrix is simply a diagonal matrix. Thus, another way to say that a matrix is diagonalizable is to say that its Jordan blocks are 1×1. In reference to the previous example, we see that if the matrix has the first Jordan canonical form, then it is diagonalizable, while if it has the second, it is nondiagonalizable.

Now suppose that the matrix A is a transition matrix for a discrete dynamical system and A is not diagonalizable. What can one say? For one thing, the Jordan canonical form can be used to recover part (1) of Theorem 5.8. Part (4) remains valid as well; the proof we gave does not depend on A being diagonalizable. Unfortunately, things are a bit more complicated as regards parts (2) and (3). In fact, they fail to be true, as the following example shows.

Example 5.15. Show that the following matrices have spectral radius 1 and determine if conclusions of parts (2) and (3) of Theorem 5.8 fail to be true in each case.

(a) $A = J_2(1)$ (b) $A = I_2$ (c) $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$

Solution. (a) The eigenvalues of A are $1, 1$, so $\rho(A) = 1$. We check that for $\mathbf{x}^{(0)} = (0, 1)$

$$A^2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix},$$

and in general,

$$A^k = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}.$$

Now take $\mathbf{x}^{(0)} = (0, 1)$, and we see that

$$\mathbf{x}^{(k)} = A^k \mathbf{x}^{(0)} = \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} k \\ 1 \end{bmatrix}.$$

It follows that the norms $\left\|\mathbf{x}^{(k)}\right\| = \sqrt{k^2 + 1}$ are not a bounded sequence, so that part (2) of the theorem fails to be true. Also, the conclusion of (3) fails.

(b) The eigenvalues of A are $1, 1$, so so $\rho(A) = 1$. Also, $\mathbf{x}^{(k)} = A^{(k)}\mathbf{x}^{(0)} = \mathbf{x}^{(0)}$, so part (2) holds but the conclusion of (3) fails ($\lambda = 1$ is not dominant.)

(c) The eigenvalues of A are $\mathrm{i}, -\mathrm{i}$, so again $\rho(A) = 1$, but one calculates that if $\mathbf{x}^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}$ then $\mathbf{x}^{(1)} = \begin{bmatrix} b \\ -a \end{bmatrix}$, $\mathbf{x}^{(2)} = \begin{bmatrix} -a \\ -b \end{bmatrix}$, $\mathbf{x}^{(3)} = \begin{bmatrix} -b \\ a \end{bmatrix}$, and $\mathbf{x}^{(4)} = \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{x}^{(0)}$. So again (2) holds but the conclusion of (3) fails. □

A helpful tool for analysis of these transition matrices is the following result, whose proof is left as an exercise.

Theorem 5.11. Jordan Block Powers If $J = J_d(\lambda)$ is a Jordan block of size $d > 1$, then one of the following hold:

(1) $|\lambda| < 1$, in which case $J^k \longrightarrow_{k\to\infty} \mathbf{0}$, i.e., each entry of J^k tends to 0 as $k \to \infty$.

(2) $|\lambda| \geq 1$, in which case the magnitude of entries of J above the diagonal tends to ∞.

In spite of Example 5.15, the news is not all negative. The problem with the preceding examples is that each matrix does not have 1 as the *dominant* eigenvalue. We show by way of the Jordan canonical form that parts (2) and (3) do hold in general provided that 1 is the dominant eigenvalue.

In Chapter 3 we defined an *stable* stochastic matrix to be a stochastic matrix with the property that all states of a Markov chain with such a transition matrix converge to a steady state, independent of the initial state. With a general discrete dynamical system we cannot hope for such a result since, for any initial state $\mathbf{x}^{(0)}$ and scalar c, if $\mathbf{x}^{(0)}$ converges to $\mathbf{x}_*$, then by linearity the initial state $c\mathbf{x}^{(0)}$ converges to $c\mathbf{x}_*$ Hence, the following theorem is the best we can do for general discrete dynamical systems:

Theorem 5.12. Stability Theorem The following are equivalent for a discrete dynamical system with transition matrix A:

(1) The matrix A has $\lambda = 1$ as the dominant eigenvalue of A.

(2) There is a vector $\mathbf{x}_*$ such that $\lim_{k\to\infty} \mathbf{x}^{(k)} = c\mathbf{x}_*$, where the scalar c is uniquely determined by the initial state $\mathbf{x}^{(0)}$.

Proof. Assume (1). According to Theorem 5.10 there is an invertible matrix S such that

$$S^{-1}AS = \begin{bmatrix} J_1(1) & 0 & \cdots & 0 \\ 0 & J_{d_2}(\lambda_2) & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & J_{d_m}(\lambda_m) \end{bmatrix} \equiv J \tag{5.7}$$

where $\lambda_1 = 1, \lambda_2, \ldots, \lambda_m$ are eigenvalues of A and the zeros are suitably sized zero matrices. Since 1 is the dominant eigenvalue of A, its Jordan block is the 1×1 matrix $[1]$ and we may assume that $|\lambda_j| < 1$ for $j = 2, \ldots, m$. Clearly $\mathbf{e}_1 = (1, 0, \ldots, 0)$ is an eigenvector of J for the eigenvalue $\lambda = 1$. Set $\mathbf{x}_* = S\mathbf{e}_1$. Given an initial vector $\mathbf{x}^{(0)}$ for a discrete dynamical system with transition matrix A, let c be the first coordinate of $S^{-1}\mathbf{x}^{(0)}$ and we calculate using Theorem 5.11:

$$\mathbf{x}^{(k)} = A^k\mathbf{x}^{(0)} = S \begin{bmatrix} J_1(1)^k & 0 & \cdots & 0 \\ 0 & J_{d_2}(\lambda_2)^k & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & J_{d_m}(\lambda_m)^k \end{bmatrix} S^{-1}\mathbf{x}^{(0)}$$
$$\xrightarrow[k\to\infty]{} cS\mathbf{e}_1 = c\mathbf{x}_*,$$

which proves (2).

Conversely, assume (2). Let $\mathbf{e_j}$, $j = 1, \ldots, n$, be the jth column of the identity. Since an arbitrary vector $\mathbf{v} = \mathbf{x}^{(0)} \in \mathbb{C}^n$ can be expressed as $\mathbf{v} = v_1\mathbf{e}_1 + \cdots + v_n\mathbf{e}_n$ and each $\mathbf{e}_j$ as an initial vector leads to a discrete dynamical system that converges to $c_j\mathbf{x}_*$ for some scalars c_j, $j = 1, \ldots n$, and d we have that

$$\mathbf{x}^{(m)} = A^m\mathbf{x}^{(0)} = \sum_{j=1}^{n} v_j A^m\mathbf{e}_j \xrightarrow[m\to\infty]{} \sum_{j=1}^{n} v_j c_j \mathbf{x}_*$$
$$= \left(\sum_{j=1}^{n} v_j c_j\right)\mathbf{x}_* = d\mathbf{x}_*.$$

In particular, if $\mathbf{x}_j^{(0)} = \mathbf{e}_j$ leads to a dynamical system converging to $d_j\mathbf{x}_*$, then

$$A^m = A^m \left[\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \ldots, \mathbf{x}_n^{(0)}\right] = \left[\mathbf{x}_1^{(m)}, \mathbf{x}_2^{(m)}, \ldots, \mathbf{x}_n^{(m)}\right]$$
$$\xrightarrow[m\to\infty]{} [d_1\mathbf{x}_*, d_2\mathbf{x}_*, \ldots, d_n\mathbf{x}_*] .$$

This limiting matrix is clearly rank one. Rank does not increase with multiplication on either side by an invertible matrix, so the limiting matrix of powers of the Jordan canonical form J as in equation (5.7) should also be rank one. According to Theorem 5.11 this cannot happen unless $|\lambda_j| < 1$, $j = 2, \ldots, n$. Therefore, condition (1) holds. □

This theorem inspires the following definition:

Definition 5.11. **Stable Matrix and DDS** A *stable discrete dynamical system*is one whose transition matrix A has $\lambda = 1$ as its dominant eigenvalue, in which case the matrix A is also called *stable*.

This line of thought brings us back to Markov chains and PageRank once more. Recall that Theorem 2.8 asserts that transition matrix $Q = \alpha P + (1 - \alpha)\mathbf{v}\mathbf{e}^T$, where P is stochastic, $\mathbf{v}$ is a distribution vector and $0 < \alpha < 1$, the equation $Q\mathbf{x} = \mathbf{x}$ has a unique distribution solution vector $\mathbf{x}$ to which all Markov chains with transition matrix Q converge. So Q is a stable stochastic matrix. This raises the question of whether or not stable stochastic matrices are stable in the sense of the previous definition. The answer follows from Theorem 5.12:

Corollary 5.3. If a stochastic matrix Q has a unique stationary distribution vector to which all Markov chains with transition matrix Q converge, then $\lambda = 1$ is the dominant eigenvalue of Q.

Proof. Let Q be $n \times n$ with stationary vector $\mathbf{x}_*$ and let $\mathbf{e_j}$, $j = 1, \ldots, n$, be the jth column of the identity. Since an arbitrary vector $\mathbf{v} = \mathbf{x}^{(0)} \in \mathbb{C}^n$ can be expressed as $\mathbf{v} = v_1\mathbf{e}_1 + \cdots + v_n\mathbf{e}_n$ and each $\mathbf{e}_j$ is a distribution vector we have that

$$\mathbf{x}^{(m)} = Q^m\mathbf{x}^{(0)} = \sum_{j=1}^{n} v_j Q^m \mathbf{e}_j \xrightarrow[m\to\infty]{} \sum_{j=1}^{n} v_j \mathbf{x}_* = \left(\sum_{j=1}^{n} v_j\right) \mathbf{x}_*.$$

It follows from Theorem 5.12 that $\lambda = 1$ is the dominant eigenvalue of Q. □

It is not sufficient in Corollary 5.3 to have a stationary distribution vector: It must be unique. We leave it as an exercise to show that every stochastic matrix has $\lambda = 1$ as an eigenvalue. (In fact, it can be shown with a bit more work that every stochastic matrix Q has a stationary distribution vector.) But this does not imply uniqueness, since I_2 is an obvious counterexample.

5.3 Exercises and Problems

Exercise 1. Find the spectral radius of each of the following matrices and determine whether there is a dominant eigenvalue.

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

Exercise 2. Find the spectral radius and dominant eigenvalue, if any.

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

Exercise 3. For initial state $\mathbf{x}^{(0)}$and transition matrix A below find an eigensystem of A and use this to produce a formula for the kth state $\mathbf{x}^{(k)}$ in the form of equation (5.6).

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

Exercise 4. Repeat Exercise 3 for these pairs $\mathbf{x}^{(0)}$, A.

(a) $\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \frac{1}{2}\begin{bmatrix} 3 & 0 \\ 8 & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

Exercise 5. If the matrices of Exercise 1 are transition matrices, for which do all $\mathbf{x}^{(k)}$ approach $\mathbf{0}$ as $k \to \infty$? Does the stability theorem apply to any of these?

Exercise 6. If the matrices of Exercise 2 are transition matrices, for which do all $\mathbf{x}^{(k)}$ remain bounded as $k \to \infty$? Are any of these matrices stable?

Exercise 7. You are given that a 5×5 matrix has eigenvalues $2, 2, 3, 3, 3$. What are the possible Jordan canonical forms for this matrix?

Exercise 8. What are the possible Jordan canonical forms for a 6×6 matrix with eigenvalues $-1, -1, -1, 4, 4, 4$?

Exercise 9. Let $A = J_3(2)$, a Jordan block. Show that the *Cayley–Hamilton theorem* is valid for A, that is, $p(A) = 0$, where $p(x)$ is the characteristic polynomial of A.

Exercise 10. Let $A = \begin{bmatrix} J_2(1) & 0 \\ 0 & J_2(1) \end{bmatrix}$. Verify that $p(A) = 0$, where $p(x)$ is the characteristic polynomial of A, and find a polynomial $q(x)$ of degree less than 4 such that $q(A) = 0$.

Exercise 11. The three-stage insect model of Example 2.21 yields a transition matrix

$$A = \begin{bmatrix} 0.2 & 0 & 0.25 \\ 0.6 & 0.3 & 0 \\ 0 & 0.6 & 0.8 \end{bmatrix}.$$

Use a technology tool to calculate the eigenvalues of this matrix. Deduce that A is diagonalizable and determine the approximate growth rate from one state to the next, given a random initial vector.

Exercise 12. The financial model of Example 2.27 gives rise to a discrete dynamical system $x^{(k+1)} = Ax^{(k)}$, where the transition matrix is

$$A = \begin{bmatrix} 1 & 0.06 & 0.12 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Use a technology tool to calculate the eigenvalues of this matrix. Deduce that A is diagonalizable and determine the approximate growth rate from one state to the next, given a random initial vector. Compare the growth rate to a constant interest rate that closely matches the model.

Exercise 13. A (two) age structured population model results in a transition matrix $A = \begin{bmatrix} 0 & f_2 \\ s_1 & 0 \end{bmatrix}$ with positive per-capita reproductive rate f_2 and survival rate s_1. There exists a positive eigenpair $(\lambda, \mathbf{p})$ for A. Assume this and use the equation $A\mathbf{p} = \lambda\mathbf{p}$ to express $\mathbf{p} = (p_1, p_2)$ in terms of p_1, and to find a polynomial equation in terms of birth and survival rates that λ satisfies.

Exercise 14. Repeat Exercise 13 for the (three) age structured model with transition matrix $A = \begin{bmatrix} 0 & f_2 & f_3 \\ s_1 & 0 & 0 \\ 0 & s_2 & 0 \end{bmatrix}$ where f_2, f_3, s_1, s_2 are all positive.

*Problem 15. Let A be a 2×2 transition matrix of a Markov chain where A is not the identity matrix.

(a) Show that A can be written in the form $A = \begin{bmatrix} 1-a & b \\ a & 1-b \end{bmatrix}$ for suitable real numbers $0 \leq a, b \leq 1$.
(b) Show that (b, a) and $(1, -1)$ are eigenvectors for A.
(c) Find a formula for the kth state $\mathbf{x}^{(k)}$ in the form of equation (5.6).

Problem 16. Let $A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ be a transition matrix for a discrete dynamical system. Show that A is not stable for any choice of $a, b \in \mathbb{R}$ with $b \neq 0$.

*Problem 17. Show that 1 is an eigenvalue for all stochastic matrices.

Problem 18. Part (3) of Theorem 5.8 suggests that two possible limiting values are possible. Use your technology tool to carry out this experiment: compute a random 2×1 vector and normalize it by dividing by its length. Let the resulting initial vector be $\mathbf{x}^{(0)} = (x_1, x_2)$ and compute the state vector $\mathbf{x}^{(20)}$ using the transition matrix A of Example 5.12. Do this for a large number of times (say 500) and keep count of the number of times $\mathbf{x}^{(20)}$ is close to $\mathbf{0}$, say $\|\mathbf{x}^{(20)}\| < 0.1$. Conclusions?

Problem 19. Use a technology tool to construct a 3×10 table whose jth column is $A^j\mathbf{x}$, where $\mathbf{x} = (1, 1, 1)$ and $A = \begin{bmatrix} 10 & 17 & 8 \\ -8 & -13 & -6 \\ 4 & 7 & 4 \end{bmatrix}$. What can you deduce about the eigenvalues of A based on inspection of this table? Give reasons. Check your claims by finding the eigenvalues of A.

Problem 20. A species of bird can be divided into three age groups: age less than 2 years for group 1, age between 2 and 4 years for group 2, and age between 4 and 6 years for the third group. Assume that these birds have at most a 6-year life span. It is estimated that the survival rates for birds in groups 1 and 2 are 50% and 75%, respectively. Also, birds in groups 1, 2, and 3 produce 0, 1, and 3 offspring on average in any biennium (period of 2 years). Model this bird population as a discrete dynamical system and analyze the long-term change in the population. If the survival rates are unknown, but the population is known to be stable, assume that survival rates for groups 2 and 3 are equal and estimate this number.

Problem 21. Show that if $U = J_d(0)$, then for $k \geq d$, $U^k = 0$ and for $k < d$, U^k is a matrix of all zeros except the kth superdiagonal, which consists of ones.

*Problem 22. Show that if $A = \lambda I + U$, where $U = J_d(0)$, then entries below the main diagonal of A^m are zero and the (j, k)th entry on or above the main diagonal is $\binom{m}{\ell} \lambda^{m-\ell}$, for $k - j = \ell < d < m$.

Problem 23. Use Problem 22 to prove Theorem 5.11. (The exponential form of powers, $a^x = e^{x \ln a}$ is helpful.)

*Problem 24. Use the Jordan Canonical Form theorem to prove the Cayley-Hamilton theorem: If $p(\lambda) = \det(\lambda I - A) = 0$ is the characteristic equation of square matrix A, then $p(A) = 0$.

5.4 Orthogonal Diagonalization

We are going to explore some very remarkable facts about Hermitian and real

Unitarily and Orthogonally Diagonalizable Matrices

symmetric matrices. These matrices are diagonalizable, and moreover, diagonalization can be accomplished by a unitary (orthogonal if A is real) matrix. This means that $P^{-1}AP = P^*AP$ is diagonal. In this situation we say that the matrix A is *unitarily (orthogonally) diagonalizable*. Unitary and orthogonal matrices are particularly attractive since the inverse calculation is essentially free and error-free as well: $P^{-1} = P^*$.

Eigenvalue of Hermitian Matrices

As a first step, we need to observe a curious property of Hermitian matrices. It turns out that their eigenvalues are guaranteed to be real, even if the matrix itself is complex. This is one reason that one might prefer to work with these matrices.

Theorem 5.13. If A is Hermitian, then the eigenvalues of A are real.

Proof. Let λ be an eigenvalue of A with corresponding nonzero eigenvector $\mathbf{x}$, so that $A\mathbf{x} = \lambda\mathbf{x}$. Form the scalar $c = \mathbf{x}^*A\mathbf{x}$. We have that

$$\overline{c} = c^* = (\mathbf{x}^*A\mathbf{x})^* = \mathbf{x}^*A^*(\mathbf{x}^*)^* = \mathbf{x}^*A\mathbf{x} = c.$$

It follows that c is a real number. However, we also have that

$$c = \mathbf{x}^*\lambda\mathbf{x} = \lambda\mathbf{x}^*\mathbf{x} = \lambda\|\mathbf{x}\|^2$$

so that $\lambda = c/\|\mathbf{x}\|^2$ is also real. □

Example 5.16. Show that Theorem 5.13 is applicable if $A = \begin{bmatrix} 1 & 1-\mathrm{i} \\ 1+\mathrm{i} & 0 \end{bmatrix}$ and verify the conclusion of the theorem.

Solution. First notice that

$$A^* = \begin{bmatrix} 1 & 1-\mathrm{i} \\ 1+\mathrm{i} & 0 \end{bmatrix}^* = \begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & 0 \end{bmatrix}^T = \begin{bmatrix} 1 & 1-\mathrm{i} \\ 1+\mathrm{i} & 0 \end{bmatrix} = A.$$

It follows that A is Hermitian and the preceding theorem is applicable. Now we compute the eigenvalues of A by solving the characteristic equation

$$\begin{aligned} 0 = \det(A - \lambda I) &= \det \begin{bmatrix} 1-\lambda & 1-\mathrm{i} \\ 1+\mathrm{i} & -\lambda \end{bmatrix} \\ &= (1-\lambda)(-\lambda) - (1+\mathrm{i})(1-\mathrm{i}) \\ &= \lambda^2 - \lambda - 2 = (\lambda+1)(\lambda-2). \end{aligned}$$

Hence, the eigenvalues of A are $\lambda = -1, 2$, which are real. □

Caution: Although the *eigenvalues* of a Hermitian matrix are guaranteed to be real, the *eigenvectors* may not be real unless the matrix in question is real.

The Principal Axes Theorem

A key fact about Hermitian matrices is the so-called *principal axes theorem*; its proof is a simple consequence of the Schur triangularization theorem which is proved in Section 5.5. We will content ourselves here with stating the theorem and supplying a proof for the case that the eigenvalues of A are distinct. This proof also shows us one way to carry out the diagonalization process.

Theorem 5.14. Principal Axes Theorem Every Hermitian matrix is unitarily diagonalizable, and every real symmetric matrix is orthogonally diagonalizable.

Proof. Let us assume that the eigenvalues of the $n \times n$ matrix A are distinct. We saw in Theorem 5.13 that the eigenvalues of A are real. Let these eigenvalues be $\lambda_1, \lambda_2, \ldots, \lambda_n$. Now find an eigenvector $\mathbf{v}_k$ for each eigenvalue λ_k. We can assume that each $\mathbf{v}_k$ is of unit length by replacing it by the vector divided by its length if necessary. We now have a diagonalizing matrix, as prescribed by Theorem 5.6 (the diagonalization theorem), namely the matrix $P = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$.

Recalling that $A\mathbf{v}_j = \lambda_j \mathbf{v}_j$, $A\mathbf{v}_k = \lambda_k \mathbf{v}_k$, and that $A^* = A$, we see that

$$\lambda_k \mathbf{v}_j^* \mathbf{v}_k = \mathbf{v}_j^* \lambda_k \mathbf{v}_k = \mathbf{v}_j^* A \mathbf{v}_k = (A\mathbf{v}_j)^* \mathbf{v}_k = (\lambda_j \mathbf{v}_j)^* \mathbf{v}_k = \lambda_j \mathbf{v}_j^* \mathbf{v}_k.$$

Now bring both terms to one side of the equation and factor out the term $\mathbf{v}_j^* \mathbf{v}_k$ to obtain

$$(\lambda_k - \lambda_j)\mathbf{v}_j^* \mathbf{v}_k = 0.$$

Thus, if $\lambda_k \neq \lambda_j$, it follows that $\mathbf{v}_j \cdot \mathbf{v}_k = \mathbf{v}_j^* \mathbf{v}_k = 0$. In other words the eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ form an orthonormal set. Therefore, the matrix P is unitary. If A is real, then so are the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and P is orthogonal in this case. □

The proof we have just given suggests a practical procedure for diagonalizing a Hermitian or real symmetric matrix. The only additional information that we need for the complete procedure is advice on what to do if the eigenvalue λ is repeated. This is a sticky point. What we need to do in this case is find an orthogonal basis of the eigenspace $\mathcal{E}_\lambda(A) = \mathcal{N}(A - \lambda I)$. It is always possible to find such a basis using the Gram–Schmidt algorithm (Theorem 4.10). For the hand calculations that we do in this chapter, the worst situation that we will encounter is that the eigenspace $\mathcal{E}_\lambda$ is two-dimensional, say with a basis $\mathbf{v}_1, \mathbf{v}_2$. In this case replace $\mathbf{v}_2$ by $\tilde{\mathbf{v}}_2 = \mathbf{v}_2 - \text{proj}_{\mathbf{v}_1} \mathbf{v}_2$. We know that $\tilde{\mathbf{v}}_2$ is orthogonal to $\mathbf{v}_1$ (see Theorem 6.4), so that $\mathbf{v}_1, \tilde{\mathbf{v}}_2$ is an orthogonal basis of $\mathcal{E}_\lambda(A)$. We illustrate the procedure with a few examples.

Example 5.17. Find an eigensystem for the matrix $A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix}$ and use this to orthogonally diagonalize A.

Solution. Notice that A is real symmetric, so diagonalizable by the principal axes theorem. First calculate the characteristic polynomial of A as

$$\begin{aligned} |A-\lambda I| &= \begin{vmatrix} 1-\lambda & 2 & 0 \\ 2 & 4-\lambda & 0 \\ 0 & 0 & 5-\lambda \end{vmatrix} \\ &= ((1-\lambda)(4-\lambda)-2\cdot 2)\,(5-\lambda) \\ &= -\lambda(\lambda-5)^2, \end{aligned}$$

so that the eigenvalues of A are $\lambda = 0, 5, 5$.

Next find eigenspaces for each eigenvalue. For $\lambda = 0$, we find the null space by row reduction,

$$A - 0I = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix} \xrightarrow[E_{21}(-2)]{} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 5 \end{bmatrix} \xrightarrow[E_2(\frac{1}{5})]{E_{23}} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

so that the null space is spanned by the vector $(-2, 1, 0)$. Normalize this vector to obtain $\mathbf{v}_1 = (-2, 1, 0)/\sqrt{5}$. Next compute the eigenspace for $\lambda = 5$ via row reductions,

$$A - 5I = \begin{bmatrix} -4 & 2 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{E_{21}(1/2)} \begin{bmatrix} -4 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{E_1(-1/4)} \begin{bmatrix} 1 & -1/2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

which gives two eigenvectors, $(1/2, 1, 0)$ and $(0, 0, 1)$. Normalize these to get $\mathbf{v}_2 = (1, 2, 0)/\sqrt{5}$ and $\mathbf{v}_3 = (0, 0, 1)$. In this case $\mathbf{v}_2$ and $\mathbf{v}_3$ are already orthogonal, so the diagonalizing matrix can be written as

$$P = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] = \frac{1}{\sqrt{5}} \begin{bmatrix} -2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & \sqrt{5} \end{bmatrix}.$$

We leave it to the reader to check that $P^T A P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$. □

Example 5.18. Let $A = \begin{bmatrix} 1 & 1-\mathrm{i} \\ 1+\mathrm{i} & 0 \end{bmatrix}$ as in Example 5.16. Unitarily diagonalize this matrix.

Solution. In Example 5.16 we computed the eigenvalues to be $\lambda = -1, 2$. Next find eigenspaces for each eigenvalue. For $\lambda = -1$, we find the null space by row reduction,

$$A + I = \begin{bmatrix} 2 & 1-\mathrm{i} \\ 1+\mathrm{i} & 1 \end{bmatrix} \xrightarrow[E_{21}(-(1+\mathrm{i})/2)]{} \begin{bmatrix} 2 & 1-\mathrm{i} \\ 0 & 0 \end{bmatrix} \xrightarrow{E_1(1/2)} \begin{bmatrix} 1 & (1-\mathrm{i})/2 \\ 0 & 0 \end{bmatrix},$$

so that the null space is spanned by the vector $((-1+\mathrm{i})/2, 1)$. A similar calculation shows that a basis of eigenvectors for $\lambda = 2$ consists of the vector $((-1-\mathrm{i})/2, 1)$. Normalize these vectors to obtain $\mathbf{u}_1 = ((-1+\mathrm{i})/2, 1)/\sqrt{3/2}$ and $\mathbf{u}_2 = (-1, (-1-\mathrm{i})/2)/\sqrt{3/2}$. So set

$$U = \sqrt{\frac{2}{3}} \begin{bmatrix} \frac{-1+\mathrm{i}}{2} & -1 \\ 1 & \frac{-1-\mathrm{i}}{2} \end{bmatrix}$$

and obtain that (the reader should check this)

$$U^{-1}AU = U^*AU = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}. \qquad \square$$

5.4 Exercises and Problems

Exercise 1. Show that the following matrices are real symmetric and find orthogonal matrices that diagonalize these matrices.

(a) $\begin{bmatrix} -2 & 2 \\ 2 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 36 \\ 36 & 23 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

Exercise 2. Show that the following matrices are Hermitian and find unitary matrices that diagonalize these matrices.

(a) $\begin{bmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & \mathrm{i} \\ -\mathrm{i} & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & \mathrm{i} \\ 0 & -\mathrm{i} & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1+\mathrm{i} & 0 \\ 1-\mathrm{i} & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

Exercise 3. Show that these matrices are orthogonal and compute their eigenvalues. Determine whether it is possible to orthogonally or unitarily diagonalize these matrices. (*Hint:* look for orthogonal sets of eigenvectors.)

(a) $\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ (b) $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Exercise 4. Show that these matrices are unitary and compute their eigenvalues. Unitarily diagonalize these matrices.

(a) $\frac{1}{\sqrt{5}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & \mathrm{i} & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \end{bmatrix}$ (c) $\frac{1}{5\sqrt{2}} \begin{bmatrix} 5 & -3+4\mathrm{i} \\ 3+4\mathrm{i} & 5 \end{bmatrix}$

Exercise 5. A square matrix A is called *normal* if $AA^* = A^*A$. Which of the matrices in Exercises 3 and 1 are normal?

Exercise 6. Which of the matrices in Exercise 4 are normal or Hermitian?

Exercise 7. Use orthogonal diagonalization to find a formula for the kth power of $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$.

Exercise 8. Use unitary diagonalization to find a formula for the kth power of $A = \begin{bmatrix} 3 & \mathrm{i} \\ -\mathrm{i} & 3 \end{bmatrix}$.

Exercise 9. Let $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & -1 \\ 0 & -1 & 2 \end{bmatrix}$. The eigenvalues of A are 1, 2, and 4. Find an orthogonal matrix P that diagonalizes A to $D = \operatorname{diag}\{1, 2, 4\}$, calculate $B = P \operatorname{diag}\left\{1, \sqrt{2}, 2\right\} P^T$, and show that B is a symmetric positive definite square root of A, that is, $B^2 = A$ and B is symmetric positive definite.

Exercise 10. Let $A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -\mathrm{i} \\ 0 & \mathrm{i} & 1 \end{bmatrix}$. The eigenvalues of A are 0, 1, and 3. Find a unitary matrix P that diagonalizes A to $D = \operatorname{diag}\{0, 1, 3\}$ and confirm that $B = P \operatorname{diag}\left\{0, 1, \sqrt{3}\right\} P^*$ is a Hermitian square root of A.

Problem 11. Show that if A is orthogonally diagonalizable, then so is A^T.

*Problem 12. Let B be a Hermitian matrix. Show that the eigenvalues of B are positive if and only if B is a positive definite matrix.

Problem 13. Show that if the real matrix A is orthogonally diagonalizable, then A is symmetric.

Problem 14. Show that if the real matrix A is skew-symmetric ($A^T = -A$), then iA is Hermitian.

Problem 15. Suppose that A is symmetric and orthogonal. Prove that the only possible eigenvalues of A are ± 1.

*Problem 16. Let A be real symmetric positive definite matrix. Show that A has a real symmetric positive definite square root, that is, there is a symmetric positive definite matrix S such that $S^2 = A$.

*Problem 17. Let A be any real matrix and show that the eigenvalues of $A^T A$ are all nonnegative.

Problem 18. Let $\mathbf{0} \neq \mathbf{u} \in \mathbb{C}^n$ and $A = \mathbf{u}\mathbf{u}^*$. Show that A is Hermitian and exhibit the single nonzero eigenvalue and corresponding eigenvector. How could you describe an eigensystem for this matrix?

5.5 *Schur Form and Applications

Recall that matrices A and B are similar if there is an invertible matrix S such that $B = S^{-1}AS$; if the transformation matrix S is unitary, then $S^{-1} = S^*$. The main object of this section is to prove a famous theorem in linear algebra that provides a nice answer to the following question: If we wish to use only orthogonal (or unitary) matrices as similarity transformation matrices, what is the simplest form to which a matrix A can be transformed? It would be nice if we could say something like "diagonal" or "Jordan canonical form." Unfortunately, neither is possible. However, upper triangular matrices are very nice special forms of matrices. In particular, we can see the eigenvalues of an upper triangular matrix at a glance. That makes the following theorem extremely attractive. Its proof is also very interesting, in that it actually suggests an algorithm for computing the so-called *Schur triangular form.*

Theorem 5.15. **Schur Triangularization** Let A be an arbitrary square matrix. Then there exists a unitary matrix U such that U^*AU is an upper triangular matrix. If A and its eigenvalues are real, then U can be chosen to be orthogonal.

Proof. To get started, take $k = 0$ and $V_0 = I$. Suppose we have reached the kth stage where we have a unitary matrix V_k such that

$$V_k^* A V_k = \begin{bmatrix} \lambda_1 & * & \cdots & * \\ \vdots & \ddots & * & \vdots \\ 0 & \cdots & \lambda_k & * \\ 0 & \cdots & 0 & B \end{bmatrix} = \begin{bmatrix} R_k & C \\ 0 & B \end{bmatrix}$$

with the submatrix R_k upper triangular. Compute an eigenvalue λ_{k+1} of B and a corresponding eigenvector $\mathbf{w}$ of unit length in the standard norm of B. If the first coordinate of $\mathbf{w}$ is not real, replace $\mathbf{w}$ by $e^{-i\theta}\mathbf{w}$ where θ is a polar argument of the first coordinate of $\mathbf{w}$. This does not affect the length of $\mathbf{w}$, and any multiple of $\mathbf{w}$ is still an eigenvector of A. Now let $\mathbf{v} = \mathbf{w} - \mathbf{e}_1$, where $\mathbf{e}_1 = (1, 0, \ldots, 0)$. Form the (possibly complex) Householder matrix $H_\mathbf{v}$. Since $\mathbf{w} \cdot \mathbf{e}_1$ is real, $\mathbf{w} \cdot \mathbf{e}_1 = \mathbf{e}_1 \cdot \mathbf{w}$ and hence $\mathbf{v}^* (\mathbf{w} + \mathbf{e}_1) = 0$. Thus, $H_\mathbf{v} (\mathbf{w} + \mathbf{e}_1) = \mathbf{w} + \mathbf{e}_1$. Use the facts that $H_\mathbf{v}\mathbf{v} = -\mathbf{v}$ and $\mathbf{w} = \frac{1}{2}\{\mathbf{v} + (\mathbf{w} + \mathbf{e}_1)\}$ to deduce that $H_\mathbf{v}\mathbf{w} = \mathbf{e}_1$. Recall that Householder matrices are unitary and Hermitian, so that $H_\mathbf{v}^* = H_\mathbf{v} = H_\mathbf{v}^{-1}$. Hence

$$H_\mathbf{v}^* B H_\mathbf{v} \mathbf{e}_1 = H_\mathbf{v} B H_\mathbf{v}^{-1} \mathbf{e}_1 = H_\mathbf{v} B \mathbf{w} = H_\mathbf{v} \lambda_1 \mathbf{w} = \lambda_1 \mathbf{e}_1.$$

Therefore, the entries under the first row and in the first column of $H_\mathbf{v}^* B H_\mathbf{v}$ are zero. Form the unitary matrix

$$V_{k+1} = \begin{bmatrix} I_k & 0 \\ 0 & H_\mathbf{v} \end{bmatrix} V_k$$

and obtain that

$$V_{k+1}^* A V_{k+1} = \begin{bmatrix} I_k & 0 \\ 0 & H_{\mathbf{v}} \end{bmatrix} V_k^* A V_k \begin{bmatrix} I_k & 0 \\ 0 & H_{\mathbf{v}} \end{bmatrix}$$
$$= \begin{bmatrix} I_k & 0 \\ 0 & H_{\mathbf{v}} \end{bmatrix} \begin{bmatrix} R_k & C \\ 0 & B \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & H_{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} R_k & CH_{\mathbf{v}} \\ 0 & H_{\mathbf{v}}^* B H_{\mathbf{v}} \end{bmatrix}.$$

This new matrix is upper triangular in the first $k+1$ columns, so we can continue in this fashion until we reach the last column, at which point we set $U = V_n$ to obtain that U^*AU is upper triangular.

Finally, notice that if the eigenvalues and eigenvectors that we calculate are real, which would be the case if A and the eigenvalues of A were real, then the Householder matrices used in the proof are all real, so that the matrix U is orthogonal. □

Of course, the upper triangular matrix T and triangularizing matrix U are not unique. Nonetheless, this is a very powerful theorem. Consider what it says in the case that A is Hermitian: the principal axes theorem is a simple special case of it.

Corollary 5.4. Principal Axes Theorem Every Hermitian matrix is unitarily (orthogonally, if the matrix is real) diagonalizable.

Proof. Let A be Hermitian. According to the Schur triangularization theorem there is a unitary matrix U such that $U^*AU = R$ is upper triangular. We check that

$$R^* = (U^*AU)^* = U^*A^*(U^*)^* = U^*AU = R.$$

Therefore, R is both upper and lower triangular. This makes R a diagonal matrix. If A is real symmetric, then A and its eigenvalues are real. By the triangularization theorem U can be chosen orthogonal. □

Normal Matrix

Another application of the Schur triangularization theorem is that we can show the real significance of *normal* matrices. This term has appeared in several exercises. Recall that a matrix A is *normal* if $A^*A = AA^*$. Clearly, every Hermitian matrix is normal, as is every unitary matrix.

Theorem 5.16. A matrix is unitarily diagonalizable if and only if it is normal.

Proof. We leave it as an exercise to show that a unitarily diagonalizable matrix is normal. Conversely, let A be normal. According to the Schur triangularization theorem there is a unitary matrix U such that $U^*AU = R$ is upper triangular. But then we have that $R^* = U^*A^*U$, so that

$$R^*R = U^*A^*UU^*AU = U^*A^*AU = U^*AA^*U = U^*AUU^*A^*U = RR^*.$$

Therefore, R commutes with R^*, which means that R is diagonal by Problem 11 at the end of this section. This completes the proof. □

Our last application extends Theorem 5.2 to rational functions.

Corollary 5.5. Let $f(x)$ and $g(x)$ be polynomials and A a square matrix such that $g(A)$ is invertible. Then the eigenvalues of the matrix $f(A)g(A)^{-1}$ are of the form $f(\lambda)/g(\lambda)$, where λ runs over the eigenvalues of A.

Proof. We sketch the proof. As a first step, we make two observations about upper triangular matrices S and T with diagonal terms $\lambda_1, \lambda_2, \ldots, \lambda_n$, and $\mu_1, \mu_2, \ldots, \mu_n$, respectively. First, ST is upper triangular with diagonal terms $\lambda_1\mu_1, \lambda_2\mu_2, \ldots, \lambda_n\mu_n$. Next, if S is invertible, then S^{-1} is also an upper triangular matrix, whose diagonal terms are $1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_n$.

Now, we have seen in Theorem 5.5 that for any invertible P of the right size, $P^{-1}f(A)P = f\left(P^{-1}AP\right)$. Similarly, if we multiply the identity $g(A)g(A)^{-1} = I$ by P^{-1} and P, we see that $P^{-1}g(A)^{-1}P = g(P^{-1}AP)^{-1}$. Thus, if P is a matrix that unitarily triangularizes A, then

$$P^{-1}f(A)\,g(A)^{-1}P = f\left(P^{-1}AP\right)g\left(P^{-1}AP\right)^{-1},$$

so that by our first observations, this matrix is upper triangular with diagonal entries of the required form. Since similar matrices have the same eigenvalues, it follows that the eigenvalues of $f(A)g(A)^{-1}$ are of the required form. □

5.5 Exercises and Problems

You may use a technology tool for the following exercises.

Exercise 1. Apply one step of Schur triangularization to the following specified eigenvalues.

(a) $\lambda = -3$, $A = \begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}$ (b) $\lambda = \sqrt{2}$, $A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & \mathrm{i} \\ 0 & -\mathrm{i} & 0 \end{bmatrix}$

Exercise 2. Apply Schur triangularization to the following matrices.

(a) $\begin{bmatrix} 4 & 4 & 1 \\ -1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} \mathrm{i} & 0 & 2 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$ (c) $= \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}$

Exercise 3. Use Schur triangularization to find eigenvalues of the following matrices.

(a) $\begin{bmatrix} 5 & 6 & 18 \\ 11 & 6 & 24 \\ -4 & -2 & -8 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 8 & 20 \\ 3 & 14 & 32 \\ -1 & -5 & -11 \end{bmatrix}$ (c) $\begin{bmatrix} 4 & 20 & 42 & 12 \\ 8 & 32 & 72 & 15 \\ -3 & -14 & -31 & -6 \\ -1 & -6 & -12 & -4 \end{bmatrix}$

Exercise 4. Find a unitary matrix that upper triangularizes the following matrices.

(a) $\begin{bmatrix} 3 & 6 & 2 \\ 1 & 4 & 2 \\ 4 & 2 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 4 & 8 & 10 \\ 3 & 14 & 0 \\ -1 & 5 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 2 & 0 & 0 \\ -2 & 2 & 0 & 0 \\ 0 & -2 & 2 & 2 \\ 0 & 0 & -2 & 1 \end{bmatrix}$

Exercise 5. Verify Corollary 5.5 in the case that $A = \begin{bmatrix} 22 & 10 \\ -50 & -23 \end{bmatrix}$, $f(x) = x^2 - 1$, and $g(x) = x^2 + 1$ by calculating the eigenvalues $f(A)/g(A)$ directly and comparing them to $f(\lambda)/g(\lambda)$, where λ runs over the eigenvalues of A.

Exercise 6. Verify that Corollary 5.5 fails in the case that $A = \begin{bmatrix} 22 & 10 \\ -50 & -23 \end{bmatrix}$, $f(x) = x - 1$, and $g(x) = x^2 + 4x + 3$ and explain why.

Problem 7. Show that every unitary matrix is normal. Give an example of a unitary matrix that is not Hermitian.

*Problem 8. Let A be an invertible matrix. Use Schur triangularization to reduce the problem $A\mathbf{x} = \mathbf{b}$ to a problem with triangular coefficient matrix.

Problem 9. A square matrix A is *skew-Hermitian* if $A^* = -A$. Show that every skew-Hermitian matrix is normal.

Problem 10. Use Corollary 5.4 to show that the eigenvalues of a Hermitian matrix must be real.

*Problem 11. Prove that if an upper triangular matrix commutes with its Hermitian transpose, then the matrix must be diagonal.

*Problem 12. Suppose that A is an $n \times n$ matrix whose only eigenvalue is 1 and whose Schur triangularization yields $U^*AU = R$. Find a formula for A^{-1} that involves only I, U, U^*, R and no explicit inverses.

5.6 *The Singular Value Decomposition

The object of this section is to develop yet one more factorization of a matrix that provides valuable information about the matrix. For simplicity, we stick with the case of a real matrix A and orthogonal matrices. However, the factorization we are going to discuss can be done with complex A and unitary matrices. This factorization is called the *singular value decomposition (SVD for short)*. It has a long history in matrix theory, but was popularized in the

1960s as a powerful computational tool. We saw in Section 4.4 that multiplication on one side by an orthogonal matrix can produce an upper triangular matrix. This is called the *QR factorization*. Here is the basic question that the SVD answers: if multiplication on one side by an orthogonal matrix can produce an upper triangular matrix, how simple a matrix can be produced by multiplying on each side by a (possibly different) orthogonal matrix? The answer, as you might guess, is a matrix that is both upper and lower triangular, that is, diagonal. However, verification of this fact is much more subtle than that of the one-sided QR factorization of Section 4.4. Here is the key result:

Theorem 5.17. Singular Value Decomposition Let A be an $m \times n$ real matrix. Then there exist an $m \times m$ orthogonal matrix U, an $n \times n$ orthogonal matrix V, and an $m \times n$ diagonal matrix Σ with diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, with $p = \min\{m, n\}$, such that $U^T A V = \Sigma$. Moreover, the numbers $\sigma_1, \sigma_2, \ldots, \sigma_p$ are uniquely determined by A.

Proof. There is no loss of generality in assuming that $n \leq m$. For if this is not the case, we can prove the theorem for A^T, and by transposing the resulting SVD for A^T, obtain a factorization for A. Form the $n \times n$ matrix $B = A^T A$. This matrix is symmetric and its eigenvalues are nonnegative (we leave these facts as exercises). Because they are nonnegative, we can write the eigenvalues of B in decreasing order of magnitude as the squares of nonnegative real numbers, say as $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2$. Now we know from the principal axes theorem that we can find an orthonormal set of eigenvectors corresponding to these eigenvalues, say $B\mathbf{v}_k = \sigma_k^2 \mathbf{v}_k$, $k = 1, 2, \ldots, n$. Let $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$. Then V is an orthogonal $n \times n$ matrix. We may assume for some index r that $\sigma_{r+1}, \sigma_{r+2}, \ldots, \sigma_n$ are zero, while $\sigma_r \neq 0$.

