7 Distance and Approximation

A straight line may be the shortest distance between two points, but it is by no means the most interesting.
—Doctor Who
In "The Time Monster"
By Robert Sloman
BBC, 1972

Although this may seem a paradox, all exact science is dominated by the idea of approximation.
—Bertrand Russell
In W. H. Auden and
L. Kronenberger, eds.
The Viking Book of Aphorisms
Viking, 1962, p. 263

7.0 Introduction: Taxicab Geometry

We live in a three-dimensional Euclidean world, and, therefore, concepts from Euclidean geometry govern our way of looking at the world. In particular, imagine stopping people on the street and asking them to fill in the blank in the following sentence: "The shortest distance between two points is a _____." They will almost certainly respond with "straight line." There are, however, other equally sensible and intuitive notions of distance. By allowing ourselves to think of "distance" in a more flexible way, we will open the door to the possibility of having a "distance" between polynomials, functions, matrices, and many other objects that arise in linear algebra.

In this section, you will discover a type of "distance" that is every bit as real as the straight-line distance you are used to from Euclidean geometry (the one that is a consequence of Pythagoras' Theorem). As you'll see, this new type of "distance" still behaves in some familiar ways.

Suppose you are standing at an intersection in a city, trying to get to a restaurant at another intersection. If you ask someone how far it is to the restaurant, that person is unlikely to measure distance "as the crow flies" (i.e., using the Euclidean version of distance). Instead, the response will be something like "It's five blocks away." Since this is the way taxicab drivers measure distance, we will refer to this notion of "distance" as ***taxicab distance.***

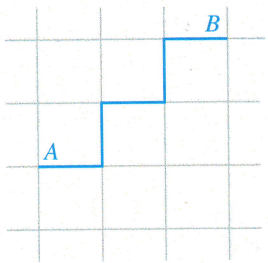


Figure 7.1
Taxicab distance

Figure 7.1 shows an example of taxicab distance. The shortest path from A to B requires traversing the sides of five city blocks. Notice that although there is more than one route from A to B, all shortest routes require three horizontal moves and two vertical moves, where a "move" corresponds to the side of one city block. (How many shortest routes are there from A to B?) Therefore, the taxicab distance from A to B is 5.

Idealizing this situation, we will assume that all blocks are unit squares, and we will use the notation $d_t(A, B)$ for the taxicab distance from A to B.

Problem 1 Find the taxicab distance between the following pairs of points:

(a) $(1, 2)$ and $(5, 5)$ (b) $(2, 4)$ and $(3, -2)$

(c) $(0, 0)$ and $(-4, -3)$ (d) $(-2, 3)$ and $(1, 3)$

(e) $(1, \frac{1}{2})$ and $(-\frac{3}{2}, \frac{3}{2})$ (f) $(2.5, 4.6)$ and $(3.1, 1.5)$

Problem 2 Which of the following is the correct formula for the taxicab distance $d_t(A, B)$ between $A = (a_1, a_2)$ and $B = (b_1, b_2)$?

(a) $d_t(A, B) = (a_1 - b_1) + (a_2 - b_2)$

(b) $d_t(A, B) = (|a_1| - |b_1|) + (|a_2| - |b_2|)$

(c) $d_t(A, B) = |a_1 - b_1| + |a_2 - b_2|$

We can define the ***taxicab norm*** of a vector $\mathbf{v}$ as

$$\|\mathbf{v}\|_t = d_t(\mathbf{v}, \mathbf{0})$$

Problem 3 Find $\|\mathbf{v}\|_t$ for the following vectors:

(a) $\mathbf{v} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$ (b) $\mathbf{v} = \begin{bmatrix} 6 \\ -4 \end{bmatrix}$

(c) $\mathbf{v} = \begin{bmatrix} -3 \\ -6 \end{bmatrix}$ (d) $\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

Problem 4 Show that Theorem 1.3 is true for the taxicab norm.

Problem 5 Verify the Triangle Inequality (Theorem 1.5), using the taxicab norm and the following pairs of vectors:

(a) $\mathbf{u} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ (b) $\mathbf{u} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$

Problem 6 Show that the Triangle Inequality is true, in general, for the taxicab norm.

In Euclidean geometry, we can define a circle of radius r, centered at the origin, as the set of all $\mathbf{x}$ such that $\|\mathbf{x}\| = r$. Analogously, we can define a ***taxicab circle*** of radius r, centered at the origin, as the set of all $\mathbf{x}$ such that $\|\mathbf{x}\|_t = r$.

Problem 7 Draw taxicab circles centered at the origin with the following radii:

(a) $r = 3$ (b) $r = 4$ (c) $r = 1$

Problem 8 In Euclidean geometry, the value of π is half the circumference of a unit circle (a circle of radius 1). Let's define ***taxicab pi*** to be the number π_t that is half the circumference of a taxicab unit circle. What is the value of π_t?

In Euclidean geometry, the perpendicular bisector of a line segment $\overline{AB}$ can be defined as the set of all points that are equidistant from A and B. If we use taxicab distance instead of Euclidean distance, it is reasonable to ask what the perpendicular bisector of a line segment now looks like. To be precise, the ***taxicab perpendicular bisector*** of $\overline{AB}$ is the set of all points X such that

$$d_t(X, A) = d_t(X, B)$$

Problem 9 Draw the taxicab perpendicular bisector of $\overline{AB}$ for the following pairs of points:

(a) $A = (2, 1), B = (4, 1)$ (b) $A = (-1, 3), B = (-1, -2)$

(c) $A = (1, 1), B = (5, 3)$ (d) $A = (1, 1), B = (5, 5)$

As these problems illustrate, taxicab geometry shares some properties with Euclidean geometry, but it also differs in some striking ways. In this chapter, we will

encounter several other types of distances and norms, each of which is useful in its own way. We will try to discover what they have in common and use these common properties to our advantage. We will also explore a variety of approximation problems in which the notion of "distance" plays an important role.