Next set $\mathbf{u}_j = \frac{1}{\sigma_j} A\mathbf{v}_j$, $j = 1, 2, \ldots, r$. These are orthonormal vectors in $\mathbb{R}^m$ since

$$\mathbf{u}_j^T \mathbf{u}_k = \frac{1}{\sigma_j \sigma_k} \mathbf{v}_j^T A^T A \mathbf{v}_k = \frac{1}{\sigma_j \sigma_k} \mathbf{v}_j^T B \mathbf{v}_k = \frac{\sigma_k^2}{\sigma_j \sigma_k} \mathbf{v}_j^T \mathbf{v}_k = \begin{cases} 0, & \text{if } j \neq k, \\ 1, & \text{if } j = k. \end{cases}$$

Now expand this set to an orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ of $\mathbb{R}^m$. This is possible by Theorem 4.7 in Section 4.3. Set $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m]$. This matrix is orthogonal. We calculate that if $k > r$, then $\mathbf{u}_j^T A \mathbf{v}_k = 0$ since $A\mathbf{v}_k = \mathbf{0}$, and if $k < r$, then

$$\mathbf{u}_j^T A \mathbf{v}_k = \sigma_k \mathbf{u}_j^T \mathbf{u}_k = \begin{cases} 0, & \text{if } j \neq k, \\ \sigma_k, & \text{if } j = k. \end{cases}$$

It follows that $U^T A V = [\mathbf{u}_j^T A \mathbf{v}_k] = \Sigma$, which is the desired SVD.

Finally, if U, V are orthogonal matrices such that $U^T A V = \Sigma$, then $A = U \Sigma V^T$ and therefore

$$B = A^T A = V \Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T,$$

so that the squares of the diagonal entries of Σ are the eigenvalues of B. It follows that the numbers $\sigma_1, \sigma_2, \ldots, \sigma_n$ are uniquely determined by A. □

A similar theorem holds for complex matrices, with "orthogonal" replaced by "unitary". The nonnegative numbers $\sigma_1, \sigma_2, \ldots, \sigma_p$ are called the *singular values* of the matrix A, the columns of U are the *left singular vectors* of A, and the columns of V are the *right singular vectors* of A.

Singular Values and Vectors

There is an interesting geometrical interpretation of this theorem from the perspective of linear transformations and change of basis as developed in Section 3.7. It can be stated as follows.

Corollary 5.6. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation with matrix A with respect to the standard bases. Then there exist orthonormal bases $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, such that the matrix of T with these bases is diagonal with nonnegative entries down the diagonal.

Proof. First observe that if $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m]$ and $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$, then U and V are the change of basis matrices from the bases $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, to the standard bases. Also, $U^{-1} = U^T$. Now apply Corollary 3.9 of Section 3.7, and the result follows. □

Corollary 5.7. Let $U^T A V = \Sigma$ be the SVD of A and suppose that $\sigma_r \neq 0$ and $\sigma_{r+1} = 0$. Then

(1) $\operatorname{rank} A = r$.
(2) $A = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \operatorname{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_r\} [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]^T$.
(3) $\mathcal{N}(A) = \operatorname{span}\{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n\}$.
(4) $\mathcal{C}(A) = \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$.
(5) If $A^\dagger$ is given by

$$A^\dagger = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r] \operatorname{diag}\{1/\sigma_1, 1/\sigma_2, \ldots, 1/\sigma_r\} [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]^T,$$

then $\mathbf{x} = A^\dagger \mathbf{b}$ is a least squares solution to $A\mathbf{x} = \mathbf{b}$.
(6) $A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T$.

Proof. Multiplication by invertible matrices does not change rank, and the rank of Σ is clearly r, so (1) follows. For (2), multiply the SVD equation by U on the left and V^T on the right to obtain

$$\begin{aligned} A = U\Sigma V^T &= [\sigma_1 \mathbf{u}_1, \sigma_2 \mathbf{u}_2, \ldots, \sigma_r \mathbf{u}_r, \mathbf{0}, \ldots, \mathbf{0}] [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]^T \\ &= \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \operatorname{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_r\} [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]^T. \end{aligned}$$

This also proves item (6). The remaining items are left as exercises. □

Item (2) is called the *compact* SVD form for A. The matrix $A^\dagger$ of (5)

Compact SVD and Pseudoinverse

is called the *pseudoinverse* of A and behaves in many ways like an inverse for matrices that need not be invertible or even square. Item (5) presents an important application of the pseudoinverse. We have only scratched the surface of the many facets of the SVD. Like most good ideas, it is rich in applications. We mention one more. It is based on item (6), which says that a matrix A of rank r can be written as a sum of r rank-one matrices. In fact, it can be shown that this representation is the most economical in the sense that the partial sums

$$\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^T, \qquad k = 1, 2, \ldots, r,$$

give the rank-k approximation to A that is closest among all rank-k approximations to A. This suggests an intriguing way to compress data in a lossy way (i.e., with some loss of data). For example, suppose A is a matrix of floating-point numbers representing a picture. We might get a reasonably good approximation to the picture using only the σ_k larger than a certain threshold. Thus, with a $1{,}000 \times 1{,}000$ matrix A that has a very small σ_{21}, we could get by with the data $\sigma_k, \mathbf{u}_k, \mathbf{v}_k$, $k = 1, 2, \ldots, 20$. Consequently, we would store only these quantities, which add up to $1{,}000 \times 40 + 20 = 40{,}020$ numbers. Contrast this with storing the full matrix of $1{,}000 \times 1{,}000 = 1{,}000{,}000$ entries, and you can see the gain in economy.

Here is a simple example of an SVD calculation. Interestingly enough, the method of calculation is to simply follow the details of the proof of Theorem 5.17.

Example 5.19. Find a singular value decomposition for $A = \begin{bmatrix} 2 & 1 \\ 1 & -2 \\ 1 & -1 \end{bmatrix}$.

Solution. First form the matrix $B = A^T A = \begin{bmatrix} 6 & -1 \\ -1 & 6 \end{bmatrix}$. Next, find the eigenvalues of B via the characteristic equation $0 = (\lambda - 6)^2 - 1 = \lambda^2 - 12\lambda + 35 = (\lambda - 5)(\lambda - 7)$. Thus, the singular values are $\sigma_1 = \sqrt{7}$ and $\sigma_2 = \sqrt{5}$. Here $B - 7I = \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix}$, so unit eigenvector $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1, -1)$ solves $(B - 7I)\mathbf{v} = \mathbf{0}$. Similarly, $B - 5I = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, so unit eigenvector $\mathbf{v}_2 = \frac{1}{\sqrt{2}}(1, 1)$ solves $(B - 5I)\mathbf{v} = \mathbf{0}$. Next we calculate $\mathbf{u}_1 = \frac{1}{\sqrt{7}} A\mathbf{v}_1 = \frac{1}{\sqrt{14}}(1, 3, 2)$, $\mathbf{u}_2 = \frac{1}{\sqrt{5}} A\mathbf{v}_2 = \frac{1}{\sqrt{10}}(3, -1, 0)$. Finally, we note that $\mathbf{e}_3$ is outside the span of $\mathbf{u}_1$ and $\mathbf{u}_2$, so an orthogonal vector to these two vectors is

$$\mathbf{w} = \mathbf{e}_3 - (\mathbf{u}_1 \cdot \mathbf{e}_3)\mathbf{u}_1 - (\mathbf{u}_1 \cdot \mathbf{e}_3)\mathbf{u}_2 = \frac{1}{7}(-1, -3, 5).$$

Let $\mathbf{u}_3 = \mathbf{w}/\|\mathbf{w}\| = \frac{1}{\sqrt{35}}(-1, -3, 5)$. Then we assemble

$$U = \begin{bmatrix} 1/\sqrt{14} & 3/\sqrt{10} & -1/\sqrt{35} \\ 3/\sqrt{14} & -1/\sqrt{10} & -3/\sqrt{35} \\ 2/\sqrt{14} & 0 & 5/\sqrt{35} \end{bmatrix}, V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \text{ and } \Sigma = \begin{bmatrix} \sqrt{7} & 0 \\ 0 & \sqrt{5} \\ 0 & 0 \end{bmatrix},$$

to obtain the SVD decomposition $U^T A V = \Sigma$. □

5.6 Exercises and Problems

Exercise 1. Exhibit a singular value decomposition for the following matrices.

(a) $\begin{bmatrix} 3 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} -2 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & -2 & 0 \\ 2 & 0 & 0 \end{bmatrix}$

Exercise 2. Calculate a singular value decomposition for the following matrices.

(a) $\begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

Exercise 3. Use a technology tool to compute an orthonormal basis for the null space and column space of the following matrices with the SVD and Corollary 5.7. You will have to decide which nearly-zero terms are really zero.

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

Exercise 4. Use the pseudoinverse to find a least squares solution $A\mathbf{x} = \mathbf{b}$, where A is a matrix from Exercise 3 with corresponding right-hand side below.

(a) $(2, 2, 6, 5)$ (b) $(2, 3, 1)$ (c) $(4, 1, 2, 3)$

*Problem 5. Prove (3) and (4) of Corollary 5.7.

Problem 6. Show that if A is invertible, then A^{-1} is the pseudoinverse of A.

*Problem 7. Prove (5) of Corollary 5.7.

5.7 *Applications and Computational Notes

Jordan Canonical Form Theorem

Although the Jordan Canonical Form theorem is not used often in practical computations due to numerical complexity and stability issues, this theorem is a fundamental theoretical tool in matrix analysis. It warrants a proof, and at this point we have sufficient machinery to provide it:

Theorem 5.18. **Jordan Canonical Form** Every matrix A is similar to a block diagonal matrix that consists of Jordan blocks down the diagonal. Moreover, these blocks are uniquely determined by A up to order.

Proof. Assume that A is $n \times n$ possibly complex matrix and define the matrix multiplication linear operator $T = T_A : \mathbb{C}^n \to \mathbb{C}^n$. We use the notation of Theorem 3.14: $U_j = \ker(T^j)$ and $W_j = \text{range}(T^j)$, $j = 1, 2, \ldots$, so that for all j, $U_j \subseteq U_{j+1}$ and $W_j \supseteq W_{j+1}$; Also, $U_n = U_k$ and $W_n = W_k$ for $k > n$. Let $m \leq n$ be the first index such that $U_m = U_{m+1}$, so that, as we saw in the proof of Theorem 3.14, all other inclusions $U_j \subseteq U_{j+1}$ are strict for $j < m$ and equality for $j \geq m$, so that $U_m = U_n$. Since $U_{m-1} \subset U_m$ we can expand a basis of U_{m-1} to a basis of U_m. Let C_m be the set of additional vectors and $V_m = \text{span}\, C_m$; so we can write $U_m = U_{m-1} \oplus V_m$.

Let's carry this one step further. First note that if the operator T is restricted to V_m, then for nonzero $\mathbf{v} \in V_m$ we have $T(\mathbf{v}) \neq \mathbf{0}$ for otherwise we would have $\mathbf{v} \in U_{m-1}$. Thus, the restricted map has kernel $\{\mathbf{0}\}$, so is one-to-one by Theorem 3.9. Moreover, $T(\mathbf{v}) \notin U_{m-2}$ else $\mathbf{v} \in U_{m-1}$. Thus, $T(C_m)$ is a linearly independent set whose span is contained in U_{m-1} but whose intersection with U_{m-2} is $\{\mathbf{0}\}$. So we can expand $T(C_m)$ to a set of linearly independent vectors C_{m-1} such that if we set $V_{m-1} = \text{span}\, C_{m-1}$, then $U_{m-1} = U_{m-2} \oplus V_{m-1}$. Now continue this process until we reach U_1. Expand $T(C_2)$ to a basis C_1 of $U_1 = \ker T$, so that $V_1 = \text{span}\, C_1 = U_1$. Observe that

$$U_m = U_{m-1} \oplus V_m = U_{m-2} \oplus V_{m-1} \oplus V_m = \ldots = V_1 \oplus V_2 \oplus \cdots \oplus V_m.$$

Thus, a basis of U_m consists of sets of vectors of the form $\left\{\mathbf{v}, T(\mathbf{v}), \ldots T^j(\mathbf{v})\right\}$, where $T^{j+1}(\mathbf{v}) = \mathbf{0}$. So we can construct an ordered basis C of $\mathbb{C}^n$ which lists all these sets of vectors in order (that gives us a basis of U_n) followed by some basis of $W_n = \text{range}(T^n)$, say of size d. Then the coordinates of elements of W_n with respect to C are 0, except for the last d coordinates. So the matrix of T with respect to this basis takes the form

$$\begin{bmatrix} J(0) & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} J_{d_1}(0) & 0 & \cdots & 0 & 0 \\ 0 & J_{d_2}(0) & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & \cdots & J_{d_m}(0) & 0 \\ 0 & 0 & 0 & 0 & Q \end{bmatrix}, \tag{5.8}$$

where Q is a matrix the size of $\dim W_n$ and the d_j's are the length of the sequences $\left\{\mathbf{v}, T(\mathbf{v}), \ldots T^j(\mathbf{v})\right\}$. Also, the restriction of T, $T : W_n \to W_n$ has kernel zero since $W_n \cap U_m = \{\mathbf{0}\}$ and its matrix with respect to C is just Q. It should be noted here that although the order of these Jordan blocks is arbitrary, the number of each of size d is just the dimension of V_d, which is uniquely determined by A, and V_d is where the starting terms of such sequences are found.

Here is how we relate this construction to nonzero eigenvalues: We find an eigenvalue λ of A. Next, we apply the construction of the preceding paragraph to the matrix $A - \lambda I$. If P is the change of basis matrix from the standard basis B to our new basis C, then we know by Theorem 5.5

$$\begin{bmatrix} J(0) & 0 \\ 0 & Q \end{bmatrix} = P^{-1}(A - \lambda I)P = P^{-1}AP - P^{-1}\lambda IP = P^{-1}AP - \lambda I.$$

Hence, we have $P^{-1}AP = \begin{bmatrix} J(\lambda) & 0 \\ 0 & Q + \lambda I \end{bmatrix}$. The matrix $Q + \lambda I$ will not have λ as an eigenvalue (else Q is singular, which implies that $T : W_n \to W_n$ has nonzero kernel.) But similar matrices have the same characteristic polynomials, so the remaining eigenvalues of A are also eigenvalues of $Q + \lambda I$ of the same multiplicity. Repeat this construction on $Q - \mu I$, where μ is an eigenvalue of $Q + \lambda I$ to obtain that $Q + \lambda I$ satisfies an equation like (5.8) so that $P_1^{-1}QP_1 = \begin{bmatrix} J(\mu) & 0 \\ 0 & Q_1 + \mu I \end{bmatrix}$. But then the matrix $\begin{bmatrix} I & 0 \\ 0 & P_1 \end{bmatrix}$ transforms A to a matrix with Jordan blocks corresponding to λ and μ on the diagonal. We can continue this process until all eigenvalues of A are accounted for, at which point A is similar to a Jordan canonical form. Ordering of these blocks is not unique, but the number of each size is, as we have noted above. This completes the proof. □

Computation of Eigensystems

Nowadays, one can use a technology tool to find a complete eigensystem for, say a 100×100 matrix, in a fraction of a second. That's pretty remarkable and, to some extent, a tribute to the fast cheap hardware commonly available to the public. But hardware is only part of the story. Bad computational algorithms can bring the fastest computer to its knees. The rest of the story concerns the remarkable developments in numerical linear algebra over the past sixty years that have given us fast reliable algorithms for eigensystem calculation. We can only scratch the surface of these developments in this brief discussion. At the outset, we rule out the methods developed in this chapter as embodied in the eigensystem algorithm (page 10). These are for simple hand calculations and theoretical purposes. In a few special cases we can derive general formulas for eigenvectors and eigenvalues. One such example is a *Toeplitz* matrix (a matrix with constant entries down each diagonal) that is also tridiagonal. We outline the approach in a problem at the end of this section, but these complete solution formulas are the exception, not the rule.

We are going to examine some iterative methods for selectively finding eigenpairs of a real matrix whose eigenvalues are real and distinct. Hence, the matrix A is diagonalizable. The hypothesis of diagonalizability may seem too constraining, but there is this curious aphorism that "numerically every matrix is diagonalizable." The reason is as follows: once you store and perform numerical calculations on the entries of A, you perturb them a small essentially random amount. This has the effect of perturbing the eigenvalues of the calculated A a small random amount. Thus, the probability that any two eigenvalues of A are numerically equal is quite small. To focus matters, consider the test matrix

$$A = \begin{bmatrix} -8 & -5 & 8 \\ 6 & 3 & -8 \\ -3 & 1 & 9 \end{bmatrix}.$$

Just for the record, the actual eigenvalues of A are 1, -2 and 5. Now we ask three questions about A:

1. How can we get a ballpark estimate of the location of the eigenvalues of A?
2. How can we estimate the *dominant* eigenpair $(\lambda, \mathbf{x})$ of A? (Recall that "dominant" means that λ is larger in absolute value than any other eigenvalue of A.)
3. Given a good estimate of any eigenvalue λ of A, how can we improve the estimate and compute a corresponding eigenvector?

An answer to question (1) is the following theorem, which predates modern numerical analysis, but has proved to be quite useful. Because it helps locate eigenvalues, it is called a "localization theorem."

Theorem 5.19. Gershgorin Circle Theorem Let $A = [a_{ij}]$ be an $n \times n$ matrix and define disks D_j in the complex plane by $r_j = \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}|$ and $D_j = \{z \,|\, |z - a_{jj}| \leq r_j\}$. Then

(1) Every eigenvalue of A is contained in some disk D_j.
(2) If k of the disks are disjoint from the others, then exactly k eigenvalues are contained in the union of these disks.

Proof. To prove (1), let λ be an eigenvalue of A and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ an eigenvector corresponding to λ. Suppose that x_j is the largest coordinate of $\mathbf{x}$ in absolute value. Divide $\mathbf{x}$ by this entry to obtain an eigenvector whose largest coordinate is $x_j = 1$. Without loss of generality, this vector is $\mathbf{x}$. Consider the jth entry of the zero vector $\lambda\mathbf{x} - A\mathbf{x}$, which is

$$(\lambda - a_{jj})1 + \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} x_k = 0.$$

Bring the sum to the right-hand side and take absolute values to obtain

$$|\lambda - a_{jj}| = \Big| \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} x_k \Big| \leq \sum_{\substack{k=1 \\ k \neq j}}^{n} |\, a_{jk} \,|\, |\, x_k \,| \leq r_j,$$

since $|x_k| \leq 1$ for each x_k. This shows that $\lambda \in D_j$, which proves (1). We will not prove (2), since it requires some complex analysis (see the Horn and Johnson text [17], page 344, for a proof.)

Example 5.20. Apply the Gershgorin circle theorem to the test matrix A and sketch the resulting Gershgorin disks.

Solution. The disks are easily seen to be

$$D_1 = \{\, z \,|\, |z+8| \le 13 \},$$
$$D_2 = \{\, z \,|\, |z-3| \le 14 \},$$
$$D_3 = \{\, z \,|\, |z-9| \le 4 \}.$$

A sketch of them is provided in Figure 5.1. □

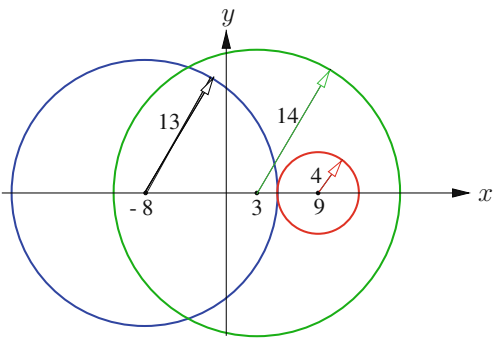


Fig. 5.1: Gershgorin disks for A.

Now we turn to question (2). One answer to it is contained in the following algorithm, known as the *power method*.

Power Method To compute an approximate eigenpair $(\lambda, \mathbf{x})$ of A with $\|\mathbf{x}\| = 1$ and λ the dominant eigenvalue:

(1) Input an initial guess $\mathbf{x}_0$ for $\mathbf{x}$
(2) For $k = 0, 1, \ldots$ until convergence of $\lambda^{(k)}$'s:
  (a) $\mathbf{y} = A\mathbf{x}_k$,
  (b) $\mathbf{x}_{k+1} = \dfrac{\mathbf{y}}{\|\mathbf{y}\|}$,
  (c) $\lambda^{(k+1)} = \mathbf{x}_{k+1}^T A \mathbf{x}_{k+1}$.

That's all there is to it! Why should this algorithm converge? The secret to this algorithm lies in a formula we saw earlier in our study of discrete dynamical systems, namely equation (5.6) which we reproduce here:

$$\mathbf{x}^{(k)} = A^k \mathbf{x}^{(0)} = c_1 \lambda_1^k \mathbf{v}_1 + c_2 \lambda_2^k \mathbf{v}_2 + \cdots + c_n \lambda_n^k \mathbf{v}_n.$$

Here it is understood that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis of eigenvectors corresponding to eigenvalues $\lambda_1 \lambda_2, \ldots, \lambda_n$, which, with no loss of generality, we can assume to be unit-length vectors. Notice that at each stage of the power method we divided the computed iterate $\mathbf{y}$ by its length to get the next $\mathbf{x}_{\mathbf{k}+1}$, and this division causes no directional change. Thus, we would get exactly the same vector if we simply set $\mathbf{x}_{k+1} = \mathbf{x}^{(k+1)} / \left\|\mathbf{x}^{(k+1)}\right\|$. Now for large k the ratios

$(\lambda_j/\lambda_1)^k$ can be made as small as we please, so we can rewrite the above equation as

$$\mathbf{x}^{(k)} = A^k\mathbf{x}^{(0)} = \lambda_1^k\left\{c_1\mathbf{v}_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathbf{v}_2 + \cdots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\mathbf{v}_n\right\} \approx \lambda_1^k c_1\mathbf{v}_1.$$

Assuming that $c_1 \neq 0$, which is likely if $\mathbf{x}_0$ is randomly chosen, we see that

$$\mathbf{x}_{k+1} = \frac{A\mathbf{x}^{(k)}}{\left\|A\mathbf{x}^{(k)}\right\|} \approx \frac{\lambda_1^k c_1\lambda_1\mathbf{v}_1}{\left|\lambda_1^k c_1\lambda_1\right|} = \pm\mathbf{v}_1,$$
$$\lambda^{(k+1)} = \mathbf{x}_{k+1}^T A\mathbf{x}_{k+1} \approx (\pm\mathbf{v}_1)^T A(\pm\mathbf{v}_1) = \lambda_1.$$

Thus, we see that the sequence of $\lambda^{(k)}$'s converges to λ_1 and the sequence of $\mathbf{x}_k$'s converges to $\pm\mathbf{v}_1$, provided that λ_1 is a dominant eigenvalue and A has a basis of eigenvectors, i.e., is diagonalizable. The argument (it isn't rigorous enough to be called a proof) we have just given shows that the oscillation in sign in the entries of $\mathbf{x}_k$ occurs in the case $\lambda < 0$. This argument doesn't require the initial guess to be real. Complex numbers are permitted if transpose is replaced by *conjugate* transpose in the algorithm.

If we apply the power method to our test problem with an initial guess of $\mathbf{x}_0 = (1, 1, 1)$, we get every third value as follows:

k	$\lambda^{(k)}$	$\mathbf{x}_k$
0		$(1, 1, 1)$
3	5.7311	$(0.54707, -0.57451, 0.60881)$
6	4.9625	$(0.57890, -0.57733, 0.57581)$
9	5.0025	$(0.57725, -0.57735, 0.57745)$
12	4.9998	$(0.57736, -0.57735, 0.57734)$

Notice that the eigenvector looks a lot like a multiple of $(1, -1, 1)$, and the eigenvalue looks a lot like 5, which is the dominant eigenvalue of our test matrix. This is an exact eigenpair, as one can check.

Finally, we turn to question (3). One answer to it is contained in the following algorithm, known as the *inverse iteration method*.

Inverse Iteration Method To compute an approximate eigenpair $(\lambda, \mathbf{x})$ of A with $\|\mathbf{x}\| = 1$ and λ an eigenvalue:

(1) Input an initial guess $\mathbf{x}_0$ for $\mathbf{x}$ and a *close* approximation $\mu = \lambda_0$ to λ.
(2) For $k = 0, 1, \ldots$ until convergence of the $\lambda^{(k)}$'s:
  (a) $\mathbf{y} = (A - \mu I)^{-1}\mathbf{x}_k$,
  (b) $\mathbf{x}_{k+1} = \dfrac{\mathbf{y}}{\|\mathbf{y}\|}$,
  (c) $\lambda^{(k+1)} = \mathbf{x}_{k+1}^T A\mathbf{x}_{k+1}$.

Notice that the inverse iteration method is simply the power method applied to the matrix $(A - \mu I)^{-1}$. In fact, it is sometimes called the inverse power method. The scalar μ is called a *shift*. Here is the secret of success for this method: we assume that μ is closer to a definite eigenvalue λ of A than to any other eigenvalue. But we don't want too much accuracy! We need $\mu \neq \lambda$. Theorem 5.2 in Section 1 of this chapter shows that the eigenvalues of the matrix $A - \mu I$ are of the form $\sigma - \mu$, where σ runs over the eigenvalues of A. Thus, the matrix $A - \mu I$ is nonsingular since no eigenvalue is zero, and Exercise 17 of Section 5.1 shows us that the eigenvalues of $(A - \mu I)^{-1}$ are of the form $1/(\sigma - \mu)$, where σ runs over the eigenvalues of A. Since μ is closer to λ than to any other eigenvalue of A, the eigenvalue $1/(\lambda - \mu)$ is the dominant eigenvalue of $(A - \mu I)^{-1}$, which is exactly what we need to make the power method work on $(A - \mu I)^{-1}$. Indeed, if μ is *very* close (but not equal!) to λ, convergence should be very rapid.

In a general situation, we could now have the Gershgorin circle theorem team up with inverse iteration. Gershgorin would put us in the right ballpark for values of μ, and inverse iteration would finish the job. Let's try this with our test matrix and choices of μ in the interval suggested by Gershgorin. Let's try $\mu = 0$. Here are the results in tabular form:

k	$\lambda^{(k)}$	$\mathbf{x}_k$ with $\mu = 0.0$
0	0.0	$(1, 1, 1)$
3	0.77344	$(-0.67759, 0.65817, -0.32815)$
6	1.0288	$(-0.66521, 0.66784, -0.33391)$
9	0.99642	$(-0.66685, 0.66652, -0.33326)$
12	1.0004	$(-0.66664, 0.66668, -0.33334)$

It appears that inverse iteration is converging to $\lambda = 1$ and the eigenvector looks suspiciously like a multiple of $(-2, 2, -1)$, which is an exact eigenpair of A.

There is much more to modern eigenvalue algorithms than we have indicated here. Central topics include deflation, the QR algorithm, numerical stability analysis, and many other issues. The interested reader might consult more advanced texts such as references [8], [9], [14] and [25], to name a few. As with the power method, complex numbers are permitted in these algorithms if transpose is replaced by *conjugate* transpose.

5.7 Exercises and Problems

Exercise 1. The matrix of (c) below may have complex eigenvalues. Use the Gershgorin circle theorem to locate eigenvalues and the iteration methods of this section to compute an approximate eigensystem.

(a) $\begin{bmatrix} 4 & -1 & 0 & 2 \\ 0 & 5 & 0 & -1 \\ -1 & -2 & 2 & 0 \\ 0 & 0 & 2 & 10 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 1 & 2 \\ 2 & 0 & 1 \\ -1 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & -2 & 0 & 0 \\ 2 & 4 & -2 & 0 \\ 0 & 2 & 4 & -2 \\ 0 & 0 & 2 & 1 \end{bmatrix}$

Exercise 2. Use the Gershgorin circle theorem to locate eigenvalues and the iteration methods of this section to compute an approximate eigensystem.

(a) $\begin{bmatrix} 3 & 1 & 0 & 0 \\ 1 & 5 & 1 & 0 \\ 0 & 1 & 7 & 1 \\ 0 & 0 & 1 & 9 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 1 & -2 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & -2 & -2 & 0 \\ 6 & -7 & 21 & -18 \\ 4 & -8 & 22 & -18 \\ 2 & -4 & 13 & -13 \end{bmatrix}$

*Problem 3. A square matrix is *strictly diagonally dominant* if in each row the sum of the absolute values of the off-diagonal entries is strictly less than the absolute value of the diagonal entry. Show that a strictly diagonally dominant matrix is invertible.

Problem 4. Let A be an $n \times n$ tridiagonal matrix with possibly complex entries a, b, c down the first subdiagonal, main diagonal, and first superdiagonal, respectively, where $a, c \neq 0$. Let $\mathbf{v} = (v_1, \ldots, v_n)$ satisfy $A\mathbf{v} = \lambda\mathbf{v}$.

(a) Show that $\mathbf{v}$ satisfies the difference equation $av_{j-1} + (b - \lambda)\, v_j + cv_{j+1} = 0$, $j = 1, \ldots, n$, with $v_0 = 0 = v_{n+1}$.

(b) Show that $v_j = Ar_1^j + Br_2^j$, where r_1, r_2 are (distinct) solutions to the auxiliary equation $a + (b - \lambda)\, r + cr^2 = 0$, is a solution to the difference equation in (a).

(c) Determine that $r_1 r_2 = a/c$, $r_1 + r_2 = (\lambda - b)/c$, and $(r_1/r_2)^{n+1} = e^{2i\pi s}$, $s = 1, \ldots, n$.

(d) Use (c) to find all r_1, r_2, and λ. (It helps to use the conditions $v_0 = 0 = v_{n+1}$ and examine the cases $j = 0$ and $j = n + 1$.)

(e) Conclude that a complete set of eigenpairs of A is specified by

$$\lambda_j = b + 2c\left(\frac{a}{c}\right)^{1/2} \cos\frac{j\pi}{n+1} \text{ and } \mathbf{v}_j = \left(\left(\frac{a}{c}\right)^{j/2} \sin\frac{sj\pi}{n+1}\right)_{s=1}^{n},\ j = 1, \ldots, n.$$

Problem 5. Determine the flop count for a single iteration of the power method applied to an $n \times n$ matrix A.

5.8 *Project Topics

Project: Finding a Jordan Canonical Form

A challenge: Use a technology tool to find the Jordan canonical form of the 10×10 matrix A, which is given *exactly* as follows.

$$A = \begin{bmatrix} 1 & 1 & 1 & -2 & 1 & -1 & 2 & -2 & 4 & -3 \\ -1 & 2 & 3 & -4 & 2 & -2 & 4 & -4 & 8 & -6 \\ -1 & 0 & 5 & -5 & 3 & -3 & 6 & -6 & 12 & -9 \\ -1 & 0 & 3 & -4 & 4 & -4 & 8 & -8 & 16 & -12 \\ -1 & 0 & 3 & -6 & 5 & -4 & 10 & -10 & 20 & -15 \\ -1 & 0 & 3 & -6 & 2 & -2 & 12 & -12 & 24 & -18 \\ -1 & 0 & 3 & -6 & 2 & -5 & 15 & -13 & 28 & -21 \\ -1 & 0 & 3 & -6 & 2 & -5 & 15 & -11 & 32 & -24 \\ -1 & 0 & 3 & -6 & 2 & -5 & 15 & -14 & 37 & -26 \\ -1 & 0 & 3 & -6 & 2 & -5 & 15 & -14 & 36 & -25 \end{bmatrix}.$$

Your main task is to devise a strategy for identifying the Jordan canonical form matrix J. Do *not* expect to find the invertible matrix S for which $J = S^{-1}AS$. However, a key fact to keep in mind is that if A and B are similar matrices, i.e., $A = S^{-1}BS$ for some invertible S, then $\operatorname{rank} A = \operatorname{rank} B$. Prove this rank fact for A and B. In particular, if S is a matrix that puts A into Jordan canonical form, then $J = S^{-1}AS$. Of course, this rank fact will also apply to $A - cI$ and its powers as well, for any scalar. Now you have the necessary machinery for determining numerically the Jordan canonical form.

As a first step, one can find the eigenvalues of A. Of course, these will only be approximate, so one has to decide how many eigenvalues are really repeated. Next, one has to determine the number of Jordan blocks of a given type. Suppose λ is an eigenvalue and find the rank of various powers of $A - \lambda I$. It would help in understanding how all this counts blocks if you first experiment with a matrix already in Jordan canonical form with different Jordan blocks.

Project: Solving Polynomial Equations

In homework problems we solve for the roots of the characteristic polynomial in order to get eigenvalues. To this end we can use algebra methods or even Newton's method for numerical approximations to the roots. This is the conventional wisdom usually proposed in introductory linear algebra. But for larger problems this method can be too slow and inaccurate. In fact, numerical methods hiding under the hood in a technology tool for finding eigenvalues are so efficient that it is better to turn this whole procedure on its head. Rather than find roots to solve linear algebra (eigenvalue) problems, we can use (numerical) linear algebra to find roots of polynomials. In this project we discuss this methodology and document it in a fairly nontrivial example.

Given a polynomial $f(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1} + x^n$, form the *companion matrix* of $f(x)$,

$$C(f) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -c_0 & -c_1 & \cdots & -c_{n-2} & -c_{n-1} \end{bmatrix}.$$

It is a key fact that the eigenvalues of $C(f)$ are precisely the roots of the equation $f(x) = 0$. Experiment with $n = 2, 3, 4$ by expansion across the bottom row of $\det(A - \lambda I)$ to confirm that this result is true. Then use a technology tool to illustrate this method by finding approximate roots of three polynomials: a cubic and quartic of your choice and then the polynomial

$$f(x) = 5 + 11x + 4x^2 + 6x^3 + x^4 - 15x^5 + 5x^6 - 3x^7 - 2x^8 + 8x^9 - 5x^{10} + x^{11}.$$

In each case use Newton's method to improve the values of some of the roots (it works with complex numbers as well as reals, provided one starts close enough to a root). Check your answers to this problem by evaluating the polynomial. Use your results to write the polynomial as a product of the linear factors $x - \lambda$, where λ is a root and check the correctness of this factorization.

Project: Classification of Quadric Forms

In order to classify *quadratic* equations in x and y one goes through roughly three steps. First, perform a rotation transformation of coordinates to get rid of mixed terms such as $2xy$ in the quadratic equation $x^2+2xy-y^2+x-3y = 4$. Second, do a translation of coordinates to put the equation in a "standard form." Third, identify the curve by your knowledge of the shape of a curve in that standard form. Standard forms are equations like $x^2/4 + y^2/2 = 1$, an ellipse with its axes along the x- and y-axes. It is the second-degree terms (x^2, $2xy$, and y^2) alone that determine the nature of a quadratic.

Now you're ready for the rest of the story. Just as with curves in x and y, the basic shape of the surface of a *quadric* equation in x, y, and z is determined by the second-degree terms. So we will focus on an example with no first-degree terms, namely,

$$Q(x, y, z) = 2x^2 + 4y^2 + 6z^2 - 4xy - 2xz + 2yz = 1.$$

The problem is this: find a change of coordinates that will make it clear what standard forms is represented by this surface. (You will need to google around a bit to find names of standard quadric forms.) First you must express the so-called *quadratic form* $Q(x, y, z)$ in matrix form as $Q(x, y, z) = [x, y, z]A[x, y, z]^T$. It is easy to find such matrices A. But not any such A will do. Next, replace A by the equivalent matrix $(A + A^T)/2$. (Check that if A specifies the quadratic form Q, then so will $(A + A^T)/2$.) The advantage of this latter matrix is that it is symmetric, so that there is an orthogonal matrix P such that P^TAP is diagonal. Next, make the linear change of variables $[x, y, z]^T = P[x', y', z']^T$ and deduce that $Q(x, y, z) = [x', y', z']P^TAP[x', y', z']^T$. If P^TAP is diagonal, we end up with squares of x', y' and z', and no mixed terms.

Find a symmetric A for this problem and use a technology tool to calculate the eigenvalues of this A. Also find unit-length eigenvectors for each eigenvalue. Put these together to form the desired orthogonal matrix P that eliminates mixed terms. From this data alone you should be able to classify

the surface represented by the above equation. An outstanding reference on this topic and many others relating to matrix analysis is the recently republished textbook [3] by Richard Bellman, widely considered to be a classic in the field.

Project: Image Compression

The object of this project is to illustrate the space savings capabilities of the SVD as applied to an image of your own choosing. Digitize a picture into a matrix of grayscale pixels. Each dimension of this matrix should be at least 400. Compute the SVD of this image matrix and display various reconstructions of the image using 20, 40, and 60 of the singular values and vector pairs. Do any of these give a good visual approximation to the picture? Find a minimal number that works. Compare the amount of storage required for these singular values and vectors to storage requirements of the full image.

Implementation Notes: First, you must convert the image to a grayscale format without layers. For this you will need an image manipulation program such as the GNU program Gimp or commercial software such as Adobe Photoshop. You will also need a technology tool capable of calculating an SVD and importing standard flattened image grayscale images (such as .png, etc.) into matrices and vice versa. The freely available R programming language and Octave, as well as commercial Matlab and others are perfectly capable of these tasks. In order to save storage space use single precision floating point arithmetic.

Report: Management of Sheep Populations

Description of the problem: You are working for the New Zealand Department of Agriculture on a project for sheep farmers. The species of sheep that these shepherds raise have a life span of 12 years. Of course, some live longer, but they are sufficiently few in number and their reproductive rate is so low that they may be ignored in your population study. Accordingly, you divide sheep into 12 age classes, namely those in the first year of life, etc. An extensive survey of the demographics of this species of sheep results in the following approximations for the demographic parameters f_i and s_i, where f_i is the per-capita reproductive rate for sheep in the ith age class and s_i is the survival rate for sheep in that age class, i.e., the fraction of sheep in that age class that survive to the $(i+1)$th class. (As a matter of fact, this table is related to real data. The interested reader might consult the article [7] in the bibliography.)

i	1	2	3	4	5	6	7	8	9	10	11	12
f_i	.000	.023	.145	.236	.242	.273	.271	.251	.234	.229	.216	.210
s_i	.845	.975	.965	.950	.926	.895	.850	.786	.691	.561	.370	-

The problem is as follows: in order to maintain a constant population of sheep, shepherds will harvest a certain number of sheep each year. Harvesting need not mean slaughter; it simply means removing sheep from the population (e.g., selling animals to other shepherds). Denote the fraction of sheep that are removed from the ith age group at the end of each growth period

(a year in our case) by h_i. If these numbers are constant from year to year, they constitute a *harvesting policy.* If, moreover, the yield of each harvest, i.e., total number of animals harvested each year, is a constant and the age distribution of the remaining populace is essentially constant after each harvest, then the harvesting policy is called *sustainable.* If all the h_i's are the same, say h, then the harvesting policy is called *uniform.* Uniform policies are simple to implement: One selects the sheep to be harvested at random.

Your problem: Find a uniform sustainable harvesting policy to recommend to shepherds, and find the resulting distribution of sheep that they can expect with this policy. Shepherds who raise sheep for sale to markets are also interested in a sustainable policy that gives a maximum yield. If you can find such a policy that has a larger annual yield than the uniform policy, then recommend it. On the other hand, shepherds who raise sheep for their wool may prefer to minimize the annual yield. If you can find a sustainable policy whose yield is smaller than that of the uniform policy, make a recommendation accordingly. In each case find the expected distribution of your harvesting policies. Do you think that there might be other economic factors that should be taken into account in this model? Organize your results for a report to be read by your supervisor and an informed public.

Procedure: Express this problem as a discrete linear dynamical system $\mathbf{x}^{(k+1)} = L\mathbf{x}^{(k)}$, where L is a so-called *Leslie matrix* of the form

$$L = \begin{bmatrix} f_1 & f_2 & f_3 & \cdots & f_{n-1} & f_n \\ s_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & s_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s_{n-1} & 0 \end{bmatrix}.$$

It is understood that $0 < s_i \leq 1$, $0 \leq f_i$, and at least one f_i is nonzero.

In regard to harvesting, let H be a diagonal matrix with the harvest fractions h_i down the diagonal. (Here $0 \leq h_i \leq 1$.) Then the population that results from this harvesting at the end of each period is given by $\mathbf{x}^{k+1} = L\mathbf{x}^k - HL\mathbf{x}^k = (I - H)L\mathbf{x}^k$. There are other theoretical tools, but all you need to do is to find a matrix H such that 1 is the dominant eigenvalue of $(I - H)L$. You can do this by trial and error, a method that is applicable to any harvesting policy, uniform or not. However, in the case of uniform policies it's simpler to note that $(I - H)L = (1 - h)L$, where h is the diagonal entry of H. Find an eigenvector corresponding to this eigenvalue and scale this vector by dividing it by the sum of its components to obtains a probability distribution vector that could be used for any population.

6

GEOMETRICAL ASPECTS OF ABSTRACT SPACES

Two basic ideas that we learn in geometry are those of length of a line segment and angle between lines. We have already seen how to extend these ideas to the standard vector spaces. The objective of this chapter is to extend these powerful ideas to general linear spaces. A surprising number of concepts and techniques that we learned in a standard setting can be carried over, almost word for word, to more general vector spaces. Once this is accomplished, we will be able to use our geometrical intuition in entirely new ways. For example, we will be able to have notions of size (length) and perpendicularity for nonstandard vectors such as functions in a function space. We will be able to give a sensible meaning to the size of the error incurred in solving a linear system with finite-precision arithmetic. We shall see that there are many more applications of this abstraction.

6.1 Normed Spaces

Definitions and Examples

The basic function of a norm is to measure length and distance, independent of any other considerations, such as angles or orthogonality. There are different ways to accomplish such a measurement. One method of measuring length might be more natural for a given problem, or easier to calculate than another. For these reasons, we would like to have the option of using different methods of length measurement. You may recognize the properties listed below from earlier in the text; they are the basic norm laws given in Section 4.1 for the standard norm. We are going to abstract the norm idea to arbitrary vector spaces.

Definition 6.1. Abstract Norm A *norm* on the vector space V is a function $\|\cdot\|$ that assigns to each vector $\mathbf{v} \in V$ a real number $\|\mathbf{v}\|$ such that for c a scalar and $\mathbf{u}, \mathbf{v} \in V$ the following hold:

(1) $\|\mathbf{u}\| \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$.
(2) $\|c\mathbf{u}\| = |c| \, \|\mathbf{u}\|$.
(3) (Triangle Inequality) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

Definition 6.2. Normed Space and Distance Function A vector space V, together with a norm $\|\cdot\|$ on the space V, is called a *normed space*. If $\mathbf{u}, \mathbf{v} \in V$, the distance between $\mathbf{u}$ and $\mathbf{v}$ is defined to be $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$.

Notice that if V is a normed space and W is any subspace of V, then W automatically becomes a normed space if we simply use the norm of V on elements of W. Obviously all the norm laws still hold, since they hold for elements of the bigger space V.

Standard Norms

Of course, we have already studied some very important examples of normed spaces, namely the standard vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$, or any subspace thereof, together with the standard norms given by

$$\begin{aligned} \|(z_1, z_2, \ldots, z_n)\| &= \sqrt{\overline{z_1} z_1 + \overline{z_2} z_2 + \cdots + \overline{z_n} z_n} \\ &= \left(|z_1|^2 + |z_2|^2 + \cdots + |z_n|^2 \right)^{1/2}. \end{aligned}$$

If the vectors are real then we can drop the conjugate bars. This norm is actually one of a family of norms that are commonly used.

Definition 6.3. p-norm Let V be one of the standard spaces $\mathbb{R}^n$ or $\mathbb{C}^n$ and $p \geq 1$ a real number. The p-norm of a vector in V is defined by the formula

$$\|(z_1, z_2, \ldots, z_n)\|_p = \left(|z_1|^p + |z_2|^p + \cdots + |z_n|^p \right)^{1/p}.$$

Notice that when $p = 2$ we have the familiar example of the standard norm. Another important case is that in which $p = 1$. The last important instance of a p-norm is one that isn't so obvious: $p = \infty$. It turns out that the value of this norm is the limit of p-norms as $p \to \infty$. To keep matters simple, we'll supply a separate definition for this norm.

Definition 6.4. ∞-norm Let V be one of the standard spaces $\mathbb{R}^n$ or $\mathbb{C}^n$. The ∞-norm of a vector in V is defined by the formula

$$\|(z_1, z_2, \ldots, z_n)\|_\infty = \max \{ |z_1|, |z_2|, \ldots, |z_n| \}.$$

That norm laws (1) and (2) hold for all p-norms is easy to see. The triangle inequality is more subtle. We verified it for $p = 2$ in Section 4.2, will verify it for $p = \infty$ in Example 6.3 and leave the case $p = 1$ as an exercise. For the

other values of p, a fact called the Minkowski inequality is required, and the interested reader can consult [18] for details.

Example 6.1. Calculate $\|\mathbf{v}\|_p$, where $p = 1$, 2, or ∞ and $\mathbf{v} = (1, -3, 2, -1) \in \mathbb{R}^4$.

Solution. We calculate:

$$\begin{aligned}\|(1,-3,2,-1)\|_1 &= |1| + |-3| + |2| + |-1| = 7\\ \|(1,-3,2,-1)\|_2 &= \sqrt{|1|^2 + |-3|^2 + |2|^2 + |-1|^2} = \sqrt{15}\\ \|(1,-3,2,-1))\|_\infty &= \max\{|1|, |-3|, |2|, |-1|\} = 3.\end{aligned}$$

□

It may seem a bit odd at first to speak of the same vector as having different lengths. You should take the point of view that choosing a norm is a bit like choosing a measuring stick. If you choose a yard stick, you won't measure the same number as you would by using a meter stick on an object.

Example 6.2. Calculate $\|\mathbf{v}\|_p$, where $p = 1$, 2, or ∞ and $\mathbf{v} = (2 - 3\mathrm{i}, 1 + \mathrm{i}) \in \mathbb{C}^2$.

Solution. We calculate:

$$\begin{aligned}\|(2-3\mathrm{i}, 1+\mathrm{i})\|_1 &= |2-3\mathrm{i}| + |1+\mathrm{i}| = \sqrt{13} + \sqrt{2}\\ \|(2-3\mathrm{i}, 1+\mathrm{i})\|_2 &= \sqrt{|2-3\mathrm{i}|^2 + |1+\mathrm{i}|^2} = \sqrt{(2)^2 + (-3)^2 + 1^2 + 1^2} = \sqrt{15}\\ \|(2-3\mathrm{i}, 1+\mathrm{i})\|_\infty &= \max\{|2-3\mathrm{i}|, |1+\mathrm{i}|\} = \max\left\{\sqrt{13}, \sqrt{2}\right\} = \sqrt{13}.\end{aligned}$$

□

Example 6.3. Verify that the norm properties are satisfied for the p-norm in the case that $p = \infty$.

Solution. Let c be a scalar, and let $\mathbf{u} = (z_1, z_2, \ldots, z_n)$, and $\mathbf{v} = (w_1, w_2, \ldots, w_n)$ be two vectors. Any absolute value is nonnegative, and any vector whose largest component in absolute value is zero must have all components equal to zero. Property (1) follows. Next, we have that

$$\begin{aligned}\|c\mathbf{u}\|_\infty &= \|(cz_1, cz_2, \ldots, cz_n)\|_\infty = \max\{|cz_1|, |cz_2|, \ldots, |cz_n|\}\\ &= |c| \max\{|z_1|, |z_2|, \ldots, |z_n|\} = |c| \ \|\mathbf{u}\|_\infty,\end{aligned}$$

which proves (2). For (3) we observe that

$$\begin{aligned}\|\mathbf{u} + \mathbf{v}\|_\infty &= \max\{|z_1| + |w_1|, |z_2| + |w_2|, \ldots, |z_n| + |w_n|\}\\ &\le \max\{|z_1|, |z_2|, \ldots, |z_n|\} + \max\{|w_1|, |w_2|, \ldots, |w_n|\}\\ &\le \|\mathbf{u}\|_\infty + \|\mathbf{v}\|_\infty.\end{aligned}$$

□

Unit Vectors

Sometimes it is convenient to deal with vectors whose length is one. Such a

Unit Vector vector is called a *unit vector*. We saw in Chapter 3 that it is easy to concoct a unit vector $\mathbf{u}$ in the same direction as a nonzero vector $\mathbf{v}$ when using the standard norm, namely take

$$\mathbf{u} = \frac{\mathbf{v}}{\|\mathbf{v}\|}. \tag{6.1}$$

The same formula holds for all norms because of norm property (2).

Example 6.4. Construct a unit vector in the direction of $\mathbf{v} = (1, -3, 2, -1)$, where the 1-norm, 2-norm, and ∞-norms are used to measure length.

Solution. We already calculated each of the norms of $\mathbf{v}$ in Example 6.1. Use these numbers in equation (6.1) to obtain unit-length vectors

$$\begin{aligned}
\mathbf{u}_1 &= \frac{1}{7}(1, -3, 2, -1) \\
\mathbf{u}_2 &= \frac{1}{\sqrt{15}}(1, -3, 2, -1) \\
\mathbf{u}_\infty &= \frac{1}{3}(1, -3, 2, -1).
\end{aligned}$$

□

From a geometric point of view there are certain sets of vectors in the vector space V that tell us a lot about distances. These are the so-called *balls about a vector (or point)* $\mathbf{v}_0$ *of radius* r, whose definition is as follows:

$$B_r(\mathbf{v}_0) = \{\mathbf{v} \in V \mid \|\mathbf{v} - \mathbf{v}_0\| \leq r\}.$$

Ball in Normed Space Sometimes these are called *closed balls*, as opposed to *open balls,* which are defined using strict inequality. Here is a situation in which these balls are very helpful: imagine trying to find the distance from a vector $\mathbf{v}_0$ to a closed (this means it contains all points on itsboundary) set S of vectors that need not be a subspace. One way to accomplish this is to start with a ball centered at $\mathbf{v}_0$ such that the ball avoids S. Then expand this ball by increasing its radius until you have found the least upper bound R of radii r such that the ball $B_r(\mathbf{v}_0)$ has empty intersection with S. Then the distance from $\mathbf{v}_0$ to this set is this number R. Actually, this is a reasonable *definition* of the distance from $\mathbf{v}_0$ to the set S. One expects these balls, for a particular norm, to have the same shape, so it is sufficient to look at the unit balls, that is, the case $r = 1$.

Example 6.5. Sketch the unit balls centered at the origin for the 1-norm, 2-norm, and ∞-norms in the space $V = \mathbb{R}^2$.

Solution. In each case it's easiest to determine the boundary of the ball $B_1(0)$, i.e., the set of vectors $\mathbf{v} = (x, y)$ such that $\|\mathbf{v}\| = 1$. These boundaries are sketched in Figure 6.1, and the ball consists of the boundaries plus

the interior of each boundary. Let's start with the familiar 2-norm. Here the boundary consists of points (x, y) such that

$$1 = \|(x, y)\|_2 = \sqrt{x^2 + y^2},$$

which is the familiar circle of radius 1 centered at the origin. Next, consider the 1-norm, in which case

$$1 = \|(x, y)\|_1 = |x| + |y| .$$

It's easier to examine this formula in each quadrant, where it becomes one of the four possibilities

$$\pm x \pm y = 1.$$

For example, in the first quadrant we get $x + y = 1$. These equations give lines that connect to form a square whose sides are diagonal lines. Finally, for the ∞-norm we have

$$1 = |(x, y)|_\infty = \max \{|x| , |y|\} ,$$

which gives four horizontal and vertical lines $x = \pm 1$ and $y = \pm 1$. These intersect to form another square. Thus, we see that the unit "balls" for the 1- and ∞-norms have corners, unlike the 2-norm. See Figure 6.1 for a picture of these balls. □

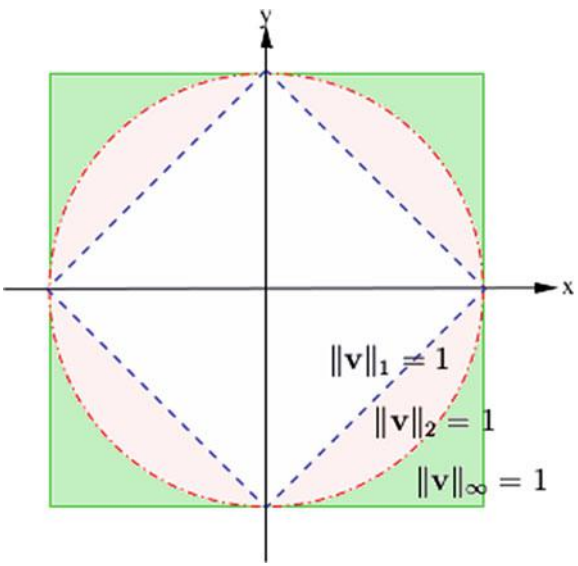


Fig. 6.1: Boundaries of unit balls in various norms.

Recall from Section 4.1 that one of the important applications of the norm concept is that it enables us to make sense out of the idea of limits and convergence of vectors. In a nutshell, $\lim_{n\to\infty} \mathbf{v}_n = \mathbf{v}$ was taken to mean that $\lim_{n\to\infty} \|\mathbf{v}_n - \mathbf{v}\| = 0$. In this case we said that the sequence $\mathbf{v}_1, \mathbf{v}_2, \ldots$ *converges* to $\mathbf{v}$. Will we have to have a different notion of limits for different norms? For *finite-dimensional* spaces, the somewhat surprising answer is no. The reason is that given any two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on a finite-dimensional vector

Equivalent Norms

space, it is always possible to find positive real constants c and d such that for any vector $\mathbf{v}$,

$$\|\mathbf{v}\|_a \leq c \cdot \|\mathbf{v}\|_b \text{ and } \|\mathbf{v}\|_b \leq d\,\|\mathbf{v}\|_a .$$

Hence, if $\|\mathbf{v}_n - \mathbf{v}\|$ tends to 0 in one norm, it will tend to 0 in the other norm. For this reason, any two norms satisfying these inequalities are called *equivalent*. It can be shown that all norms on a finite-dimensional vector space are equivalent (see Section 6.5). Indeed, it can be shown that the condition that $\|\mathbf{v}_n - \mathbf{v}\|$ tends to 0 in any one norm is equivalent to the condition that each coordinate of $\mathbf{v}_n$ converges to the corresponding coordinate of $\mathbf{v}$. We will verify the limit fact in the following example.

Example 6.6. Verify that $\lim_{n\to\infty} \mathbf{v}_n$ exists and is the same with respect to both the 1-norm and 2-norm, where

$$\mathbf{v}_n = \begin{bmatrix} (1-n)/n \\ e^{-n}+1 \end{bmatrix}.$$

Which norm is easier to work with?

Solution. First we have to know what the limit will be. Let's examine the limit in each coordinate. We have

$$\lim_{n\to\infty} \frac{1-n}{n} = \lim_{n\to\infty} \frac{1}{n} - 1 = 0 - 1 = -1 \text{ and } \lim_{n\to\infty} e^{-n} + 1 = 0 + 1 = 1.$$

So we try to use $\mathbf{v} = (-1, 1)$ as the limiting vector. Now calculate

$$\mathbf{v} - \mathbf{v}_n = \begin{bmatrix} -1 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1-n}{n} \\ e^{-n}+1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{n} \\ -e^{-n} \end{bmatrix},$$

so that

$$\|\mathbf{v} - \mathbf{v}_n\|_1 = \left|-\frac{1}{n}\right| + \left|-e^{-n}\right| \underset{n\to\infty}{\longrightarrow} 0$$

and

$$\|\mathbf{v} - \mathbf{v}_n\| = \sqrt{\left(\frac{1}{n}\right)^2 + \left(e^{-n}\right)^2} \underset{n\to\infty}{\longrightarrow} 0,$$

which shows that the limits are the same in either norm. In this case the 1-norm appears to be easier to work with, since no squaring and square roots are involved. In fact, if we are dealing only with nonnegative vectors such as distribution vectors computing 1-norms amounts to adding up the coordinates. This explains why it appeared in the analysis of PageRank matrices in Section 2.5 of Chapter 2. □

Here are two examples of norms defined on nonstandard vector spaces:

Definition 6.5. Frobenius Norm Let $V = \mathbb{R}^{m,n}$ (or $\mathbb{C}^{m,n}$). The Frobenius norm of an $m \times n$ matrix $A = [a_{ij}]$ is defined by the formula

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}.$$

We leave verification of the norm laws as an exercise.

Definition 6.6. Uniform (Infinity) Norm on Function Space The uniform (or infinity) *norm* on $C[a, b]$ is defined by $\|f\|_\infty = \max_{a \le x \le b} |f(x)|$.

This norm is well defined by the extreme value theorem, which guarantees that the maximum value of a continuous function on a closed interval exists. We leave verification of the norm laws as an exercise.

6.1 Exercises and Problems

Exercise 1. Find the 1-, 2-, and ∞-norms of each of the following real vectors and the distance between these pairs in each norm.
(a) $(2, 1, 3)$, $(-3, 1, -1)$ (b) $(1, -2, 0, 1, 3)$, $(2, 2, -1, -1, -2)$

Exercise 2. Find the 1-, 2-, and ∞-norms of each of the following complex vectors and the distance between these pairs in each norm.
(a) $(1 + \mathrm{i}, -1, 0, 1)$, $(1, 1, 2, -4)$ (b) $(\mathrm{i}, 0, 3 - 2\mathrm{i})$, $(\mathrm{i}, 1 + \mathrm{i}, 0)$

Exercise 3. Find unit vectors in the direction of each of the following vectors with respect to the 1-, 2-, and ∞-norms.
(a) $(1, -3, -1)$ (b) $(3, 1, -1, 2)$ (c) $(2, 1, 3 + \mathrm{i})$

Exercise 4. Find a unit vector in the direction of $f(x) \in C[0, 1]$ with respect to the uniform norm, where $f(x)$ is one of the following.
(a) $\sin(\pi x)$ (b) $x(x - 1)$ (c) e^x

Exercise 5. Verify the norm laws for the 1-norm in the case that $c = -2$, $\mathbf{u} = (0, 2, 3, 1)$, and $\mathbf{v} = (1, -3, 2, -1)$ in $V = \mathbb{R}^4$.

Exercise 6. Verify the norm laws for the Frobenius norm in the case that $c = -4$, $\mathbf{u} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 0 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} -2 & 0 & 2 \\ 1 & 0 & -3 \end{bmatrix}$ in $V = \mathbb{R}^{2,3}$.

Exercise 7. Find the distance from the point $\left(-1, -\frac{1}{2}\right)$ to the line $x + y = 2$ using the ∞-norm by sketching a picture of the ball centered at that point that touches the line.

Exercise 8. Find the constant function that is nearest the function $f(x) = 4x(1-x) \in V = C[0,1]$ with the infinity norm. (*Hint:* examine a graph of $f(x)$ and a constant function.)

Exercise 9. Describe in words the unit ball $B_1\left([1,1,1]^T\right)$ in the normed space $V = \mathbb{R}^3$ with the infinity norm.

Exercise 10. Describe in words the unit ball $B_1(g(x))$ in the normed space $V = C[0,1]$ with the uniform norm and $g(x) = 2$.

Exercise 11. Verify that $\lim_{n\to\infty} \mathbf{v}_n$ exists and is the same with respect to both the 1- and 2-norms in $V = \mathbb{R}^2$, where $\mathbf{v}_n = ((1-n)/n, e^{-n}+1)$.

Exercise 12. Calculate $\lim_{n\to\infty} f_n$ using the uniform norm on $V = C[0,1]$, where $f_n(x) = (x/2)^n + 1$.

*Problem 13. Given the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, find the largest possible value of $||A\mathbf{x}||_\infty$, where $\mathbf{x}$ ranges over the vectors whose ∞-norm is 1.

*Problem 14. Verify that the 1-norm satisfies the definition of a norm.

*Problem 15. Show that the Frobenius norm satisfies the norm properties.

Problem 16. Show that the infinity norm on $C[0,1]$ satisfies the norm properties.

Problem 17. Show that if A is a nonsingular $n \times n$ matrix and $\|\cdot\|$ is a norm on one of the standard spaces $\mathbb{R}^n$ or $\mathbb{C}^n$, then the formula $\|\mathbf{x}\|_A = \|A\mathbf{x}\|$ defines another norm $\|\cdot\|_A$ on that space.

Problem 18. Determine whether or not the formula for $f(x) \in C[0,1]$ given by $\|f\|_{max} = \max_{0\le x\le 1} f(x)^2$ defines a norm on the vector space $C[0,1]$.

6.2 Inner Product Spaces

Definitions and Examples

We saw in Section 4.2 that the notion of a dot product of two vectors had many handy applications, including the determination of the angle between two vectors. This dot product amounted to the "standard" inner product of the two standard vectors. We now extend this idea to a setting that allows for abstract vector spaces.

Definition 6.7. Abstract Inner Product and Inner Product Space An (abstract) *inner product* on the vector space V is a function $\langle \cdot, \cdot \rangle$ that assigns to each pair of vectors $\mathbf{u}, \mathbf{v} \in V$ a scalar $\langle \mathbf{u}, \mathbf{v} \rangle$ such that for c a scalar and $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ the following hold:

(1) $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ with $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$.
(2) $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$
(3) $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$
(4) $\langle \mathbf{u}, c\mathbf{v} \rangle = c \langle \mathbf{u}, \mathbf{v} \rangle$

A vector space V, together with an inner product $\langle \cdot, \cdot \rangle$ on the space V, is called an *inner product space*.

Notice that in the case of the more common vector spaces over *real scalars*, property (2) becomes a commutative law: $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$. Also observe that if V is an inner product space and W is any subspace of V, then W automatically becomes an inner product space if we simply use the inner product of V on elements of W. For all the inner product laws still hold, since they hold for elements of the larger space V.

Of course, we have the standard examples of inner products, namely the dot products on $\mathbb{R}^n$ and $\mathbb{C}^n$. Here is an example of a nonstandard inner product on a standard space that is useful in certain engineering problems.

Nonstandard Inner Products

Example 6.7. For vectors $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ in $V = \mathbb{R}^2$, define an inner product by the formula

$$\langle \mathbf{u}, \mathbf{v} \rangle = 2u_1v_1 + 3u_2v_2.$$

Show that this formula satisfies the inner product laws.

Solution. First we see that

$$\langle \mathbf{u}, \mathbf{u} \rangle = 2u_1^2 + 3u_2^2,$$

so the only way for this sum to be 0 is for $u_1 = u_2 = 0$. Hence, (1) holds. For (2) calculate

$$\langle \mathbf{u}, \mathbf{v} \rangle = 2u_1v_1 + 3u_2v_2 = 2v_1u_1 + 3v_2u_2 = \langle \mathbf{v}, \mathbf{u} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle},$$

since all scalars in question are real. For (3) let $\mathbf{w} = (w_1, w_2)$ and calculate

$$\begin{aligned} \langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle &= 2u_1 (v_1 + w_1) + 3u_2 (v_2 + w_2) \\ &= 2u_1v_1 + 3u_2v_2 + 2u_1w_1 + 3u_2 = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle . \end{aligned}$$

For the last property, check that for a scalar c,

$$\langle \mathbf{u}, c\mathbf{v} \rangle = 2u_1cv_1 + 3u_2cv_2 = c (2u_1v_1 + 3u_2v_2) = c \langle \mathbf{u}, \mathbf{v} \rangle . \qquad \square$$

It follows that this "weighted" inner product is indeed an inner product according to our definition. In fact, we can do a whole lot more with even less effort. Consider this example, of which the preceding is a special case.

Example 6.8. Let A be an $n \times n$ Hermitian matrix ($A = A^*$) and define the product $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^* A \mathbf{v}$ for all $\mathbf{u}, \mathbf{v} \in V$, where V is $\mathbb{R}^n$ or $\mathbb{C}^n$. Show that this product satisfies inner product laws (2), (3), and (4) and that if, in addition, A is positive definite, then the product satisfies (1) and is an inner product.

Solution. As usual, let $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and let c be a scalar. For (2), remember that for a 1×1 scalar quantity q, $q^* = \overline{q}$, so we calculate

$$\langle \mathbf{v}, \mathbf{u} \rangle = \mathbf{v}^* A \mathbf{u} = (\mathbf{u}^* A \mathbf{v})^* = \langle \mathbf{u}, \mathbf{v} \rangle^* = \overline{\langle \mathbf{u}, \mathbf{v} \rangle}.$$

For (3), we calculate

$$\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \mathbf{u}^* A (\mathbf{v} + \mathbf{w}) = \mathbf{u}^* A \mathbf{v} + \mathbf{u}^* A \mathbf{w} = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle .$$

For (4), we have that

$$\langle \mathbf{u}, c\mathbf{v} \rangle = \mathbf{u}^* A c \mathbf{v} = c \mathbf{u}^* A \mathbf{v} = c \langle \mathbf{u}, \mathbf{v} \rangle .$$

Finally, if we suppose that A is also positive definite, then by definition,

$$\langle \mathbf{u}, \mathbf{u} \rangle = \mathbf{u}^* A \mathbf{u} > 0, \text{ for } \mathbf{u} \neq \mathbf{0},$$

which shows that inner product property (1) holds. Hence, this product defines an inner product. □

We leave it to the reader to check that if we take $A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$, then the inner product defined by this matrix is exactly the inner product of Example 6.7.

The previous example demonstrates that a vector space may have more than one inner product on it. In particular, $V = \mathbb{R}^2$ could have the standard inner product, i.e., dot product or something else like the previous example. The space V, together with each one of these inner products, provides us with two separate inner product spaces.

Here is a rather more exotic example of an inner product involving a nonstandard vector space.

Example 6.9. Let $V = C[a, b]$, the space of continuous functions on the interval $[a, b]$ with the usual function addition and scalar multiplication. Show that the formula

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, dx$$

defines an inner product on the space V.

Solution. Certainly, $\langle f, g\rangle$ is a real number. Now if $f(x)$ is a continuous function then $f(x)^2$ is nonnegative on $[a, b]$ and therefore $\int_a^b f(x)^2 dx = \langle f, f\rangle \geq 0$. Furthermore, if $f(x)$ is nonzero, then the area under the curve $y = f(x)^2$ must also be positive since $f(x)$ will be positive and bounded away from 0 on some subinterval of $[a, b]$. This establishes property (1) of inner products.

Now let $f(x), g(x), h(x) \in V$. For property (2), notice that

$$\langle f, g\rangle = \int_a^b f(x)g(x)dx = \int_a^b g(x)f(x)dx = \langle g, f\rangle .$$

Also,

$$\begin{aligned}\langle f, g + h\rangle &= \int_a^b f(x)(g(x) + h(x))dx \\ &= \int_a^b f(x)g(x)dx + \int_a^b f(x)h(x)dx = \langle f, g\rangle + \langle f, h\rangle ,\end{aligned}$$

which establishes property (3). Finally, we see that for a scalar c,

$$\langle f, cg\rangle = \int_a^b f(x)cg(x)\, dx = c\int_a^b f(x)g(x)\, dx = c\,\langle f, g\rangle ,$$

which shows that property (4) holds. □

We shall refer to this inner product on a function space as the *standard inner product* on the function space $C[a, b]$. (Most of our examples and exercises involving function spaces will deal with polynomials, so we remind the reader of the integration formula $\int_a^b x^m\, dx = \frac{1}{m+1}\left(b^{m+1} - a^{m+1}\right)$ and special case $\int_0^1 x^m\, dx = \frac{1}{m+1}$ for $m \geq 0$.)

Function Space
Standard Inner Product

Following are a few simple facts about inner products that we will use frequently. The proofs are left to the exercises.

Theorem 6.1. Let V be an inner product space with inner product $\langle \cdot, \cdot \rangle$. Then we have that for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and scalars a,

(1) $\langle \mathbf{u}, \mathbf{0}\rangle = 0 = \langle \mathbf{0}, \mathbf{u}\rangle$,
(2) $\langle \mathbf{u} + \mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{u}, \mathbf{w}\rangle + \langle \mathbf{v}, \mathbf{w}\rangle$,
(3) $\langle a\mathbf{u}, \mathbf{v}\rangle = \overline{a}\langle \mathbf{u}, \mathbf{v}\rangle$

Induced Norms and the CBS Inequality

It is a striking fact that we can accomplish all the goals we set for the standard inner product using general inner products: we can introduce the ideas of angles, orthogonality, projections, and so forth. We have already seen much

of the work that has to be done, though it was stated in the context of the standard inner products. As a first step, we want to point out that every inner product has a "natural" norm associated with it.

Definition 6.8. Induced Norm Let V be an inner product space. For vectors $\mathbf{u} \in V$, the norm defined by the equation

$$\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$$

is called the *norm induced by the inner product* $\langle \cdot, \cdot \rangle$ *on* V.

As a matter of fact, this idea is not really new. Recall that we introduced the standard inner product on $V = \mathbb{R}^n$ or $\mathbb{C}^n$ with an eye toward the standard norm. At the time it seemed like a nice convenience that the norm could be expressed in terms of the inner product. It is, and so much so that we have turned this cozy relationship into a definition. Just calling the induced norm a norm doesn't make it so. Is the induced norm really a norm? We have some work to do. The first norm property is easy to verify for the induced norm: from property (1) of inner products we see that $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, with equality if and only if $\mathbf{u} = 0$. This confirms norm property (1). Norm property (2) isn't too hard either: let c be a scalar and check that

$$\|c\mathbf{u}\| = \sqrt{\langle c\mathbf{u}, c\mathbf{u} \rangle} = \sqrt{c\bar{c} \langle \mathbf{u}, \mathbf{u} \rangle} = \sqrt{|c|^2} \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle} = |c| \, \|\mathbf{u}\| .$$

Norm property (3), the triangle inequality, remains. This one isn't easy to verify from first principles. We need a tool that we have seen before, the Cauchy–Bunyakovsky–Schwarz (CBS) inequality. We restate it below as the next theorem. Indeed, the very same proof that is given in Theorem 4.2 carries over word for word to general inner products over real vector spaces. We need only replace dot products $\mathbf{u} \cdot \mathbf{v}$ by abstract inner products $\langle \mathbf{u}, \mathbf{v} \rangle$. We can also replace dot products by inner products in Problem 18 of Chapter 4, which establishes CBS for complex inner products. Similarly, the proof of the triangle inequality as given in Example 4.10 of Section 4.2, carries over to establish the triangle inequality for abstract inner products. Hence, property (3) of norms holds for any induced norm.

Theorem 6.2. CBS Inequality Let V be an inner product space. For $\mathbf{u}, \mathbf{v} \in V$, if we use the inner product of V and its induced norm, then

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \, \|\mathbf{v}\| .$$

Henceforth, when the norm sign $\|\cdot\|$ is used in connection with an inner product, it is understood that this norm is the induced norm of this inner product, unless otherwise stated.

Just as with the standard dot products, we can formulate the following definition thanks to the CBS inequality.

Definition 6.9. Angle Between Vectors For vectors $\mathbf{u}, \mathbf{v} \in V$, a real inner product space, we define the *angle* between $\mathbf{u}$ and $\mathbf{v}$ to be any angle θ satisfying

$$\cos\theta = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \, \|\mathbf{v}\|}.$$

We know that $|\langle \mathbf{u}, \mathbf{v} \rangle| \, / \, (\|\mathbf{u}\| \, \|\mathbf{v}\|) \leq 1$, so that this formula for $\cos\theta$ makes sense.

Example 6.10. Let $\mathbf{u} = (1, -1)$ and $\mathbf{v} = (1, 1)$ be vectors in $\mathbb{R}^2$. Compute an angle between these two vectors using the inner product of Example 6.7. Compare this to the angle found when one uses the standard inner product in $\mathbb{R}^2$.

Solution. According to 6.7 and the definition of angle, we have

$$\cos\theta = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \, \|\mathbf{v}\|} = \frac{2 \cdot 1 \cdot 1 + 3 \cdot (-1) \cdot 1}{\sqrt{2 \cdot 1^2 + 3 \cdot (-1)^2}\sqrt{2 \cdot 1^2 + 3 \cdot 1^2}} = \frac{-1}{5}.$$

Hence, the angle in radians is

$$\theta = \arccos\left(\frac{-1}{5}\right) \approx 1.7722.$$

On the other hand, if we use the standard norm, then

$$\langle \mathbf{u}, \mathbf{v} \rangle = 1 \cdot 1 + (-1) \cdot 1 = 0,$$

from which it follows that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal and $\theta = \pi/2 \approx 1.5708$. □

In the previous example, it shouldn't be too surprising that we can arrive at two different values for the "angle" between two vectors. Using different inner products to measure angle is somewhat like measuring length with different norms. Next, we extend the perpendicularity idea to arbitrary inner product spaces.

Definition 6.10. Orthogonal Vectors Two vectors $\mathbf{u}$ and $\mathbf{v}$ in the same inner product space are *orthogonal* if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Note that if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, then $\langle \mathbf{v}, \mathbf{u} \rangle = \overline{\langle \mathbf{u}, \mathbf{v} \rangle} = 0$. Also, this definition makes the zero vector orthogonal to every other vector. It also allows us to speak of things like "orthogonal functions." One has to be careful with new ideas like this. Orthogonality in a function space is not something that can be as easily visualized as orthogonality of geometrical vectors. Inspecting the graphs of two functions may not be quite enough. If, however, graphical data is tempered with a little understanding of the particular inner product in use, orthogonality can be detected.

Example 6.11. Show that $f(x) = x$ and $g(x) = x - \frac{2}{3}$ are orthogonal elements of $C[0,1]$ with the inner product of Example 6.9 and provide graphical evidence of this fact.

Solution. According to the definition of inner product in this space,

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx = \int_0^1 x\left(x - \frac{2}{3}\right)dx = \left(\frac{x^3}{3} - \frac{x^2}{3}\right)\Big|_0^1 = 0.$$

It follows that f and g are orthogonal to each other. For graphical evidence, sketch $f(x)$, $g(x)$, and $f(x)g(x)$ on the interval $[0,1]$ as in Figure 6.2. The graphs of f and g are not especially enlightening; but we can see in the graph that the area below $f \cdot g$ and above the x-axis to the right of $(2/3, 0)$ seems to be about equal to the area to the left of $(2/3, 0)$ above $f \cdot g$ and below the x-axis. Therefore, the integral of the product on the interval $[0,1]$ might be expected to be zero, which is indeed the case. □

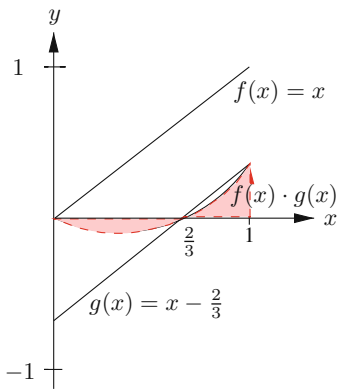


Fig. 6.2: Graphs of f, g, and $f \cdot g$ on the interval $[0,1]$.

Some of the basic ideas from geometry that fuel our visual intuition extend very elegantly to the inner product space setting. One such example is the famous Pythagorean theorem, which takes the following form in an inner product space.

Theorem 6.3. **Pythagorean Theorem** Let $\mathbf{u}, \mathbf{v}$ be orthogonal vectors in an inner product space V. Then $\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \|\mathbf{u} + \mathbf{v}\|^2$.

Proof. Compute

$$\begin{aligned}\|\mathbf{u} + \mathbf{v}\|^2 &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v}\rangle \\ &= \langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{u}, \mathbf{v}\rangle + \langle \mathbf{v}, \mathbf{u}\rangle + \langle \mathbf{v}, \mathbf{v}\rangle \\ &= \langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{v}, \mathbf{v}\rangle = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2.\end{aligned}$$

□

Here is an example of another standard geometrical fact that fits well in

the abstract setting. This is equivalent to the parallelogram equality, which says that the sum of the squares of the diagonals of a parallelogram is equal to the sum of the squares of all four sides.

Parallelogram Equality

Example 6.12. Use properties of inner products to show that if we use the induced norm, then

$$\|\mathbf{u}+\mathbf{v}\|^2+\|\mathbf{u}-\mathbf{v}\|^2=2\left(\|\mathbf{u}\|^2+\|\mathbf{v}\|^2\right).$$

Solution. The key to proving this fact is to relate induced norm to inner product. Specifically,

$$\|\mathbf{u}+\mathbf{v}\|^2=\langle\mathbf{u}+\mathbf{v},\mathbf{u}+\mathbf{v}\rangle=\langle\mathbf{u},\mathbf{u}\rangle+\langle\mathbf{u},\mathbf{v}\rangle+\langle\mathbf{v},\mathbf{u}\rangle+\langle\mathbf{v},\mathbf{v}\rangle,$$

while

$$\|\mathbf{u}-\mathbf{v}\|^2=\langle\mathbf{u}-\mathbf{v},\mathbf{u}-\mathbf{v}\rangle=\langle\mathbf{u},\mathbf{u}\rangle-\langle\mathbf{u},\mathbf{v}\rangle-\langle\mathbf{v},\mathbf{u}\rangle+\langle\mathbf{v},\mathbf{v}\rangle.$$

Now add these two equations and obtain by using the definition of induced norm again that

$$\|\mathbf{u}+\mathbf{v}\|^2+\|\mathbf{u}-\mathbf{v}\|^2=2\langle\mathbf{u},\mathbf{u}\rangle+2\langle\mathbf{v},\mathbf{v}\rangle=2\left(\|\mathbf{u}\|^2+\|\mathbf{v}\|^2\right),$$

which is what was to be shown. □

It would be nice to think that every norm on a vector space is induced from some inner product. Unfortunately, this is not true, as the following example shows.

Example 6.13. Use the result of Example 6.12 to show that the infinity norm on $V=\mathbb{R}^2$ is not induced by any inner product on V.

Solution. Suppose the infinity norm were induced by some inner product on V. Let $\mathbf{u}=(1,0)$ and $\mathbf{v}=(0,1/2)$. Then we have

$$\|\mathbf{u}+\mathbf{v}\|_\infty^2+\|\mathbf{u}-\mathbf{v}\|_\infty^2=\|(1,1/2)\|_\infty^2+\|(1,-1/2)\|_\infty^2=2,$$

while

$$2\left(\|\mathbf{u}\|^2+\|\mathbf{v}\|^2\right)=2\left(1+1/4\right)=5/2.$$

This contradicts the parallelogram equality of Example 6.12, so that the infinity norm cannot be induced from an inner product. □

One last example of a geometrical idea that generalizes to inner product spaces is the notion of projections of one vector along another. The projection formula for vectors in Section 4.2 works perfectly well for general inner products. Since the proof of this fact amounts to replacing dot products by inner products in the original formulation of the theorem (see Theorem 4.3), we omit it and simply state the result.

Theorem 6.4. Projection Formula for Vectors Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in an inner product space with $\mathbf{v} \neq \mathbf{0}$. Define the projection of $\mathbf{u}$ along $\mathbf{v}$ as

$$\text{proj}_{\mathbf{v}}\,\mathbf{u} = \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}\mathbf{v}$$

and let $\mathbf{p} = \text{proj}_{\mathbf{v}}\,\mathbf{u}$, $\mathbf{q} = \mathbf{u} - \mathbf{p}$. Then $\mathbf{p}$ is parallel to $\mathbf{v}$, $\mathbf{q}$ is orthogonal to $\mathbf{v}$, and $\mathbf{u} = \mathbf{p} + \mathbf{q}$.

As with the standard inner product, it is customary to call the vector $\text{proj}_{\mathbf{v}}\,\mathbf{u}$ of this theorem the *(parallel) projection of* $\mathbf{u}$ *along* $\mathbf{v}$. In summary, we have the two vector and one scalar quantities:

Projection

$$\text{proj}_{\mathbf{v}}\,\mathbf{u} = \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}\mathbf{v},$$

Orthogonal Projection

$$\text{orth}_{\mathbf{v}}\,\mathbf{u} = \mathbf{u} - \text{proj}_{\mathbf{v}}\,\mathbf{u},$$

Component

$$\text{comp}_{\mathbf{v}}\,\mathbf{u} = \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\|\mathbf{v}\|}.$$

Orthogonal Sets of Vectors

We have already seen the development of the ideas of orthogonal sets of vectors and bases in Chapter 4. Much of this development can be abstracted easily to general inner product spaces, simply by replacing dot products by inner products. Accordingly, we can make the following definition.

Definition 6.11. Orthogonal and Orthonormal Set of Vectors The set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ in an inner product space is said to be an *orthogonal set* if $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ whenever $i \neq j$. If, in addition, each vector has unit length, i.e., $\langle \mathbf{v}_i, \mathbf{v}_i \rangle = 1$ for all i, then the set of vectors is said to be an *orthonormal set* of vectors.

The proof of the following key fact and its corollary are the same as those of Theorem 4.6 in Section 4.3. All we have to do is replace dot products by inner products. The observations that followed the proof of this theorem are valid for general inner products as well. Again we omit the proofs and refer the reader to Chapter 4.

Theorem 6.5. Orthogonal Coordinates Formula Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be an orthogonal set of nonzero vectors and suppose that $\mathbf{v} \in \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Then $\mathbf{v}$ can be expressed uniquely (up to order) as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, namely

$$\mathbf{v} = \frac{\langle \mathbf{v}_1, \mathbf{v} \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 + \frac{\langle \mathbf{v}_2, \mathbf{v} \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2 + \cdots + \frac{\langle \mathbf{v}_n, \mathbf{v} \rangle}{\langle \mathbf{v}_n, \mathbf{v}_n \rangle} \mathbf{v}_n.$$

Some useful corollaries, whose proofs are left as exercises:

Corollary 6.1. Every orthogonal set of nonzero vectors is linearly independent.

Corollary 6.2. If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is an orthogonal set of vectors and $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$, then

$$\|\mathbf{v}\|^2 = c_1^2 \|\mathbf{v}_1\|^2 + c_2^2 \|\mathbf{v}_2\|^2 + \cdots + c_n^2 \|\mathbf{v}_n\|^2 .$$

Example 6.14. Find an orthogonal basis of $V = \mathbb{R}^2$ with respect to the inner product of Example 6.7 that includes $\mathbf{v}_1 = (1, -1)$. Calculate the coordinates of $\mathbf{v} = (1, 1)$ with respect to this basis and verify the formula of Corollary 6.2.

Solution. Recall that the inner product is given by $\langle \mathbf{u}, \mathbf{v} \rangle = 2u_1v_1 + 3u_2v_2$. Use the induced norm for $\|\cdot\|$. Let $\mathbf{w}$ be a nonzero solution to the equation

$$0 = \langle \mathbf{v}_1, \mathbf{w} \rangle = 2 \cdot 1 \cdot w_1 + 3 \cdot (-1)\, w_2,$$

say $\mathbf{w} = (3, 2)$. Then $\mathbf{v}_1$ and $\mathbf{v}_2 = \mathbf{w}$ are orthogonal, hence linearly independent and a basis of the two-dimensional space V. Now $\|\mathbf{v}_1\|^2 = 2 \cdot 1^2 + 3 \cdot (-1)^2 = 5$ and $\|\mathbf{v}_2\|^2 = 2 \cdot 3^2 + 3 \cdot 2^2 = 30$. The coordinates of $\mathbf{v}$ are easily calculated:

$$c_1 = \frac{\langle \mathbf{v}_1, \mathbf{v} \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} = \frac{1}{5}\left(2 \cdot 1 \cdot 1 + 3(-1) \cdot 1\right) = \frac{-1}{5}$$

$$c_2 = \frac{\langle \mathbf{v}_2, \mathbf{v} \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} = \frac{1}{30}\left(2 \cdot 3 \cdot 1 + 3 \cdot 2 \cdot 1\right) = \frac{2}{5}.$$

From the definition we have that $\|\mathbf{v}\|^2 = 2 \cdot 1^2 + 3 \cdot 1^2 = 5$. Similarly, we calculate that

$$c_1^2 \|\mathbf{v}_1\|^2 + c_2^2 \|\mathbf{v}_2\|^2 = \left(\frac{-1}{5}\right)^2 5 + \left(\frac{2}{5}\right)^2 30 = 5 = \|\mathbf{v}\|^2 . \qquad \square$$

Note 6.1. In all exercises of this chapter except those of Section 6.6, use the standard inner products and induced norms for $\mathbb{R}^n$ and $C[a, b]$ unless otherwise specified.

6.2 Exercises and Problems

Exercise 1. Verify the Cauchy–Bunyakovsky–Schwarz inequality and calculate the angle between the vectors for the following pairs of vectors $\mathbf{u}$, $\mathbf{v}$ and specified inner product.
(a) $\mathbf{u} = (2, 3)$, $\mathbf{v} = (-1, 2)$, inner product $\langle (x, y), (w, z) \rangle = 4xw + 9yz$ on $\mathbb{R}^2$.
(b) $\mathbf{u} = x$, $\mathbf{v} = x^3$, inner product of Example 6.9 on $C[0, 1]$.

Exercise 2. Verify the CBS inequality and calculate the inner product and angle between the vectors for the following pairs of vectors $\mathbf{u}, \mathbf{v}$.
(a) $(1, -1, 1)$, $(-1, 2, 3)$, inner product $\langle (x, y, z), (u, v, w) \rangle = xu + 2yv + zw$.
(b) $(2, 3)$, $(-1, 2)$, inner product $\langle (x, y), (w, z) \rangle = 2xw + xz + yw + yz$.

Exercise 3. For each of the pairs $\mathbf{u}, \mathbf{v}$ of vectors in Exercise 1, calculate the projection, component, and orthogonal projection of $\mathbf{u}$ to $\mathbf{v}$ using the specified inner product.

Exercise 4. For each of the pairs $\mathbf{u}, \mathbf{v}$ of vectors in Exercise 2, calculate the projection, component, and orthogonal projection of $\mathbf{u}$ to $\mathbf{v}$ using the specified inner product.

Exercise 5. Find an equation for the hyperplane defined by $\langle \mathbf{a}, \mathbf{x} \rangle = 2$ in $\mathbb{R}^3$ with inner product of Exercise 2(a) and $\mathbf{a} = (4, -1, 2)$.

Exercise 6. Find an equation for the hyperplane defined by $\langle f, g \rangle = 2$ in $\mathcal{P}_3$ with the standard inner product of $C[0, 1]$, $f(x) = x + 3$, and $g(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$.

Exercise 7. The formula $\left\langle [x_1, x_2]^T, [y_1, y_2]^T \right\rangle = 3x_1 y_1 - 2x_2 y_2$ fails to define an inner product on $\mathbb{R}^2$. What laws fail?

Exercise 8. Do any inner product laws fail for the formula $\langle (x_1, x_2), (y_1, y_2) \rangle = x_1 y_1 - x_1 y_2 - x_2 y_1 + 2x_2 y_2$ on $\mathbb{R}^2$. (*Hint:* $\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$.)

Exercise 9. Which of the following are orthogonal or orthonormal sets?
(a) $(2, -1, 2)$, $(2, 2, 0)$ in $\mathbb{R}^3$ with the inner product of Exercise 2(a).
(b) 1, x, x^2 as vectors in $C[-1, 1]$ with the standard inner product.
(c) $\frac{1}{5}(-2, 1)$, $\frac{1}{30}(9, 8)$ in $\mathbb{R}^2$ with the inner product of Exercise 1(a).

Exercise 10. Determine whether the following sets of vectors are linearly independent, orthogonal, or orthonormal.
(a) $\frac{1}{10}(3, 4)$, $\frac{1}{10}(4, -3)$ in $\mathbb{R}^2$ with inner product $\langle (x, y), (w, z) \rangle = 4xw + 4yz$.
(b) 1, $\cos(x)$, $\sin(x)$ in $C[-\pi, \pi]$ with the standard inner product.
(c) $(2, 4)$, $(1, 0)$ in $\mathbb{R}^2$ with inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \mathbf{y}$.