7.1 Inner Product Spaces

In Chapter 1, we defined the dot product $\mathbf{u} \cdot \mathbf{v}$ of vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, and we have made repeated use of this operation throughout this book. In this section, we will use the properties of the dot product as a means of defining the general notion of an *inner product*. In the next section, we will show that inner products can be used to define analogues of "length" and "distance" in vector spaces other than $\mathbb{R}^n$.

The following definition is our starting point; it is based on the properties of the dot product proved in Theorem 1.2.

Definition An ***inner product*** on a vector space V is an operation that assigns to every pair of vectors $\mathbf{u}$ and $\mathbf{v}$ in V a real number $\langle \mathbf{u}, \mathbf{v} \rangle$ such that the following properties hold for all vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in V and all scalars c:

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
2. $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$
3. $\langle c\mathbf{u}, \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$
4. $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ and $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$

A vector space with an inner product is called an ***inner product space.***

Remark Technically, this definition defines a *real* inner product space, since it assumes that V is a real vector space and since the inner product of two vectors is a real number. There are *complex* inner product spaces too, but their definition is somewhat different. (See Exploration: Vectors and Matrices with Complex Entries at the end of this section.)

Example 7.1 $\mathbb{R}^n$ is an inner product space with $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v}$. Properties (1) through (4) were verified as Theorem 1.2.

The dot product is not the only inner product that can be defined on $\mathbb{R}^n$.

Example 7.2 Let $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be two vectors in $\mathbb{R}^2$. Show that

$$\langle \mathbf{u}, \mathbf{v} \rangle = 2u_1v_1 + 3u_2v_2$$

defines an inner product.

Solution We must verify properties (1) through (4). Property (1) holds because

$$\langle \mathbf{u}, \mathbf{v} \rangle = 2u_1v_1 + 3u_2v_2 = 2v_1u_1 + 3v_2u_2 = \langle \mathbf{v}, \mathbf{u} \rangle$$

Next, let $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. We check that

$$\begin{aligned} \langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle &= 2u_1(v_1 + w_1) + 3u_2(v_2 + w_2) \\ &= 2u_1v_1 + 2u_1w_1 + 3u_2v_2 + 3u_2w_2 \\ &= (2u_1v_1 + 3u_2v_2) + (2u_1w_1 + 3u_2w_2) \\ &= \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle \end{aligned}$$

which proves property (2).

If c is a scalar, then

$$\begin{aligned} \langle c\mathbf{u}, \mathbf{v} \rangle &= 2(cu_1)v_1 + 3(cu_2)v_2 \\ &= c(2u_1v_1 + 3u_2v_2) \\ &= c\langle \mathbf{u}, \mathbf{v} \rangle \end{aligned}$$

which verifies property (3).

Finally,

$$\langle \mathbf{u}, \mathbf{u} \rangle = 2u_1u_1 + 3u_2u_2 = 2u_1^2 + 3u_2^2 \geq 0$$

and it is clear that $\langle \mathbf{u}, \mathbf{u} \rangle = 2u_1^2 + 3u_2^2 = 0$ if and only if $u_1 = u_2 = 0$ (that is, if and only if $\mathbf{u} = \mathbf{0}$). This verifies property (4), completing the proof that $\langle \mathbf{u}, \mathbf{v} \rangle$, as defined, is an inner product.



Example 7.2 can be generalized to show that if $w_1, \ldots, w_n$ are *positive* scalars and

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

are vectors in $\mathbb{R}^n$, then

$$\langle \mathbf{u}, \mathbf{v} \rangle = w_1u_1v_1 + \cdots + w_nu_nv_n \tag{1}$$

defines an inner product on $\mathbb{R}^n$, called a ***weighted dot product.*** If any of the weights w_i is negative or zero, then Equation (1) does not define an inner product. (See Exercises 13 and 14.)

Recall that the dot product can be expressed as $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v}$. Observe that we can write the weighted dot product in Equation (1) as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^TW\mathbf{v}$$

where W is the $n \times n$ diagonal matrix

$$W = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

The next example further generalizes this type of inner product.

Example 7.3

Let A be a symmetric, positive definite $n \times n$ matrix (see Section 5.5) and let $\mathbf{u}$ and $\mathbf{v}$ be vectors in $\mathbb{R}^n$. Show that

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T A \mathbf{v}$$

defines an inner product.

Solution We check that

$$\begin{aligned} \langle \mathbf{u}, \mathbf{v} \rangle &= u^T A \mathbf{v} = \mathbf{u} \cdot A\mathbf{v} = A\mathbf{v} \cdot \mathbf{u} \\ &= A^T \mathbf{v} \cdot \mathbf{u} = (\mathbf{v}^T A)^T \cdot \mathbf{u} = \mathbf{v}^T A \mathbf{u} = \langle \mathbf{v}, \mathbf{u} \rangle \end{aligned}$$

Also,

$$\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \mathbf{u}^T A(\mathbf{v} + \mathbf{w}) = \mathbf{u}^T A \mathbf{v} + \mathbf{u}^T A \mathbf{w} = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$$

and

$$\langle c\mathbf{u}, \mathbf{v} \rangle = (c\mathbf{u})^T A \mathbf{v} = c(\mathbf{u}^T A \mathbf{v}) = c\langle \mathbf{u}, \mathbf{v} \rangle$$

Finally, since A is positive definite, $\langle \mathbf{u}, \mathbf{u} \rangle = \mathbf{u}^T A \mathbf{u} > 0$ for all $\mathbf{u} \neq \mathbf{0}$, so $\langle \mathbf{u}, \mathbf{u} \rangle = \mathbf{u}^T A \mathbf{u} = \mathbf{0}$ if and only if $\mathbf{u} = \mathbf{0}$. This establishes the last property.

To illustrate Example 7.3, let $A = \begin{bmatrix} 4 & -2 \\ -2 & 7 \end{bmatrix}$. Then

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T A \mathbf{v} = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} 4 & -2 \\ -2 & 7 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 4u_1v_1 - 2u_1v_2 - 2u_2v_1 + 7u_2v_2$$

The matrix A is positive definite, by Theorem 5.22, since its eigenvalues are 3 and 8. Hence, $\langle \mathbf{u}, \mathbf{v} \rangle$ defines an inner product on $\mathbb{R}^2$.

We now define some inner products on vector spaces other than $\mathbb{R}^n$.

Example 7.4

In $\mathcal{P}_2$, let $p(x) = a_0 + a_1x + a_2x^2$ and $q(x) = b_0 + b_1x + b_2x^2$. Show that

$$\langle p(x), q(x) \rangle = a_0b_0 + a_1b_1 + a_2b_2$$

defines an inner product on $\mathcal{P}_2$. (For example, if $p(x) = 1 - 5x + 3x^2$ and $q(x) = 6 + 2x - x^2$, then $\langle p(x), q(x) \rangle = 1 \cdot 6 + (-5) \cdot 2 + 3 \cdot (-1) = -7$.)

Solution Since $\mathcal{P}_2$ is isomorphic to $\mathbb{R}^3$, we need only show that the dot product in $\mathbb{R}^3$ is an inner product, which we have already established.

Example 7.5

Let f and g be in $\mathscr{C}[a, b]$, the vector space of all continuous functions on the closed interval $[a, b]$. Show that

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,dx$$

defines an inner product on $\mathscr{C}[a, b]$.

Solution We have

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,dx = \int_a^b g(x)f(x)\,dx = \langle g, f \rangle$$

Also, if h is in $\mathscr{C}[a, b]$, then

$$\begin{aligned}\langle f, g + h \rangle &= \int_a^b f(x)(g(x) + h(x))\,dx \\ &= \int_a^b (f(x)g(x) + f(x)h(x))\,dx \\ &= \int_a^b f(x)g(x)\,dx + \int_a^b f(x)h(x)\,dx \\ &= \langle f, g \rangle + \langle f, h \rangle\end{aligned}$$

If c is a scalar, then

$$\begin{aligned}\langle cf, g \rangle &= \int_a^b cf(x)g(x)\,dx \\ &= c\int_a^b f(x)g(x)\,dx \\ &= c\langle f, g \rangle\end{aligned}$$

Finally, $\langle f, f \rangle = \int_a^b (f(x))^2\,dx \geq 0$, and it follows from a theorem of calculus that, since f is continuous, $\langle f, f \rangle = \int_a^b (f(x))^2\,dx = 0$ if and only if f is the zero function. Therefore, $\langle f, g \rangle$ is an inner product on $\mathscr{C}[a, b]$.

Example 7.5 also defines an inner product on any *subspace* of $\mathscr{C}[a, b]$. For example, we could restrict our attention to polynomials defined on the interval $[a, b]$. Suppose we consider $\mathscr{P}[0, 1]$, the vector space of all polynomials on the interval $[0, 1]$. Then, using the inner product of Example 7.5, we have

$$\begin{aligned}\langle x^2, 1 + x \rangle &= \int_0^1 x^2(1 + x)\,dx = \int_0^1 (x^2 + x^3)\,dx \\ &= \left[\frac{x^3}{3} + \frac{x^4}{4}\right]_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}\end{aligned}$$

Properties of Inner Products

The following theorem summarizes some additional properties that follow from the definition of inner product.

Theorem 7.1 Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be vectors in an inner product space V and let c be a scalar.

a. $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$
b. $\langle \mathbf{u}, c\mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$
c. $\langle \mathbf{u}, \mathbf{0} \rangle = \langle \mathbf{0}, \mathbf{v} \rangle = 0$

Proof We prove property (a), leaving the proofs of properties (b) and (c) as Exercises 23 and 24. Referring to the definition of inner product, we have

$$\begin{aligned} \langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle &= \langle \mathbf{w}, \mathbf{u} + \mathbf{v} \rangle && \text{by (1)} \\ &= \langle \mathbf{w}, \mathbf{u} \rangle + \langle \mathbf{w}, \mathbf{v} \rangle && \text{by (2)} \\ &= \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle && \text{by (1)} \end{aligned}$$

Length, Distance, and Orthogonality

In an inner product space, we can define the length of a vector, distance between vectors, and orthogonal vectors, just as we did in Section 1.2. We simply have to replace every use of the dot product $\mathbf{u} \cdot \mathbf{v}$ by the more general inner product $\langle \mathbf{u}, \mathbf{v} \rangle$.

Definition Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in an inner product space V.

1. The ***length*** (or ***norm***) of $\mathbf{v}$ is $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$.
2. The ***distance*** between $\mathbf{u}$ and $\mathbf{v}$ is $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$.
3. $\mathbf{u}$ and $\mathbf{v}$ are ***orthogonal*** if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Note that $\|\mathbf{v}\|$ is always defined, since $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ by the definition of inner product, so we can take the square root of this nonnegative quantity. As in $\mathbb{R}^n$, a vector of length 1 is called a ***unit vector.*** The ***unit sphere*** in V is the set S of all unit vectors in V.



Example 7.6 Consider the inner product on $\mathscr{C}[0, 1]$ given in Example 7.5. If $f(x) = x$ and $g(x) = 3x - 2$, find

(a) $\|f\|$ (b) $d(f, g)$ (c) $\langle f, g \rangle$

Solution (a) We find that

$$\langle f, f \rangle = \int_0^1 f^2(x)\,dx = \int_0^1 x^2\,dx = \frac{x^3}{3}\Big]_0^1 = \frac{1}{3}$$

so $\|f\| = \sqrt{\langle f, f \rangle} = 1/\sqrt{3}$.