Exercise 11. Let $\mathbf{v}_1 = (1, 3, 2)$ and $\mathbf{v}_2 = (-4, 1, -1)$. Show that $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal with respect to the inner product of Exercise 2(a) and use this to determine whether the following vectors $\mathbf{v}$ belong to $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ by checking whether Theorem 6.5 is satisfied.

(a) $(11, 7, 8)$ (b) $(5, 1, 3)$ (c) $(5, 2, 3)$

Exercise 12. Confirm that $p_1(x) = x$ and $p_2(x) = 3x^2 - 1$ are orthogonal elements of $C[-1, 1]$ with the standard inner product and determine whether the following polynomials belong to $\text{span}\{p_1(x), p_2(x)\}$ using Theorem 6.5.

(a) x^2 (b) $1 + x - 3x^2$ (c) $1 + 3x - 3x^2$

Exercise 13. Let $\mathbf{v}_1 = (1, 0, 0)$, $\mathbf{v}_2 = (-1, 2, 0)$, $\mathbf{v}_3 = (1, -2, 3)$. Let $V = \mathbb{R}^3$ with inner product defined by the formula $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T A \mathbf{y}$, where $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$. Verify that $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ form an orthogonal basis of V and find the coordinates of the following vectors with respect to this basis.

(a) $(3, 1, 1)$ (b) $(0, 0, 1)$ (c) $(0, 2, 0)$

Exercise 14. Let $\mathbf{v}_1 = (1, 3, 2)$, $\mathbf{v}_2 = (-4, 1, -1)$, and $\mathbf{v}_3 = (10, 7, -26)$. Verify that $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ form an orthogonal basis of $\mathbb{R}^3$ with the inner product of Exercise 2(a). Convert this basis to an orthonormal basis and find the coordinates of the following vectors with respect to this orthonormal basis.

(a) $(1, 1, 0)$ (b) $(2, 1, 1)$ (c) $(0, 2, 2)$ (d) $(0, 0, -1)$

Exercise 15. Let $\mathbf{x} = (a, b)$ and $\mathbf{y} = (c, d)$. Let $V = \mathbb{R}^2$ with inner product defined by the formula $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T A \mathbf{y}$, where $A = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}$. Calculate a formula for $\langle \mathbf{x}, \mathbf{y} \rangle$ in terms of coordinates a, b, c, d.

Exercise 16. Let $f(x) = a + bx$ and $g(x) = c + dx$. Let $V = \mathcal{P}_1$, the space of linear polynomials, with the standard function space inner product in $C[0, 1]$. Calculate a formula for $\langle f, g \rangle$ in terms of coordinates a, b, c, d. Compare with Exercise 15. Conclusions?

*Problem 17. Show that any inner product on $\mathbb{R}^2$ can be expressed as $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T A \mathbf{v}$ for some symmetric positive definite matrix A.

*Problem 18. Show that $\|\cdot\|_1$ is not an induced norm on $\mathbb{R}^2$.

*Problem 19. Let $V = \mathbb{R}^n$ or $\mathbb{C}^n$ and let $\mathbf{u}, \mathbf{v} \in V$. Let A be a fixed $n \times n$ *nonsingular* matrix. Show that the matrix A defines an inner product by the formula $\langle \mathbf{u}, \mathbf{v} \rangle = (A\mathbf{u})^* A\mathbf{v}$.

*Problem 20. Prove Theorem 6.1.

Problem 21. Prove Corollary 6.2.

*Problem 22. Let V be a real inner product space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Prove the *polarization identity*, which recovers the inner product from its induced norm:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{4} \left\{ \|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2 \right\}.$$

*Problem 23. Let $V = C^1[0,1]$, the space of continuous functions with a continuous derivative on the interval $[0,1]$ (see Exercise 24 of Section 3.2). Show that the formula

$$\langle f, g \rangle = \int_0^1 f'(x) g'(x) dx + \int_0^1 f(x) g(x) dx$$

defines an inner product on V (called the *Sobolev* inner product).

Problem 24. Let $V = C[0,1]$, the space of continuous functions on the interval $[0,1]$, and define a candidate for inner product on V by this formula: For $f, g \in V$ $\langle f, g \rangle = \max_{0 \le x \le 1} f(x)\, g(x)$. Does this define an inner product on V? If so, prove it and if not, show which parts of the inner product definition fail.

6.3 Orthogonal Vectors and Projection

We have seen that orthogonal bases have some very pleasant properties, such as easy coordinate calculations. In this section we generalize the Gram–Schmidt algorithm to arbitrary inner product spaces. In fact the proof of this version of the Gram–Schmidt algorithm is exactly the same as the one given in Theorem 4.10: simply replace any occurrence of $\mathbf{v} \cdot \mathbf{w}$ with $\langle \mathbf{v}, \mathbf{w} \rangle$.

Description of the Algorithm

Theorem 6.6. **Gram–Schmidt Algorithm** Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ be linearly independent vectors in the inner product space V. Define vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ recursively by the formula

$$\mathbf{v}_k = \mathbf{w}_k - \frac{\langle \mathbf{v}_1, \mathbf{w}_k \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \frac{\langle \mathbf{v}_2, \mathbf{w}_k \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2 - \cdots - \frac{\langle \mathbf{v}_{k-1}, \mathbf{w}_k \rangle}{\langle \mathbf{v}_{k-1}, \mathbf{v}_{k-1} \rangle} \mathbf{v}_{k-1}, \quad k = 1, \ldots, n.$$

Then

(1) The vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ form an orthogonal set.
(2) For each index $k = 1, \ldots, n$,

$$\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k\} = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}.$$

The Gram–Schmidt formula is conceptually simple: subtract from the vector $\mathbf{w}_k$ all of the projections of $\mathbf{w}_k$ along the directions $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}$ to obtain the vector $\mathbf{v}_k$.

Example 6.15. Let $C[0,1]$ be the space of continuous functions on the interval $[0,1]$ with the usual function addition and scalar multiplication, and (standard) inner product given by

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx$$

as in Example 6.9. Let $V = \mathcal{P}_2 = \text{span}\{1, x, x^2\}$ and apply the Gram–Schmidt algorithm to the basis $1, x, x^2$ to obtain an orthogonal basis for the space of quadratic polynomials.

Solution. Set $\mathbf{w}_1 = 1$, $\mathbf{w}_2 = x$, $\mathbf{w}_3 = x^2$ and calculate the Gram–Schmidt formulas:

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{w}_1 = 1,\\
\mathbf{v}_2 &= \mathbf{w}_2 - \frac{\langle \mathbf{v}_1, \mathbf{w}_2\rangle}{\langle \mathbf{v}_1, \mathbf{v}_1\rangle}\mathbf{v}_1 = x - \frac{1/2}{1}1 = x - \frac{1}{2},\\
\mathbf{v}_3 &= \mathbf{w}_3 - \frac{\langle \mathbf{v}_1, \mathbf{w}_3\rangle}{\langle \mathbf{v}_1, \mathbf{v}_1\rangle}\mathbf{v}_1 - \frac{\langle \mathbf{v}_2, \mathbf{w}_3\rangle}{\langle \mathbf{v}_2, \mathbf{v}_2\rangle}\mathbf{v}_2\\
&= x^2 - \frac{1/3}{1}1 - \frac{1/12}{1/12}(x - \frac{1}{2}) = x^2 - x + \frac{1}{6}.
\end{aligned}$$

□

Had we used $C[-1,1]$ and "normalized" the resulting polynomials by requiring that they have value 1 at $x = 1$, the same calculations would have given us the first of three well-known functions called *Legendre polynomials*: $P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}\left(3x^2 - 1\right)$. These polynomials are used extensively in approximation theory and applied mathematics.

Legendre Polynomials

As usual, if we prefer to have an orthonormal basis rather than an orthogonal basis, then, as a final step in the orthogonalizing process, simply replace each vector $\mathbf{v}_k$ by the normalized vector $\mathbf{u}_k = \mathbf{v}_k / \|\mathbf{v}_k\|$.

Application to Projections

We can use the machinery of orthogonal vectors to give a nice solution to a very practical and important question that can be phrased as follows (see Figure 6.3 for a graphical interpretation of it):

The Projection Problem: Given a finite-dimensional subspace V of a real inner product space W, together with a vector $\mathbf{b} \in W$, to find the vector $\mathbf{v} \in V$ which is closest to $\mathbf{b}$ in the sense that $\|\mathbf{b} - \mathbf{v}\|^2$ is minimized.

Observe that the quantity $\|\mathbf{b}-\mathbf{v}\|^2$ will be minimized exactly when $\|\mathbf{b}-\mathbf{v}\|$ is minimized, since the latter is always nonnegative. The squared term has the virtue of avoiding square roots that computing sometimes $\|\mathbf{b}-\mathbf{v}\|$ requires.

The projection problem looks vaguely familiar. It reminds us of the least squares problem of Chapter 4, which was to minimize the quantity $\|\mathbf{b}-A\mathbf{x}\|^2$, where A is an $m \times n$ real matrix and $\mathbf{b}, \mathbf{x}$ are standard vectors. Recall that $\mathbf{v} = A\mathbf{x}$ is a typical element in the column space of A. Therefore, the quantity to be minimized is

$$\|\mathbf{b}-A\mathbf{x}\|^2 = \|\mathbf{b}-\mathbf{v}\|^2,$$

where on the left-hand side $\mathbf{x}$ runs over all standard n-vectors and on the right-hand side $\mathbf{v}$ runs over all vectors in the space $V = \mathcal{C}(A)$. The difference between least squares and the projection problem is this: In the least squares problem we want to know the vector $\mathbf{x}$ of coefficients of $\mathbf{v}$ as a linear combination of columns of A, whereas in the projection problem we are interested only in $\mathbf{v}$. Knowing $\mathbf{v}$ doesn't tell us what $\mathbf{x}$ is, but knowing $\mathbf{x}$ easily gives $\mathbf{v}$ since $\mathbf{v} = A\mathbf{x}$.

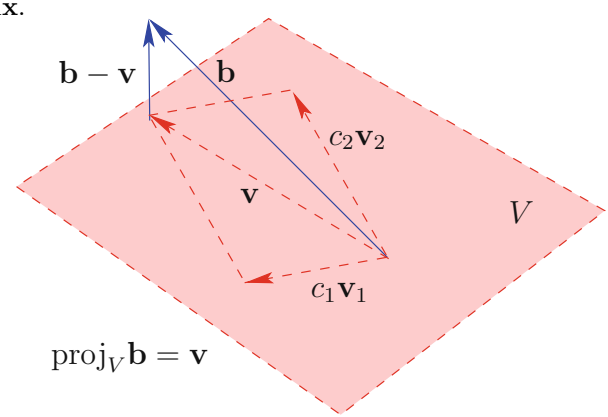


Fig. 6.3: Projection $\mathbf{v}$ of $\mathbf{b}$ into the subspace V spanned by the orthogonal vectors $\mathbf{v}_1, \mathbf{v}_2$.

To solve the projection problem we need the following key concept.

Definition 6.12. Projection Formula for Subspaces Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be an orthogonal basis for the subspace V of the inner product space W. For any $\mathbf{b} \in \mathbf{W}$, the *(parallel) projection of* $\mathbf{b}$ *into the subspace* V is the vector

$$\operatorname{proj}_V \mathbf{b} = \frac{\langle \mathbf{v}_1, \mathbf{b}\rangle}{\langle \mathbf{v}_1, \mathbf{v}_1\rangle}\mathbf{v}_1 + \frac{\langle \mathbf{v}_2, \mathbf{b}\rangle}{\langle \mathbf{v}_2, \mathbf{v}_2\rangle}\mathbf{v}_2 + \cdots + \frac{\langle \mathbf{v}_n, \mathbf{b}\rangle}{\langle \mathbf{v}_n, \mathbf{v}_n\rangle}\mathbf{v}_n.$$

Notice that in the case of $n = 1$ the definition amounts to a familiar friend, the projection of $\mathbf{b}$ along the vector $\mathbf{v}_1$.

It appears that the definition of $\operatorname{proj}_V \mathbf{b}$ depends on the basis vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, but we see from the next theorem that this is not the case.

Theorem 6.7. Projection Theorem Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be an orthogonal basis for the subspace V of the inner product space W. For any $\mathbf{b} \in \mathbf{W}$, the vector $\mathbf{v} = \text{proj}_V \mathbf{b}$ is the unique vector in V that minimizes $\|\mathbf{b} - \mathbf{v}\|^2$.

Proof. Let $\mathbf{v}$ be a solution to the projection problem and $\mathbf{p}$ the projection of $\mathbf{b} - \mathbf{v}$ to any vector in V. Use the Pythagorean theorem to obtain that

$$\|\mathbf{b} - \mathbf{v}\|^2 = \|\mathbf{b} - \mathbf{v} - \mathbf{p}\|^2 + \|\mathbf{p}\|^2 .$$

However, $\mathbf{v} + \mathbf{p} \in V$, so that $\|\mathbf{b} - \mathbf{v}\|$ cannot be the minimum distance from $\mathbf{b}$ to a vector in V unless $\|\mathbf{p}\| = 0$. It follows that $\mathbf{b} - \mathbf{v}$ is orthogonal to any vector in V. Now let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be an *orthogonal* basis of V and express the vector $\mathbf{v}$ in the form

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n.$$

Then for each $\mathbf{v}_k$ we must have

$$\begin{aligned} 0 = \langle \mathbf{v}_k, \mathbf{b} - \mathbf{v} \rangle &= \langle \mathbf{v}_k, \mathbf{b} - c_1\mathbf{v}_1 - c_2\mathbf{v}_2 - \cdots - c_n\mathbf{v}_n \rangle \\ &= \langle \mathbf{v}_k, \mathbf{b} \rangle - c_1 \langle \mathbf{v}_k, \mathbf{v}_1 \rangle - c_2 \langle \mathbf{v}_k, \mathbf{v}_2 \rangle - \cdots c_n \langle \mathbf{v}_k, \mathbf{v}_n \rangle \\ &= \langle \mathbf{v}_k, \mathbf{b} \rangle - c_k \langle \mathbf{v}_k, \mathbf{v}_k \rangle , \end{aligned}$$

from which we deduce that $c_k = \langle \mathbf{v}_k, \mathbf{b} \rangle / \langle \mathbf{v}_k, \mathbf{v}_k \rangle$. It follows that

$$\mathbf{v} = \frac{\langle \mathbf{v}_1, \mathbf{b} \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 + \frac{\langle \mathbf{v}_2, \mathbf{b} \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle}\mathbf{v}_2 + \cdots + \frac{\langle \mathbf{v}_n, \mathbf{b} \rangle}{\langle \mathbf{v}_n, \mathbf{v}_n \rangle}\mathbf{v}_n = \text{proj}_V \mathbf{b}.$$

This proves that there can be only one solution to the projection problem, namely the one given by the projection formula above. To finish the proof one has to show that $\text{proj}_V \mathbf{b}$ actually solves the projection problem. This is left to the exercises. □

The projection has the same nice properties that we observed in the case of standard inner products, namely, $\mathbf{p} = \text{proj}_V \mathbf{b} \in V$ and $\mathbf{b} - \mathbf{p}$ is orthogonal to every $\mathbf{v} \in V$. For the latter assertion, notice that for any j,

$$\langle \mathbf{v}_j, \mathbf{b} - \mathbf{p} \rangle = \langle \mathbf{v}_j, \mathbf{b} \rangle - \sum_{k=1}^{n} \left\langle \mathbf{v}_j, \frac{\langle \mathbf{v}_k, \mathbf{b} \rangle}{\langle \mathbf{v}_k, \mathbf{v}_k \rangle}\mathbf{v}_k \right\rangle = \langle \mathbf{v}_j, \mathbf{b} \rangle - \frac{\langle \mathbf{v}_j, \mathbf{v}_j \rangle}{\langle \mathbf{v}_j, \mathbf{v}_j \rangle} \langle \mathbf{v}_j, \mathbf{b} \rangle = 0.$$

One checks that the same is true if $\mathbf{v}_j$ is replaced by a $\mathbf{v} \in V$. In analogy with the standard inner products, we define the *orthogonal projection of* $\mathbf{b}$ to V by the formula

$$\text{orth}_V \mathbf{b} = \mathbf{b} - \text{proj}_V \mathbf{b}.$$

Orthogonal Projection

Let's specialize to standard real vectors and inner products and take a closer look at the formula for the projection operator in the case that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is an orthonormal set. We then have $\langle \mathbf{v}_j, \mathbf{v}_j \rangle = 1$, so

$$\begin{aligned}\operatorname{proj}_V \mathbf{b} &= \langle \mathbf{v}_1, \mathbf{b}\rangle \mathbf{v}_1 + \langle \mathbf{v}_2, \mathbf{b}\rangle \mathbf{v}_2 + \cdots + \langle \mathbf{v}_n, \mathbf{b}\rangle \mathbf{v}_n \\ &= \left(\mathbf{v}_1^T \mathbf{b}\right) \mathbf{v}_1 + \left(\mathbf{v}_2^T \mathbf{b}\right) \mathbf{v}_2 + \cdots + \left(\mathbf{v}_n^T \mathbf{b}\right) \mathbf{v}_n \\ &= \mathbf{v}_1 \mathbf{v}_1^T \mathbf{b} + \mathbf{v}_2 \mathbf{v}_2^T \mathbf{b} + \cdots + \mathbf{v}_n \mathbf{v}_n^T \mathbf{b} \\ &= \left(\mathbf{v}_1 \mathbf{v}_1^T + \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \mathbf{v}_n \mathbf{v}_n^T\right) \mathbf{b} = P\mathbf{b},\end{aligned}$$

where the matrix P is defined as

Projection Matrix Formula

$$P = \mathbf{v}_1 \mathbf{v}_1^T + \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \mathbf{v}_n \mathbf{v}_n^T.$$

The significance of this expression for projections in standard spaces over the reals with the standard inner product is as follows: computing the projection of a vector into a subspace amounts to multiplying the vector by a matrix P that can be computed from V. Even in the one-dimensional case this gives us a new slant on projections:

$$\operatorname{proj}_V \mathbf{u} = (\mathbf{v}\mathbf{v}^T)\mathbf{u} = P\mathbf{u}.$$

Similarly, we see that the orthogonal projection has a matrix representation

$$\operatorname{orth}_V \mathbf{u} = \mathbf{u} - P\mathbf{u} = (I - P)\,\mathbf{u}.$$

The general projection matrix P has some interesting properties. It is symmetric, i.e., $P^T = P$, and idempotent, i.e., $P^2 = P$. Therefore, this notation is compatible with the definition of projection matrix introduced in earlier exercises (see Exercise 11 of Section 4.3). Symmetry follows from the fact that $\left(\mathbf{v}_k \mathbf{v}_k^T\right)^T = \mathbf{v}_k \mathbf{v}_k^T$. For idempotence, notice that

$$\left(\mathbf{v}_j \mathbf{v}_j^T\right)\left(\mathbf{v}_k \mathbf{v}_k^T\right) = \left(\mathbf{v}_j^T \mathbf{v}_k\right)\left(\mathbf{v}_k \mathbf{v}_j^T\right) = \delta_{j,k} \mathbf{v}_k \mathbf{v}_j^T.$$

It follows that $P^2 = P$. One can show that the converse is true: if P is real symmetric and idempotent, then it is the projection matrix for the subspace $\mathcal{C}(P)$ (see Problem 14 at the end of this section.)

Example 6.16. Find the projection matrix for the subspace of $\mathbb{R}^3$ (with the standard inner product) spanned by the orthonormal vectors $\mathbf{v}_1 = (1/\sqrt{2})[1, -1, 0]^T$ and $\mathbf{v}_2 = (1/\sqrt{3})[1, 1, 1]^T$ and use it to solve the projection problem with $V = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\mathbf{b} = [2, 1, -3]^T$.

Solution. Use the formula developed above for the projection matrix

$$P = \mathbf{v}_1 \mathbf{v}_1^T + \mathbf{v}_2 \mathbf{v}_2^T = \frac{1}{2}\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}\begin{bmatrix} 1 & -1 & 0 \end{bmatrix} + \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 5 & -1 & 2 \\ -1 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix}.$$

Thus, the solution to the projection problem for $\mathbf{b}$ is

$$\mathbf{v} = P\mathbf{b} = \frac{1}{6}\begin{bmatrix} 5 & -1 & 2 \\ -1 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{bmatrix}. \qquad \square$$

The projection problem is closely related to another problem that we have seen before, namely the least squares problem of Section 4.2 in Chapter 4. Recall that the least squares problem amounted to minimizing the function $f(x) = \|\mathbf{b} - A\mathbf{x}\|^2$, which in turn led to the normal equations. Here A is an $m \times n$ real matrix. Now consider the projection problem for the subspace $V = \mathcal{C}(A)$ of $\mathbb{R}^m$, where $\mathbf{b} \in \mathbb{R}^m$. We know that elements of $\mathcal{C}(A)$ can be written in the form $\mathbf{v} = A\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^n$. Therefore, $\|\mathbf{b} - A\mathbf{x}\|^2 = \|\mathbf{b} - \mathbf{v}\|^2$, where $\mathbf{v}$ ranges over elements of V. It follows that when we solve a least squares problem, we are really solving a projection problem as well in the sense that the vector $A\mathbf{x}$ is the element of $\mathcal{C}(A)$ closest to the right-hand-side vector $\mathbf{b}$.

Least Squares as Projection Problem

The normal equations give us another way to generate projection matrices in the case of standard vectors and inner products. As above, let $V = \mathcal{C}(A) \subseteq \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^m$ and P the projection matrix for V. Assume that the columns of A are linearly independent, i.e., that A has full column rank. Then, as we have seen in Theorem 4.5, the matrix A^TA is invertible and the normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$ have the unique solution

$$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}.$$

Consequently, the solution to the projection problem is

$$\mathbf{v} = A\mathbf{x} = A(A^TA)^{-1}A^T\mathbf{b} = P\mathbf{b}.$$

Since this holds for all vectors $\mathbf{b}$, it follows that the projection matrix for this subspace is given by the formula

Column Space Projection Formula

$$P = A(A^TA)^{-1}A^T.$$

Example 6.17. Find the projection matrix for the subspace $V = \text{span}\{\mathbf{w}_1, \mathbf{w}_2\}$ of $\mathbb{R}^3$ with $\mathbf{w}_1 = (1, -1, 0)$ and $\mathbf{w}_2 = (2, 0, 1)$.

Solution. Let $A = [\mathbf{w}_1, \mathbf{w}_2]$, so that $A^TA = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix}$.

Thus

$$P = A(A^TA)^{-1}A^T = \begin{bmatrix} 1 & 2 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}\frac{1}{6}\begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} 1 & -1 & 0 \\ 2 & 0 & 1 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 5 & -1 & 2 \\ -1 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix}. \qquad \square$$

Curiously, this is exactly the same matrix as the projection matrix found in the preceding example. What is the explanation? Notice that $\mathbf{w}_1 = \sqrt{2}\mathbf{v}_1$ and $\mathbf{w}_2 = \sqrt{2}\mathbf{v}_1 + \sqrt{3}\mathbf{v}_2$, so that $V = \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2\} = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$. Hence, the subspaces of both examples, though specified by different bases, are the same subspace and we should expect the projection operators to be the same.

6.3 Exercises and Problems

Exercise 1. Find the projection matrix for the column space of each of the following matrices using the projection matrix formula (you will need an orthonormal basis).

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

Exercise 2. Redo Exercise 1 using the column space projection formula (remember to use a matrix of full column rank for this formula, so you may have to discard columns).

Exercise 3. Let $V = \operatorname{span}\{(1,-1,1),(1,1,0)\} \subseteq \mathbb{R}^3$. Compute $\operatorname{proj}_V \mathbf{w}$ and $\operatorname{orth}_V \mathbf{w}$ for the following $\mathbf{w}$.
(a) $(4,-1,2)$ (b) $(1,1,1)$ (c) $(0,0,1)$

Exercise 4. Repeat Exercise 3 using the inner product $\langle (x,y,z),(u,v,w)\rangle = 2xu - xv - yu + 3yv + zw$.

Exercise 5. Find the projection of the polynomial $f(x) = x^3$ into the subspace $V = \operatorname{span}\{1, x\}$ of $C[0,1]$ with the standard inner product and calculate $\|f - \operatorname{proj}_V f\|$.

Exercise 6. Repeat Exercise 5 using the Sobolev inner product of Problem 23, Section 6.2.

Exercise 7. Use the Gram–Schmidt algorithm to expand the orthogonal vectors $\mathbf{w}_1 = (-1,1,1,-1)$ and $\mathbf{w}_2 = (1,1,1,1)$ to an orthogonal basis of $\mathbb{R}^4$ (you will need to supply additional vectors).

Exercise 8. Apply the Gram–Schmidt algorithm to the following vectors using the specified inner product:
(a) $(1,-2,0)$, $(0,1,1)$, $(1,0,2)$, inner product of Exercise 4.
(b) $(1,0,0)$, $(1,1,0)$, $(1,1,1)$, inner product of Exercise 13, Section 6.2.

Exercise 9. Show that the matrices $A = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 4 & 2 \\ 1 & 1 & 8 \end{bmatrix}$ and $B = \begin{bmatrix} 4 & 3 & 1 \\ 5 & 7 & 0 \\ 2 & -5 & 3 \end{bmatrix}$ have the same column space by computing the projection matrices into these column spaces.

Exercise 10. Use projection matrices to determine whether the row spaces of the matrices $A = \begin{bmatrix} 3 & -4 & 7 & 2 \\ 0 & 5 & -5 & -1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & -3 & 4 & 1 \\ 3 & 1 & 2 & 1 \end{bmatrix}$ are equal; if not, exhibit vectors in one space but not the other, if possible.

Problem 11. Show that if P is a projection matrix, then so is $I - P$.

*Problem 12. Show that if $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ is an orthonormal basis of $\mathbb{R}^3$, then $\mathbf{u}_1\mathbf{u}_1^T + \mathbf{u}_2\mathbf{u}_2^T + \mathbf{u}_3\mathbf{u}_3^T = I_3$.

Problem 13. Assume A has full column rank. Verify directly that if $P = A(A^TA)^{-1}A^T$, then P is symmetric and idempotent.

*Problem 14. Show that if P is an $n \times n$ projection matrix, then for every $\mathbf{v} \in \mathbb{R}^n$, $P\mathbf{v} \in \mathcal{C}(P)$ and $\mathbf{v} - P\mathbf{v}$ is orthogonal to every element of $\mathcal{C}(P)$.

Problem 15. Write out a proof of the Gram–Schmidt algorithm (Theorem 6.6) in the case that $n = 3$.

*Problem 16. Complete the proof of the Projection theorem (Theorem 6.7) by showing that $\text{proj}_V \mathbf{b}$ solves the projection problem.

Problem 17. How does the projection matrix formula on page 414 have to be changed if the vectors in question are complex? Illustrate your answer with the orthonormal vectors $\mathbf{v}_1 = ((1+\mathrm{i})/2, 0, (1+\mathrm{i})/2)$, $\mathbf{v}_2 = (0, 1, 0)$ in $\mathbb{C}^2$.

*Problem 18. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ be an orthonormal basis for the subspace V of the inner product space W, and $\mathbf{b} \in W$ and $\mathbf{p} = \text{proj}_V \mathbf{b}$. Show that

$$\|\mathbf{b}\|^2 = |\langle \mathbf{u}_1, \mathbf{b}\rangle|^2 + |\langle \mathbf{u}_2, \mathbf{b}\rangle|^2 + \cdots + |\langle \mathbf{u}_n, \mathbf{b}\rangle|^2 + \|\mathbf{b} - \mathbf{p}\|^2$$

Problem 19. Let $W = C[-1, 1]$ with the standard function space inner product. Suppose V is the subspace of linear polynomials and $\mathbf{b} = e^x$.

(a) Find an orthogonal basis for V.
(b) Find the projection $\mathbf{p}$ of $\mathbf{b}$ into V.
(c) Compute the "mean error of approximation" $\|\mathbf{b} - \mathbf{p}\|$, and compare it to the mean error of approximation $\|\mathbf{b} - \mathbf{q}\|$, where $\mathbf{q}$ is the first-degree Taylor series of $\mathbf{b}$ centered at 0.
(d) Use a technology tool to plot $\mathbf{b} - \mathbf{p}$ and $\mathbf{b} - \mathbf{q}$.

6.4 Linear Systems Revisited

Once again we revisit our old friend, $A\mathbf{x} = \mathbf{b}$, where A is an $m \times n$ matrix. The notions of orthogonality can shed still more light on the nature of this system of equations, especially in the case of a homogeneous system $A\mathbf{x} = \mathbf{0}$. The kth entry of the column vector $A\mathbf{x}$ is simply the kth row of A multiplied by the column vector $\mathbf{x}$. Designate this row by $\mathbf{r}_k^T$, and we see that

$$\mathbf{r}_k \cdot \mathbf{x} = 0, \qquad k = 1, \ldots, n.$$

In other words, $A\mathbf{x} = 0$, that is, $\mathbf{x} \in \mathcal{N}(A)$, precisely when $\mathbf{x}$ is orthogonal (with the standard inner product) to every row of A. We will see in Theorem 6.10 below that this means that $\mathbf{x}$ will be orthogonal to any linear combination of the rows of A. Thus, we could say

$$\mathcal{N}(A) = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{r} \cdot \mathbf{x} = 0 \text{ for every } \mathbf{r} \in \mathcal{R}(A)\}\,. \tag{6.2}$$

We are going to digress and put this equation in a more general context. Then we will return to linear systems with a new perspective on their meaning.

Orthogonal Complements and Homogeneous Systems

Definition 6.13. Orthogonal Complement Let V be a subspace of an inner product space W. Then the *orthogonal complement* of V in W is the set

$$V^\perp = \{\mathbf{w} \in W \,|\, \langle \mathbf{v}, \mathbf{w} \rangle = 0 \text{ for all } \mathbf{v} \in V\}\,.$$

We can see from the subspace test that $V^\perp$ is a subspace of W. Recall that if U and V are two subspaces of the vector space W, then two other subspaces that we can construct are the *intersection* and *sum* of these subspaces. The former is just the set intersection of the two subspaces, and the latter is the set of elements of the form $\mathbf{u} + \mathbf{v}$, where $\mathbf{u} \in U$ and $\mathbf{v} \in V$. One can use the subspace test to verify that these are indeed subspaces of W (see Problem 19 of Section 3.2). In fact, it isn't too hard to see that $U + V$ is the smallest space containing all elements of both U and V. Basic facts about the orthogonal complement of V are summarized as follows.

Theorem 6.8. Let V be a subspace of the finite-dimensional inner product space W. Then the following are true:

(1) $V^\perp$ is a subspace of W.
(2) $V \cap V^\perp = \{\mathbf{0}\}$.
(3) $V + V^\perp = W$.
(4) $\dim V + \dim V^\perp = \dim W$.
(5) $\left(V^\perp\right)^\perp = V$.

Proof. We leave (1) and (2) as exercises. To prove (3), we notice that $V + V^\perp \subseteq W$ since W is closed under sums. Now suppose that $\mathbf{w} \in W$. Let $\mathbf{v} = \text{proj}_V \mathbf{w}$. We know that $\mathbf{v} \in V$ and $\mathbf{w} - \mathbf{v}$ is orthogonal to every element of V. It follows that $\mathbf{w} - \mathbf{v} \in V^\perp$. Therefore, every element of W can be expressed as a sum of an element in V and an element in $V^\perp$. This shows that $W \subseteq V + V^\perp$, from which it follows that $V + V^\perp = W$.

To prove (4), let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ be a basis of V and $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s$ a basis of $V^\perp$. Certainly, the union of the two sets spans V because of (3). Now if there were an equation of linear dependence, we could gather all terms involving $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ on one side of the equation, those involving $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s$ on the other side, and deduce that each is equal to zero separately, in view of (2). It follows that the union of these two bases must be an independent set. Therefore, it forms a basis of W. It follows that $\dim W = r + s = \dim V + \dim V^\perp$.

Finally, apply (4) to $V^\perp$ in place of V and obtain that $\dim \left(V^\perp\right)^\perp = \dim W - \dim V^\perp$. But (4) implies directly that $\dim V = \dim W - \dim V^\perp$, so that $\dim \left(V^\perp\right)^\perp = \dim V$. Now if $\mathbf{v} \in V$, then $\langle \mathbf{w}, \mathbf{v} \rangle = 0$ for all $\mathbf{w} \in V^\perp$. Hence, $V \subseteq \left(V^\perp\right)^\perp$. Since these two spaces have the same dimension, they must be equal, which proves (5). □

Using the notation of Definition 3.15 we can summarize (1)–(3) as follows: $W = V \oplus V^\perp$. Orthogonal complements of the sum and intersections of two subspaces have an interesting relationship to each other, whose proofs we leave as exercises.

Theorem 6.9. Let U and V be subspaces of the inner product space W. Then

(1) $(U \cap V)^\perp = U^\perp + V^\perp$.
(2) $(U + V)^\perp = U^\perp \cap V^\perp$.

The following fact greatly simplifies the calculation of an orthogonal complement. It says that a vector is orthogonal to every element of a vector space if and only if it is orthogonal to every element of a spanning set of the space.

Theorem 6.10. Let $V = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a subspace of the inner product space W. Then

$$V^\perp = \{\mathbf{w} \in W \mid \langle \mathbf{w}, \mathbf{v}_j \rangle = 0,\ j = 1, 2, \ldots, n\}.$$

Proof. Let $\mathbf{v} \in V$, so that for some scalars $c_1, c_2, \ldots, c_n$,

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n.$$

Take the inner product of both sides with a vector $\mathbf{w}$. We see by the linearity of inner products that

$$\langle \mathbf{w}, \mathbf{v} \rangle = c_1\langle \mathbf{w}, \mathbf{v}_1 \rangle + c_2\langle \mathbf{w}, \mathbf{v}_2 \rangle + \cdots + c_n\langle \mathbf{w}, \mathbf{v}_n \rangle,$$

so that if $\langle \mathbf{w}, \mathbf{v}_j \rangle = 0$ for each j then $\langle \mathbf{w}, \mathbf{v} \rangle = 0$. Conversely, if $\langle \mathbf{w}, \mathbf{v}_j \rangle = 0$, for $j = 1, 2, \ldots, n$, then clearly $\langle \mathbf{w}, \mathbf{v}_j \rangle = 0$, which proves the theorem. □

Example 6.18. Compute $V^\perp$, where

$$V = \operatorname{span}\{(1,1,1,1),(1,2,1,0)\} \subseteq \mathbb{R}^4$$

with the standard inner product on $\mathbb{R}^4$.

Solution. Form the matrix A with the two spanning vectors of V as rows. According to Theorem 6.10, $V^\perp$ is the null space of this matrix. We have

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 0 \end{bmatrix} \xrightarrow{E_{21}(-1)} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & -1 \end{bmatrix} \xrightarrow{E_{12}(-1)} \begin{bmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & -1 \end{bmatrix},$$

from which it follows that the null space of A consists of vectors of the form

$$\begin{bmatrix} -x_3 - 2x_4 \\ x_4 \\ x_3 \\ x_4 \end{bmatrix} = x_3 \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -2 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

Therefore, $V^\perp = \operatorname{span}\{(-1,0,1,0),(-2,1,0,1)\}$. □

Nothing prevents us from considering more exotic inner products as well. The arithmetic may be a bit more complicated, but the underlying principles are the same. Here is such an example.

Example 6.19. Let $V = \operatorname{span}\{1, x\} \subset \mathcal{P}_2[0,1] = W$, where the space $\mathcal{P}_2[0,1]$ of polynomial functions of degree at most 2 on the interval $[0,1]$ has the standard inner product of $C[0,1]$. Compute $V^\perp$ and use this to verify that $\dim V + \dim V^\perp = \dim W$.

Solution. According to Theorem 6.10, $V^\perp$ consists of those polynomials $p(x) = c_0 + c_1 x + c_2 x^2$ for which

$$0 = \langle p, 1 \rangle = \int_0^1 \left(c_0 + c_1 x + c_2 x^2\right) 1 \, dx = c_0 \int_0^1 1 \, dx + c_1 \int_0^1 x \, dx + c_2 \int_0^1 x^2 \, dx,$$

$$0 = \langle p, x \rangle = \int_0^1 \left(c_0 + c_1 x + c_2 x^2\right) x \, dx = c_0 \int_0^1 x \, dx + c_1 \int_0^1 x^2 \, dx + c_2 \int_0^1 x^3 \, dx.$$

Integrate, and we obtain the system of equations

$$\begin{aligned} c_0 + \frac{1}{2} c_1 + \frac{1}{3} c_2 &= 0, \\ \frac{1}{2} c_0 + \frac{1}{3} c_1 + \frac{1}{4} c_2 &= 0. \end{aligned}$$

Solve this system to obtain $c_0 = \frac{1}{6} c_2$, $c_1 = -c_2$, and c_2 is free. Therefore, $V^\perp$ consists of polynomials of the form

$$p(x) = \frac{1}{6}c_2 - c_2 x + c_2 x^2 = c_2\left(\frac{1}{6} - x + x^2\right).$$

It follows that $V^\perp = \operatorname{span}\left\{\frac{1}{6} - x + x^2\right\}$ and $\dim V^\perp = 1$. Since $\{1, x\}$ is a basis of V, $\dim V = 2$. Therefore, $\dim V + \dim V^\perp = \dim W$. □

Finally, we return to solutions to the homogeneous system $A\mathbf{x} = 0$. We have seen that the null space of A consists of elements that are orthogonal to the rows of A. One could turn things around and ask what we can say about a vector that is orthogonal to every element of the null space of A. This question has a surprisingly simple answer. In fact, there is a fascinating interplay between row spaces, column spaces, and null spaces that can be summarized in the following theorem:

Theorem 6.11. Orthogonal Complements Theorem For a matrix A,

(1) $\mathcal{R}(A)^\perp = \mathcal{N}(A)$.
(2) $\mathcal{N}(A)^\perp = \mathcal{R}(A)$.
(3) $\mathcal{N}(A^T)^\perp = \mathcal{C}(A)$.

Proof. We have already seen item (1) in the discussion at the beginning of this section, where it was stated in equation (6.2). For item (2) we take orthogonal complements of both sides of (1) and use part (5) of Theorem 6.8 to obtain

$$\mathcal{N}(A)^\perp = \left(\mathcal{R}(A)^\perp\right)^\perp = \mathcal{R}(A),$$

which proves (2). Finally, for (3) we observe that $\mathcal{R}(A^T) = \mathcal{C}(A)$. Apply (2) with A^T in place of A and the result follows. □

The connections spelled out by this theorem are powerful ideas. Here is one example of how they can be used. Consider the following problem: suppose we are given subspaces U and V of the standard space $\mathbb{R}^n$ with the standard inner product (the dot product) in some concrete form, and we want to compute a basis for the subspace $U \cap V$. How do we proceed? One answer is to use part (1) of Theorem 6.9 to see that $(U \cap V)^\perp = U^\perp + V^\perp$. Now use part (5) of Theorem 6.8 to obtain that

$$U \cap V = (U \cap V)^{\perp\perp} = (U^\perp + V^\perp)^\perp.$$

The strategy that this equation suggests is this: Express U and V as row spaces of matrices and compute bases for the null spaces of each. Put these bases together to obtain a spanning set for $U^\perp + V^\perp$. Use this spanning set as the rows of a matrix B. Then the complement of this space is, on the one hand, $U \cap V$, but by part (1) of the orthogonal complements theorem, it is also $\mathcal{N}(B)$. Therefore, $U \cap V = \mathcal{N}(B)$, so all we have to do is calculate a basis for $\mathcal{N}(B)$, which we know how to do.

Example 6.20. Find a basis for $U \cap V$, where these subspaces of $\mathbb{R}^4$ are given as follows:

$$U = \text{span}\{(1,2,1,2),(0,1,0,1)\}$$
$$V = \text{span}\{(1,1,1,1),(1,2,1,0)\}.$$

Solution. We have already determined in Example 6.18 that $V^\perp$ has a basis $(-1,0,1,0)$ and $(-2,1,0,1)$. Form the matrix A with the two spanning vectors of U as rows. By Theorem 6.10, $U^\perp = \mathcal{N}(A)$. We have

$$A = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 0 & 1 & 0 & 1 \end{bmatrix} \xrightarrow{E_{12}(-2)} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix},$$

from which it follows that the null space of A consists of vectors of the form

$$\begin{bmatrix} -x_3 \\ -x_4 \\ x_3 \\ x_4 \end{bmatrix} = x_3 \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}.$$

Therefore, $U^\perp$ has basis $(-1,0,1,0)$ and $(0,-1,0,1)$. The vector $(-1,0,1,0)$ of this basis is repeated in the basis of $V^\perp$, so we only to need list it once. Form the matrix B whose rows are $(-1,0,1,0)$, $(-2,1,0,1)$, and $(0,-1,0,1)$, then calculate the reduced row echelon form of B:

$$B = \begin{bmatrix} -1 & 0 & 1 & 0 \\ -2 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 \end{bmatrix} \xrightarrow[E_1(-1)]{E_{21}(-2)} \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$
$$\xrightarrow{E_{32}(1)} \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & -2 & 2 \end{bmatrix} \xrightarrow[E_{13}(1)]{E_3(-1/2)\quad E_{23}(2)} \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

It follows that $\mathcal{N}(B)$ consists of vectors of the form

$$\begin{bmatrix} x_4 \\ x_4 \\ x_4 \\ x_4 \end{bmatrix} = x_4 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Therefore, $U \cap V = \mathcal{N}(B)$ is a one-dimensional space spanned by the vector $(1,1,1,1)$. □

Our last application of the orthogonal complements theorem is another Fredholm alternative theorem (compare this to Corollary 2.5.)

Corollary 6.3. Fredholm Alternative If $A\mathbf{x} = \mathbf{b}$ is a real linear system with $\mathbf{b} \neq 0$, then either the system is consistent or there is a solution $\mathbf{y}$ to the homogeneous system $A^T\mathbf{y} = \mathbf{0}$ such that $\mathbf{y}^T\mathbf{b} \neq 0$.

Proof. Let $V = \mathcal{C}(A)$. By (3) of Theorem 6.8, $\mathbb{R}^n = V + V^\perp$, where $\mathbb{R}^n$ has the standard inner product. From (3) of the orthogonal complements theorem,

$\mathcal{C}(A) = \mathcal{N}(A^T)^\perp$. Take complements again and use (5) of Theorem 6.8 to deduce that $V^\perp = \mathcal{N}(A^T)$. Now the system either has a solution or does not. If the system has no solution, then by Theorem 3.15, $\mathbf{b}$ does not belong to $V = \mathcal{C}(A)$. Since $\mathbf{b} \notin V$, we can write $\mathbf{b} = \mathbf{v} + \mathbf{y}$, where $\mathbf{y} \neq 0$, $\mathbf{y} \in V^\perp$ and $\mathbf{v} \in V$. It follows that

$$\langle \mathbf{y}, \mathbf{b} \rangle = \mathbf{y} \cdot \mathbf{b} = \mathbf{y} \cdot (\mathbf{v} + \mathbf{y}) = 0 + \mathbf{y} \cdot \mathbf{y} \neq 0.$$

On the other hand, if the system has a solution $\mathbf{x}$, then for any vector $\mathbf{y} \in \mathcal{N}(A)$ we have $\mathbf{y}^T A\mathbf{x} = \mathbf{y}^T\mathbf{b}$. It follows that if $\mathbf{y}^T A = \mathbf{0}$, then $\mathbf{y}^T\mathbf{b} = 0$. This completes the proof. □

6.4 Exercises and Problems

Exercise 1. Let $V = \text{span}\{(1, -1, 2, 0), (2, 0, -1, 1)\} \subset \mathbb{R}^4 = W$. Compute $V^\perp$ and use it to verify that $V + V^\perp = \mathbb{R}^4$.

Exercise 2. Let $V = \text{span}\{(1, -1, 2)\} \subset \mathbb{R}^3 = W$. Compute $V^\perp$ and use it to verify that $V \cap V^\perp = \{\mathbf{0}\}$.

Exercise 3. Let $V = \text{span}\{1 + x, x^2\} \subset W = \mathcal{P}_2$, where the space $\mathcal{P}_2$ of polynomials of degree at most 2 has the standard inner product of $C[0, 1]$. Compute $V^\perp$.

Exercise 4. Let $V = \text{span}\{1 + x + x^3\} \subset W = \mathcal{P}_3$, where $\mathcal{P}_3$ has the standard inner product of $C[0, 1]$ and compute $V^\perp$.

Exercise 5. Let $V = \text{span}\{(1, 0, 2), (0, 2, 1)\} \subset \mathbb{R}^3 = W$. Compute $V^\perp$and verify that $\left(V^\perp\right)^\perp = V$.

Exercise 6. Let $V = \text{span}\{(4, 1, -2)\} \subset \mathbb{R}^3 = W$, where W has the weighted inner product $\langle (x, y, z), (u, v, w) \rangle = 2xu + 3yv + zw$. Compute $V^\perp$ and verify that $\left(V^\perp\right)^\perp = V$.

Exercise 7. Carry out the method of computing $U \cap V$ discussed on page 421 with $U = \text{span}\{(1, 2, 1), (2, 1, 0)\}$, $V = \text{span}\{(1, 1, 1), (1, 1, 3)\}$ and $W = \mathbb{R}^3$.

Exercise 8. Repeat Exercise 7 with $U = \text{span}\{(1, 2, 1, 1), (2, 7, 5, 3), (1, 2, 2, 3))\}$, $V = \text{span}\{(1, -3, 2, 4), (1, 4, -4, -5), (0, -1, 2, 2)\}$ and $W = \mathbb{R}^4$.

Problem 9. Show that if V is a subspace of the inner product space W, then so is $V^\perp$.

Problem 10. Show that if V is a subspace of the inner product space W, then $V \cap V^\perp = \{\mathbf{0}\}$.

*Problem 11. Let U and V be subspaces of the inner product space W. Prove the following.
(a) $(U \cap V)^\perp = U^\perp + V^\perp$ (b) $(U + V)^\perp = U^\perp \cap V^\perp$

*Problem 12. Use the Fredholm alternative of this section to prove that the normal equations $A^T A\mathbf{x} = A^T\mathbf{b}$ are consistent for any matrix A.

6.5 *Operator Norms

The object of this section is to develop a useful notion of the norm of a matrix. For simplicity, we stick with real matrices, but all of the results in this section carry over to complex matrices. In Chapters 4 and 5 we studied the concept of a vector norm, which gave us a way of thinking about the "size" of a vector. We could easily extend this to matrices, just by thinking of a matrix as a vector that had been chopped into segments of equal length and re-stacked as a matrix. Thus, every vector norm on the space $\mathbb{R}^{mn}$ of vectors of length mn gives rise to a vector norm on the space $\mathbb{R}^{m,n}$ of $m \times n$ matrices. Experience has shown that this is not the best way to look for norms of matrices because it may not connect well to the operation of matrix multiplication. One exception is the standard norm, from which follows the Frobenius norm, It would be too much to expect norms to distribute over products. The following definition takes a middle ground that has proved to be useful for many applications.

Definition 6.14. Matrix Norm A vector norm $\|\cdot\|$ that is defined on the vector space $\mathbb{R}^{m,n}$ of $m \times n$ matrices, for all pairs m, n, is said to be a *matrix norm* if for all pairs of matrices A, B that are conformable for multiplication,

$$\|AB\| \leq \|A\| \, \|B\| \, .$$

Our first example of such a norm is the Frobenius norm; it is the one exception that we mentioned above.

Theorem 6.12. The Frobenius norm is a matrix norm.

Proof. Let A and B be matrices conformable for multiplication and suppose that the rows of A are $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_m^T$, while the columns of B are $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$. Then we have that $AB = \left[\mathbf{a}_i^T \mathbf{b}_j\right]$, so that by applying the definition and the CBS inequality, we obtain that

$$\begin{aligned} \|AB\|_F &= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} \left| \mathbf{a}_i^T \mathbf{b}_j \right|^2 \right)^{1/2} \leq \left(\sum_{i=1}^{m} \sum_{j=1}^{n} \|\mathbf{a}_i\|^2 \, \|\mathbf{b}_j\|^2 \right)^{1/2} \\ &\leq \left(\|A\|_F^2 \, \|B\|_F^2 \right)^{1/2} = \|A\|_F \, \|B\|_F \, . \end{aligned}$$

□

The most common multiplicative norms come from a rather general notion. Just as every inner product "induces" a norm in a natural way, every norm on the standard spaces induces a norm on matrices in a natural way. First recall that an *upper bound* for a set of real numbers is a number greater than or equal to any number in the set, and the *supremum* of a set of reals is the least (smallest) upper bound. We abbreviate this to "sup." For example, the sup of the open interval $(0, 1)$ is 1.

Supremum

Definition 6.15. Operator Norm The *operator norm* induced on matrices by a norm on the standard spaces is defined by the formula

$$\|A\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

A useful fact about these norms is the following equivalence:

$$\|A\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\mathbf{x}\neq\mathbf{0}} \left\| A\frac{\mathbf{x}}{\|\mathbf{x}\|} \right\| = \sup_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|.$$

We can see from this that the intuitive content of the operator norm of matrix A is that it is a measure of the largest expansion of a unit "sphere" caused by multiplication by A.

The reason for the term "operator" in the preceding definition is that the matrix A is being measured by its operator action on vectors. We could just as well have used the preceding definition to define the norm $\|T_A\|$ of the operator T_A. The notion of norm of a general linear operator is studied extensively in a branch of mathematical analysis known as *functional analysis*.

Operator norms are also are special cases of matrix norms.

Theorem 6.13. Every operator norm is a matrix norm.

Proof. For a matrix A, clearly $\|A\| \geq 0$ with equality if and only if $A\mathbf{x} = 0$ for all vectors $\mathbf{x}$, which is equivalent to $A = 0$. The remaining two norm properties are left as exercises. Finally, if A and B are conformable for multiplication, then $B\mathbf{x} = \mathbf{0}$ implies $\|AB\mathbf{x}\| = 0$, so we have

$$\|AB\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|AB\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{B\mathbf{x}\neq\mathbf{0}} \frac{\|AB\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{B\mathbf{x}\neq\mathbf{0}} \frac{\|AB\mathbf{x}\|}{\|B\mathbf{x}\|} \cdot \frac{\|B\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\| \cdot \|B\|. \quad \square$$

Incidentally, one difference between the Frobenius norm and operator norms is how the identity I_n is handled. Notice that $\|I_n\|_F = \sqrt{n}$, while with any operator norm $\|\cdot\|$ we have from the definition that $\|I_n\| = 1$.

How do we compute these norms? Here are three common cases:

Theorem 6.14. If $A = [a_{ij}]_{m,n}$, then

(1) $\|A\|_\infty = \max_{1\leq i\leq m} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}$
(2) $\|A\|_1 = \max_{1\leq j\leq n} \{ \sum_{i=1}^{m} |a_{ij}| \}$
(3) $\|A\|_2 = \rho(A^T A)^{1/2}$

Proof. Items (2) and (3) are left as exercises. For the proof of (1), use the fact that $\|A\|_\infty = \sup_{\|\mathbf{v}\|_\infty=1} \|A\mathbf{v}\|_\infty$. Now a vector has infinity norm 1 if and only if the largest absolute value of its coordinates is 1. Notice that we can make the ith entry of $A\mathbf{v}$ as large as possible simply by choosing $\mathbf{v}$ so

that the jth coordinate of $\mathbf{v}$ is ± 1 and agrees with the sign of a_{ij}. Thus, the infinity norm of $A\mathbf{v}$ is the maximum of the row sums of the absolute values of the entries of A. □

One of the more important applications of the idea of a matrix norm is the famous Banach lemma. Essentially, it amounts to a matrix version of the familiar geometric series.

Theorem 6.15. Banach Lemma If M is a square matrix such that $\|M\| < 1$ for some operator norm $\|\cdot\|$, then $I - M$ is invertible. Moreover, $\left\|(I-M)^{-1}\right\| \leq 1/(1-\|M\|)$ and

$$(I-M)^{-1} = I + M + M^2 + \cdots + M^k + \cdots .$$

Proof. We have seen this trick before: Form the telescoping series

$$(I-M)\left(I + M + M^2 + \cdots + M^k\right) = I - M^{k+1},$$

so that

$$I - (I-M)\left(I + M + M^2 + \cdots + M^k\right) = M^{k+1}.$$

Now by the multiplicative property of matrix norms and fact that $\|M\| < 1$,

$$\left\|M^{k+1}\right\| \leq \|M\|^{k+1} \to 0, \text{ as } k \to \infty.$$

It follows that the matrix $\lim_{k\to\infty}\left(I + M + M^2 + \cdots + M^k\right) = B$ exists and that $I - (I-M)B = 0$, from which it follows that $B = (I-M)^{-1}$. Finally, note that

$$\begin{aligned}\left\|I + M + M^2 + \cdots + M^k\right\| &\leq \|I\| + \|M\| + \left\|M^2\right\| + \cdots + \left\|M^k\right\| \\ &\leq 1 + \|M\| + \|M\|^2 + \cdots + \|M\|^k \\ &\leq \frac{1}{1-\|M\|}.\end{aligned}$$

Now take the limit as $k \to \infty$ to obtain the desired result. □

A fundamental idea in numerical linear algebra is the notion of the **Condition Number** *condition number* of a matrix A. Roughly speaking, the condition number measures the degree to which changes in A lead to changes in solutions of systems $A\mathbf{x} = \mathbf{b}$. A large condition number means that small changes in A or $\mathbf{b}$ may lead to large changes in $\mathbf{x}$. In the case of an invertible matrix A, the condition number of A is defined to be

$$\operatorname{cond}(A) = \|A\|\left\|A^{-1}\right\|.$$

Of course this quantity is norm dependent. In the case of a p-norm, the condition number is denoted by $\operatorname{cond}_p(A)$. For an operator norm $\|\cdot\|$ on $\mathbb{R}^{m,n}$, the Banach lemma has a nice application, whose proof is left as an exercise.

Corollary 6.4. If $A = I + N$, where $\|N\| < 1$, then

$$\text{cond}(A) \leq \frac{1 + \|N\|}{1 - \|N\|}.$$

We conclude with a very fundamental result for numerical linear algebra. Here is the scenario: we desire to solve the linear system $A\mathbf{x} = \mathbf{b}$, where A is invertible. Due to arithmetic error or possibly input data error, we end up with a value $\mathbf{x}+\delta\mathbf{x}$ that solves exactly a "nearby" system $(A+\delta A)(\mathbf{x}+\delta\mathbf{x}) = \mathbf{b}+\delta\mathbf{b}$. (It can be shown using an idea called "backward error analysis" that this is really what happens when many algorithms are used to solve a linear system.) The question is, what is the size of the relative error $\|\delta\mathbf{x}\| / \|\mathbf{x}\|$? As long as the perturbation matrix $\|\delta A\|$ is reasonably small, there is a very elegant answer.

Theorem 6.16. Perturbation Theorem Suppose that A is invertible, $\mathbf{b} \neq \mathbf{0}$, $A\mathbf{x} = \mathbf{b}$, $(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$, and $\left\|A^{-1}\delta A\right\| = c < 1$ with respect to some operator norm. Then $A + \delta A$ is invertible and

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\text{cond}\,(A)}{1 - c}\left\{\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right\}.$$

Proof. That the matrix $I + A^{-1}\delta A$ is invertible follows from hypothesis and the Banach lemma. Since A is invertible by hypothesis, $A\left(I + A^{-1}\delta A\right) = A + \delta A$ is also invertible. Expand the perturbed equation to obtain

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = A\mathbf{x} + \delta A\mathbf{x} + A\delta\mathbf{x} + \delta A\,\delta\mathbf{x} = \mathbf{b} + \delta\mathbf{b}.$$

Now subtract the terms $A\mathbf{x} = \mathbf{b}$ from each side and solve for $\delta\mathbf{x}$ to obtain

$$(A + \delta A)\delta\mathbf{x} = A(I + A^{-1}\delta A)\delta\mathbf{x} = -\delta A\mathbf{x} + \delta\mathbf{b},$$

so that

$$\delta\mathbf{x} = (I + A^{-1}\delta A)^{-1}A^{-1}\left(-\delta A \cdot \mathbf{x} + \delta\mathbf{b}\right).$$

Take norms, use the additive and multiplicative properties of matrix norms and also the Banach lemma to obtain

$$\|\delta\mathbf{x}\| \leq \frac{\left\|A^{-1}\right\|}{1 - c}\left\{\|\delta A\|\,\|\mathbf{x}\| + \|\delta\mathbf{b}\|\right\}.$$

Next divide both sides by $\|\mathbf{x}\|$ to obtain

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\left\|A^{-1}\right\|}{1 - c}\left\{\|\delta A\| + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{x}\|}\right\}.$$

Finally, notice that $\|\mathbf{b}\| \leq \|A\|\,\|\mathbf{x}\|$. Therefore, $1/\|\mathbf{x}\| \leq \|A\| / \|\mathbf{b}\|$. Replace $1/\|\mathbf{x}\|$ in the right hand side by $\|A\| / \|\mathbf{b}\|$ and factor out $\|A\|$ to obtain

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|A\|\left\|A^{-1}\right\|}{1-c}\left\{\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right\},$$

which completes the proof, since by definition, $\operatorname{cond} A = \|A\|\left\|A^{-1}\right\|$. □

If we believe that the inequality in the perturbation theorem can be sharp (it can!), then it becomes clear how the condition number of the matrix A is a direct factor in how relative error in the solution vector is amplified by perturbations in the coefficient matrix.

Example 6.21. Suppose we wish to solve the nonsingular system $A\mathbf{x} = \mathbf{b}$ exactly, where the coefficient matrix $A = \begin{bmatrix} 1 & 10 \\ 10 & 101 \end{bmatrix}$ is known, but the right-hand-side vector $\mathbf{b}$ is determined from measured data whose error of measurement is such that the ratio of the largest error in any coordinate of $\mathbf{b}$ to the largest coordinate of $\mathbf{b}$ (this ratio is called the *relative error*) is no more than 0.01 in absolute value. Estimate the size of the relative error in the solution.

Solution. Let the correct value of the right-hand side be $\mathbf{b}$ and the measured value of the right-hand side be $\widetilde{\mathbf{b}}$, so that the error of measurement is the vector $\delta\mathbf{b} = \widetilde{\mathbf{b}} - \mathbf{b}$. Rather than solving the system $A\mathbf{x} = \mathbf{b}$, we end up solving the system $A\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b}$, where $\widetilde{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$. The relative error in data is the quantity $\|\delta\mathbf{b}\|_\infty / \|\mathbf{b}\|_\infty \leq 0.01$, while the relative error in the computed solution is $\|\delta\mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty$. We have $\operatorname{cond}_\infty(A) = 12321$ (left as an exercise), so the relative error in the solution satisfies the inequality

$$\frac{\|\delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \frac{\operatorname{cond}_\infty(A)}{1-0} \cdot \frac{\|\delta\mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty} = 12321 \cdot 0.01 = 123.21.$$

In other words, the relative error in our computed solution could be as large as 12321% which, of course, would make it quite worthless. □

Here is one more useful observation about operator norms that can be couched in very general terms.

Definition 6.16. **Equivalent Norms** Two norms $\|\cdot\|$ and $|||\cdot|||$ on the vector space V are said to be *equivalent* if there exist positive constants C, D such that for all $\mathbf{x} \in V$,

$$C\|\mathbf{x}\| \leq |||\mathbf{x}||| \leq D\|\mathbf{x}\|.$$

It is easily seen that this relation is symmetric, for we deduce from the definition that

$$\frac{1}{D}|||\mathbf{x}||| \leq \|\mathbf{x}\| \leq \frac{1}{C}|||\mathbf{x}|||.$$

Similarly, one checks that equivalence is a transitive relation, that is, if norm $\|\cdot\|_a$ is equivalent to $\|\cdot\|_b$ and $\|\cdot\|_b$ is equivalent to $\|\cdot\|_c$, then $\|\cdot\|_a$ is equivalent to $\|\cdot\|_c$. Roughly speaking, the definition says that equivalent norms yield the same value up to fixed upper and lower scale factors. The significance of equivalence of norms is that convergence of a sequence of vectors in one norm

implies convergence in the other equivalent norm. In general, a vector space can have inequivalent norms. However, in order to do so, the space must be infinite-dimensional. The following theorem applies to all finite-dimensional vector spaces, so it certainly applies to the space of $n \times n$ matrices $\mathbb{R}^{n,n}$ with an operator norm. Thus, all operator norms are equivalent in the above sense.

Theorem 6.17. All norms on a finite-dimensional space are equivalent.

We sketch a proof. Let V be a finite-dimensional vector space. We know that there is an arithmetic preserving one-to-one correspondence between elements $\mathbf{x}$ of V and their coordinate vectors with respect to some basis of V, so that elements of V are identified with some $\mathbb{R}^n$. Without loss of generality $V = \mathbb{R}^n$. Now let $\|\cdot\|$ be any norm on V. First, one establishes that $\|\cdot\| : V \to \mathbb{R}$ is a continuous function by proving that for all $x, y \in V$, $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\|$. The proof of Problem 22, Section 4.1, shows this inequality.

Next, one observes that the boundary of unit ball $B_1(\mathbf{0})$ in the infinity norm in V is a closed and bounded set that does not contain the origin. By the extreme value theorem of analysis, the function $\|\cdot\|$ assumes its maximum and minimum values on the ball, and these must be positive, say C, D. Thus, for all nonzero vectors $\mathbf{x}$ we have

$$C \leq \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty} \right\| \leq D.$$

Multiply through by $\|\mathbf{x}\|_\infty$, and we see that $C\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\| \leq D\|\mathbf{x}\|_\infty$, which proves the equivalence of the given norm to the infinity norm. It follows from transitivity of the equivalence property that all norms are equivalent to each other. □

6.5 Exercises and Problems

Exercise 1. Compute the Frobenius, 1-, and ∞-norms of the following matrices.

(a) $\begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 2 & 2 & 0 \\ 1 & -3 & 0 & -1 \\ 1 & 1 & -2 & 0 \\ -2 & 1 & 6 & 1 \end{bmatrix}$

Exercise 2. Compute the condition number of each matrix in Exercise 1 using the infinity norm.

Exercise 3. Let $A = \begin{bmatrix} 2 & 7 \\ 3 & 10 \end{bmatrix}$, $\mathbf{b} = [5.7, 8.2]^T$ and solve the system $A\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ with a technology tool. Next, let $\delta\mathbf{b} = [0.096, -0.025]^T$ and $\mathbf{x} + \delta\mathbf{x}$ be the solution to $A(\mathbf{x} + \delta\mathbf{x}) = (\mathbf{b} + \delta\mathbf{b})$. Compute $\|\delta\mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty$ and compare it to $\text{cond}_\infty(A)\, \|\delta\mathbf{b}\|_\infty / \|\mathbf{b}\|_\infty$.

Exercise 4. Repeat Exercise 3 using the ∞-norm with $A = \begin{bmatrix} 1 & 10 \\ 10 & 101 \end{bmatrix}$, $\mathbf{b} = [0.985, 9.95]^T$, and $\delta\mathbf{b} = [-0.0995, 0.00985]^T$. How good is the solution if this error is introduced into the right-hand side?

Exercise 5. Verify that the perturbation theorem is valid for $A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & -2 \\ 0 & -2 & 1 \end{bmatrix}$, $\mathbf{b} = [-5, 1, -3]^T$, $\delta A = 0.05A$, and $\delta\mathbf{b} = 0.5\mathbf{b}$. Use the 2-norm.

Exercise 6. Verify the inequality of Corollary 6.4 using the infinity norm and $N = \frac{1}{3}\begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}$.

Problem 7. Show that if A is a stochastic matrix, then $\|A\|_1 = 1$.

*Problem 8. Prove Corollary 6.4.

*Problem 9. Show that if A is invertible and $\|A^{-1}\delta A\| < 1$, then so is $A + \delta A$.

Problem 10. Prove that $\|A\|_1 = \max_{1 \le j \le n} \{\sum_{i=1}^{m} |a_{ij}|\}$.

Problem 11. Prove that $\|A\|_2 = \rho(A^T A)^{1/2}$.

Problem 12. Suppose we want to approximately solve a system of the form $A\mathbf{x} = \mathbf{b}$, where $A = I - M$ and $\|M\| < 1$ for some operator norm. Use the Banach lemma to devise such a scheme involving only a finite number of matrix additions and multiplications.

*Problem 13. Show that for any any operator norm $\|\cdot\|$, $\rho(A) \le \|A\|$.

*Problem 14. Show that a square matrix A is *power bounded*, that is, $\|A^m\|_2 \le C$ for all positive m and some constant C independent of m, if every eigenvalue of A is either strictly less than 1 in absolute value or of absolute value equal to 1 and simple.

Problem 15. Does it follow from Problem 14 and the equivalence of operator norms that power bounded in one operator norm implies power bounded in any other? Justify your answer.

*Problem 16. Let A be a real matrix and U, V orthogonal matrices.

(a) Show from the definition that $\|U^T A V\|_2 = \|A\|_2$.
(b) Determine $\|\Sigma\|_2$ if Σ is a diagonal matrix with nonnegative entries.
(c) Use (a) and (b) to express $\|A\|_2$ in terms of the singular values of A.

*Problem 17. Show that if A is a real invertible $n \times n$ matrix, then using $\|\cdot\|_2$ yields cond $(A) = \sigma_1/\sigma_n$, the ratio of largest to smallest singular values of A.

Problem 18. Example 6.21 gives an upper bound on the error propagated to the solution of a system due to right-hand-side error. How pessimistic is it? Experiment with various random different erroneous right-hand-sides with your choice of error tolerance and compare actual error with estimated error.

6.6 *Applications and Computational Notes

Introduction to Fourier Analysis

In the early 1800's Joseph Fourier, in his historic monograph "Théorie analytique de la chaleur" (The Analytical Theory of Heat), introduced the idea that solutions to heat problems that he introduced and, for that matter, any "arbitrary function" could be expressed as a infinite sum of trigonometric functions. As it turns out, this was not exactly correct. It was the lack of precision in the definition of "arbitrary function" that generated early criticisms of Fourier's ideas. Nonetheless his extraordinary insight has reverberated throughout the mathematical and scientific world for the past 200 years. The resulting Fourier analysis has had deep applications in diverse areas such as applied mathematics, electrical engineering, physics, quantum theory, signal analysis and many other fields.

So let's set the stage for our study of Fourier analysis. First, we will extend our notion of function to include complex values. Henceforth, we redefine $C[a,b]$ to be the set of all continuous complex-valued functions $f(t)$, $a \leq t \leq b$. Such functions can be described as $f(t) = f_R(t) + \mathrm{i} f_I(t)$, where f_R and f_I are real-valued continuous functions. We've seen such things before, e.g., the trigonometric function $f(t) = e^{it} = \cos t + \mathrm{i} \sin t$. As such, derivatives, integrals and limits are perfectly straightforward:

Complex-valued Function

$$\begin{aligned} f'(t) &= f_R'(t) + \mathrm{i} f_I'(t) \\ \int_a^b f(t)\, dt &= \int_a^b f_R(t)\, dt + \mathrm{i} \int_a^b f_I(t)\, dt \\ \lim_{h \to 0} f(t+h) &= \lim_{h \to 0} f_R(t+h) + \mathrm{i} \lim_{h \to 0} f_I(t+h) \end{aligned}$$

In the same way that we showed in Example 3.4 of Chapter 3 that $C[0,1]$ is a vector space, so is our new version of $C[a,b]$. Moreover, there is an inner product that we can attach to this space:

$$\langle f, g \rangle = \int_a^b g(t)\, \overline{f(t)}\, dt \tag{6.3}$$

Be aware that most analysis texts use $\langle f, g \rangle = \int_a^b f(t)\, \overline{g(t)}\, dt$. We conjugate the first term so that condition (4) of Definition 6.7 is satisfied. We leave it as an exercise to show that the inner product laws are satisfied.

Next, we will expand our universe of functions, but first a bit of reminder from calculus: Recall that if a function $f(t)$ is defined on an open interval with the point t_0 in the interval or on the boundary, then we have one-sided limits $f(t_0^+) = \lim_{h \to 0^+} f(t_0 + h)$ or $f(t_0^-) = \lim_{h \to 0^-} f(t_0 + h)$ which may or may not exist. (Recall that the expression $h \to 0^+$ means that h is allowed to approach 0 from the right side, i.e.,

One-sided Limits

$h > 0$ and similarly $h \to 0^-$ means we only allow $h < 0$. As usual, the expression $h \to 0$ means that there is no restriction on the sign of h.) We have already made implicit use of these ideas early on in calculus: A function is continuous on the closed interval $[a,b]$ if at every interior point t_0, the two-sided limit satisfies $f(t_0) = \lim_{h\to 0} f(t_0+h)$ and at the boundaries $f(a) = \lim_{h\to 0^+} f(a+h)$ and $f(b) = \lim_{h\to 0^-} f(b+h)$.

Definition 6.17. PWC Function A function $f(t)$ defined on the interval $[a,b]$ is *piecewise continuous* (PWC) on the interval $[a,b]$ if it has at most a finite number of discontinuities and at each such point, as well as the endpoints of the interval, one-sided limits exist and are finite. The set of all such functions is denoted by $C_{PW}[a,b]$.

We need one more refinement of function spaces, a subspace of $C_{PW}[a,b]$:

Definition 6.18. PWS Function A function $f(t)$ defined on the interval $[a,b]$ is *piecewise smooth* (PWS) on the interval $[a,b]$ if both $f(t)$ and $f'(t)$ are piecewise continuous on $[a,b]$. The set of all such functions is denoted by $C^1_{PW}[a,b]$.

Clearly sums and products of functions of piecewise smooth functions are themselves piecewise smooth. Also we integrate functions $f, g \in C_{PW}[a,b]$ or $C^1_{PW}[a,b]$ as well as their sums and products over the interval $[a,b]$ by adding up all the integrals over successive subintervals of $[a,b]$ on which the function is continuous and using one-sided limits to define the boundaries of integration on each subinterval.

Integration of PWC Function

Theorem 6.18. $C_{PW}[a,b]$ with the usual pointwise addition and multiplication is an inner product space with inner product given by equation (6.3).

Note that as a subspace of $C_{PW}[a,b]$, the space $C^1_{PW}[a,b]$ is also an inner product space with the inherited inner product.

As we have learned, any inner product space is automatically also a normed linear space with the norm induced from the inner product. Thus, in the case of $f \in C_{PW}[a,b]$ or $f \in C^1_{PW}[a,b]$ the norm of this function is given by

$$\|f\|_2 = \sqrt{\int_a^b f(t)\,\overline{f(t)}\,dt} = \sqrt{\int_a^b |f(t)|^2\,dt}.$$

In anticipation of our exploration of DSP, we will search for a basis of the inner product space $C^1_{PW}[0,T]$, $T > 0$. That search will lead us into Fourier analysis. Certainly, this space is not finite dimensional since it contains the space of all polynomial functions on the interval $[0,T]$ (see Example 2.8 of Chapter 3.) In fact, matters are even worse than that: Unlike the space of polynomials, which has a basis that can be listed, namely $\{1, t, t^2, \ldots, t^n, \ldots\}$,

the space of piecewise smooth functions on $[0,T]$ has no such basis. Fourier's key discovery was that it does have such a basis in a looser sense. Enter the trigonometric functions:

Theorem 6.19. Trigonometric Polynomials Let $\omega = 2\pi/T$ and define $e_n(t) = e^{\mathrm{i}\omega nt}$, $n \in \mathbb{Z}$. Then

(1) $e_n(t) \in C^1_{PW}[0,T]$.
(2) $e_n'(t) = \mathrm{i}\omega n e_n(t)$.
(3) $\{e_n(t)\}_{n=-\infty}^{\infty}$ is an orthogonal set of functions in $C^1_{PW}[0,T]$.
(4) The best approximation $x_N(t)$ to $x(t) \in C^1_{PW}[0,T]$ from the subspace spanned by $\{e_n(t)\}_{n=-N}^{N}$ (in the sense of minimizing $\|x_N - x\|_2$) is

$$x_N(t) = \sum_{n=-N}^{N} c_n e_n(t), \qquad c_n = \frac{1}{T}\int_0^T x(t)\, e^{-\mathrm{i}n\omega t}\, dt, \ -N \le n \le N.$$

Proof. For (1) note that $e_n(t) = e^{\mathrm{i}n\omega t} = \cos(n\omega t) + \mathrm{i}\sin(n\omega t)$, so $e_n(t)$ has continuous derivatives of all orders. Hence, $e_n(t) \in C^1_{PW}[0,T]$. For (2) calculate

$$\begin{aligned} e_n'(t) &= (\cos(n\omega t) + \mathrm{i}\sin(n\omega t))' = n\omega(-\sin(n\omega t) + \mathrm{i}\cos(n\omega t)) \\ &= \mathrm{i}n\omega(\mathrm{i}\sin(n\omega t) + \cos(n\omega t)) = \mathrm{i}n\omega e^{\mathrm{i}n\omega} = \mathrm{i}n\omega e_n(t) \end{aligned}$$

To prove (3), use (2) and the fact that for $m \neq n$, $e^{\mathrm{i}\omega(n-m)L} = e^{\mathrm{i}2\pi(n-m)} = 1$, to calculate

$$\langle e_m, e_n\rangle = \int_0^T e^{\mathrm{i}\omega nt}\overline{e^{\mathrm{i}\omega mt}}\, dt = \int_0^T e^{\mathrm{i}\omega(n-m)t}\, dt = \frac{e^{\mathrm{i}\omega(n-m)t}}{\mathrm{i}\omega(n-m)}\bigg|_{t=0}^{T} = 0.$$

For $m = n$, $\langle e_n, e_n\rangle = \int_0^T e^{\mathrm{i}\omega 0t}\, dt = \int_0^T 1\, dt = T$. Hence, the $e_n(t)$'s form an orthogonal set. Finally, according to the projection theorem (Theorem 6.7) the best approximation to $x(t)$ from the subspace spanned by $\{e_n(t)\}_{n=-N}^{N}$ is the partial sum

Partial Fourier Sum

$$x_N(t) = \sum_{n=-N}^{N} c_n e_n(t), \text{ with } c_n = \frac{\langle e_n, x\rangle}{\langle e_n, e_n\rangle} = \frac{1}{T}\int_0^T x(t)\, e^{-\mathrm{i}n\omega t}\, dt.$$

□

In light of this theorem it is tempting to use limits and conclude that

Fourier Series

$$x(t) = \lim_{N\to\infty}\sum_{n=N}^{N} c_n e^{in\omega t} = \sum_{n=-\infty}^{\infty} c_n e^{in\omega t}, \tag{6.4}$$

where $c_n = \frac{1}{T}\int_0^T x(t)\, e^{-in\omega t}\, dt, n \in \mathbb{Z}$. This is what Fourier did with a somewhat vague definition of "function." In other words, even though the trigonometric polynomials did not form a basis for a function space in the usual sense

of basis, they were somehow sufficiently "dense" in the function space to be a basis for all in some limiting sense. The unanswered question was "What function space?" and generations of mathematicians have worked to make Fourier's powerful insight precise. The trigonometric series above is called the Fourier series of $x(t)$ in his honor. The interested reader can pursue these developments in detail in the text [11] by C. Gasquet and P. Witomski. For our purposes the following theorem, which is really a special case of Dirichlet's Theorem, is sufficient (see [11, p. 43] for a full description and proof):

Theorem 6.20. Dirichlet Theorem Let $x(t) \in C^1_{PW}[0,T]$. Then the Fourier series $F(x(t)) = \sum_{n=-\infty}^{\infty} c_n e^{in\omega t}$ is convergent for $t \in [0,T]$ with values as follows:

(1) If t is a point of continuity of $x(t)$, then $F(x(t)) = x(t)$.
(2) If t is a point of discontinuity of $x(t)$, then
$F(x(t)) = \frac{1}{2}(x(t^-) + x(t^+))$.

For the record, note that the mapping $F : C^1_{PW}[0,T] \to C^1_{PW}[0,T]$ given by $F(x(t)) = \sum_{n=-\infty}^{\infty} c_n e^{in\omega t}$ is a linear operator from this vector space into itself by Theorem 6.20. Were we to require that at every point t_0 of discontinuity of $x(t) \in C^1_{PW}[0,T]$, $x(t_0) = \frac{1}{2}\left(x\left(t_0^-\right) + x\left(t_0^+\right)\right)$ (a reasonable requirement), then F would actually be the identity mapping!

Also note that that choice of interval $[0,T]$ is simply a matter of convenience. In the case of a function $x(t) \in C_{PW}[a,b]$ we can extend the definition of $x(t)$ to a periodic function of period $T = b - a$ over all of $\mathbb{R}$. The restriction of this extended function to the interval $[0,T]$ is a member of $C^1_{PW}[0,T]$. Moreover, periodicity implies that its integral over any interval of length T yields the same result. Thus, the Fourier series for $x(t)$ on the real line can be written as

$$F(x(t)) = \sum_{n=-\infty}^{\infty} c_n e^{in\omega t}, \quad c_n = \frac{1}{T}\int_a^b x(t)\, e^{-in\omega t}\, dt.$$

Real Fourier Series

The case of a real-valued function deserves special consideration. In fact, the original thesis of Fourier was that an "arbitrary function" of real values on a finite interval could be represented as an infinite sum of sine and cosine functions. Given a real-valued function $x(t) \in C^1_{PW}[0,T]$, we can express its Fourier series in the following form:

$$F(x(t)) = \sum_{n=-\infty}^{\infty} c_n e^{in\omega t} = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(n\omega t) + b_n \sin(n\omega t)) \tag{6.5}$$

where for $n \geq 0$,

$$a_n = \frac{2}{T}\int_0^T x(t)\cos(n\omega t)\, dt \text{ and } b_n = \frac{2}{T}\int_0^T x(t)\sin(n\omega t)\, dt. \tag{6.6}$$

We leave the details of these formulas as an exercise.

Even and Odd Functions

For the following examples, it is helpful to recall that a function $f(t)$ is *even* if $f(-t) = f(t)$ and *odd* if $f(-t) = -f(t)$. If $f(t)$ is odd, it follows that the integral $\int_{-M}^{M} f(t)\,dt = \int_{0}^{M} (f(t) + f(-t))\,dt = 0$. Likewise, if If $f(t)$ is even, then $\int_{-M}^{M} f(t)\,dt = \int_{0}^{M} (f(t) + f(-t))\,dt = 2\int_{0}^{M} f(t)\,dt$.

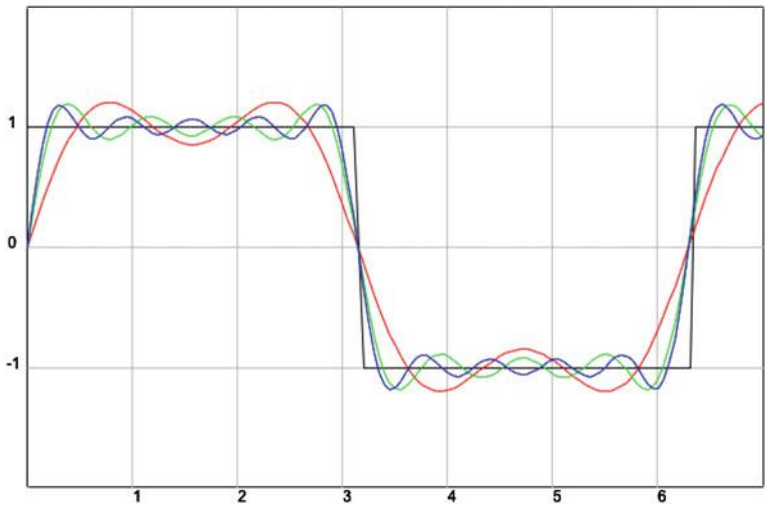


Fig. 6.4: Graph of $x(t)$ (——), Fourier sums $N = 3$ (——), $N = 7$ (——) and $N = 11$ (——) from Example 6.22.

Example 6.22. Find the Fourier series of the function $x(t) \in C^1_{PW}[0, 2\pi]$ defined by $x(t) = 1$ for $0 < t < \pi$, $x(t) = -1$ for $\pi < t < 2\pi$ and $x(0) = x(\pi) = x(2\pi) = 0$, so that $x(t)$ can be viewed as a periodic function of period 2π defined on $\mathbb{R}$. Graph this function along with the partial Fourier sums for $N = 3, 7, 11$.

Solution. Here $T = 2\pi$ and $\omega = 1$. Clearly $x(t)$ is an odd function and $\cos(nt)$ is even, so their product is odd. By periodicity $\int_0^{2\pi} x(t)\cos(n\omega t)\,dt = \int_{-\pi}^{\pi} x(t)\cos(n\omega t)\,dt$. So this integral is 0. Hence, for all n, $a_n = 0$. On the other hand $\sin(nt)$ is odd, so its product with $x(t)$ is even and thus $\int_0^{2\pi} x(t)\sin(n\omega t)\,dt = \int_{-\pi}^{\pi} x(t)\sin(n\omega t)\,dt = 2\int_0^{\pi} x(t)\sin(n\omega t)\,dt$. We calculate with $n \geq 1$ that

$$b_n = \frac{2}{2\pi} \cdot 2\int_0^{\pi} 1 \cdot \sin(nt)\,dt = \frac{2}{\pi} \cdot \frac{-\cos(nt)}{n}\bigg|_0^{\pi} = \frac{2}{n\pi}\left(1 - (-1)^n\right).$$

Thus, the Fourier series for $x(t)$ is

$$\frac{4}{\pi}\left\{\sin(t)+\frac{\sin(3t)}{3}+\frac{\sin(5t)}{5}+\frac{\sin(7t)}{7}+\cdots\right\}=\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\sin((2n-1)t)}{2n-1}.$$

See Figure 6.4 for a graph of $x(t)$ and partial Fourier sums with $N=3,7$ and 11 which correspond to upper limits 2, 4 and 6 in the above series. □

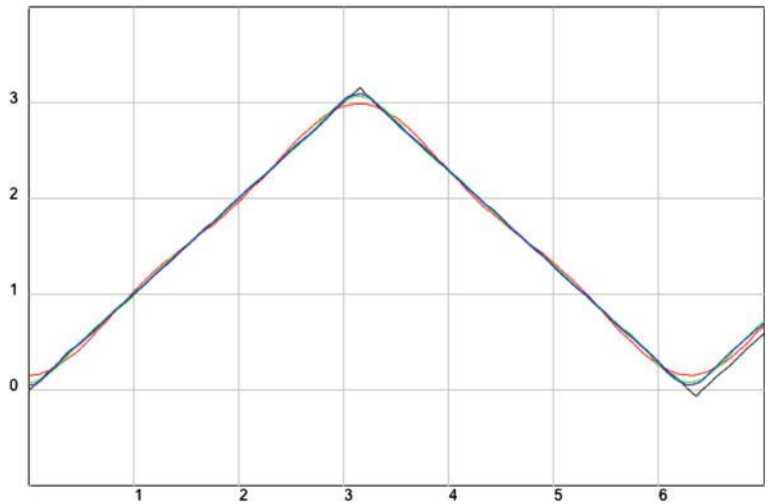


Fig. 6.5: Graph of $x(t)$ (——), Fourier sums$N=3$ (——), $N=7$ (——) and $N=11$ (——) from Example 6.23.

Example 6.23. Find the Fourier series of the function $x(t)\in C^1_{PW}[0,2\pi]$ defined by $x(t)=t$ for $0\le t<\pi$, $x(t)=2\pi-t$ for $\pi<t\le 2\pi$. Graph this function along with the partial Fourier sums for $N=3,7,11$.

Solution. Here again $T=2\pi$ and $\omega=1$. Clearly $x(t)$ is an even function and $\sin(nt)$ is odd, so their product is odd. By periodicity $\int_0^{2\pi}x(t)\sin(n\omega t)\,dt=\int_{-\pi}^{\pi}x(t)\sin(n\omega t)\,dt$. So this integral is 0. Hence, for all n, $b_n=0$. On the other hand $\cos(nt)$ is even, so its product with $x(t)$ is even and thus $\int_0^{2\pi}x(t)\cos(n\omega t)\,dt=\int_{-\pi}^{\pi}x(t)\cos(n\omega t)\,dt=2\int_0^{\pi}x(t)\cos(n\omega t)\,dt$. We calculate with $n\ge 1$ and integration by parts that

$$\begin{aligned}a_n&=\frac{2}{2\pi}\cdot 2\int_0^{\pi}t\cdot\cos(nt)\,dt\\&=\frac{2}{\pi}\cdot\left(t\frac{\sin(nt)}{n}\bigg|_0^{\pi}-\int_0^{\pi}\frac{\sin(nt)}{n}\,dt\right)=\frac{2}{n^2\pi}\left((-1)^n-1\right).\end{aligned}$$

The case $n=0$ yields $a_0=\frac{2}{\pi}\frac{\pi^2}{2}=\pi$. Thus, the Fourier series for $x(t)$ is

$$\frac{\pi}{2}-\frac{4}{\pi}\left\{\frac{\cos(t)}{1}+\frac{\cos(3t)}{9}+\frac{\cos(5t)}{25}+\cdots\right\}=\frac{\pi}{2}-\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\cos((2n-1)t)}{(2n-1)^2}.$$

See Figure 6.5 for a graph of $x(t)$ and partial Fourier sums with $N = 3, 7, 11$ which correspond to upper limits 2, 4 and 6 in the above series. □

Digital Signal Processing and Fourier Series

We are only going to examine signal sampling of finite length, and therefore sampled function $x(t) \in C^1_{PW}[0,T]$ of finite duration (who's going to sample a signal infinitely many times in the real world?). So let's say the duration interval of time (or space) t is the interval $[0,T]$. As a matter of convenience we can extend the definition of the signal over all $t \in \mathbb{R}$ and think of the sampling function $x(t)$ as defined over all t but periodic of period T. So the function has frequency $f = 1/T$ and angular frequency $\omega = 2\pi/T$. We restrict our sampling functions to the inner product space $C^1_{PW}[0,T]$, so that Theorem 6.20 applies.

First there is the matter of a filter in general:

Definition 6.19. Discrete Filter A *discrete filter* is a sequence of complex numbers $\mathbf{h} = \{h_n\}_{n=-\infty}^{\infty}$.

We can associate a Fourier series with each discrete filter (N.B.: we make no claims of convergence here):

Definition 6.20. DTFT If $\mathbf{h} = \{h_n\}_{n=-\infty}^{\infty}$ is a discrete filter, then the *discrete time Fourier transform* (DTFT) of $\mathbf{h}$ is the Fourier series $X(\mathbf{h})(\zeta) = \sum_{n=-\infty}^{\infty} h_n e^{i\zeta}$, $\zeta \in \mathbb{R}$.

We are particularly interested in this type of filter:

Definition 6.21. FIR Filter The filter $\mathbf{h} = \{h_n\}_{n=-\infty}^{\infty}$ is a *finite impulse response (FIR) filter* if there exists a positive integer L such that h_0 and h_L are nonzero and $h_n = 0$ for $n > L$ or $n < 0$. In this case we say $\mathbf{h}$ has length L and express it in the form $\mathbf{h} = \{h_n\}_{n=0}^{L}$.