(b) Since $d(f, g) = \|f - g\| = \sqrt{\langle f - g, f - g\rangle}$ and

$$f(x) - g(x) = x - (3x - 2) = 2 - 2x = 2(1 - x)$$

we have

$$\langle f - g, f - g\rangle = \int_0^1 (f(x) - g(x))^2\,dx = \int_0^1 4(1 - 2x + x^2)\,dx$$

$$= 4\left[x - x^2 + \frac{x^3}{3}\right]_0^1 = \frac{4}{3}$$

Combining these facts, we see that $d(f, g) = \sqrt{4/3} = 2/\sqrt{3}$.

(c) We compute

$$\langle f, g\rangle = \int_0^1 f(x)g(x)\,dx = \int_0^1 x(3x - 2)\,dx = \int_0^1 (3x^2 - 2x)\,dx = [x^3 - x^2]_0^1 = 0$$

Thus, f and g are orthogonal.

It is important to remember that the "distance" between f and g in Example 7.6 does *not* refer to any measurement related to the graphs of these functions. Neither does the fact that f and g are orthogonal mean that their graphs intersect at right angles. We are simply applying the definition of a particular inner product. However, in doing so, we should be guided by the corresponding notions in $\mathbb{R}^2$ and $\mathbb{R}^3$, where the inner product is the dot product. The geometry of Euclidean space can still guide us here, even though we cannot visualize things in the same way.

Example 7.7

Using the inner product on $\mathbb{R}^2$ defined in Example 7.2, draw a sketch of the unit sphere (circle).

Solution If $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$, then $\langle \mathbf{x}, \mathbf{x}\rangle = 2x^2 + 3y^2$. Since the unit sphere (circle) consists of all $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$, we have

$$1 = \|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} = \sqrt{2x^2 + 3y^2} \quad \text{or} \quad 2x^2 + 3y^2 = 1$$

This is the equation of an ellipse, and its graph is shown in Figure 7.2.

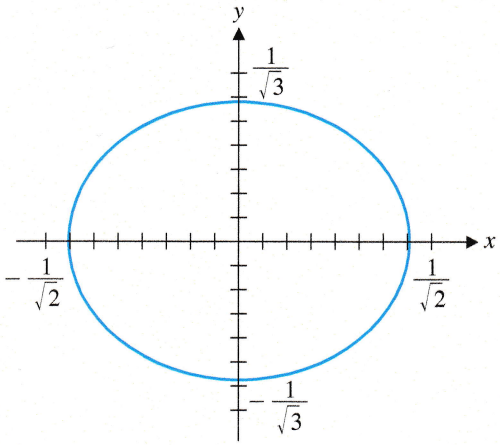


Figure 7.2
A unit circle that is an ellipse

We will discuss properties of length, distance, and orthogonality in the next section and in the exercises. One result that we will need in this section is the generalized version of Pythagoras' Theorem, which extends Theorem 1.6.

Theorem 7.2 **Pythagoras' Theorem**

Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in an inner product space V. Then $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$$

Proof As you will be asked to prove in Exercise 32, we have

$$\|\mathbf{u} + \mathbf{v}\|^2 = \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2$$

It follows immediately that $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$ if and only if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Orthogonal Projections and the Gram-Schmidt Process

In Chapter 5, we discussed orthogonality in $\mathbb{R}^n$. Most of this material generalizes nicely to general inner product spaces. For example, an ***orthogonal set*** of vectors in an inner product space V is a set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of vectors from V such that $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ whenever $\mathbf{v}_i \neq \mathbf{v}_j$. An ***orthonormal set*** of vectors is then an orthogonal set of *unit* vectors. An ***orthogonal basis*** for a subspace W of V is just a basis for W that is an orthogonal set; similarly, an ***orthonormal basis*** for a subspace W of V is a basis for W that is an orthonormal set.

In $\mathbb{R}^n$, the Gram-Schmidt Process (Theorem 5.15) shows that every subspace has an orthogonal basis. We can mimic the construction of the Gram-Schmidt Process to show that every finite-dimensional subspace of an inner product space has an orthogonal basis—all we need to do is replace the dot product by the more general inner product. We illustrate this approach with an example. (Compare the steps here with those in Example 5.13.)

Example 7.8 Construct an orthogonal basis for $\mathcal{P}_2$ with respect to the inner product

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)\,dx$$

by applying the Gram-Schmidt Process to the basis $\{1, x, x^2\}$.

Solution Let $\mathbf{x}_1 = 1$, $\mathbf{x}_2 = x$, and $\mathbf{x}_3 = x^2$. We begin by setting $\mathbf{v}_1 = \mathbf{x}_1 = 1$. Next we compute

$$\langle \mathbf{v}_1, \mathbf{v}_1 \rangle = \int_{-1}^{1} dx = x \Big]_{-1}^{1} = 2 \quad \text{and} \quad \langle \mathbf{v}_1, \mathbf{x}_2 \rangle = \int_{-1}^{1} x\,dx = \frac{x^2}{2} \Big]_{-1}^{1} = 0$$



Science Source/Photo Researchers

Adrien Marie Legendre (1752–1833) was a French mathematician who worked in astronomy, number theory, and elliptic functions. He was involved in several heated disputes with Gauss. Legendre gave the first published statement of the law of quadratic reciprocity in number theory in 1765. Gauss, however, gave the first rigorous proof of this result in 1801 and claimed credit for the result, prompting understandable outrage from Legendre. Then in 1806, Legendre gave the first published application of the method of least squares in a book on the orbits of comets. Gauss published on the same topic in 1809 but claimed he had been using the method since 1795, once again infuriating Legendre.

Therefore,

$$\mathbf{v}_2 = \mathbf{x}_2 - \frac{\langle \mathbf{v}_1, \mathbf{x}_2 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 = x - \frac{0}{2}(1) = x$$

To find $\mathbf{v}_3$, we first compute

$$\langle \mathbf{v}_1, \mathbf{x}_3 \rangle = \int_{-1}^{1} x^2\,dx = \frac{x^3}{3}\Big]_{-1}^{1} = \frac{2}{3}, \qquad \langle \mathbf{v}_2, \mathbf{x}_3 \rangle = \int_{-1}^{1} x^3\,dx = \frac{x^4}{4}\Big]_{-1}^{1} = 0,$$

$$\langle \mathbf{v}_2, \mathbf{v}_2 \rangle = \int_{-1}^{1} x^2\,dx = \frac{2}{3}$$

Then

$$\mathbf{v}_3 = \mathbf{x}_3 - \frac{\langle \mathbf{v}_1, \mathbf{x}_3 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 - \frac{\langle \mathbf{v}_2, \mathbf{x}_3 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle}\mathbf{v}_2 = x^2 - \frac{\frac{2}{3}}{2}(1) - \frac{0}{\frac{2}{3}}x = x^2 - \frac{1}{3}$$

It follows that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is an orthogonal basis for $\mathcal{P}_2$ on the interval $[-1, 1]$. The polynomials

$$1, \quad x, \quad x^2 - \tfrac{1}{3}$$

are the first three ***Legendre polynomials.*** If we divide each of these polynomials by its length relative to the same inner product, we obtain ***normalized Legendre polynomials*** (see Exercise 41).

Just as we did in Section 5.2, we can define the ***orthogonal projection*** $\text{proj}_W(\mathbf{v})$ of a vector $\mathbf{v}$ onto a subspace W of an inner product space. If $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthogonal basis for W, then

$$\text{proj}_W(\mathbf{v}) = \frac{\langle \mathbf{u}_1, \mathbf{v} \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 + \cdots + \frac{\langle \mathbf{u}_k, \mathbf{v} \rangle}{\langle \mathbf{u}_k, \mathbf{u}_k \rangle}\mathbf{u}_k$$

Then the ***component of*** $\mathbf{v}$ ***orthogonal to*** W is the vector

$$\text{perp}_W(\mathbf{v}) = \mathbf{v} - \text{proj}_W(\mathbf{v})$$

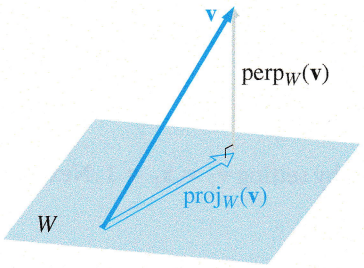


Figure 7.3

As in the Orthogonal Decomposition Theorem (Theorem 5.11), $\text{proj}_W(\mathbf{v})$ and $\text{perp}_W(\mathbf{v})$ are orthogonal (see Exercise 43), and so, schematically, we have the situation illustrated in Figure 7.3.

We will make use of these formulas in Sections 7.3 and 7.5 when we consider approximation problems—in particular, the problem of how best to approximate a

given function by "nice" functions. Consequently, we will defer any examples until then, when they will make more sense. Our immediate use of orthogonal projection will be to prove an inequality that we first encountered in Chapter 1.

The Cauchy-Schwarz and Triangle Inequalities

The proofs of identities and inequalities involving the dot product in $\mathbb{R}^n$ are easily adapted to give corresponding results in general inner product spaces. Some of these are given in Exercises 31–36. In Section 1.2, we first encountered the Cauchy-Schwarz Inequality, which is important in many branches of mathematics. We now give a proof of this result for inner product spaces.

Theorem 7.3 **The Cauchy-Schwarz Inequality**

Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in an inner product space V. Then

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|$$

with equality holding if and only if $\mathbf{u}$ and $\mathbf{v}$ are scalar multiples of each other.

Proof If $\mathbf{u} = \mathbf{0}$, then the inequality is actually an equality, since

$$|\langle \mathbf{0}, \mathbf{v} \rangle| = 0 = \|\mathbf{0}\| \|\mathbf{v}\|$$

If $\mathbf{u} \neq \mathbf{0}$, then let W be the subspace of V spanned by $\mathbf{u}$. Since $\text{proj}_W(\mathbf{v}) = \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}\mathbf{u}$ and $\text{perp}_W \mathbf{v} = \mathbf{v} - \text{proj}_W(\mathbf{v})$ are orthogonal, we can apply Pythagoras' Theorem to obtain

$$\begin{aligned} \|\mathbf{v}\|^2 &= \|\text{proj}_W(\mathbf{v}) + (\mathbf{v} - \text{proj}_W(\mathbf{v}))\|^2 = \|\text{proj}_W(\mathbf{v}) + \text{perp}_W(\mathbf{v})\|^2 \\ &= \|\text{proj}_W(\mathbf{v})\|^2 + \|\text{perp}_W(\mathbf{v})\|^2 \end{aligned} \tag{2}$$

It follows that $\|\text{proj}_W(\mathbf{v})\|^2 \leq \|\mathbf{v}\|^2$. Now

$$\|\text{proj}_W(\mathbf{v})\|^2 = \left\langle \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}\mathbf{u}, \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}\mathbf{u} \right\rangle = \left(\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \right)^2 \langle \mathbf{u}, \mathbf{u} \rangle = \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\|\mathbf{u}\|^2}$$

so we have

$$\frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\|\mathbf{u}\|^2} \leq \|\mathbf{v}\|^2 \quad \text{or, equivalently,} \quad \langle \mathbf{u}, \mathbf{v} \rangle^2 \leq \|\mathbf{u}\|^2 \|\mathbf{v}\|^2$$

Taking square roots, we obtain $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|$.