Now suppose we are given an FIR filter $\mathbf{h} = \{h_n\}_{n=0}^{L}$ and we sample the signal $x(t)$ in sampling periods T_s, i.e., at $t_n = nT_s$, yielding the discrete signal $x_n = x(t_n)$, $n \in \mathbb{Z}$. Suppose we use the filter to transform the signal via the difference equation formula as we did in Sections 2.8 and 4.4:

$$y_n = h_0 x_n + h_1 x_{n-1} + \cdots + h_L x_{n-L}, \quad n \in \mathbb{Z}$$

Rather than use this formula globally let's use the facts that $x(t)$ is a Fourier series by Theorem 6.20 and that the signal transformation is linear. Thus, it is sufficient to examine the filter's effect on a Fourier mode of $x(t)$ with a possible phase shift ϕ of the form $x_m(t) = c_m e^{im(\omega t+\phi)}$ which yields sampling signals $x_{m,n} = c_m e^{im\omega(nT_s+\phi)}$, $n \in \mathbb{Z}$. Thus, for fixed m and arbitrary $n \in \mathbb{Z}$ we have

$$\begin{aligned} y_{m,n} &= h_0 c_m e^{im\omega(nT_s+\phi)} + h_1 c_m e^{im\omega((n-1)T_s+\phi)} + \cdots + h_L c_m e^{im\omega((n-L)T_s+\phi)} \\ &= \left(h_0 \cdot 1 + h_1 e^{-im\omega T_s} + \cdots + h_L e^{-iLm\omega T_s}\right) c_m e^{im\omega(nT_s+\phi)} \\ &= H\left(-m\omega T_s\right) x_{m,n}. \end{aligned}$$

Here $H(\zeta) = h_0 + h_1 e^{i\zeta} + \cdots + h_L e^{iL\zeta}$, $\zeta \in \mathbb{R}$ is the DTFT of $\mathbf{h}$.

It follows that $|y_{m,n}| \leq |H(-m\omega T_s)|\,|x_{m,n}|$, which is why the we define **Gain and Phase Rotation** the *gain* or *attenuation* of this transformation as $G(\zeta) = |H(\zeta)|$ and *phase rotation* as $\Theta(\zeta) = \theta$, where $H(\zeta) = |H(\zeta)|\, e^{i\theta}$. So $G(\zeta)$ measures how the amplitude of this mode of the signal $x(t)$ is changed by the filter and $\Theta(\zeta)$ measures how the argument of this mode is shifted. Note that $H(\zeta)$ is periodic of period 2π and and if the filter $\mathbf{h}$ is real, then $|H(-\zeta)| = \left|\overline{H(\zeta)}\right| = |H(\zeta)|$.

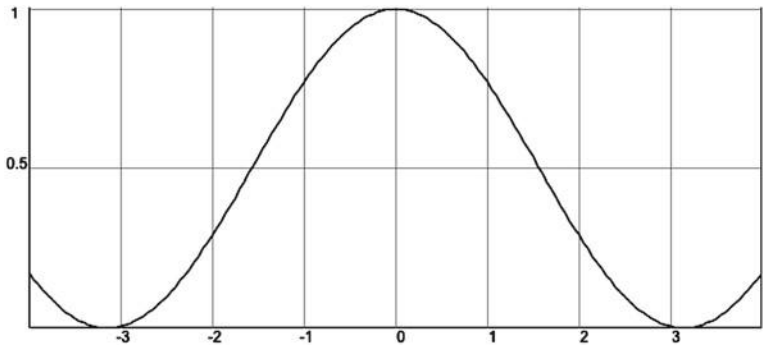


Fig. 6.6: Gain function $G(\zeta) = (1 + \cos\zeta)/2$, $-4 \leq \zeta \leq 4$, for filter of Example 2.71.

OK, now let's take a closer look at an example in light of earlier discussion. We will illustrate these ideas as applied to Example 2.71 of Section 2.8. That example computed as output a weighted average of samples by using what we can now describe as the FIR filter $\mathbf{h} = \left\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right\}$.

Example 6.24. The function $f(t) = \cos(\pi t)$, $-1 \leq t \leq 1$ defines the exact signal that we want to sample, but we actually sample this signal plus noise, namely the function $g(t) = \cos(\pi t) + \frac{1}{5}\sin(24\pi t) + \frac{1}{4}\cos(30\pi t)$. Assume that sampling is at the equally spaced points $t_k = -1 + \frac{2}{64}k$, $k = 0, 1, \ldots, 64$, yielding data points $x_k = g(t_k)$. How effective is the length two FIR filter $\mathbf{h} = \left\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right\}$ in removing noise? Compute the DTFT of the filter of Example 2.71 and use it to explain the behavior displayed in Figure 2.10 (page 170).

Solution. We saw from the graph of the exact data, noisy data and filtered data in Figure 2.10 that, although it is somewhat crude, it appears to do a decent job of filtering out the noise in the sampled signal $g(t)$. Specialize the general argument above to $c_m e^{im\omega t}$ with $x_{m,n} = c_m e^{im\omega nT_s}$, so that $T = 2$, $\omega = 2\pi/2 = \pi$, $T_s = 2/64 = 1/32$ and $\zeta = m\omega T_s = m\pi/32$. Use the fact that $\cos\zeta = 2\cos^2\left(\frac{\zeta}{2}\right) - 1$ to obtain that the DTFT of this filter is

$$\begin{aligned} H(\zeta) =& \frac{1}{4} + \frac{1}{2}e^{i\zeta} + \frac{1}{4}e^{i2\zeta} = \left(\frac{1}{2}\left(1 + e^{i\zeta}\right)\right)^2 \\ =& \left(e^{i\zeta/2}\left(\frac{e^{-i\zeta/2} + e^{i\zeta/2}}{2}\right)\right)^2 = e^{i\zeta}\cos^2(\zeta/2) \\ =& e^{i\zeta}\frac{(1+\cos\zeta)}{2}. \end{aligned}$$

If follows from this that the gain or attenuation of this transformation is $G(\zeta) = |H(\zeta)| = (1+\cos\zeta)/2$ and phase rotation as $\Theta(\zeta) = \zeta$. See Figure 6.6 for a plot of $G(\zeta)$. We use these functions to describe the effects of this filter on our sampled data: One calculates that $G(\pi/32) \doteq 0.9976$, $G(24\pi/32) \doteq 0.1464$, and $G(30\pi/32) \doteq 0.0096$. Thus, the amplitude of the low frequency component of the signal generated by $g(t)$ is not changed by much, while the amplitudes of the high frequency components of this signal are considerably damped. In addition, the phase rotation of $\Theta(\pi/32) \doteq 0.098$ explains the slight forward shift of filtered low frequency values observed in Figure 2.10. □

This discussion inspires a first attempt at a formal definition of lowpass and highpass FIR filters:

Definition 6.22. The FIR filter $\mathbf{h} = \{h_k\}_{k=0}^{L}$ with discrete time Fourier transform $H(\zeta)$ is a *lowpass filter* if $|H(0)| = 1$ and $|H(\pi)| = 0$; $\mathbf{h}$ is a *highpass filter* if $|H(0)| = 0$ and $|H(\pi)| = 1$.

The idea behind this definition is that a lowpass filter will not modify the amplitude of sinusoidal components of low frequency (ζ near 0) modes by much, whereas it will significantly dampen the amplitude of high frequency (ζ near π) modes. Likewise, highpass filters tend to do the opposite. However, things are a bit tricky here: High frequency modes don't have to be near π. In fact, they can be too high, e.g., near 2π, in which case they will not be damped out since $H(2\pi) = H(0) = 1$. Thus, we're confronted with an important issue: How high is "too high"? There is a beautiful answer to this question, known as the Nyquist-Shannon sampling theorem, that asserts roughly in our situation that if a function $x(t)$ contains no frequencies higher than F then it can be completely determined by any sampling frequency $f_s > 2F$. The frequency $2F$ is called the *Nyquist sampling rate*. If this condition fails, then the corresponding signal will contain imperfections such as aliasing, in which a high frequency mode is misinterpreted as a lower frequency mode.

Nyquist Rate

Example 6.25. Explain what the Nyquist-Shannon sampling theorem says about Example 6.24.

Solution. The frequencies of the three modes in the noisy signal $g(t)$ from low to high are 1/2, 12 and 15. So we can take $F = 15$. On the other hand, the sampling rate that we used had a sampling period of $T_s = 1/32$, hence

sampling frequency $f_s = 1/T_s = 32$. Since $2F = 30 < 32 = f_s$, the sampling rate is sufficient to recover the noisy signal, so we can be confident that the source of errors in the filtered signal is not due to aliasing. □

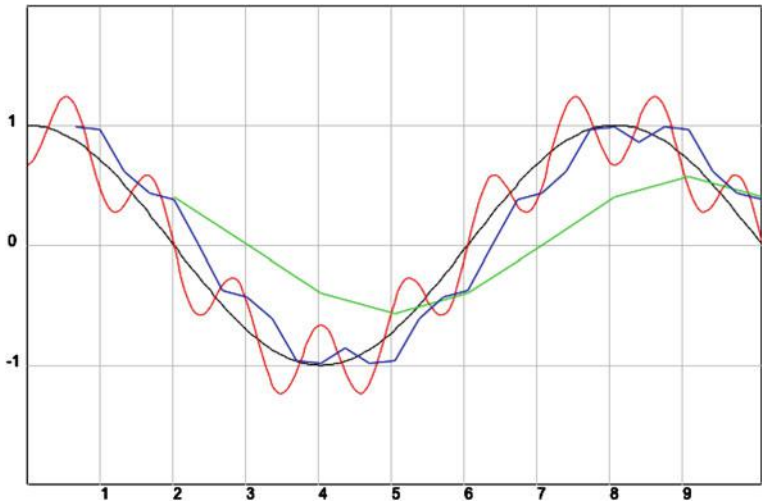


Fig. 6.7: Graph of data from Example 6.26: Exact (——), noisy (——), filtered (——) with $T_s = 1$ and filtered (——) with $T_s = \frac{1}{3}$.

Our last example is one in which the phenomenon of aliasing is an issue that leads to a curious result.

Example 6.26. Repeat the calculations of Example 2.71 with exact signal $f(t) = \cos\left(\frac{\pi}{4}t\right)$, $0 \leq t \leq 10$, noisy signal $g(t) = \cos\left(\frac{\pi}{4}t\right) - \frac{1}{3}\cos\left(\frac{7\pi}{4}t\right)$, sampling at the equally spaced points $t_k = k$, $k = 0, 1, \ldots, 10$ and also at sampling points $t_k = \frac{k}{3}$, $k = 0, 1, \ldots, 30$. Interpret the results.

Solution. A graph of the exact data, noisy data and the two filtered datasets is given in Figure 6.7. That the graph with with $T_s = 1$ is much less accurate than that with $T_s = \frac{1}{3}$ is no surprise. What is curious about the graph of the coarser sampling $T_s = 1$ is that it is rather smooth and has a shape similar to that of the exact signal if it were dampened and shifted slightly to the right, while the finer sampling is more accurate but has the oscillations one would expect from an imperfect attempt to filter out high frequencies.

The reason for this curiosity becomes clear upon examination of the low and high frequency curves in Figure 6.8. Notice that at the coarse sampling points with $T_s = 1$, the graphs of $f(t) = \cos\left(\frac{\pi}{4}t\right)$ and $h(t) = \cos\left(\frac{7\pi}{4}t\right)$ intersect. Thus, they cannot be distinguished with this sampling, which is exactly the aliasing phenomenon alluded to earlier. This explains the apparent smoothness the filtered data of the coarser sampling exhibits. These results are consistent with the Nyquist-Shannon theorem, since the noisy data has

maximum frequency $F = \frac{7}{8}$ while the sampling frequency $f_s = \frac{1}{T_s} = 1$ yields $f_s < \frac{7}{4} = 2F$. On the other hand, the sampling period of $T_s = \frac{1}{3}$ yields a sampling frequency of $3 > \frac{7}{4} = 2F$, so the error in that filtered data is not due to aliasing. □

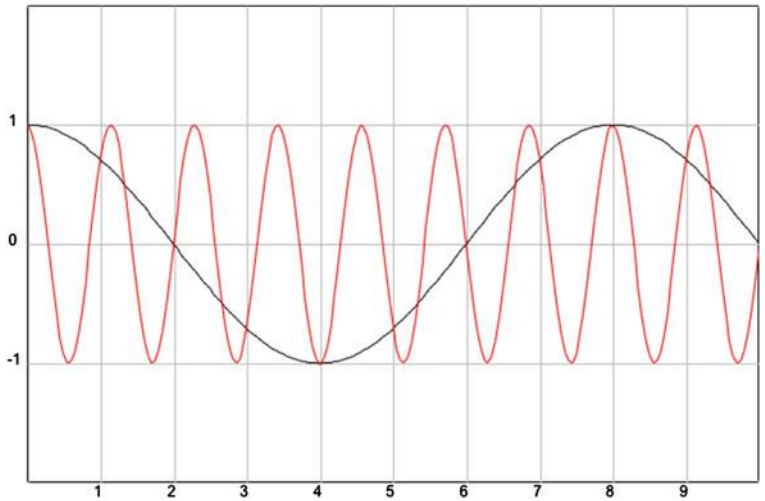


Fig. 6.8: Graphs of $f(t) = \cos\left(\frac{\pi}{4}t\right)$ (——) and $h(t) = \cos\left(\frac{7\pi}{4}t\right)$ (——).

In addition to [10] another excellent source for further study of the connections between Fourier analysis and digital filtering is the text [11] by C. Gasquet and P. Witomski.

6.6 Exercises and Problems

Exercise 1. Compute the DTFT for the FIR filter $\mathbf{h} = \left\{\frac{1}{2}, \frac{1}{2}\right\}$ and confirm that $\mathbf{h}$ is a lowpass filter.

Exercise 2. Compute the DTFT for the FIR filter $\mathbf{h} = \left\{-\frac{1}{2}, \frac{1}{2}\right\}$ and confirm that $\mathbf{h}$ is a highpass filter.

Exercise 3. Compute the Fourier series for $x(t) \in C^1_{PW}[-\pi, \pi]$, where $x(t) = t^2/\pi$, $-\pi \leq t \leq \pi$ and graph $x(t)$ and the partial Fourier sums with $N = 3, 6$.

Exercise 4. Compute the Fourier series for $x(t) \in C^1_{PW}[0, 2\pi]$, where $x(t) = t^3/6 - 1$, $0 \leq t < 2\pi$, and $x(2\pi) = -1$ and graph results as in Exercise 3.

Exercise 5. Apply the filter of Exercise 1 to the sampling problem of Example 6.24 and graph the results as in Figure 2.10. Is the filter effective?

Exercise 6. Apply the filter of Exercise 2 to the sampling problem of Example 6.24 and graph the results as in Figure 2.11. Is the filter effective?

Exercise 7. Apply the filter of Example 6.25 to the sampling problem of that example with sampling rates of $T_s = 15, 30, 45$ and compute the infinity norm of the vector of differences between exact and filtered noisy values in each case.

Exercise 8. Apply the filter of Example 6.25 to the sampling problem of Example 6.26 with sampling rates of $T_s = 3/7, 8/7, 10/7$ and compute the infinity norm of the vector of differences between exact and filtered noisy values in each case.

Problem 9. Verify Theorem 6.18.

*Problem 10. Deduce the form of the Fourier series of a real-valued function $x(t) \in C^1_{PW}[0,T]$ in equation (6.5) from the general form of equation (6.4).

Problem 11. Let $x(t) \in C_{PW}[0,T]$ have Fourier series $\sum_{n=-\infty}^{\infty} c_n e^{in\omega t}$. Use Corollary 6.2 and Problem 18 of Section 6.3 to prove the following:

(Bessel's inequality) For any integer $N > 0$, $\sum_{n=-N}^{N} |c_n|^2 \leq \frac{1}{T}\int_0^T |x(t)|^2\, dt$.

Problem 12. Let $x(t) \in C^1_{PW}[0,T]$ have Fourier series $\sum_{n=-\infty}^{\infty} c_n e^{in\omega t}$. Use Corollary 6.2, Problem 18 of Section 6.3 and Dirichlet's theorem to prove the following:

(Parseval's equality) $\sum_{n=-\infty}^{\infty} |c_n|^2 = \frac{1}{T}\int_0^T |x(t)|^2\, dt$.

6.7 *Projects and Reports

Project: Testing Least Squares Solvers

The object of this project is to test the quality of the solutions of three different methods for solving least squares problems $A\mathbf{x} = \mathbf{b}$ using a technology tool:

1. Solution by solving the normal equations by Gaussian elimination.
2. Solution by reduced QR factorization obtained by Gram–Schmidt.
3. Solution by full QR factorization by Householder matrices.

Here is the test problem: suppose we want to approximate the curve $f(x) = e^{\sin(6x)}$, $0 \leq x \leq 1$, by a tenth-degree polynomial. The input data will be the sampled values of $f(x)$ at equally spaced nodes $x_k = kh$, $k = 0, 1, \ldots, 20$, $h = 0.05$. This gives 21 equations $f(x_k) = c_0 + c_1 x_k + \cdots + c_{10} x_k^{10}$ for the 11 unknown coefficients c_k, $k = 0, 1, \ldots, 20$. The coefficient matrix that results from this problem is called a Vandermonde matrix.

Procedure: First set up the system matrix A and right-hand-side vector $\mathbf{b}$. The built-in procedure for computing a QR factorization will very likely be Householder matrices, which will take care of (3). You will need to check

the documentation to verify this. The Gram–Schmidt method of finding QR factorization may have to be programmed by you.

Once you have solved the system by these three methods, make out a table that has the computed coefficients for each of the three methods. Then make plots of the difference between the function $f(x)$ and the computed polynomial for each method. Relate these results to the condition numbers of the matrices constructed. Discuss your results.

There are a number of good texts that discuss numerical methods for least squares; see, e.g., references [8], [9], [14]. More advanced treatments can be found in references [1] and [15]. Or you can read from the master himself in [12] (Gauss's original text conveniently translated from the Latin with a very enlightening supplement by G. W. Stewart).

Report: Approximation Theory

Suppose you work for a manufacturer of calculators, and are involved in the design of a new calculator. The problem is this: As one of the "features" of this calculator, the designers decided that it would be nice to have a key that calculated a transcendental function, namely, $f(x) = \sin(x)$. Your job is to come up with an adequate way of calculating $f(x)$, say with an error no worse than 0.0001.

Polynomials are a natural idea for approximating functions. From a designer's point of view they are particularly attractive because they are so easy to implement. Given the coefficients of a polynomial, it is easy to design a very efficient and compact algorithm for calculating values of the polynomial. Such an algorithm would fit nicely into a small ROM for the calculator, or could even be microcoded into the chip.

Your task is to find a low-degree polynomial that approximates $\sin(x)$ to within the specified accuracy on a suitable interval. Beyond that interval you may use properties of the sin function to finish the job. For comparison, find a Taylor polynomial of lowest degree for $\sin x$ that gives sufficient accuracy. Next, use the projection problem idea to project the function $\sin x$, which is viewed as a member of a suitable function space with the standard inner product, into the subspace $\mathcal{P}_n$ of polynomials functions on that interval of degree at most n. Is it of lower degree than the best Taylor polynomial approximation?

Use a technology tool to do the computations and graphics. Then report on your findings. Include graphs that will be helpful in interpreting your conclusions. Also, give suggestions on how to compute this polynomial efficiently.

Report: Fourier Analysis

In the first part of this report you will write a brief introduction to Fourier analysis in which you exhibit formulas for the Fourier coefficients of a real-valued function $f(t)$ and explain the form and meaning of the projection formula in this setting.

In the second part you will explore the quality of these approximations for the following test functions. The functions are specified on the interval $(-\pi, \pi)$, given common values at the endpoints, and then each graph is replicated on

adjacent intervals of length 2π, yielding periodic functions:

$$(1)\ h(t) = \frac{t^2}{\pi}, \quad (2)\ g(t) = \frac{e^t}{\pi}, \quad (3)\ f(t) = t^3/6 - 1.$$

Notice that the second and third functions are discontinuous periodic functions.

For each test function you should compute explicit formulas for the Fourier series of the function, if possible. Then prepare a graph that includes the test function and at least two of its partial Fourier sums. You will need a technology tool to carry out some of the calculations and the graphs. Discuss the quality of the approximations and report on any conclusions that you can draw from this data. In particular, note the curious behavior of these partial Fourier sums near a discontinuity. There is a name for this phenomenon, the "Gibbs phenomenon". Research this topic and confirm it with your graphs. Finally, consider adding a few comments about the curious connection between Fourier series and the epicycles of Ptolomy as a cultural footnote to your report.

Symbols

Symbol	Meaning	Reference
$\emptyset$	Empty set	Page 12
$\in$	Member symbol	Page 12
$\subseteq$	Subset symbol	Page 12
$\subset$	Proper subset symbol	Page 12
$\cap$	Intersection symbol	Page 12
$\cup$	Union symbol	Page 12
$\otimes$	Tensor symbol	Page 161
$\oplus$	Direct sum symbol	Page 234
$\approx$	Approximate equality sign	Page 92
$\overrightarrow{PQ}$	Displacement vector	Page 183
$\mid z \mid$	Absolute value of complex z	Page 15
$\mid A \mid$	determinant of matrix A	Page 141
$\mid\mid \mathbf{u} \mid\mid$	Norm of vector $\mathbf{u}$	Page 278
$\mid\mid \mathbf{u} \mid\mid_p$	p-norm of vector $\mathbf{u}$	Page 392
$\mathbf{u} \cdot \mathbf{v}$	Standard inner product	Page 282
$\langle \mathbf{u}, \mathbf{v} \rangle$	Inner product	Page 399
A_{cof}	Cofactor matrix of A	Page 149
$\operatorname{adj} A$	Adjoint of matrix A	Page 149
A^*	Conjugate (Hermitian) transpose of matrix A	Page 107
A^T	Transpose of matrix A	Page 107
$\mathcal{C}(A)$	Column space of matrix A	Page 220
$\operatorname{cond}(A)$	Condition number of matrix A	Page 426
$C[a, b]$	Function space	Page 187
$\mathbb{C}$	Complex numbers $a + bi$	Page 14
$\mathbb{C}^n$	Standard complex vector space	Page 185
$\operatorname{comp}_{\mathbf{v}} \mathbf{u}$	Component	Page 291
$\overline{z}$	Complex conjugate of z	Page 16
δ_{ij}	Kronecker delta	Page 75
$\dim V$	Dimension of space V	Page 232
$\det A$	Determinant of A	Page 141
$\operatorname{domain}(T)$	Domain of operator T	Page 226
$\operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$	Diagonal matrix with $\lambda_1, \lambda_2, \ldots, \lambda_n$ along diagonal	Page 120
E_{ij}	Elementary operation switch ith and jth rows	Page 29
$E_i(c)$	Elementary operation multiply ith row by c	Page 29
$E_{ij}(d)$	Elementary operation add d times jth row to ith row	Page 29

Symbol	Meaning	Reference
$\mathcal{E}_\lambda(A)$	Eigenspace	Page 334
$H_\mathbf{v}$	Householder matrix	Page 306
I, I_n	Identity matrix, $n \times n$ identity	Page 75
id_V	Identity function for V	Page 192
$\Im(z)$	Imaginary part of z	Page 14
$\ker(T)$	Kernel of operator T	Page 225
$M_{ij}(A)$	Minor of A	Page 143
$M(A)$	Matrix of minors of A	Page 149
$\max\{a_1, a_2, \ldots, a_m\}$	Maximum value	Page 45
$\min\{a_1, a_2, \ldots, a_m\}$	Minimum value	Page 45
$\mathcal{N}(A)$	Null space of matrix A	Page 221
$\mathbb{N}$	Natural numbers $1, 2, \ldots$	Page 13
$\mathrm{null}A$	Nullity of matrix A	Page 44
$\mathcal{P}$	Space of polynomials of any degree	Page 200
$\mathcal{P}_n$	Space of polynomials of degree $\leq n$	Page 200
$\mathrm{proj}_\mathbf{v}\mathbf{u}$	Projection vector along a vector	Page 291
$\mathrm{proj}_V\mathbf{u}$	Projection vector into subspace	Page 412
$\mathbb{Q}$	Rational numbers a/b	Page 13
$\Re(z)$	Real part of z	Page 14
$\mathcal{R}(A)$	Row space of matrix A	Page 221
$R(\theta)$	Rotation matrix	Page 215
$\mathbb{R}$	Real numbers	Page 14
$\mathbb{R}^n$	Standard real vector space	Page 184
$\mathbb{R}^{m,n}$	Space of $m \times n$ real matrices	Page 187
T_A	Matrix operator associated with A	Page 83
$\mathrm{range}(T)$	Range of operator T	Page 226
$\mathrm{rank}A$	Rank of matrix A	Page 44
$\rho(A)$	Spectral radius of A	Page 354
$\mathrm{span}\{S\}$	Span of vectors in S	Page 201
$\sup\{\mathrm{E}\}$	Supremum of set E of reals	Page 424
$\mathrm{target}(T)$	Target of operator T	Page 226
$[T]_{B,C}$	Matrix of operator T	Page 249
$V^\perp$	Orthogonal complement of V	Page 418
$\mathbb{Z}$	Integers $0, \pm1, \pm2, \ldots$	Page 13

Solutions to Selected Exercises

Section 1.1, Page 10

1 (a) $x = -1$, $y = 1$ (b) $x = 2$, $y = -2$, $z = 1$ (c) $x = 2$, $y = 1$

3 (a) linear, $x - y - z = -2$, $3x - y = 4$ (b) nonlinear (c) linear, $x + 4y = 0$, $2x - y = 0$, $x + y = 2$

5 (a) $m = 3$, $n = 3$, $a_{11} = 1$, $a_{12} = -2$, $a_{13} = 1$, $b_1 = 2$, $a_{21} = 0$, $a_{22} = 1$, $a_{23} = 0$, $b_2 = 1$, $a_{31} = -1$, $a_{32} = 0$, $a_{33} = 1$, $b_3 = 1$ (b) $m = 2$, $n = 2$, $a_{11} = 1$, $a_{12} = -3$, $b_1 = 1$, $a_{21} = 0$, $a_{22} = 1$, $b_2 = 5$

7 $\frac{47}{25}y_1 - y_2 = 0$, $-y_1 + \frac{47}{25}y_2 - y_3 = 0$, $-y_2 + \frac{47}{25}y_3 - y_4 = 0$, $-y_3 + \frac{47}{25}y_4 = 50$

9 $p_1 = 0.2p_1 + 0.1p_2 + 0.4p_3$, $p_2 = 0.3p_1 + 0.3p_2 + 0.2p_3$, $p_3 = 0.1p_1 + 0.2p_2 + 0.1p_3$

11 $x_1 = x_2 = x_3 = x_4 = 0$

13 One solution is $x_1 = 1$, $2x_2 = -2$, $3x_1 = 3$.

18 Counting inflow as positive, the equation for vertex v_1 is $x_1 - x_4 - x_5 = 0$.

Section 1.2, Page 22

1 (a) $\{0, 1\}$ (b) $\{x \,|\, x \in \mathbb{Z} \text{ and } x > 1\}$ (c) $\{x \,|\, x \in \mathbb{Z} \text{ and } x \leq -1\}$ (d) $\{0, 1, 2, \ldots\}$ (e) A

3 (a) $\mathrm{e}^{3\pi\mathrm{i}/2}$ (b) $\sqrt{2}\mathrm{e}^{\pi\mathrm{i}/4}$ (c) $2\mathrm{e}^{2\pi\mathrm{i}/3}$ (d) $E^{0\mathrm{i}}$or 1 (e) $2\sqrt{2}\mathrm{e}^{7\pi\mathrm{i}/4}$ (f) $2\mathrm{e}^{\pi\mathrm{i}/2}$ (g) $\pi\mathrm{e}^{0\mathrm{i}}$

5 (a) $1 + 8\mathrm{i}$ (b) $10 + 10i$ (c) $\frac{3}{5} + \frac{4}{5}\mathrm{i}$ (d) $-\frac{3}{5} - \frac{4}{5}\mathrm{i}$ (e) $42 + 7\mathrm{i}$

7 (a) $\frac{6}{5} - \frac{8}{5}\mathrm{i}$, (b) $\pm\sqrt{2} \pm \mathrm{i}\sqrt{2}$, (c) $z = 1$ (d) $z = -1$, $\pm\mathrm{i}$

9 (a) $\frac{1}{2} + \frac{1}{2}\mathrm{i} = \frac{1}{2}\sqrt{2}\mathrm{e}^{\pi\mathrm{i}/4}$ (b) $-1 - \mathrm{i}\sqrt{3} = 2\mathrm{e}^{4\pi\mathrm{i}/3}$ (c) $-1 + \mathrm{i}\sqrt{3} = 2\mathrm{e}^{2\pi\mathrm{i}/3}$ (d) $-\frac{1}{2}\mathrm{i} = \frac{1}{2}\mathrm{e}^{3\pi\mathrm{i}/2}$ (e) $\mathrm{i}\mathrm{e}^{\pi/4} = \mathrm{e}^{\pi/4}\mathrm{e}^{\pi\mathrm{i}/2}$

11 (a) $z = \frac{-1}{2} \pm \frac{\sqrt{11}}{2}i$, (b) $z = \pm\frac{\sqrt{3}}{2} + \frac{1}{2}\mathrm{i}$ (c) $z = 1 \pm (\frac{-\sqrt{2\sqrt{2}+2}}{2} - \frac{\sqrt{2\sqrt{2}-2}}{2}\mathrm{i})$ (d) $\pm 2\mathrm{i}$

13 (a) Circle of radius 2, center at origin (b) $\Re(z) = 0$, the imaginary axis (c) Interior of circle of radius 1, center at $z = 2$.

15 $\overline{2 + 4\mathrm{i}} + \overline{1 - 3\mathrm{i}} = 2 - 4\mathrm{i} + 1 + 3\mathrm{i} = 3 - \mathrm{i}$ and $\overline{(2 + 4\mathrm{i}) + (1 - 3\mathrm{i})} = \overline{3 + \mathrm{i}} = 3 - \mathrm{i}$

17 $z = 1 \pm \mathrm{i}$, $(z - (1 + \mathrm{i}))\,(z - (1 - \mathrm{i})) = z^2 - 2z + 2$

19 (a) $g(x,y) = (x-2)^2 - y^2$, $h(x,y) = 2(x-2)y$, (b) $g(x,y) = x^3 - 3xy^2 - 2x + 1$, $h(x,y) = 3x^2y - y^3 - 2y$

23 Use $|z|^2 = z\bar{z}$ and $\overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$.

26 Write $p(w) = a_0 + a_1 w + \cdots + a_n w^n = 0$ and conjugate both sides.

Section 1.3, Page 34

1 (a) Size 2×4, $a_{11} = a_{14} = a_{23} = a_{24} = 1$, $a_{12} = -1$, $a_{21} = -2$, $a_{13} = a_{22} = 2$ (b) Size 3×2, $a_{11} = 0$, $a_{12} = 1$, $a_{21} = 2$, $a_{22} = -1$, $a_{31} = 0$, $a_{32} = 2$ (c) Size 2×1 , $a_{11} = -2$, $a_{21} = 3$ (d) Size 1×1, $a_{11} = 1 + \mathrm{i}$

3 (a) 2×3 augmented matrix $\begin{bmatrix} 2 & 3 & 7 \\ 1 & 2 & -2 \end{bmatrix}$, $x = 20$, $y = -11$ (b) 3×4 augmented matrix $\begin{bmatrix} 3 & 6 & -1 & -4 \\ -2 & -4 & 1 & 3 \\ 0 & 0 & 1 & 1 \end{bmatrix}$, $x_1 = -1 - 2x_2$, x_2 free, $x_3 = 1$, (c) 3×3 augmented matrix $\begin{bmatrix} 1 & 1 & -2 \\ 5 & 2 & 5 \\ 1 & 2 & -7 \end{bmatrix}$, $x_1 = 3$, $x_2 = -5$

5 (a) $x_1 = 1 - x_2$, $x_3 = -1$, x_2 free (b) $x_1 = -1 - 2x_2$, $x_3 = -2$, $x_4 = 3$, x_2 free (c) $x_1 = 3 - 2x_3$, $x_2 = -1 - x_3$, x_3 free (d) $x_1 = 1 + \frac{2}{3}\mathrm{i}$, $x_2 = 1 - \frac{1}{3}\mathrm{i}$ (e) $x_1 = \frac{7}{11}x_4$, $x_2 = \frac{-2}{11}x_4$, $x_3 = \frac{6}{11}x_4$, x_4 free

7 (a) $x_1 = 4$, $x_3 = -2$, x_2 free (b) $x_1 = 1$, $x_2 = 2$, $x_3 = 2$ (c) Inconsistent system (d) $x_1 = 1$, x_2 and x_3 free

9 Augmented matrix is $\begin{bmatrix} 1 & -1 & -1 & 0 & 0 \\ -1 & 1 & -1 & 0 & 0 \\ -1 & 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{bmatrix}$ with RREF $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$. Hence the only solution is the trivial solution, which ranks all pages as equal in importance (not likely a useful ranking).

11 (a) $x_1 = \frac{2}{3}b_1 + \frac{1}{3}b_2$, $x_2 = \frac{-1}{3}b_1 + \frac{1}{3}b_2$ (b) Inconsistent if $b_2 \neq 2b_1$, otherwise solution is $x_1 = b_1 + x_2$ and x_2 arbitrary. (c) $x_1 = \frac{1}{4}(2b_1 + b_2)(1 - \mathrm{i})$, $x_2 = \frac{1}{4}(\mathrm{i}b_2 - 2b_1)(1 - \mathrm{i})$

13 Augmented matrix with three right-hand sides reduces to $\begin{bmatrix} 1 & 0 & 2 & -1 & 1 \\ 0 & 1 & 1 & -1 & -1 \end{bmatrix}$ given solutions (a) $x_1 = 2$, $x_2 = 1$ (b) $x_1 = -1$, $x_2 = -1$ (c) $x_1 = 1$, $x_2 = -1$.

15 The only solution is the trivial solution $p_1 = 0$, $p_2 = 0$, and $p_3 = 0$, which is not meaningful since entries are not positive.

17 (a) $x = 0$, $y = 0$ or divide by xy and get $y = -8/5$, $x = 4/7$ (b) Either two of x, y, z are zero and the other arbitrary or all are nonzero, divide by xyz and obtain $x = -2z$, $y = z$, and z is arbitrary nonzero.

19 With polynomial $p(x) = x_1 + x_2 x + x_3 x^2$ coefficient matrix is $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$, augmented matrix is $\begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 1 & 2 \\ 1 & 2 & 4 & 4 \end{bmatrix}$, solution polynomial is $p(x) = 2 - x + x^2$.

21 Suppose not and consider such a solution (x, y, z, w). At least one variable is positive and largest. Now examine the equation corresponding to that variable.

23 (a) Equation for $x_2 = 1/2$ is $a + b \cdot 1/2 + c \cdot (1/2)^2 = e^{1/2}$.

26 Let x_i be the importance of location i, $i = 1, 2, 3, 4, 5$ and the result of solving the system is that x_5 is free, so take it to be 1 and deduce that $x_1 = 2/3$ and $x_2 = x_3 = x_4 = 4/3$.

Section 1.4, Page 48

1 (a), (e), (f), and (h) are in reduced row echelon form, (b) and (d) are in reduced row form. Leading entries (a) $(1,2)$ (b) $(1,1)$, $(2,2)$, $(3,4)$ (c) $(1,2)$, $(2,1)$ (d) $(1,1)$, $(2,2)$ (e) $(1,1)$ (f) $(1,1)$, $(2,2)$, $(3,3)$ (g) $(1,1)$, $(3,3)$ (h) $(1,1)$.

3 (a) 3 (b) 0 (c) 3, (d) 1 (e) 1

5 (a) $E_{21}(-1)$, $E_{31}(-2)$, $E_{32}(-1)$, $E_2\left(\frac{1}{4}\right)$, $E_{12}(1)$, $\begin{bmatrix} 1 & 0 & \frac{5}{2} \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}$, rank 2, nullity 1 (b) $E_{21}(1)$, $E_{23}(-15)$, $E_{13}(-9)$, $E_{12}(-1)$, $E_1\left(\frac{1}{3}\right)$, $\begin{bmatrix} 1 & 0 & 0 & \frac{17}{3} \\ 0 & 1 & 0 & -33 \\ 0 & 0 & 1 & 2 \end{bmatrix}$, rank 3, nullity 1 (c) E_{12}, $E_1\left(\frac{1}{2}\right)$, $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$, rank 2, nullity 2 (d) $E_1\left(\frac{1}{2}\right)$, $E_{21}(-4)$, $E_{31}(-2)$, $E_{32}(1)$, $E_{12}(-2)$, $\begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$, rank 2, nullity 1 (e) E_{12}, $E_{21}(-2)$, $E_2\left(\frac{1}{9}\right)$, $E_{12}(2)$ $\begin{bmatrix} 1 & 1 & 0 & \frac{22}{9} \\ 0 & 0 & 1 & \frac{2}{9} \end{bmatrix}$, rank 2, nullity 2 (f) E_{12}, $E_{21}(-2)$, $E_{31}(-1)$, E_{23}, $E_2(-1)$, $E_{32}(3)$, $E_3\left(\frac{-1}{4}\right)$, $E_{23}(1)$, $E_{13}(-1)$, $E_{12}(-2)$, $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, rank 3, nullity 0

7 Systems are not equivalent since (b) has trivial solution, (a) does not. (a) $\operatorname{rank}\left(\widetilde{A}\right) = 2$, $\operatorname{rank}(A) = 2$, $\{(-1 + x_3 + x_4, 3 - 2x_2, x_3, x_4) \mid x_3, x_4 \in \mathbb{R}\}$ (b) $\operatorname{rank}\left(\widetilde{A}\right) = 3$, $\operatorname{rank}(A) = 3$, $\{(-2x_2, x_2, 0, 0) \mid x_2 \in \mathbb{R}\}$

9 $0 \leq \operatorname{rank}(A) < 3$

11 (a) Infinitely many solutions for all c (b) Inconsistent for all c (c) Inconsistent if $c = -2$, infinitely many solutions if $c = 1$, unique solution otherwise.

13 Rank of augmented matrix equals rank of coefficient matrix independently of right-hand side, so system is always consistent. Solution is $x_1 = -a + 2b - c + 4x_4$, $x_2 = -b + a + \frac{1}{2}c - 2x_4$, $x_3 = \frac{1}{2}c - x_4$, x_4 free.

15 (a) 3 (b) solution set (c) $E_{23}(-5)$ (d) 0 or 1

17 Two elementary row ops yield reduced row form $\begin{bmatrix} 1 & 2 & 1 & 0 \\ 0 & -3 & -3 & 0 \\ 0 & 0 & 0 & b + \frac{a}{3} \end{bmatrix}$. If $b = -\frac{a}{3}$, there are infinitely many solutions, otherwise there are no solutions.

19 (a) false, $0x = 1$ (b) true, has trivial solution (c) false, could be inconsistent (d) false, $[1,0]$ and $[2,0]$ (e) false, $\begin{bmatrix} 1 \\ 2 \end{bmatrix} x = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

22 Consider what you need to do to go from reduced row form to reduced row echelon form.

Section 1.5, Page 59

1 Flop count is 6.

3 Answer will depend on your calculator. Octave and ALAMA calculator yield $4.7740e{-}15$, which is close to, but not equal to the answer 0.

6 Work of jth stage: $2 + 2(n - j)$. Add them up.

8 Implicit Euler as linear equation: $-\sigma y_{i-1,j+1} + (1 + 2\sigma) y_{i,j+1} - \sigma y_{i+1,j+1} = y_{i,j} + kf(x_i, t_{j+1})$.

Section 2.1, Page 70

1 (a) $\begin{bmatrix} -2 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 4 \\ -1 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 8 \\ 6 & 3 \end{bmatrix}$ (d) not possible (e) $\begin{bmatrix} 7 & 4 & -1 \\ 10 & 4 & 4 \\ 2 & 4 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} x-2+4y \\ 3x-2+y \\ -1 \end{bmatrix}$

3 (a) not possible (b) $\begin{bmatrix} -1 & -3 & -2 \\ -4 & -1 & 4 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & 2 \end{bmatrix}$ (d) not possible (e) $\begin{bmatrix} 5 & 8 & 3 \\ 13 & 5 & -6 \end{bmatrix}$

5 (a) $x\begin{bmatrix} 1 \\ 2 \end{bmatrix} + y\begin{bmatrix} 2 \\ 0 \end{bmatrix} + z\begin{bmatrix} 0 \\ -1 \end{bmatrix}$ (b) $x\begin{bmatrix} 1 \\ 2 \end{bmatrix} + y\begin{bmatrix} -1 \\ 3 \end{bmatrix}$ (c) $x\begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} + y\begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} + z\begin{bmatrix} 0 \\ -1 \\ 5 \end{bmatrix}$ (d) $x\begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix} + y\begin{bmatrix} -3 \\ 0 \\ 2 \end{bmatrix} + z\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$

7 $a = \frac{-2}{3}$, $b = \frac{2}{3}$, $c = \frac{-4}{3}$

9 $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = a\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + c\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + d\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

11 $A + (B+C) = \begin{bmatrix} -1 & 2 & -3 \\ 5 & 1 & 5 \end{bmatrix} = (A+B)+C$, $A+B = \begin{bmatrix} 0 & 2 & -2 \\ 4 & 2 & 5 \end{bmatrix} = B+A$

13 $R = \begin{bmatrix} 1 & -1 & 0 & -3 & 4 \\ 0 & 0 & 1 & 1 & 2 \end{bmatrix}$

18 Solve for A in terms of B with the first equation and deduce $B = \frac{1}{4}I$.

Section 2.2, Page 80

1 (a) $[11+3i]$, (b) $\begin{bmatrix} 6 & 8 \\ 3 & 4 \end{bmatrix}$, (c) impossible (d) impossible (e) $\begin{bmatrix} 15+3\mathrm{i} & 20+4\mathrm{i} \\ -3 & -4 \end{bmatrix}$ (f) impossible (g) [10] (h) impossible

3 (a) $\begin{bmatrix} 1 & -2 & 4 & 0 \\ 0 & 1 & -1 & 0 \\ -1 & 0 & 0 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$

(b) $\begin{bmatrix} 1 & -1 & -3 \\ 2 & 2 & 4 \\ -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 10 \\ 3 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & -3 \\ 0 & 2 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$

5 $\begin{bmatrix} 10 & -1 & 1 \\ 2 & -4 & -2 \\ 4 & 2 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$

7 (a) $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 3 \\ 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ -4 \\ -3 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 1 \\ 3 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ 2\mathrm{i} \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 3 \\ 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ -3x_2 \\ x_3 \end{bmatrix}$ or $\begin{bmatrix} 1 & 0 & 1 \\ 1 & -3 & 3 \\ 0 & -3 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

9 $f(A) = \begin{bmatrix} 3 & 4 \\ 2 & 5 \end{bmatrix}$, $g(A) = \begin{bmatrix} 1 & -2 \\ -1 & 0 \end{bmatrix}$, $h(A) = \begin{bmatrix} -1 & -6 \\ -3 & -4 \end{bmatrix}$

11 $A^2 = \begin{bmatrix} -1 & -8 \\ 4 & 7 \end{bmatrix}$, $BA = \begin{bmatrix} 6 & 8 \end{bmatrix}$, $AC = \begin{bmatrix} -9 \\ 16 \end{bmatrix}$, $AD = \begin{bmatrix} 3 & -1 & -2 \\ -2 & 9 & 3 \end{bmatrix}$, $BC = [22]$, $CB = \begin{bmatrix} 2 & 4 \\ 10 & 20 \end{bmatrix}$, $BD = \begin{bmatrix} -2 & 14 & 4 \end{bmatrix}$

13 (b) is not nilpotent, the others are.

15 $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ are nilpotent, $A + B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ is not nilpotent.