Clearly this last inequality is an equality if and only if $\|\text{proj}_W(\mathbf{v})\|^2 = \|\mathbf{v}\|^2$. By Equation (2) this is true if and only if $\text{perp}_W(\mathbf{v}) = \mathbf{0}$ or, equivalently,

$$\mathbf{v} = \text{proj}_W(\mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}\mathbf{u}$$

This inequality was discovered by several different mathematicians, in several different contexts. It is no surprise that the name of the prolific Cauchy is attached to it. The second name associated with this result is that of Karl Herman Amandus Schwarz (1843–1921), a German mathematician who taught at the University of Berlin. His version of the inequality that bears his name was published in 1885 in a paper that used integral equations to study surfaces of minimal area. A third name also associated with this important result is that of the Russian mathematician Viktor Yakovlevitch Bunyakovsky (1804–1889). Bunyakovsky published the inequality in 1859, a full quarter-century before Schwarz's work on the same subject. Hence, it is more proper to refer to the result as the Cauchy-Bunyakovsky-Schwarz Inequality.

If this is so, then $\mathbf{v}$ is a scalar multiple of $\mathbf{u}$. Conversely, if $\mathbf{v} = c\mathbf{u}$, then

$$\text{perp}_W(\mathbf{v}) = \mathbf{v} - \text{proj}_W(\mathbf{v}) = c\mathbf{u} - \frac{\langle \mathbf{u}, c\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}\mathbf{u} = c\mathbf{u} - \frac{c\langle \mathbf{u}, \mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}\mathbf{u} = \mathbf{0}$$

so equality holds in the Cauchy-Schwarz Inequality.

For an alternative proof of this inequality, see Exercise 44. We will investigate some interesting consequences of the Cauchy-Schwarz Inequality and related inequalities in Exploration: Geometric Inequalities and Optimization Problems, which follows this section. For the moment, we use it to prove a generalized version of the Triangle Inequality (Theorem 1.5).

Theorem 7.4 The Triangle Inequality

Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in an inner product space V. Then

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$$

Proof Starting with the equality you will be asked to prove in Exercise 32, we have

$$\begin{aligned}
\|\mathbf{u} + \mathbf{v}\|^2 &= \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|^2 \\
&\leq \|\mathbf{u}\|^2 + 2|\langle \mathbf{u}, \mathbf{v}\rangle| + \|\mathbf{v}\|^2 \\
&\leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| + \|\mathbf{v}\|^2 \qquad \text{by Cauchy-Schwarz} \\
&= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2
\end{aligned}$$

Taking square roots yields the result.

Exercises 7.1

In Exercises 1–4, let $\mathbf{u} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$.

1. $\langle \mathbf{u}, \mathbf{v}\rangle$ is the inner product of Example 7.2. Compute

(a) $\langle \mathbf{u}, \mathbf{v}\rangle$ **(b)** $\|\mathbf{u}\|$ **(c)** $d(\mathbf{u}, \mathbf{v})$

2. $\langle \mathbf{u}, \mathbf{v}\rangle$ is the inner product of Example 7.3 with $A = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}$. Compute

(a) $\langle \mathbf{u}, \mathbf{v}\rangle$ **(b)** $\|\mathbf{u}\|$ **(c)** $d(\mathbf{u}, \mathbf{v})$

3. In Exercise 1, find a nonzero vector orthogonal to $\mathbf{u}$.

4. In Exercise 2, find a nonzero vector orthogonal to $\mathbf{u}$.

In Exercises 5–8, let $p(x) = 3 - 2x$ and $q(x) = 1 + x + x^2$.

5. $\langle p(x), q(x)\rangle$ is the inner product of Example 7.4. Compute

(a) $\langle p(x), q(x)\rangle$ **(b)** $\|p(x)\|$ **(c)** $d(p(x), q(x))$

6. $\langle p(x), q(x)\rangle$ is the inner product of Example 7.5 on the vector space $\mathcal{P}_2[0, 1]$. Compute

(a) $\langle p(x), q(x)\rangle$ **(b)** $\|p(x)\|$ **(c)** $d(p(x), q(x))$

7. In Exercise 5, find a nonzero vector orthogonal to $p(x)$.

8. In Exercise 6, find a nonzero vector orthogonal to $p(x)$.

In Exercises 9 and 10, let $f(x) = \sin x$ *and* $g(x) = \sin x + \cos x$ *in the vector space* $\mathcal{C}[0, 2\pi]$ *with the inner product defined by Example 7.5.*

9. Compute

(a) $\langle f, g\rangle$ **(b)** $\|f\|$ **(c)** $d(f, g)$

10. Find a nonzero vector orthogonal to f.

11. Let a, b, and c be distinct real numbers. Show that

$$\langle p(x), q(x)\rangle = p(a)q(a) + p(b)q(b) + p(c)q(c)$$

defines an inner product on $\mathcal{P}_2$. [*Hint:* You will need the fact that a polynomial of degree n has at most n zeros. See Appendix D.]

12. Repeat Exercise 5 using the inner product of Exercise 11 with $a = 0$, $b = 1$, $c = 2$.

In Exercises 13–18, determine which of the four inner product axioms do not hold. Give a specific example in each case.

13. Let $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ in $\mathbb{R}^2$. Define $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1$.

14. Let $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ in $\mathbb{R}^2$. Define $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 - u_2 v_2$.

15. Let $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ in $\mathbb{R}^2$. Define $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_2 + u_2 v_1$.

16. In $\mathcal{P}_2$, define $\langle p(x), q(x) \rangle = p(0)q(0)$.

17. In $\mathcal{P}_2$, define $\langle p(x), q(x) \rangle = p(1)q(1)$.

18. In M_{22}, define $\langle A, B \rangle = \det(AB)$.

In Exercises 19 and 20, $\langle \mathbf{u}, \mathbf{v} \rangle$ defines an inner product on $\mathbb{R}^2$, where $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$. Find a symmetric matrix A such that $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T A \mathbf{v}$.