17 $\mathbf{uv} = \begin{bmatrix} -1 & 1 & 1 \\ 0 & 0 & 0 \\ -2 & 2 & 2 \end{bmatrix} \xrightarrow[E_{31}(-2)]{E_1(-1)} \begin{bmatrix} 1 & -1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, so rank $\mathbf{uv} = 1$

19 B must be a 2×3 matrix by size check, and third column is $\begin{bmatrix} 3 \\ -3 \end{bmatrix}$

21 $A(BC) = \begin{bmatrix} 4 & 8 \\ 1 & 2 \end{bmatrix} = (AB)C$, $c(AB) = \begin{bmatrix} 0 & 16 \\ 0 & 4 \end{bmatrix} = (cA)B = A(cB)$

26 Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and try simple B like $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

30 Let $A_{m\times n} = [a_{ij}]$ and $B_{m\times n} = [b_{ij}]$. If $\mathbf{b} = [1, 0, \ldots, 0]^T$, $\begin{bmatrix} a_{11} & 0 \cdots 0 \\ \vdots & \vdots \\ a_{m1} & 0 \cdots 0 \end{bmatrix} = \begin{bmatrix} b_{11} & 0 \cdots 0 \\ \vdots & \vdots \\ b_{m1} & 0 \cdots 0 \end{bmatrix}$ so $a_{11} = b_{11}$, etc. By similar computations, you can show that for each i, j, $a_{ij} = b_{ij}$.

Section 2.3, Page 99

1 x-axis, y-axis, and points $(\pm 1, \pm 1)$ map to (a) x-axis, $-y$-axis, $(\pm 1, \mp 1)$ (b) $y = \frac{4}{3}x$, $y = -\frac{3}{4}x$, $\pm\left(\frac{7}{5}, \frac{1}{5}\right)$, $\pm\left(\frac{-1}{5}, \frac{7}{5}\right)$ (c) $-y$-axis, $-x$-axis, $\pm(1, 1)$, $\pm(1, -1)$ (d) x-axis, $y = -x$, $\pm(2, -1)$, $\pm(0, 1)$

3 T_A: (a) $A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 4 & -1 \end{bmatrix}$ (b) nonlinear (c) $A = \begin{bmatrix} 0 & 0 & 2 \\ -1 & 0 & 0 \end{bmatrix}$ (d) $A = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$

5 T_A, $A = \begin{bmatrix} \sqrt{3} & -2 \\ 1 & 2\sqrt{3} \end{bmatrix}$, reverse T_B, $B = \begin{bmatrix} \sqrt{3} & -1 \\ 2 & 2\sqrt{3} \end{bmatrix}$.

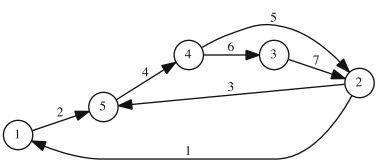
7 $S = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ and $H = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$.

9 (d) is the only candidate and the only fixed point is $(0, 0, 0)$.

11 (a), (b) and (c) are Markov. First and second states are (a) $(0.2, 0.2, 0.6)$, $(0.08, 0.68, 0.24)$ (b) $\frac{1}{2}(0, 1, 1)$, $\frac{1}{2}(1, 1, 0)$ (c) $(0.4, 0.3, 0.4)$, $(0.26, 0.08, 0.66)$ (d) $(0, 0.25, 0.25)$, $(0.225, 0, 0.15)$

13 (a) The first column says that 50% of the immature become mature and 50% of the immature remain immature in one time period. The second column says that none of the mature survive, but each mature individual produces one immature in one time period. (b) The total populations after $0, 3, 6, 9, 18$ time periods is a constant 130, and populations tend to approximately $(86.667, 43.333)$.

15 Powers of vertices 1–5 are 2, 4, 3, 5, 3, respectively. Graph is dominance directed, adjacency matrix is $\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$ and picture:



17 (a) $\begin{bmatrix} y_{k+1} \\ y_{k+2} \\ y_{k+3} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\frac{5}{2} & 2 & -\frac{3}{2} \end{bmatrix} \begin{bmatrix} y_k \\ y_{k+1} \\ y_{k+2} \end{bmatrix}$
(b) $\begin{bmatrix} y_{k+1} \\ y_{k+2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} y_k \\ y_{k+1} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

19 (a) $\begin{bmatrix} y_{k+1} \\ y_{k+2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} y_k \\ y_{k+1} \end{bmatrix}$
(b) The first ten terms are $0, 1, 1, 2, 3, 5, 8, 13, 21, 34$.

21 Points on a non-vertical line through the origin have the form (x, mx).

23 Use Exercise 30 of Section 2 and the definition of matrix operator.

26 The jth column of $\alpha A + (1-\alpha)\mathrm{B}$ is the sum of the jth column of αA, which sums to α and the jth column of $(1-\alpha)\mathrm{B}$, which sums to $1-\alpha$.

Section 2.4, Page 114

1 (a) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$
(d) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$ (e) $E_{12}(3)$ (f) $E_{31}(-a)$
(g) $E_2(3)$ (h) $E_{31}(2)$

3 (a) add 3 times third row to second (b) switch first and third rows (c) multiply third row by 2 (d) add -1 times second row to third (e) add 3 times second row to first (f) add $-a$ times first row to third (g) multiply second row by 3 (h) add 2 times first row to third

5 (a) $I_2 = E_{12}(-2)E_{21}(-1)\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} = E_{12}(-1)E_{32}(-2)\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix}$
(c) $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = E_2\left(\frac{-1}{2}\right)E_{21}(-1)\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & -2 \end{bmatrix}$
(d) $\begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & \frac{1+\mathrm{i}}{2} \end{bmatrix} = E_2\left(\frac{1}{1+\mathrm{i}}\right)E_{12}\begin{bmatrix} 0 & 1+\mathrm{i} & \mathrm{i} \\ 1 & 0 & -2 \end{bmatrix}$

7 (a) strictly upper triangular, tridiagonal (b) upper triangular (c) upper and lower triangular, scalar (d) upper and lower triangular, diagonal (e) lower triangular, tridiagonal.

9 $A = \begin{bmatrix} 0 & 2I_3 \\ C & D \end{bmatrix}$ with $C = [4, 1]$, $D = [2, 1, 3]$, $B = \begin{bmatrix} 0 & -I_2 \\ E & F \end{bmatrix}$ with $E = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 1 & 1 \end{bmatrix}$ and $F = \begin{bmatrix} 1 & 2 \\ -1 & 1 \\ 3 & 2 \end{bmatrix}$,
$AB = \begin{bmatrix} 0 + 2I_3E & 0(-I_2) + 2I_3F \\ C\,0 + DE & C(-I_2) + DF \end{bmatrix} = \begin{bmatrix} 2E & 2F \\ DE & -C + DF \end{bmatrix} = \begin{bmatrix} 0 & 0 & 2 & 4 \\ 4 & 4 & -2 & 2 \\ 2 & 2 & 6 & 4 \\ 5 & 5 & 6 & 10 \end{bmatrix}$

11 $\begin{bmatrix} 1 & 0 & 2 \end{bmatrix}^T \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$

13 (a) $(1, -3, 2)$, $(1, -3, 2)$, not symmetric or Hermitian (b) $\begin{bmatrix} 2 & 0 & 1 \\ 1 & 3 & -4 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 & 1 \\ 1 & 3 & -4 \end{bmatrix}$, not symmetric or Hermitian (c) $\begin{bmatrix} 1 & -\mathrm{i} \\ \mathrm{i} & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & \mathrm{i} \\ -\mathrm{i} & 2 \end{bmatrix}$, Hermitian, not symmetric (d) $\begin{bmatrix} 1 & 1 & 3 \\ 1 & 0 & 0 \\ 3 & 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 3 \\ 1 & 0 & 0 \\ 3 & 0 & 2 \end{bmatrix}$, symmetric and Hermitian

15 (a) true (b) false (c) false (d) true (e) false

17 $Q(x, y, z) = \mathbf{x}^T A\mathbf{x}$ with $\mathbf{x} = [x, y, z]^T$ and $A = \begin{bmatrix} 2 & 2 & -6 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$

19 $A^*A = \begin{bmatrix} 4 & -2+4\mathrm{i} \\ -2-4\mathrm{i} & 14 \end{bmatrix} = (A^*A)^*$ and $AA^* = \begin{bmatrix} 9 & 3-6\mathrm{i} \\ 3+6\mathrm{i} & 9 \end{bmatrix} = (AA^*)^*$

21 Since A is $m \times p$ and each $\mathbf{b}_j$ is $p \times 1$, $A\mathbf{b}_j$ is defined and this blocking is correct for multiplication. In particular, we have that

$$AB = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 0 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 3 \\ 1 & -4 \end{bmatrix} =$$

$$\left[\begin{bmatrix} 1 & 2 & 0 \\ 3 & 0 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 0 \\ 3 & 0 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ -4 \end{bmatrix} \right] =$$

$$\left[\begin{bmatrix} 2 \\ 10 \end{bmatrix}, \begin{bmatrix} 7 \\ -13 \end{bmatrix} \right] = \begin{bmatrix} 2 & 7 \\ 10 & -13 \end{bmatrix}.$$

24 Since A and C are square, you can confirm that block multiplication applies and use it to square M.

29 Block Q into its columns and compute the product PQ.

32 Compare (i, j)th entries of each side.

35 Substitute the expressions for A into the right-hand sides and simplify them.

36 In terms of the edge set $E = \{((v_1, w_1), (v_2, w_2), \ldots, (v_m, w_m))\}$ of the digraph G, the edge set of its reverse digraph H is

$$F = \{((w_1, v_1), (w_2, v_2), \ldots, (w_m, v_m))\}$$

which means that whenever the edge (v_k, w_k) contributes 1 to the (i, j)th entry of the adjacency matrix A of G, the edge (w_k, v_k) contributes 1 to the (j, i)th entry of the adjacency matrix B of H. Hence $B = A^T$.

Section 2.5, Page 136

1 (a) $\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$ (b) $\begin{bmatrix} 1 & \frac{-i}{4} \\ 0 & \frac{1}{4} \end{bmatrix}$ (c) does not exist (DNE) (d) $\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 1 & -1 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
(e) $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$

3 (a) $\begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix}, \begin{bmatrix} 20 \\ -11 \end{bmatrix}$
(b) $\begin{bmatrix} 3 & 6 & -1 \\ -2 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \frac{1}{15} \begin{bmatrix} 1 & -6 & 7 \\ 2 & 3 & -1 \\ 0 & 0 & 15 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$
(c) $\begin{bmatrix} 1 & 1 \\ 5 & 2 \end{bmatrix}, \frac{1}{3} \begin{bmatrix} -2 & 1 \\ 5 & -1 \end{bmatrix}, \begin{bmatrix} 3 \\ -5 \end{bmatrix}$

5 (a) $E_{21}(-3)$ (b) $E_2(-1/2)$
(c) $E_{21}(-1)\,E_{13}$ (d) $E_{12}(1)\,E_{23}(1)$
(e) $E_2\left(\frac{1}{3}\right) E_1(-1)\,E_{21}$ (i)

7 $\begin{bmatrix} 1 & -3 & -3 & 1 \\ 0 & 0 & -6 & -4 \\ 0 & -1 & -5 & -3 \end{bmatrix}$

9 (a) No two-sided inverse, so no right or left inverse, (b) no left or two-sided inverse, $\begin{bmatrix} 0 & \frac{1}{2} \\ 1 & -\frac{1}{2} \\ 0 & 0 \end{bmatrix}$ is a right inverse, (c) no right or two-sided inverse, $[1, 0, 0]$ is a left inverse.

11 Both sides give $\frac{1}{4} \begin{bmatrix} 2 & 1 & -2 \\ 2 & -1 & 2 \\ -2 & 1 & 2 \end{bmatrix}$.

13 Both sides give $\frac{1}{12} \begin{bmatrix} 18 & 12 & -9 \\ 0 & 2 & -1 \\ -6 & 0 & 3 \end{bmatrix}$.

15 (a) any k, $\begin{bmatrix} -1 & -k \\ 0 & 1 \end{bmatrix}$
(b) $k \neq 1$, $\frac{1}{k-1} \begin{bmatrix} -1 & 0 & 1 \\ -k & k-1 & 1 \\ k & 0 & -1 \end{bmatrix}$
(c) $k \neq 0$, $\begin{bmatrix} 1 & 0 & 0 & \frac{-1}{k} \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \frac{-1}{6} & 0 \\ 0 & 0 & 0 & \frac{1}{k} \end{bmatrix}$

17 Let $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$, so both invertible, but $A + B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, not invertible.

19 (a) $N = \begin{bmatrix} 0 & 1 & -2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$, $I + N + N^2 + N^3 = \begin{bmatrix} 1 & 1 & -3 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $N = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$, $I + N = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$

21 Solution vector is $\mathbf{x} = (95, 133, 189, 75, 123, 123)/738$ exactly, $\mathbf{x} = (0.129, 0.180, 0.256, 0.102, 0.167, 0.167)$ approximately.

23 Surfing matrix for this digraph is $\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}$.

25 $\mathbf{x} = (x, y)$, $\mathbf{x}^{(9)} \approx \begin{bmatrix} 1.00001 \\ -0.99999 \end{bmatrix}$, $\mathbf{F}\left(\mathbf{x}^{(9)}\right) \approx 10^{-6} \begin{bmatrix} -1.3422 \\ 2.0226 \end{bmatrix}$, $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x^2 + \sin(\pi xy) - 1 \\ x + y^2 + e^{x+y} - 3 \end{bmatrix}$, $J_{\mathbf{F}}(\mathbf{x}) = \begin{bmatrix} 2x + \cos(\pi xy), \pi y \cos(\pi xy) \, \pi x \\ 1 + e^{x+y}, \quad 2y + e^{x+y} \end{bmatrix}$

27 The quadratic function $f(x, y) = x^2 + y$ results in the second equation of the example being $1 = 0$, so $f(x, y)$ has no critical points.

29 Stationary states are $\mathbf{x} = \frac{1}{29}(8, 8, 9, 4, 0, 0)$ and $\mathbf{x} = \frac{1}{2}(0, 0, 0, 0, 1, 1)$.

31 Move constant term to right-hand side and factor A on left.

34 Multiplication by elementary matrices does not change rank.

39 Assume M^{-1} has the same form as M and solve for the blocks in M using $MM^{-1} = I$.

Section 2.6, Page 156

1 (a) $A_{11} = -1$, $A_{12} = -2$, $A_{21} = -2$, $A_{22} = 1$ (b) $A_{11} = 1$, $A_{12} = 0$, $A_{21} = -3$, $A_{22} = 1$ (c) $A_{22} = 4$, all others are 0 (d)$A_{11} = 1$, $A_{12} = 0$, $A_{21} = -1 + \mathrm{i}$, $A_{22} = 1$

3 All except (c) are invertible. (a) 3, (b) $1 + \mathrm{i}$, (c) 0, (d) -70, (e) $2\mathrm{i}$

5 Determinants of A and A^T are (a) -5 (b) 5 (c) 1 (d) 1

7 (a) $a \neq 0$ and $b \neq 1$(b) $c \neq 1$ (c) any θ

9 (a) $\begin{bmatrix} -2 & -2 & 2 \\ 4 & 4 & -4 \\ -3 & -3 & 3 \end{bmatrix}$, $0_{3,3}$ (b) $\begin{bmatrix} -1 & 0 & -3 \\ 0 & -4 & 0 \\ -1 & 0 & 1 \end{bmatrix}$, $-4I_3$ (c) $\begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix}$, $5I_2$ (d) $0_{4,4}$, $0_{4,4}$

11 (a) $\begin{bmatrix} 4 & -1 \\ -3 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -1 & 0 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} -1 & -4 & -2 \\ 0 & -1 & -1 \\ 1 & 1 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} -1 & \mathrm{i} \\ -2\mathrm{i} & -1 \end{bmatrix}$

13 (a) $x = 5$, $y = 1$ (b) $x_1 = \frac{1}{4}(b_1 + b_2)$, $x_2 = \frac{1}{2}(b_1 - b_2)$ (c) $x_1 = \frac{-7}{6}$, $x_2 = \frac{5}{3}$, $x_3 = \frac{11}{2}$

15 (a) $\det M = \det A \cdot \det D = \begin{vmatrix} -1 & 1 & 1 \\ 0 & 1 & 2 \\ 3 & 0 & 2 \end{vmatrix} \begin{vmatrix} 1 & 1+\mathrm{i} \\ 1-\mathrm{i} & -1 \end{vmatrix} = 1 \cdot (-1 - 2) = -3$.

(b) $\det M = \det A \cdot \det D = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & 3 \end{vmatrix} \begin{vmatrix} 0 & 0 & 2 \\ 1 & 7 & 0 \\ 3 & 2 & -2 \end{vmatrix} = 6 \cdot 2 \cdot (2 - 21) = -228$.

20 Proceed by induction on n. Use elementary row, then column operations to clear out the first column, then the first row of V_n to reduce the problem to one of a Vandermonde matrix of size $n-1$ by factoring common terms out of rows.

22 Take determinants of both sides of the identity $AA^{-1} = I$.

24 Factor a term of -2 out of each row. What remains?

26 Use the fact that $\begin{bmatrix} A & 0 \\ C & D \end{bmatrix}^T = \begin{bmatrix} A^T & C^T \\ 0 & D^T \end{bmatrix}$ and part (1) of Theorem 2.10

28 Write J_n as a product of row exchanges to get $J_n^2 = I_n$ and deduce from $J_n^2 = I_n$ that $(\det J_n)^2 = 1$.

Section 2.7, Page 165

1 $\begin{bmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 4 & -2 & 4 & -2 & 2 & -1 \\ 2 & 0 & 2 & 0 & 1 & 0 \\ 2 & -1 & 0 & 0 & 2 & -1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 & 0 & -1 & 0 & 0 \\ 4 & 4 & 2 & -2 & -2 & -1 \\ 2 & 0 & 2 & -1 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$

and $\frac{1}{2}\begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 \\ -2 & 4 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & -1 \\ 1 & -2 & -1 & 2 & 1 & -2 \\ 0 & -2 & 0 & 0 & 0 & 2 \\ 2 & -4 & 0 & 0 & -2 & 4 \end{bmatrix}$ for (a), (b), (c), with (d) same as (c)

3 $\begin{bmatrix} 3 & 0 & 0 & 1 & 0 & 0 \\ 2 & 4 & 1 & 0 & 1 & 0 \\ 1 & 0 & 3 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 2 & 2 & 1 \\ 0 & 0 & -1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \\ -1 \\ 0 \\ 3 \end{bmatrix}$

5 For (1) note that $(A+B)\otimes C$ has $(a_{ij}+b_{ij})\,C$ as its (i,j)th block. For (4) note that $(A\otimes B)^T$ has $a_{ji}B^T$ as its (i,j)th block.

7 If B is singular, then there is a nonzero vector $\mathbf{x}$ such that $B\mathbf{x} = 0$ by Theorem 2.6. Let $\mathbf{1}_m$ be a vector of m ones and form the product $(A\otimes B)(\mathbf{1}_m\otimes\mathbf{x}) = A\mathbf{1}_m\otimes B\mathbf{x} = A\mathbf{1}_m\otimes 0 = 0$.

10 For matrices M, N, block arithmetic gives $MN = [M\mathbf{n}_1, \ldots, M\mathbf{n}_n]$. Use this to show that $\operatorname{vec}(MN) = (I\otimes M)\operatorname{vec}(N)$. Also, $M\mathbf{n}_j = n_{1j}\mathbf{m}_1 + \cdots + n_{pj}\mathbf{m}_p$. Use this to show that $\operatorname{vec}(MN) = \left(N^T\otimes I\right)\operatorname{vec}(M)$. Then apply these to $AXB = A(XB)$.

Section 2.8, Page 176

1 Calculate LU to obtain $\begin{bmatrix} 2 & -1 & 1 \\ 2 & 3 & -2 \\ 4 & 2 & -2 \end{bmatrix}$.

3 (a) $\mathbf{x} = (1,-2,2)$ (b) $\mathbf{x} = \frac{1}{4}(3,-6,-4)$ (c) $\mathbf{x} = \frac{1}{4}(3,-2,-4)$ (d) $\mathbf{x} = \frac{1}{8}(3,6,8)$

5 $L = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ and $U = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix}$.

7 Suppose $A = LU$ is nonsingular and check the value of a_{11}.

8 G_1 inner, G_2 outer ordering matrix, rounded to three decimal places:

	A	B	C	D	E
1	0.058	0.075	0.058	0.098	0.046
2	0.058	0.075	0.058	0.098	0.046
3	0.058	0.075	0.058	0.098	0.046

Matchings are same as the example.

Section 3.1, Page 195

1 (a)$(-2,3,1)$ (b) $(6,4,9)$

3 V is a vector space.

5 V is not a vector space because it is not closed under scalar multiplication.

7 V is a vector space.

9 V is not a vector space because it is not closed under vector addition or scalar multiplication.

11 V is a vector space.

13 (a) $T = T_A$, $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & -4 \end{bmatrix}$, linear with range $\mathcal{C}(A) = \mathbb{R}^2$, equal to target (b) not linear (c) $T = T_A$, $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, linear with range $\mathcal{C}(A) = \operatorname{span}\{(1,1)\}$, not equal to target (d) not linear (e) not linear

15 (a) linear, range not V (b) not linear, (c) linear, range is V (d) linear, range not V

17 (a) identity operator is linear and invertible, $(\mathrm{id}_V)^{-1} = \mathrm{id}_V$

19 $M\mathbf{x} = (x_1 + 2, x_2 - 1, x_3 + 3, 1)$, so action of M is to translate the point in direction of vector $(2,-1,3)$. $M^{-1} = \begin{bmatrix} I_3 & -\mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ (think inverse action)

21 Write $c\mathbf{0} = c(\mathbf{0}+\mathbf{0}) = c\mathbf{0} + c\mathbf{0}$ by identity and distributive laws. Add $-(c\mathbf{0})$ to both sides.

29 Use the fact that $T_A \circ T_B = T_{AB}$ and $T_I = \mathrm{id}$

30 Use the fact that $T_A \circ T_B = T_{AB}$ and do matrix arithmetic.

Section 3.2, Page 204

1 W is not a subspace of V because W is not closed under addition and scalar multiplication.

3 W is a subspace.

5 W is a subspace.

7 Not a subspace, since W doesn't contain the zero element.

9 W is a subspace of V.

11 $\operatorname{span}\{(1,0),(0,1)\} = \mathbb{R}^2$ and $\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}\mathbf{x} = \mathbf{b}$ always has solution since coefficient matrix is invertible. So $\operatorname{span}\{(1,0),(-2,1)\} = \mathbb{R}^2$ and spans agree.

13 Write $ax^2 + bx + c = c_1 + c_2 x + c_3 x^2$ as matrix system $A\mathbf{c} = (a,b,c)$ by equating coefficients and see whether A is invertible, or use an ad hoc argument. (a) Spans $\mathcal{P}_2$. (b) Does not span $\mathcal{P}_2$ (can't get 1). (c) Spans $\mathcal{P}_2$. (d) Does not span $\mathcal{P}_2$ (can't get x).

15 (a) Add the vector $(1,0)$ to get $(0,1) = \frac{1}{2}((1,2)-(1,0))$ and $\{(1,0),(0,1)\}$ span $\mathbb{R}^2$. (b) No additions needed since $1 = x - (x-1)$ $x^3 = (x^3+1) - 1$ and $x^2 = (x^2-1)+1$.

17 $\mathbf{u} + \mathbf{w} = (4,0,4)$ and $\mathbf{v} - \mathbf{w} = (-2,0,-2)$ so $\operatorname{span}\{\mathbf{u}+\mathbf{w}, \mathbf{v}-\mathbf{w}\} = \operatorname{span}\{(1,0,1)\} \subset \operatorname{span}\{\mathbf{u},\mathbf{v},\mathbf{w}\}$, since $\mathbf{u}+\mathbf{v}, \mathbf{v}-\mathbf{w} \in \operatorname{span}\{\mathbf{u},\mathbf{v},\mathbf{w}\}$. $\mathbf{u}$ is not a multiple of $(1,0,1)$, so spans are not equal.

19 The zero vector is in all three subsets. (a) If $\mathbf{x},\mathbf{y} \in U$ and $\mathbf{x},\mathbf{y} \in V$, $\mathbf{x},\mathbf{y} \in U \cap V$. Then $c\mathbf{x} \in U$ and $c\mathbf{x} \in V$ so $c\mathbf{x} \in U \cap V$, and $\mathbf{x}+\mathbf{y} \in U$ and $\mathbf{x}+\mathbf{y} \in V$ so $\mathbf{x}+\mathbf{y} \in U \cap V$. (b) Let $\mathbf{u}_1+\mathbf{v}_1, \mathbf{u}_2+\mathbf{v}_2 \in U+V$, where $\mathbf{u}_1, \mathbf{u}_2 \in U$ and $\mathbf{v}_1, \mathbf{v}_2 \in V$. Then $c\mathbf{u}_1 \in U$ and $c\mathbf{v}_1 \in V$ so $c(\mathbf{u}_1+\mathbf{v}_1) = c\mathbf{u}_1 + c\mathbf{v}_1 \in U+V$, and similarly for sums.

21 Let A and B be $n \times n$ diagonal matrices. Then cA is diagonal matrix and $A+B$ is diagonal matrix so the set of

diagonal matrices is closed under matrix addition and scalar multiplication.

22 (a) If $A = [a_{ij}]$, $\text{vec}(A) = (a_{11}, a_{21}, a_{12}, a_{22})$ so for A there exists only one $\text{vec}(A)$. If $\text{vec}(A) = (a_{11}, a_{21}, a_{12}, a_{22})$, $A = [a_{ij}]$ so for $\text{vec}(A)$ there exists only one A. Thus vec operation establishes a one-to-one correspondence between matrices in V and vectors in $\mathbb{R}^4$.

Section 3.3, Page 217

1 (a) none (b) $(1,2,1)$, $(2,1,1)$, $(3,3,2)$ (c) every vector redundant (d) none

3 (a) linearly independent (b) linearly independent, (c) every vector redundant (d) linearly independent

5 (a) $\left(\frac{1}{4}, \frac{-3}{4}\right)$ (b) $\left(\frac{1}{2}, 1, \frac{3}{2}\right)$ (c) (b, a, c) (d) $\left(\frac{1}{2} - \mathrm{i}, 1 - \frac{3}{2}\mathrm{i}\right)$

7 (a) $\mathbf{v} = 3\mathbf{u}_1 - \mathbf{u}_2 \in \text{span}\{\mathbf{u}_1, \mathbf{u}_2\}$, (b) $\mathbf{u}_1$, $\mathbf{u}_2$, $(1,0,-1)$

9 With the given information, $\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_3\}$ are possible minimal spanning sets.

11 All values except $c = 0$, 2 or $-\frac{7}{3}$

13 e_{11}

15 (c) $W = -2$, polynomials are linearly independent (d) $W = 4$, polynomials are linearly independent

17 $\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}$

23 Assume $\mathbf{v}_i = \mathbf{v}_j$. Then there exists $c_i = -c_j \neq 0$ such that $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_i\mathbf{v}_i + \cdots + c_j\mathbf{v}_j + \cdots + c_n\mathbf{v}_n = \mathbf{0}$.

25 Start with a nontrivial linear combination of the functions that sums to 0 and differentiate it.

27 Domain and range elements $\mathbf{x}$ and $\mathbf{y}$ are given in terms of old coordinates. Express them in terms of new coordinates $\mathbf{x}'$, $\mathbf{y}'$ ($\mathbf{x} = P\mathbf{x}'$ and $\mathbf{y} = P\mathbf{y}'$.)

Section 3.4, Page 227

1 (a) $\left\{\left(-\frac{3}{2}, 0, 3, 1\right), \left(\frac{1}{2}, 1, 0, 0\right)\right\}$ (b) $\{(-4, 1)\}$ (c) $\{(-3, 1, 1)\}$ (d) $\{\,\}$

3 (a) $\{(2,4), (0,1)\}$ (b) $\{(1,-1)\}$ (c) $\{(1,-2,1), (1,-1,2)\}$ (d) $\{(2,4,1), (-1,-2,1), (0,1,-1)\}$

5 (a) $\{(2,-1,0,3), (4,-2,1,3)\}$ (b) $\{(1,4)\}$ (c) $\{(1,1,2), (2,1,5)\}$ (d) $\{(2,-1,0), (4,-2,1), (1,1,-1)\}$

7 (a) $\text{span}\{(2,2,1)\}$, $\left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)$, yes (b) $\text{span}\{(1,1)\}$, $\left(\frac{1}{2}, \frac{1}{2}\right)$, no

9 (a) kernel $\text{span}\{(1,0,-1)\}$, range $\text{span}\{(1,1,2), (-2,1,-1)\}$, not onto or one-to-one (b) kernel $\left\{a + bx + cx^2 \mid a + b + c = 0\right\}$, range $\mathbb{R}$ onto but not one-to-one

11 $\ker T = \text{span}\{\mathbf{v}_1 - \mathbf{v}_2 + \mathbf{v}_3\}$, $\text{range}\, T = \mathbb{R}^2$, T is onto but not one-to-one, hence not an isomorphism.

15 Calculate $T(\mathbf{0}) = T(\mathbf{0} + \mathbf{0})$ using linearity.

17 Use definition of isomorphism, Theorem 3.9 and for onto, solve $c_1x + c_2(x-1) + c_3x^2 = a + bx + cx^2$ for c_i's.

19 Since A is nilpotent, there exists m such that $A^m = \mathbf{0}$ so $\det(A^m) = (\det A)^m = 0$ and $\det A = 0$. Also since A is nilpotent, by Exercise 19 of Section 2.4, $(I-A)^{-1} = I + A + A^2 + \cdots + A^{m-1}$.

22 Turn rows into columns by using the transpose operator.

Section 3.5, Page 236

1 (a) None (b) Any subset of three vectors (c) Any two of $\{(2,-3,1),(4,-2,-3),(0,-4,5)\}$ and $(1,0,0)$

3 $\mathbf{w}_1$ could replace $\mathbf{v}_2$.

5 $\mathbf{w}_1$ could replace $\mathbf{v}_2$ or $\mathbf{v}_3$, $\mathbf{w}_2$ could replace any of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ and $\mathbf{w}_1, \mathbf{w}_2$ could replace any two of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$.

7 $(0,1,1)$, $(1,0,0)$, $(0,1,0)$ is one choice among many.

9 $V_1 = \text{span}\{(1,0,0),(0,1,0)\}$ and $V_2 = \text{span}\{(0,0,1),(0,1,0)\}$ are two complements and $V_1 \cap V_2 = \text{span}\{(0,1,0)\}$, so they are distinct.

11 (a) true (b) false (c) true (d) true (e) true (f) true

13 Here $\ker T_A = \text{span}\{\mathbf{e}_1\}$ and $\text{range}\, T_A = \text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ so sum is not direct. Also, $A^3 = 0_{3,3}$, so $\text{range}(T_A^3) = \{\mathbf{0}\}$, $\ker(T_A^3) = \mathbb{R}^3$ and $\mathbb{R}^3 = \mathbb{R}^3 \oplus \{\mathbf{0}\}$.

16 Suffices to note that according to definition

$$\begin{aligned} T(c_1\{y_k\} + c_2\{z_k\}) &= T(\{c_1 y_k + c_2 z_k\}) \\ &= (c_1 y_0 + c_2 z_0, \ldots, c_1 y_{m-1} + c_2 z_{m-1}) \\ &= c_1(y_0, \ldots, y_{m-1}) + c_2(z_0, \ldots, z_{m-1}) \\ &= c_1 T(\{y_k\}) + c_2 T(\{z_k\}). \end{aligned}$$

18 Suppose not and form a nontrivial linear combination of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r, \mathbf{w}$. Could the coefficient of $\mathbf{w}$ be nonzero?

19 If $c_{1,1}e_{1,1} + \cdots + c_{n,n}e_{n,n} = 0$, $c_{a,b} = 0$ for each a, b because $e_{a,b}$ is the only matrix with a nonzero entry in the (a,b)th position.

20 The union of bases for U and V will work. The fact that if $\mathbf{u} + \mathbf{v} = \mathbf{0}$, $\mathbf{u} \in U$, $\mathbf{v} \in V$, then $\mathbf{u} = \mathbf{v} = \mathbf{0}$, helps.

22 Dimension of the space is $n(n+1)/2$.

27 $\left\{I, A, A^2, \ldots, A^{n^2}\right\}$ must be linearly dependent since $\dim(\mathbb{R}^{n,n}) = n^2$. Examine a nontrivial linear combination summing to zero.

33 Use the isomorphism between $\mathcal{P}_2$ and $\mathbb{R}^3$ to turn T into a matrix operator T_A.

Section 3.6, Page 246

1 Bases for row, column, and null spaces: $\{(1,0,3,0,2),(0,1,-2,1,-1)\}$, $\{(3,1,2),(5,2,3)\}$, $\{(-3,2,1,0,0),(0,-1,0,1,0),(-2,1,0,0,1)\}$

3 Bases by row and column algorithms: (a) $\{(1,0,1),(0,1,-1)\}$, $\{(0,-1,1),(2,1,1)\}$
(b) $\left\{\left(1,0,\frac{1}{2}\right),(0,1,0)\right\}$, $\{(2,-1,1),(2,0,1)\}$ (c) $\{(1,0),(0,1)\}$, $\{(1,-1),(2,2)\}$ (d) $\left\{1+x^2, x-5x^2\right\}$, $\left\{1+x^2, -2-x+3x^2\right\}$

5 Bases for row, column, and null spaces: (a) $\{(2,0,-1)\}$, $\{1\}$, $\left\{\left(\frac{1}{2},0,1\right),(0,1,0)\right\}$ (b) $\{(1,2,0,0,1),(0,0,1,1,0)\}$, $\{(1,1,3),(0,1,2)\}$, $\{(-2,1,0,0,0),(0,0,-1,1,0),(-1,0,0,0,1)\}$ (c) $\{(1,0,-10,8,0),(0,1,5,-2,0),(0,0,0,0,1)\}$, $\{(1,1,2,2),(2,3,3,4),(0,1,0,1)\}$, $\{(10,-5,1,0,0),(-8,2,0,1,0)\}$ (d) $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$, $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$, $\{\,\}$

7 (a) $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4 = \mathbf{0}$, where $c_1 = -2c_3 - 2c_4$, $c_2 = -c_3$, and c_3, c_4 are free, $\dim \text{span}\{\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3,\mathbf{v}_4\} = 2$
(b) $c_1 x + c_2\left(x^2+x\right) + c_3\left(x^2-x\right) = 0$ where $c_1 = 2c_3$, $c_2 = -c_3$, and c_3 is free, $\dim \text{span}\left\{x, x^2+x, x^2-x\right\} = 2$
(c) $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4 = \mathbf{0}$, where $c_1 = -c_3$, $c_2 = \frac{1}{2}c_3$, $c_4 = 0$ and c_3 is free, $\dim \text{span}\{\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3,\mathbf{v}_4\} = 3$

9 (a) $\dim \mathcal{C}(A) = 2$, $\dim \mathcal{C}(B) = 2$ (b) $\dim \mathcal{C}\left(\begin{bmatrix} A & B \end{bmatrix}\right) = 3$ (c) $\dim \mathcal{C}(A) \cap \mathcal{C}(B) = 2 + 2 - 3 = 1$

11 $\mathcal{C}(A) \cap \mathcal{C}(B) = \text{span}\{(1, 1, -1)\}$

13 $[A\ I]$ has RREF $\begin{bmatrix} 1 & 0 & 2 & 1 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 2 & 0 & -1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & -\frac{1}{2} \end{bmatrix}$, so a basis of $\mathcal{C}(A)$ is $B = \{(0, 1, 2), (1, 0, 2)\}$ and can be expanded to the basis $\{(0, 1, 2), (1, 0, 2), (1, 0, 0)\}$ of $\mathbb{R}^3$ according to the column space algorithm.

15 Since $A\mathbf{x} = \mathbf{b}$ is consistent, $\mathbf{b} \in \mathcal{C}(A)$. If $\{\mathbf{a}_i\}$, the set of columns of A, has redundant vectors in it, $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = \mathbf{0}$ for some nontrivial $\mathbf{c}$.

17 What does $\mathbf{b} \notin \mathcal{C}(A)$ tell you about r and m?

Section 3.7, Page 253

1 $\begin{bmatrix} 1 & 2 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$, range (T) = span $\{(1, 1, 0), (2, -1, 1), (0, 0, 1)\}$, $\ker(T) = \{\mathbf{0}\}$

3 (a) $P = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, $Q = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$

(b) $[\text{id}]_{B',B} = Q^{-1}I_2P = \begin{bmatrix} -1 & 2 \\ 1 & -1 \end{bmatrix}$

(c) $[\mathbf{w}]_{B'} = [\text{id}]_{B',B} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \end{bmatrix}$

5 $\frac{1}{25} \begin{bmatrix} 7 & 6 & -6 \\ 1 & 8 & -8 \end{bmatrix}$

9 Let B' be any other basis and use the chain of operators $V_{B'} \overset{\text{id}_V}{\to} V_B \overset{T}{\to} V_B \overset{\text{id}_V}{\to} V_{B'}$.

Section 3.8, Page 254

1 Minimize $C = \mathbf{c}^T\mathbf{x}$ subject to the constraints $B\mathbf{x} = \mathbf{d}$, $\mathbf{x} \geq \mathbf{0}$, where $\mathbf{c} = (1, 2, 1, 0, 0, 0)$ $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$, $B = \begin{bmatrix} 1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$ and $\mathbf{d} = (4, 6, 1)$.

3 Geometric method yields corners $(0, 0)$, $(0, 2)$, $(2, 4)$, $(5, 1)$ and $(5, 0)$ with objective values 0, 4, 10, 7 and 5, resp. Simplex method yields initial augmented standard matrix $\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 6 \\ -1 & 1 & 0 & 1 & 0 & 2 \\ 1 & 0 & 0 & 0 & 1 & 5 \\ -1 & -2 & 0 & 0 & 0 & 0 \end{bmatrix}$ and final augmented standard matrix $\begin{bmatrix} 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 2 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 & 4 \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} & 1 & 3 \\ 0 & 0 & \frac{3}{2} & \frac{1}{2} & 0 & 10 \end{bmatrix}$. Both yield solution $x_1 = 2$, $x_2 = 4$ and maximum $P = x_1 + 2x_2 = 10$.

5 Feasible set is the quadrilateral bounded by the points $(15, 10)$, $(20, 0)$, $(18, 0)$ and $\left(\frac{9}{2}, \frac{27}{2}\right)$ with values of the objective function P as 65, 20, 18 and $\frac{117}{2}$. Hence the maximum value is $P = \frac{117}{2}$ at the point $x_1 = \frac{9}{2}$, $x_2 = \frac{27}{2}$.

7 The initial augmented standard matrix for this problem is $B = \begin{bmatrix} -1 & 1 & 1 & 0 & 4 \\ -1 & 2 & 0 & 1 & 10 \\ -3 & -1 & 0 & 0 & 0 \end{bmatrix}$. According to the Caution on Page 266 the objective function $P = 3x_1 + x_2$ is unbounded in the feasible set. The dual problem of minimizing $C = 4x_1 + 10x_2$ has constraints $x_1, x_2 \geq 0$, $-x_1 - x_2 \geq 3$ and $x_1 + 2x_2 \geq 1$. But the second is equivalent to $x_1 + x_2 \leq -3$, so

there is no feasible solution to this problem.

9 Standard form is to maximize objective $\mathbf{c}^T\mathbf{x}$ where $\mathbf{c} = (35, 20, 40, 25, 0, 0)$ and $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$, subject to constraints $B\mathbf{x} = \mathbf{d}$, $\mathbf{x} \geq \mathbf{0}$ with $B = \begin{bmatrix} 2&1&3&2&1&0 \\ 3&2&2&2&0&1 \end{bmatrix}$ and $\mathbf{d} = (900, 1200)$. Initial augmented matrix for this problem is $\begin{bmatrix} 2&1&3&2&1&0&900 \\ 3&2&2&2&0&1&1200 \\ -35&-20&-40&-25&0&0&0 \end{bmatrix}$ and final form is $\begin{bmatrix} 0&-\frac{1}{5}&1&\frac{2}{5}&\frac{3}{5}&-\frac{2}{5}&60 \\ 1&\frac{4}{5}&0&\frac{2}{5}&-\frac{2}{5}&\frac{3}{5}&360 \\ 0&0&0&5&10&5&15000 \end{bmatrix}$. Hence the solution is to set production levels of K_1, K_2, K_3, K_4 at $360, 0, 60, 0$, respectively, for a profit of $35 \cdot 360 + 40 \cdot 60 = 15000$.

11 The initial augmented standard matrix with problem converted to maximization with artificial variable x_6 is $\begin{bmatrix} 1&1&2&1&0&0&40 \\ 1&1&0&0&-1&1&10 \\ 6&1&4&0&0&0&0 \\ 0&0&0&0&0&1&0 \end{bmatrix}$ which results in matrix (after deleting last row and sixth column is $\begin{bmatrix} 0&0&2&1&1&30 \\ 1&1&0&0&-1&10 \\ 5&0&4&0&1&-10 \end{bmatrix}$, so the minimum value is $C = -P = 10$ at $x_1 = 0$, $x_2 = 10$ and $x_3 = 0$.

13 The initial augmented matrix for the dual max problem is $\begin{bmatrix} 3&1&1&0&3 \\ 1&2&0&1&2 \\ -6&-6&0&0&0 \end{bmatrix}$ which leads to final augmented standard matrix $\begin{bmatrix} 1&0&\frac{2}{5}&-\frac{1}{5}&\frac{4}{5} \\ 0&1&-\frac{1}{5}&\frac{3}{5}&\frac{3}{5} \\ 0&0&\frac{6}{5}&\frac{12}{5}&\frac{42}{5} \end{bmatrix}$. Thus $x_1 = \frac{6}{5}$, $x_2 = \frac{12}{5}$ yields a minimum value for the min problem of $C = \frac{42}{5}$.

16 Any elementary operation on the first m rows of $\tilde{B}$ can be expressed in the form $A = \begin{bmatrix} F & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$, where F is an elementary operation of B, whereas any elementary operation of adding the c times the jth of one of the first m rows to the last has the form $B = \begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix}$ with $\mathbf{v}$ a vector of size m, c as its jth entry and zeros elsewhere. Now check that the product of any two matrices of the form $\begin{bmatrix} E & \mathbf{0} \\ \mathbf{v}^T & 1 \end{bmatrix}$ is another matrix of the same form.

18 Let $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{z} \geq \mathbf{0}$ be two such solutions. Then optimal value Q satisfies $Q = \mathbf{c}^T\mathbf{y} = \mathbf{c}^T\mathbf{z}$. Now show that any convex combination $\mathbf{w}_\alpha = \alpha\mathbf{y} + (1-\alpha)\,\mathbf{z}$ with $0 \leq \alpha \leq 1$ is also optimal feasible. Hence there are infinitely many.

Section 4.1, Page 286

1 (a) -14, $\sqrt{34}$, $2\sqrt{5}$ (b) 7, $\sqrt{6}$, $\sqrt{14}$ (c) 8, $\sqrt{10}$, $\sqrt{26}$ (d) $12 - 6\mathrm{i}$, $\sqrt{10}$, $\sqrt{26}$ (e) 4, $\sqrt{30}$, $\sqrt{6}$ (f) -4, $2\sqrt{3}$, $\sqrt{30}$

3 (a) $-\sqrt{145}/145$ (b) 0 (c) $\sqrt{21}/6$ (d) $\sqrt{10}/10$

5 (a) $36\mathbf{k}$ (b) $-5\mathbf{i} - \mathbf{j} + 5\mathbf{k}$ (c) $(-2, -2, 4)$

7 $\|\mathbf{u}\| = \sqrt{30}$, $\|c\mathbf{u}\| = 3\sqrt{30}$, $\|\mathbf{v}\| = 4$, $\|\mathbf{u} + \mathbf{v}\| = \sqrt{30}$, $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\| = 4 + \sqrt{30}$

9 $\mathbf{u} \times \mathbf{v} = (-6, 4, -8)$, $\mathbf{v} \times \mathbf{u} = (6, -4, 8)$, $(c\mathbf{u}) \times \mathbf{v} = c\,(\mathbf{u} \times \mathbf{v}) = \mathbf{u} \times (c\mathbf{v}) = (12, -8, 16)$, $\mathbf{u} \times \mathbf{w} = (4, 1, -2)$, $\mathbf{v} \times \mathbf{w} = -(6, 7, 8)$ $\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = (-2, 5, -10)$, $(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = -(2, 6, 10)$

11 We have $\|\mathbf{u}\|^2 = 7$, $\|\mathbf{v}\|^2 = 18$, $\|\mathbf{u} + \mathbf{v}\|^2 = 37$. and $\|\mathbf{u} - \mathbf{v}\|^2 = 13$, so it checks out: $50 = 50$.

13 $\mathbf{u}_n = \left(\frac{2}{n}, \frac{\frac{1}{n^2} + 1}{2 + \frac{3}{n} + \frac{5}{n^2}} \right) \to \left(0, \frac{1}{2}\right)$

15 Let $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$, $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Then $(c\mathbf{u}) \cdot \mathbf{v} = (cu_1)v_1 + \cdots + (cu_n)v_n$ and $\mathbf{v} \cdot (c\mathbf{u}) = v_1(cu_1) + \cdots + v_n(cu_n)$ so $(c\mathbf{u}) \cdot \mathbf{v} = \mathbf{v} \cdot (c\mathbf{u})$. Similarly, show $(c\mathbf{u}) \cdot \mathbf{v} = \mathbf{v} \cdot (c\mathbf{u}) = c(\mathbf{v} \cdot \mathbf{u}) = c(\mathbf{u} \cdot \mathbf{v})$.

17 Compute A^2, A^3 in terms of $\mathbf{u}$ and the general formula will be clear.

20 $\|c\mathbf{v}\| = |c| \, \|\mathbf{v}\|$ by basic norm law (2). Since $c \in \mathbb{R}$ and $c > 0$, $\|c\mathbf{v}\| = c\|\mathbf{v}\|$. So a unit vector in direction of $c\mathbf{v}$ is $c\mathbf{v}/c\|\mathbf{v}\| = \mathbf{v}/\|\mathbf{v}\|$.

22 Apply the triangle inequality to $\mathbf{u} + (\mathbf{v} - \mathbf{u})$ and $\mathbf{v} + (\mathbf{u} - \mathbf{v})$.

24 Let $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$. Then the Law of Cosines is $\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta$ where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. Expand terms and show

$$2\|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta = 2u_1v_1 + 2u_2v_2 + 2u_3v_3,$$

from which the result follows.

27 Use (7) and equation 4.2 to obtain

$$\begin{aligned}\|\mathbf{u} \times \mathbf{v}\|^2 &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - (\mathbf{u} \cdot \mathbf{v})^2 \\ &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 - \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \cos^2\theta \\ &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \left(1 - \cos^2\theta\right) \\ &= \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \sin^2\theta.\end{aligned}$$

Section 4.2, Page 299

1 (a) 2.1176 (b) 1.6383 (c) 1.0018

3 (a) $(-2,-1)$, $-\sqrt{5}$ (b) $\frac{10}{9}(2,2,1)$, $\frac{10}{3}$ (c) $\frac{-1}{2}(1,1,1,1)$, -1

5 (a) $|\mathbf{u} \cdot \mathbf{v}| = 1 \leq \|\mathbf{u}\| \, \|\mathbf{v}\| = \sqrt{15}$ (b) $|\mathbf{u} \cdot \mathbf{v}| = 19 \leq \|\mathbf{u}\| \, \|\mathbf{v}\| = 2\sqrt{165}$ (c) $|\mathbf{u} \cdot \mathbf{v}| = 26 \leq \|\mathbf{u}\| \, \|\mathbf{v}\| = 26$

7 (a) $(M\mathbf{u}) \cdot (M\mathbf{v}) = 1$, no (b) $(M\mathbf{u}) \cdot (M\mathbf{v}) = 0$, yes (c) $(M\mathbf{u}) \cdot (M\mathbf{v}) = -13$, no

9 (a) $x + y - 4z = -6$ (b) $x - 2z = -4$

11 (a) $\mathbf{x} = \left(3, -\frac{2}{3}\right)$, $\mathbf{b} - A\mathbf{x} = \mathbf{0}$, $\|\mathbf{b} - A\mathbf{x}\| = 0$, yes (b) $\mathbf{x} = \frac{1}{21}(9,-14)$, $\mathbf{b} - A\mathbf{x} = \frac{1}{21}(-4,-16,8)$, $\|\mathbf{b} - A\mathbf{x}\| = \frac{\sqrt{336}}{21}$, no (c) $\mathbf{x} = \left(x_3 + \frac{12}{13}, -x_3 + \frac{23}{26}, x_3\right)$ where x_3 is free, $\mathbf{b} - A\mathbf{x} = \frac{1}{26}(32,-21,1,22)$, $\|\mathbf{b} - A\mathbf{x}\| = \frac{\sqrt{1950}}{26}$, no

13 $b = 0.3$, $a + b = 1.1$, $2a + b = 2$, $3a + b = 3.5$, $3.5a + b = 3.6$, least squares solution $a \approx 1.00610$, $b \approx 0.18841$, residual norm is $\|\mathbf{b} - A\mathbf{x}\| \approx 0.39962$

15 For triangle ADC an outward normal is $-4\mathbf{k}$, so unit normal is $-\mathbf{k}$. For triangle ABC, an outward normal is $-12\mathbf{i} + 8\mathbf{j} + 4\mathbf{k}$, so a unit normal is $\frac{\sqrt{14}}{14}(-3\mathbf{i} + 2\mathbf{j} + \mathbf{k})$. For triangle BDC, an outward normal is $\overrightarrow{BD} \times \overrightarrow{DC} = 12\mathbf{i} + 8\mathbf{j} + 4\mathbf{k}$, so a unit normal is $\frac{\sqrt{14}}{14}(3\mathbf{i} + 2\mathbf{j} + \mathbf{k})$. For triangle DBA, an outward normal is $\overrightarrow{DB} \times \overrightarrow{BA} = -16\mathbf{j} + 4\mathbf{k}$, so a unit normal is $\frac{\sqrt{17}}{17}(-4\mathbf{j} + \mathbf{k})$.

17 Express each norm in terms of dot products, expand and cancel the terms $\mathbf{u} \cdot \mathbf{u}$ and $\mathbf{v} \cdot \mathbf{v}$ from both sides.

23 Use Example 3.43.

25 Examine the proof of Theorem 4.3 for points where real and complex dots might differ.

Section 4.3, Page 312

1 (a) orthogonal, linearly independent (b) linearly independent (c) orthonormal, orthogonal, linearly independent

3 $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, $\mathbf{v}_1 \cdot \mathbf{v}_3 = 0$, and $\mathbf{v}_2 \cdot \mathbf{v}_3 = 0$ so $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is an orthogonal basis. $\mathbf{v}_1 \cdot \mathbf{v}_1 = 2$, $\mathbf{v}_2 \cdot \mathbf{v}_2 = 3$, and $\mathbf{v}_3 \cdot \mathbf{v}_3 = \frac{3}{2}$. Coordinates with respect to $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ are (a) $\left(\frac{3}{2}, \frac{-1}{3}, \frac{-5}{3}\right)$ (b) $\left(\frac{1}{2}, \frac{-1}{3}, \frac{1}{3}\right)$ (c) $\left(\frac{1}{2}, \frac{-5}{3}, \frac{11}{3}\right)$

5 (a) orthogonal, $\frac{1}{5}\begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}$ (b) not orthogonal (c) not orthogonal (d) not orthogonal (e) unitary, $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 & -\mathrm{i} \\ 0 & -\sqrt{2}\mathrm{i} & 0 \\ 1 & 0 & \mathrm{i} \end{bmatrix}$ (f) unitary, $\frac{1}{\sqrt{3}}\begin{bmatrix} 1-\mathrm{i} & -\mathrm{i} \\ -\mathrm{i} & 1+\mathrm{i} \end{bmatrix}$

7 $H_{\mathbf{v}} = \frac{1}{3}\begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{bmatrix}$, $H_{\mathbf{v}}\mathbf{u} = (3,0,0)$, $H_{\mathbf{v}}\mathbf{w} = (1,2,-2)$

9 (a) $\begin{bmatrix} \frac{\sqrt{6}}{6} & -\frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{6}}{3} & \frac{\sqrt{3}}{3} & 0 \\ -\frac{\sqrt{6}}{6} & \frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} \end{bmatrix}$ (b) $\frac{1}{5}\begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 \\ 0 & \mathrm{i} \end{bmatrix}$

11 Calculate both sides of each equation defining projection matrix and check that $(I-2P)^T = I-2P$ and $(I-2P)^T(I-2P) = I$.

13 (a) $(1,1,-1)$, $\frac{1}{3}(-2,7,5)$, $\frac{1}{13}(8,-2,6)$ (b) $(1,0,1)$, $\frac{1}{2}(1,8,-1)$ (c) $(1,1)$, $\frac{1}{2}(-1,1)$

16 Let $\mathbf{u}, \mathbf{v}$ be columns of P, calculate $\left|e^{\mathrm{i}\theta}\mathbf{u}\right|$ and $\left(e^{\mathrm{i}\theta}\mathbf{u}\right) \cdot \left(e^{\mathrm{i}\theta}\mathbf{v}\right)$.

Section 4.4, Page 325

1 Q, R, $\mathbf{x}$: (a) $\begin{bmatrix} \frac{3}{5} & \frac{4}{5\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \\ \frac{4}{5} & \frac{-3}{5\sqrt{2}} \end{bmatrix}$, $\begin{bmatrix} 5 & 2 \\ 0 & \sqrt{2} \end{bmatrix}$, $\begin{bmatrix} \frac{9}{5} \\ \frac{-5}{2} \end{bmatrix}$ (b) Caution: this matrix is rank deficient. $\begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{6}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{bmatrix}$, $\begin{bmatrix} \sqrt{5} & 0 & \frac{-10}{\sqrt{5}} \\ 0 & \sqrt{6} & \frac{12}{\sqrt{6}} \end{bmatrix}$, $\begin{bmatrix} 2x_3 - 3 \\ -2x_3 + 2 \\ x_3 \end{bmatrix}$, x_3 free (c) $\frac{1}{2}\begin{bmatrix} 1 & 0 & \frac{5}{3} \\ 1 & \sqrt{2} & \frac{-1}{3} \\ -1 & \sqrt{2} & \frac{1}{3} \\ -1 & 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 & \frac{3}{2} \\ 0 & \sqrt{2} & \frac{3}{2}\sqrt{2} \\ 0 & 0 & \frac{3}{2} \end{bmatrix}$, $\begin{bmatrix} \frac{-1}{2} \\ \frac{9}{2} \\ \frac{-5}{3} \end{bmatrix}$

3 $\frac{1}{2}W_m A W_n^T = \begin{bmatrix} 33 & 55 & 4 & 27 & 5 & -4 \\ 60 & 63 & 75 & -10 & 13 & -25 \\ 27 & 5 & 4 & 33 & 55 & -4 \\ -20 & 57 & 15 & 50 & 7 & 35 \end{bmatrix}$, so $B = \begin{bmatrix} 33 & 55 & 4 \\ 60 & 63 & 75 \end{bmatrix}$, $V = \begin{bmatrix} 27 & 5 & -4 \\ -10 & 13 & -25 \end{bmatrix}$, $H = \begin{bmatrix} 27 & 5 & 4 \\ -20 & 57 & 15 \end{bmatrix}$, $D = \begin{bmatrix} 33 & 55 & -4 \\ 50 & 7 & 35 \end{bmatrix}$.

7 If A is $n \times n$ then $\mathbf{x}$must be $n \times 1$ and the calculation of $A\mathbf{x}$ amounts to the inner product of $\mathbf{x}$ with each row of A, hence $2n \cdot n = 2n^2$ flops, counting addition and multiplications. Write $H_{\mathbf{v}}\mathbf{x} = \mathbf{x} - \frac{2\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}}\mathbf{v}$, and calculate total flops of this to be order n.

Section 5.1, Page 341

1 (a) -3, 2 (b) -1, -1, -1 (c) 2, 2, 3 (d) -2, 2 (e) $-2\mathrm{i}$, $2\mathrm{i}$

3 Eigenvalue, algebraic multiplicity, geometric multiplicity, basis: (a) $\lambda = -3$, 1, 1, $\{(2,1)\}$, $\lambda = 2$, 1, 1, $\{(1,1)\}$ (b) $\lambda = -1$, 3, 1, $\{(0,0,1)\}$, (c) $\lambda = 2$, 2, 2, $\{(1,0,0),(0,-1,1)\}$, $\lambda = 3$, 1, 1, $\{(1,1,0)\}$ (d) $\lambda = -2$, 1, 1, $\{(-1,1)\}$, $\lambda = 2$, 1, 1, $\{(1,1)\}$ (e) $\lambda = -2\mathrm{i}$, 1, 1, $\{(\mathrm{i},-1)\}$, $\lambda = 2\mathrm{i}$, 1, 1, $\{(\mathrm{i},1)\}$

5 $B = 3I - 5A$, so eigensystem for B consists of eigenpairs $\{-2,(1,1)\}$ and $\{8,(1,-1)\}$.

7 (a) $\mathrm{tr}A = 7 - 8 = -1 = -3 + 2$, (b) $\mathrm{tr}A = -1-1-1 = -3$, (c) $\mathrm{tr}A = 7 = 2 + 2 + 3$ (d) $\mathrm{tr}A = 0 + 0 = 0 = -1 + 1$ (e) $\mathrm{tr}A = 0 + 0 = 0 = -2\mathrm{i} + 2\mathrm{i}$

9 Eigenvalues of A^T same as Exercise 1.

11 (a) No (b) No (c) No (d) Yes (e) No

13 Eigenvalues of A are $1, 2$. Eigenvalues of B are $\frac{1}{2}\left(3 \pm \sqrt{5}\right)$. Eigenvalues of $A+B$ are $3 \pm \sqrt{3}$. Eigenvalues of AB are $3 \pm \sqrt{7}$. (a) Deny $-3 + \sqrt{3}$ not sum of 1 or 2 plus $\frac{1}{2}\left(3 \pm \sqrt{5}\right)$. (b) Deny $-3 + \sqrt{7}$ not product of 1 or 2 times $\frac{1}{2}\left(3 \pm \sqrt{5}\right)$.

17 If A is invertible, $\lambda \neq 0$, then $A^{-1}A\mathbf{v} = A^{-1}\lambda\mathbf{v}$.

19 For λ eigenvalue of A with eigenvector $\mathbf{v}$, $(I - A)\mathbf{v} = I\mathbf{v} - A\mathbf{v} = \mathbf{v} - \lambda\mathbf{v} = (1 - \lambda)\mathbf{v}$. Since $|\lambda| < 1$, $1 - \lambda > 0$.

20 Use part (1) of Theorem 5.1.

23 Deal with the 0 eigenvalue separately. If λ is an eigenvalue of AB, multiply the equation $AB\mathbf{x} = \lambda\mathbf{x}$ on the left by B.

Section 5.2, Page 351

1 All except (d) have distinct eigenvalues, so are diagonalizable. For $\lambda = 1$ (d) has eigenspace of dimension two, so is not diagonalizable.

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

5 True in every case. (a) and (c) satisfy $q(A) = 0$ and are diagonalizable, (b) and (d) do not satisfy $q(A) = 0$ and are not diagonalizable.

7 $\mathcal{E}_\lambda(J_2(\lambda)) = \mathrm{span}\{(1,0)\}$, so $J_2(\lambda)$ is not diagonalizable (not enough eigenvectors). $J_2(\lambda)^2 = \begin{bmatrix} \lambda^2 & 2\lambda \\ 0 & \lambda^2 \end{bmatrix}$, $J_2(\lambda)^3 = \begin{bmatrix} \lambda^3 & 3\lambda^2 \\ 0 & \lambda^3 \end{bmatrix}$, $J_2(\lambda)^4 = \begin{bmatrix} \lambda^4 & 4\lambda^3 \\ 0 & \lambda^4 \end{bmatrix}$, which suggests $J_2(\lambda)^k = \begin{bmatrix} \lambda^k & k\lambda^{k-1} \\ 0 & \lambda^k \end{bmatrix}$.

9 $P = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$, $Q = \begin{bmatrix} -3 & 1 \\ 2 & 0 \end{bmatrix}$, $S = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, $S^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$

11 $\sin\left(\frac{\pi}{6}A\right) = \begin{bmatrix} \frac{1}{2}\sqrt{3} & \frac{4}{5} + \frac{2}{5}\sqrt{3} \\ 0 & -1 \end{bmatrix}$, $\cos\left(\frac{\pi}{6}A\right) = \begin{bmatrix} \frac{1}{2} & \frac{2}{5} \\ 0 & 0 \end{bmatrix}$

13 Use the fact that if $\mathbf{x}$ is an eigenvector, then so is $c\mathbf{x}$ for any nonzero c.

14 Let $A = \lambda I + U$ be such and recall that similar matrices have the same eigenvalues.

16 Examine $DB = BD$, with D diagonal and no repeated diagonal entries.

18 You will find Corollary 5.1 helpful.

20 (a) Use $f_{k+2} = f_{k+1} + f_k$, $f_{k+1} = f_{k+1}$ (b) $f_n = \left(\frac{1+\sqrt{5}}{2}\right)^n \left(\frac{5+\sqrt{5}}{10}\right) + \left(\frac{1-\sqrt{5}}{2}\right)^n \left(\frac{5-\sqrt{5}}{10}\right)$.

22 In one direction use the fact that diagonal matrices commute. In the other direction, prove it for a diagonal A first, then use the diagonalization theorem to prove it for general A.

Section 5.3, Page 363

1 (a) 2, dominant eigenvalue 2 (b) 0, no dominant eigenvalue (c) 0, no dominant eigenvalue (d) 1, dominant eigenvalue -1 (e) $\frac{1}{2}$, dominant eigenvalue $\frac{-1}{2}$

3 (a) $\mathbf{x}^{(k)} = \begin{bmatrix} 2\left(\frac{1}{2}\right)^k - \left(\frac{-1}{2}\right)^k \\ -2\left(\frac{1}{2}\right)^k + 2\left(\frac{-1}{2}\right)^k \end{bmatrix}$ (b) $\mathbf{x}^{(k)} = \begin{bmatrix} 2^k \\ 3^{k+1} - 2^k \\ 2^k \end{bmatrix}$ (c) $\mathbf{x}^{(k)} = \begin{bmatrix} 13 \cdot 2^k - 10 \cdot 3^k \\ -13 \cdot 2^k + 15 \cdot 3^k \end{bmatrix}$

5 (b), (c), and (e) give matrices for which all $\mathbf{x}^{(k)} \to \mathbf{0}$ as $k \to \infty$. Stability theorem only applies to (d).

7 $\operatorname{diag}\{A, B\}$, where possibilities for A are $\operatorname{diag}\{J_1(2), J_1(1)\}$, $J_2(2)$ and possibilities for B are $\operatorname{diag}\{J_1(3), J_1(3), J_1(3)\}$, $\operatorname{diag}\{J_1(3), J_2(3)\}$, $\operatorname{diag}\{J_3(3)\}$

9 Characteristic polynomial for $J_3(2)$ is $(\lambda-2)^3$ and $(J_3(2) - 2I_3)^3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}^3 = 0.$

11 Eigenvalues are $\lambda \approx 0.1636 \pm 0.3393\text{i}$, 0.973 with absolute values 0.3766 and 0.973. So population will decline at rate of approximately 2.7% per time period.

13 $\lambda^2 = s_1 f_2$, $\mathbf{p} = p_1\left(1, \sqrt{s_1/f_2}\right)$

15 (a) Sum of each column is 1. (c) Since a and b are nonnegative, (a, b) and $(1, -1)$ are linearly independent eigenvectors. Use diagonalization theorem.

17 Show that $(1, 1, \ldots, 1)$ is a *left* eigenvector.

22 Note that λI commutes with all matrices of the same size, so one can apply the binomial formula to $(\lambda I + U)^m$.

24 Similar matrices have the same characteristic equation, so A has the same characteristic equation as its Jordan canonical form. Use block arithmetic and apply $p(\lambda)$ to each block.

Section 5.4, Page 370

1 A is real and $A = A^T$ in each case. (a) $\frac{1}{\sqrt{5}}\begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$ (b) $\frac{1}{5}\begin{bmatrix} -4 & 3 \\ 3 & 4 \end{bmatrix}$ (c) $\frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \end{bmatrix}$ (d) $\begin{bmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{6} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{6} & \frac{\sqrt{3}}{3} \\ 0 & -\frac{\sqrt{6}}{3} & \frac{\sqrt{3}}{3} \end{bmatrix}$

3 $P^T P = I$ in each case. (a) Unitarily diagonalizable by $\frac{1}{\sqrt{2}}\begin{bmatrix} 0 & -\text{i} & \text{i} \\ 0 & 1 & 1 \\ \sqrt{2} & 0 & -1 \end{bmatrix}$ (b) Unitarily diagonalizable by $\frac{1}{\sqrt{2}}\begin{bmatrix} -\text{i} & \text{i} \\ 1 & 1 \end{bmatrix}$ (c) Orthogonally diagonalizable by $\frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \end{bmatrix}$

5 All of these matrices are normal.

7 Orthogonalize by $\begin{bmatrix} -\frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{3} & 0 \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} \end{bmatrix}$, $= \begin{bmatrix} \frac{2}{3}+\frac{\sqrt{2}}{2} & \frac{1}{3} & \frac{-2}{3}+\frac{\sqrt{2}}{2} \\ & \frac{1}{3} & \frac{5}{3} & \frac{-1}{3} \\ \frac{-2}{3}+\frac{\sqrt{2}}{2} & \frac{-1}{3} & \frac{2}{3}+\frac{\sqrt{2}}{2} \end{bmatrix}$, which is symmetric positive definite and $B^2 = A$.

let $a = (-1)^k + 2^{k+1}$, $b = (-1)^{k-1} + 2^k$, $c = (-1)^k + 2^{k-1}$ and $A^k = \frac{1}{3}\begin{bmatrix} a & b & b \\ b & c & c \\ b & c & c \end{bmatrix}$

12 Use orthogonal diagonalization and change of variable $\mathbf{x} = P\mathbf{y}$ for a general B to reduce the problem to one of a diagonal matrix.

16 First show it for a diagonal matrix with positive diagonal entries. Then use Problem 12 and the principal axes theorem.

9 $P = \begin{bmatrix} -\frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} & \frac{-\sqrt{6}}{6} \\ \frac{\sqrt{3}}{3} & 0 & \frac{-\sqrt{6}}{3} \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{6} \end{bmatrix}$,
$B = P\text{diag}\left\{1, \sqrt{2}, 2\right\} P^T$

17 $A^T A$ is symmetric and square. Now calculate $\|A\mathbf{x}\|^2$ for an eigenvector $\mathbf{x}$ of $A^T A$.

Section 5.5, Page 374

1 (a) $\begin{bmatrix} -3 & 0 & 0 \\ 0 & -2.5764 & -1.5370 \\ 0 & -1.5370 & 2.5764 \end{bmatrix}$

(b) $\begin{bmatrix} 1.41421 & 0 & 0 \\ 0 & -1.25708 & 0.44444\text{i} \\ 0 & -0.44444\text{i} & -0.15713 \end{bmatrix}$

3 (a) $-2, 3, 2$ (b) $3, 1, 2$ (c) $2, -1, \pm\sqrt{2}$

5 Eigenvalues of A are $2, -3$ and eigenvalues of $f(A)/g(A)$ are $0.6, 0.8$.

8 Do a change of variables $\mathbf{x} = P\mathbf{y}$, where P upper triangularizes A.

11 Equate $(1,1)$th coefficients of the equation $R^*R = RR^*$ and see what can be gained from it. Proceed to the $(2,2)$th coefficient, etc.

12 Use Problem 37 of Section 2.5.

Section 5.6, Page 379

1 (a) $U = E_2(-1)$, $\Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $V = I_3$ (b) $U = \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{-\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix}$, $V = I_2$ (c) $U = E_{12}E_{13}$, $\Sigma = \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $V = \begin{bmatrix} 0 & 1 & 0 \\ -\frac{\sqrt{5}}{5} & 0 & \frac{2\sqrt{5}}{5} \\ \frac{2\sqrt{5}}{5} & 0 & \frac{\sqrt{5}}{5} \end{bmatrix}$

(d) $U = E_{12}E_2(-1)$, $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}$, $V = I_3$

3 Calculate U, Σ, V; null space, column space bases: (a) First three columns of U, $\{\}$ (b) First two columns of U, third column of V (c) First four columns of U, fifth column of V

5 For (3), use a change of variables $\mathbf{x} = V\mathbf{y}$.

7 Use a change of variables $\mathbf{x} = V\mathbf{y}$ and check that $\|\mathbf{b} - A\mathbf{x}\| = \left\|U^T(\mathbf{b} - A\mathbf{x})\right\| = \left\|U^T\mathbf{b} - U^TAV\mathbf{y}\right\|$.

Section 5.7, Page 385

1 Eigenvalues (a) 10.0083, 4.8368, 4.1950, 1.9599 (b) −0.48119, 3.17009, 1.3111 (c) 3.3123 ± 2.8466i, 1.6877 ± 0.8466i

3 Use Gershgorin to show that 0 is not an eigenvalue of the matrix.

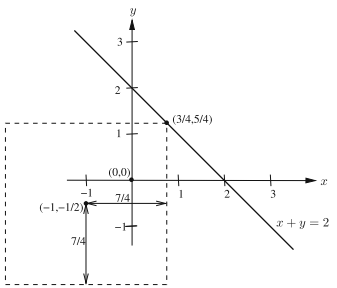
Section 6.1, Page 397

1 (a) 1-norms 6, 5, 2-norms $\sqrt{14}$, $\sqrt{11}$, ∞-norms 3, 3, distance ($\|(-5,0,-4)\|$) in each norm 9, $\sqrt{41}$, 5 (b) 1-norms 7, 8, 2-norms $\sqrt{15}$, $\sqrt{14}$, ∞-norms 3, 2, distance ($\|(1,4,-1,-2,-5)\|$) in each norm 13, $\sqrt{47}$, 5

3 (a) $\frac{1}{5}(1,-3,-1)$, $\frac{1}{\sqrt{11}}(1,-3,-1)$, $\frac{1}{3}(1,-3,-1)$ (b) $\frac{1}{7}(3,1,-1,2)$, $\frac{1}{\sqrt{15}}(3,1,-1,2)$, $\frac{1}{3}(3,1,-1,2)$ (c) $\frac{1}{3+\sqrt{10}}(2,1,3+\mathrm{i})$, $\frac{1}{\sqrt{15}}(2,1,3+\mathrm{i})$, $\frac{1}{\sqrt{10}}(2,1,3+\mathrm{i})$

5 $\|\mathbf{u}\|_1 = 6$, $\|\mathbf{v}\|_1 = 7$, (1) $\|\mathbf{u}\|_1 > 0$, $\|\mathbf{v}\|_1 > 0$ (2) $\|-2(0,2,3,1)\|_1 = 12 = |-2|\,6$ (3) $\|(0,2,3,1)+(1,-3,2,-1)\|_1 = 7 \le 6+7$

7 Ball of radius 7/4 touches line, so distance from point to line in ∞-norm is 7/4.



9 Unit ball $B_1((1,1,1))$ in $\mathbb{R}^3$ with infinity norm is set of points (x,y,z) which are between the pairs of planes (1) $x=0$, $x=2$, (2) $y=0$, $y=2$ and (3) $z=0$, $z=2$.

11 Set $\mathbf{v}=(-1,1)$, $\mathbf{v}-\mathbf{v}_n = \left(\frac{-1}{n}, -e^{-n}\right)$ so $\|\mathbf{v}-\mathbf{v}_n\|_1 = \left(\frac{1}{n}+e^{-n}\right) \overrightarrow{n\to\infty} 0$ and $\|\mathbf{v}-\mathbf{v}_n\|_2 = \sqrt{(\frac{1}{n})^2+(e^{-n})^2} \to 0$, as $n\to\infty$. So $\lim_{n\to\infty}\mathbf{v}_n$ is the same in both norms.

13 Answer: $\max\{|a|+|b|, |c|+|d|\}$. Note that a vector of length one has one coordinate equal to ±1 and the other at most 1 in absolute value.

14 Let $\mathbf{u} = (u_1,\ldots,u_n)$, $\mathbf{v} = (v_1,\ldots,v_n)$, so $|u_1|+\cdots+|u_n| \ge 0$. Also $|cu_1|+\cdots+|cu_n| = |c|\,|u_1|+\cdots+|c|\,|u_n|$ and $|u_1+v_1|+\cdots+|u_n+v_n| \le |u_1|+\cdots+|u_n|+|v_1|+\cdots+|v_n|$.

15 Observation that $\|A\|_F = \|\mathrm{vec}(A)\|_2$ enables you to use known properties of the 2-norm.

Section 6.2, Page 408

1 (a) $|\langle\mathbf{u},\mathbf{v}\rangle| = 46$, $\|\mathbf{u}\| = \sqrt{97}$, $\|\mathbf{v}\| = \sqrt{40}$ and $46 \le \sqrt{97}\sqrt{40} \approx 62.29$ (b) $|\langle\mathbf{u},\mathbf{v}\rangle| = \frac{1}{5}$, $\|\mathbf{u}\| = \frac{1}{\sqrt{3}}$, $\|\mathbf{v}\| = \frac{1}{\sqrt{7}}$ and $\frac{1}{5} = 0.2 \le \frac{1}{\sqrt{3}}\frac{1}{\sqrt{7}} \approx 0.2182$

3 $\mathrm{proj}_{\mathbf{v}}\mathbf{u}$, $\mathrm{comp}_{\mathbf{v}}\mathbf{u}$, $\mathrm{orth}_{\mathbf{v}}\mathbf{u}$: (a) $\left(\frac{-23}{20}, \frac{23}{10}\right)$, $\frac{46}{\sqrt{40}}$, $\left(\frac{63}{20}, \frac{7}{10}\right)$ (b) $\frac{7}{5}x^3$, $\frac{\sqrt{7}}{5}$, $x-\frac{7}{5}x^3$

5 If $\mathbf{x}=(x,y,z)$, equation is $4x-2y+2z=2$.

7 Only (1), since if, e.g., $\mathbf{x} = (0, 1)$, then $\langle \mathbf{x}, \mathbf{x} \rangle = -2 < 0$.

9 (a) orthogonal (b) not orthogonal or orthonormal (c) orthonormal

11 $1(-4) + 2 \cdot 3 \cdot 1 + 2(-1) = 0$. For each $\mathbf{v}$ calculate $\frac{\langle \mathbf{v}_1, \mathbf{v} \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 + \frac{\langle \mathbf{v}_2, \mathbf{v} \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2$. (a) $(11, 7, 8)$, $(11, 7, 8) \in V$ (b) $\left(\frac{2255}{437}, \frac{486}{437}, \frac{1129}{437}\right)$, $(5, 1, 3) \notin V$ (c) $(5, 2, 3)$, $(5, 2, 3) \in V$

13 $\mathbf{v}_i^T A \mathbf{v}_j = 0$ for $i \neq j$. Coordinate vectors: (a) $\left(\frac{7}{2}, \frac{5}{6}, \frac{1}{3}\right)$ (b) $\left(0, \frac{1}{3}, \frac{1}{3}\right)$ (c) $(1, 1, 0)$

15 $ac + \frac{1}{2}(ad + bc) + \frac{1}{3}bd$

17 Express $\mathbf{u}$ and $\mathbf{v}$ in terms of the standard basis $\mathbf{e}_1, \mathbf{e}_2$ and calculate $\langle \mathbf{u}, \mathbf{v} \rangle$.

18 Use the same technique as in Example 6.13 with a suitable choice of specific $\mathbf{u}$ and $\mathbf{v}$.

19 Follow Example 6.8 and use the fact that $\|A\mathbf{u}\|^2 = (A\mathbf{u})^* A\mathbf{u}$.

20 (1) Calculate $\langle \mathbf{u}, \mathbf{0} + \mathbf{0} \rangle$. (2) Use norm law (2), (3) and (2) on $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle$.

22 Express $\|\mathbf{u} + \mathbf{v}\|^2$ and $\|\mathbf{u} - \mathbf{v}\|^2$ in terms of inner products and add.

23 Imitate the steps of Example 6.9.

Section 6.3, Page 408

1 (a) $\frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (c) $\frac{1}{15}\begin{bmatrix} 14 & 1 & -2 & 3 \\ 1 & 14 & 2 & -3 \\ -2 & 2 & 11 & 6 \\ 3 & -3 & 6 & 6 \end{bmatrix}$ (d) $\frac{1}{9}\begin{bmatrix} 5 & 2 & 4 \\ 2 & 8 & -2 \\ 4 & -2 & 5 \end{bmatrix}$

3 $\text{proj}_V \mathbf{w}$, $\text{orth}_V \mathbf{w}$: (a) $\frac{1}{6}(23, -5, 14)$, $\frac{1}{6}(1, -1, -2)$ (b) $\frac{1}{3}(4, 2, 1)$, $\frac{1}{3}(-1, 1, 2)$ (c) $\frac{1}{3}(1, -1, 1)$, $\frac{1}{3}(-1, 1, 2)$

5 $\text{proj}_V x^3 = \frac{1}{10}(9x - 2)$, $\left\|x^3 - \frac{1}{10}(9x - 2)\right\| = \frac{3}{10\sqrt{7}}$

7 Use Gram–Schmidt algorithm on $\mathbf{w}_1 = (-1, 1, 1, -1)$, $\mathbf{w}_2 = (1, 1, 1, 1)$, $\mathbf{w}_3 = (1, 0, 0, 0)$, $\mathbf{w}_4 = (0, 0, 1, 0)$ to obtain orthogonal basis $\mathbf{v}_1 = (-1, 1, 1, -1)$, $\mathbf{v}_2 = (1, 1, 1, 1)$, $\mathbf{v}_3 = \left(\frac{1}{2}, 0, 0, \frac{-1}{2}\right)$, $\mathbf{v}_4 = \left(0, \frac{-1}{2}, \frac{1}{2}, 0\right)$.

9 Use Gram–Schmidt on columns of A and normalize to obtain orthonormal $\frac{1}{\sqrt{3}}(1, 1, 1)$ and $\frac{1}{\sqrt{42}}(1, 2, -5)$, then projection matrix $\frac{1}{14}\begin{bmatrix} 5 & 6 & 3 \\ 6 & 10 & -2 \\ 3 & -2 & 13 \end{bmatrix}$. Use Gram–Schmidt on columns of B and normalize to obtain orthonormal $\frac{1}{3\sqrt{5}}(4, 5, 2)$, $\frac{1}{3\sqrt{70}}(-1, 10, -23)$, then obtain same projection matrix.

12 If a vector $\mathbf{x} \in \mathbb{R}^3$ is projected into $\mathbb{R}^3$, the result is $\mathbf{x}$.

14 Use matrix arithmetic to calculate $\langle P\mathbf{u}, \mathbf{v} - P\mathbf{v} \rangle$.

16 For any $\mathbf{v} \in V$, write $\mathbf{b} - \mathbf{v} = (\mathbf{b} - \mathbf{p}) + (\mathbf{p} - \mathbf{v})$, note that $\mathbf{b} - \mathbf{p}$ is orthogonal to $\mathbf{p} - \mathbf{v}$, which belongs to V, and take norms.

18 Use the Pythagorean theorem and projection formula.

Section 6.4, Page 423

1 $V^\perp = \text{span}\left\{\left(\frac{1}{2}, \frac{5}{2}, 1, 0\right), \left(\frac{-1}{2}, \frac{-1}{2}, 0, 1\right)\right\}$ and if A consists of the columns $\left(\frac{1}{2}, \frac{5}{2}, 1, 0\right)$, $\left(\frac{-1}{2}, \frac{-1}{2}, 0, 1\right)$, $(1, -1, 2, 0)$, $(2, 0, -1, 1)$, then $\det A = 18$ which shows that the columns of A are linearly independent, hence a basis of $\mathbb{R}^4$.

3 $V^\perp = \text{span}\left\{\frac{3}{14} - \frac{38}{35}x + x^2\right\}$

5 $V^{\perp} = \text{span}\left\{\left(-2, \frac{-1}{2}, 1\right)\right\}$ and $\left(V^{\perp}\right)^{\perp} = \text{span}\left\{\left(\frac{1}{2}, 0, 1\right), \left(\frac{-1}{4}, 1, 0\right)\right\}$ which is V since $(1,0,2) = 2\left(\frac{1}{2}, 0, 1\right)$ and $(0,2,1) = \left(\frac{1}{2}, 0, 1\right) + 2\left(\frac{-1}{4}, 1, 0\right)$.

7 $U^{\perp} = \text{span}\{(1,-2,3)\}$, $V^{\perp} = \text{span}\{(-1,1,0)\}$ and $U \cap V = (U^{\perp} + V^{\perp})^{\perp} = \text{span}\{(3,3,1)\}$.

11 (a) Inclusion $U^{\perp} + V^{\perp} \subset (U \cap V)^{\perp}$ follows from the definition and inclusion $U \cap V \subset U + V$. For the converse, show that $(\mathbf{v} - \text{proj}_U \mathbf{v})$ is orthogonal to all $\mathbf{u} \in U$. (b) Use (a) on $U^{\perp}$, $V^{\perp}$.

12 Show that if $A^T A\mathbf{y} = \mathbf{0}$, then $A\mathbf{y} = \mathbf{0}$.

Section 6.5, Page 429

1 Frobenius, 1-, and ∞-norms:
(a) $\sqrt{14}$, 3, 5 (b) $3\sqrt{3}$, 5, 5 (c) $2\sqrt{17}$, 10, 10

3 $\mathbf{x} = (0.4, 0.7)$, $\|\delta\mathbf{x}\|_{\infty} / \|\mathbf{x}\|_{\infty} = 1.6214$, $\text{cond}(A)\, \|\delta\mathbf{b}\|_{\infty} / \|\mathbf{b}\|_{\infty} = 2.5873$

5 Calculate $c = \left\|A^{-1}\delta A\right\| = 0.05\, \|I_3\| = 0.05 < 1$, $\frac{\|\delta A\|}{\|A\|} = 0.05$, $\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = 0.5$, $\text{cond}(A) \approx 6.7807$. Hence, $\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \approx 0.42857 < \frac{\text{cond}(A)}{1-c}\left[\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right] \approx 1.7844$.

8 Use the triangle inequality on A and and Banach lemma on A^{-1}.

9 Factor out A and use Banach lemma.

13 Examine $\|A\mathbf{x}\|$ with $\mathbf{x}$ an eigenvector belonging to λ with $\rho(A) = |\lambda|$ and use definition of matrix norm.

14 If eigenvalue λ satisfies $|\lambda| > 1$, consider $\|A^m\mathbf{x}\|$ with $\mathbf{x}$ an eigenvector belonging to λ. For the rest, use the Jordan canonical form theorem.

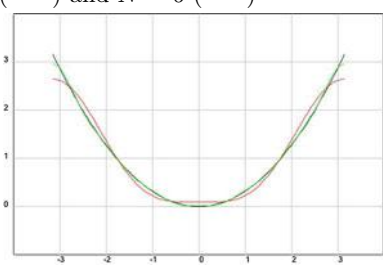
16 (a) Make change of variables $\mathbf{x} = V\mathbf{y}$ and note $\left\|U^T AV\mathbf{x}\right\|_2 = \|A\mathbf{y}\|_2$, $\|\mathbf{x}\| = \|V\mathbf{y}\|$. (c) Use SVD of A.

17 Use the fact that if $U^T AV = \Sigma$, then $A = U\Sigma V^T$ and $A^{-1} = V\Sigma^{-1}U^T$.

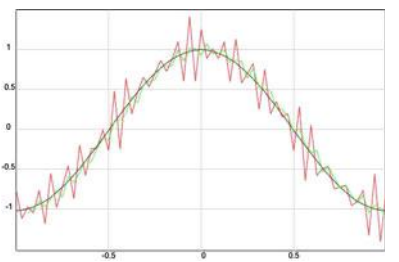
Section 6.6, Page 441

1 $H(\zeta) = e^{i\zeta/2}\cos(\zeta/2)$, so $|H(0)| = 1$ and $|H(\pi)| = 0$.

3 $a_0 = \pi^2/3$, and for $n > 0$, $b_n = 0$ and $a_n = 4(-1)^n/n^2$.
Graph of $x(t)$ (——), Fourier sums $N = 2$ (——) and $N = 6$ (——)



5 From the graph, filter is fairly effective.
Graph of data: Exact (——), noisy (——) and filtered (——).



7 For sampling rates of $T_s = 15, 30, 45$, differences are ≈ 0.6312, 0.5413, 0.2136, resp.

10 Assume $f(t)$ is real and deduce that $c_n = \frac{1}{2}(a_n - ib_n)$ and $c_{-n} = \frac{1}{2}(a_n + ib_n)$ for $n \neq 0$. Next, group terms and write the Fourier series as $c_0 + \sum_{j=1}^{\infty}\left(c_n e^{in\omega t} + c_{-n}e^{-in\omega t}\right)$. Simplify this expression to get the result.

References

1. Åke Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA, 1996.
2. Tomas Akenine-Möller and Eric Haines. *Real-Time Rendering*. A K Peters, Ltd, Natick, MA, 2002.
3. Richard Bellman. *Introduction to Matrix Analysis*. SIAM, Philadelphia, PA, 1997.
4. D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, Nashua, NH, 1997.
5. Kurt Bryan and Tanya Leise. The $25,000,000,000 eigenvector: The linear algebra behind google. *SIAM Rev.*, 48:569–581, 2006.
6. Hal Caswell. *Matrix Population Models*. Sinaur Associates, Sunderland, MA, 2001.
7. G. Caughley. Parameters for seasonally breeding populations. *Ecology*, 48:834–839, 1967.
8. Biswa Nath Datta. *Numerical Linear Algebra and Applications*. Brooks/Cole, New York, 1995.
9. James W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.
10. Patrick J. Van Fleet. *Discrete Wavelet Transformations: An Elementary Approach with Applications*. John Wiley and S, Hoboken, New Jersey, 2008.
11. C. Gasquet and transl. R. Ryan P. Witomski. *Fourier Analysis and Applications*. Springer, New York, 1998.
12. C. F. Gauss. *Theory of the Combination of Observations Least Subject to Errors, Part 1. Part 2, Supplement,* G. W. Stewart. SIAM, Philadelphia, PA, 1995.
13. David Gleich. Pagerank beyond the web. *SIAM Rev.*, 57:321–363, 2015.
14. G. H. Golub and C. F. Van Loan. *Matrix Computations*. McGraw-Hill, Baltimore, Maryland, 1983.
15. Per Christian Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, Philadelphia, PA, 1998.
16. F. S. Hillier and G. J. Lieberman. *Introduction to Operations Research*. Johns Hopkins University Press, Boston, 2010.
17. R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

18. E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley, Hoboken, NJ, 1989.
19. P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, Orlando, Florida, 1985.
20. J. David Logan. *Applied Partial Differential Equations*. Springer, New York, 2015.
21. J. Xu R. Singh and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. U. S. A.*, 105:12763–12768, 2008.
22. K. Shoemaker. Animating rotation with quaternion curves. volume 19, pages 245–254, July 1985.
23. I. Stakgold and M. Holst. *Green's Functions and Boundary Problems, 3rd Ed.* Wiley, Hoboken, NJ, 2011.
24. J. W. Thomas. *Numerical Partial Differential Equations*. Springer, New York, 1998.
25. Lloyd Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.

Index