19. $\langle \mathbf{u}, \mathbf{v} \rangle = 4u_1v_1 + u_1v_2 + u_2v_1 + 4u_2v_2$

20. $\langle \mathbf{u}, \mathbf{v} \rangle = u_1v_1 + 2u_1v_2 + 2u_2v_1 + 5u_2v_2$

In Exercises 21 and 22, sketch the unit circle in $\mathbb{R}^2$ for the given inner product, where $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$.

21. $\langle \mathbf{u}, \mathbf{v} \rangle = u_1v_1 + \frac{1}{4}u_2v_2$

22. $\langle \mathbf{u}, \mathbf{v} \rangle = 4u_1v_1 + u_1v_2 + u_2v_1 + 4u_2v_2$

23. Prove Theorem 7.1(b).

24. Prove Theorem 7.1(c).

In Exercises 25–29, suppose that $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are vectors in an inner product space such that

$$\langle \mathbf{u}, \mathbf{v} \rangle = 1, \quad \langle \mathbf{u}, \mathbf{w} \rangle = 5, \quad \langle \mathbf{v}, \mathbf{w} \rangle = 0$$
$$\|\mathbf{u}\| = 1, \quad \|\mathbf{v}\| = \sqrt{3}, \quad \|\mathbf{w}\| = 2$$

Evaluate the expressions in Exercises 25–28.

25. $\langle \mathbf{u} + \mathbf{w}, \mathbf{v} - \mathbf{w} \rangle$

26. $\langle 2\mathbf{v} - \mathbf{w}, 3\mathbf{u} + 2\mathbf{w} \rangle$

27. $\|\mathbf{u} + \mathbf{v}\|$

28. $\|2\mathbf{u} - 3\mathbf{v} + \mathbf{w}\|$

29. Show that $\mathbf{u} + \mathbf{v} = \mathbf{w}$. [*Hint:* How can you use the properties of inner product to verify that $\mathbf{u} + \mathbf{v} - \mathbf{w} = \mathbf{0}$?]

30. Show that, in an inner product space, there cannot be unit vectors $\mathbf{u}$ and $\mathbf{v}$ with $\langle \mathbf{u}, \mathbf{v} \rangle < -1$.

In Exercises 31–36, $\langle \mathbf{u}, \mathbf{v} \rangle$ is an inner product. In Exercises 31–34, prove that the given statement is an identity.

31. $\langle \mathbf{u} + \mathbf{v}, \mathbf{u} - \mathbf{v} \rangle = \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2$

32. $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2$

33. $\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \frac{1}{2}\|\mathbf{u} + \mathbf{v}\|^2 + \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2$

34. $\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{4}\|\mathbf{u} + \mathbf{v}\|^2 - \frac{1}{4}\|\mathbf{u} - \mathbf{v}\|^2$

35. Prove that $\|\mathbf{u} + \mathbf{v}\| = \|\mathbf{u} - \mathbf{v}\|$ if and only if $\mathbf{u}$ and $\mathbf{v}$ are orthogonal.

36. Prove that $d(\mathbf{u}, \mathbf{v}) = \sqrt{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2}$ if and only if $\mathbf{u}$ and $\mathbf{v}$ are orthogonal.

In Exercises 37–40, apply the Gram-Schmidt Process to the basis $\mathcal{B}$ to obtain an orthogonal basis for the inner product space V relative to the given inner product.

37. $V = \mathbb{R}^2$, $\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$, with the inner product in Example 7.2

38. $V = \mathbb{R}^2$, $\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$, with the inner product immediately following Example 7.3

39. $V = \mathcal{P}_2$, $\mathcal{B} = \{1, 1 + x, 1 + x + x^2\}$, with the inner product in Example 7.4

40. $V = \mathcal{P}_2[0, 1]$, $\mathcal{B} = \{1, 1 + x, 1 + x + x^2\}$, with the inner product in Example 7.5

41. **(a)** Compute the first three normalized Legendre polynomials. (See Example 7.8.)

(b) Use the Gram-Schmidt Process to find the fourth normalized Legendre polynomial.

42. If we multiply the Legendre polynomial of degree n by an appropriate scalar we can obtain a polynomial $L_n(x)$ such that $L_n(1) = 1$.

(a) Find $L_0(x)$, $L_1(x)$, $L_2(x)$, and $L_3(x)$.

(b) It can be shown that $L_n(x)$ satisfies the recurrence relation

$$L_n(x) = \frac{2n - 1}{n} x L_{n-1}(x) - \frac{n - 1}{n} L_{n-2}(x)$$

for all $n \geq 2$. Verify this recurrence for $L_2(x)$ and $L_3(x)$. Then use it to compute $L_4(x)$ and $L_5(x)$.

43. Verify that if W is a subspace of an inner product space V and $\mathbf{v}$ is in V, then $\text{perp}_W(\mathbf{v})$ is orthogonal to all $\mathbf{w}$ in W.

44. Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in an inner product space V. Prove the Cauchy-Schwarz Inequality for $\mathbf{u} \neq \mathbf{0}$ as follows:

(a) Let t be a real scalar. Then $\langle t\mathbf{u} + \mathbf{v}, t\mathbf{u} + \mathbf{v} \rangle \geq 0$ for all values of t. Expand this inequality to obtain a quadratic inequality of the form

$$at^2 + bt + c \geq 0$$

What are a, b, and c in terms of $\mathbf{u}$ and $\mathbf{v}$?

(b) Use your knowledge of quadratic equations and their graphs to obtain a condition on a, b, and c for which the inequality in part (a) is true.

(c) Show that, in terms of $\mathbf{u}$ and $\mathbf{v}$, your condition in part (b) is equivalent to the Cauchy-Schwarz Inequality.

Explorations

Vectors and Matrices with Complex Entries

In this book, we have developed the theory and applications of real vector spaces, the most basic example of which is $\mathbb{R}^n$. We have also explored the finite vector spaces $\mathbb{Z}_p^n$ and their applications. The set $\mathbb{C}^n$ of n-tuples of complex numbers is also a vector space, with the complex numbers $\mathbb{C}$ as scalars. The vector space axioms (Section 6.1) all hold for $\mathbb{C}^n$, and concepts such as linear independence, basis, and dimension carry over from $\mathbb{R}^n$ without difficulty.

The first notable difference between $\mathbb{R}^n$ and $\mathbb{C}^n$ is in the definition of dot product. If we define the dot product in $\mathbb{C}^n$ as in $\mathbb{R}^n$, then for the nonzero vector $\mathbf{v} = \begin{bmatrix} i \\ 1 \end{bmatrix}$ we have

$$\|\mathbf{v}\| = \sqrt{\mathbf{v}\cdot\mathbf{v}} = \sqrt{i^2 + 1^2} = \sqrt{-1 + 1} = \sqrt{0} = 0$$

This is clearly an undesirable situation (a nonzero vector whose length is zero) and violates Theorems 1.2(d) and 1.3. We now generalize the real dot product to $\mathbb{C}^n$ in a way that avoids this type of difficulty.

Definition If $\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$ are vectors in $\mathbb{C}^n$, then the ***complex dot product*** of $\mathbf{u}$ and $\mathbf{v}$ is defined by

$$\mathbf{u}\cdot\mathbf{v} = \overline{u}_1 v_1 + \cdots + \overline{u}_n v_n$$

The norm (or length) of a complex vector $\mathbf{v}$ is defined as in the real case: $\|\mathbf{v}\| = \sqrt{\mathbf{v}\cdot\mathbf{v}}$. Likewise, the distance between two complex vectors $\mathbf{u}$ and $\mathbf{v}$ is still defined as $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|$.

1. Show that, for $\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$ in $\mathbb{C}^n$, $\|\mathbf{v}\| = \sqrt{|v_1|^2 + |v_2|^2 + \cdots + |v_n|^2}$.

2. Let $\mathbf{u} = \begin{bmatrix} i \\ 1 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 2 - 3i \\ 1 + 5i \end{bmatrix}$. Find:

(a) $\mathbf{u} \cdot \mathbf{v}$ (b) $\|\mathbf{u}\|$ (c) $\|\mathbf{v}\|$ (d) $d(\mathbf{u}, \mathbf{v})$ (e) a nonzero vector orthogonal to $\mathbf{u}$
(f) a nonzero vector orthogonal to $\mathbf{v}$

The complex dot product is an example of the more general notion of a complex inner product, which satisfies the same conditions as a real inner product with two exceptions. Problem 3 provides a summary.

3. Prove that the complex dot product satisfies the following properties for all vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in $\mathbb{C}^n$ and all complex scalars.

(a) $\mathbf{u} \cdot \mathbf{v} = \overline{\mathbf{v} \cdot \mathbf{u}}$
(b) $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$
(c) $(c\mathbf{u}) \cdot \mathbf{v} = \bar{c}\,(\mathbf{u} \cdot \mathbf{v})$ and $\mathbf{u} \cdot (c\mathbf{v}) = c(\mathbf{u} \cdot \mathbf{v})$
(d) $\mathbf{u} \cdot \mathbf{u} \geqslant 0$ and $\mathbf{u} \cdot \mathbf{u} = 0$ if and only if $\mathbf{u} = \mathbf{0}$.

For matrices with complex entries, addition, multiplication by complex scalars, transpose, and matrix multiplication are all defined exactly as we did for real matrices in Section 3.1, and the algebraic properties of these operations still hold. (See Section 3.2.) Likewise, we have the notion of the inverse and determinant of a square complex matrix just as in the real case, and the techniques and properties all carry over to the complex case. (See Sections 3.3 and 4.2.)

The notion of transpose is, however, less useful in the complex case than in the real case. The following definition provides an alternative.

Definition If A is a complex matrix, then the ***conjugate transpose*** of A is the matrix A^* defined by

$$A^* = \overline{A}^T$$

In the preceding definition, $\overline{A}$ refers to the matrix whose entries are the complex conjugates of the corresponding entries of A; that is, if $A = [a_{ij}]$, then $\overline{A} = [\bar{a}_{ij}]$.

4. Find the conjugate transpose A^* of the given matrix:

(a) $A = \begin{bmatrix} i & 2i \\ -i & 3 \end{bmatrix}$

(b) $A = \begin{bmatrix} 2 & 5 - 2i \\ 5 + 2i & -1 \end{bmatrix}$

(c) $A = \begin{bmatrix} 2 - i & 1 + 3i & -2 \\ 4 & 0 & 3 - 4i \end{bmatrix}$

(d) $A = \begin{bmatrix} 3i & 0 & 1 + i \\ 1 - i & 4 & i \\ 1 + i & 0 & -i \end{bmatrix}$

Properties of the complex conjugate (Appendix C) extend to matrices, as the next problem shows.

5. Let A and B be complex matrices, and let c be a complex scalar. Prove the following properties:

(a) $\overline{\overline{A}} = A$
(b) $\overline{A + B} = \overline{A} + \overline{B}$
(c) $\overline{cA} = \bar{c}\,\overline{A}$
(d) $\overline{AB} = \overline{A}\,\overline{B}$
(e) $(\overline{A})^T = (\overline{A^T})$

The properties in Problem 5 can be used to establish the following properties of the conjugate transpose, which are analogous to the properties of the transpose for real matrices (Theorem 3.4).

6. Let A and B be complex matrices, and let c be a complex scalar. Prove the following properties:

(a) $(A^*)^* = A$
(b) $(A + B)^* = A^* + B^*$
(c) $(cA)^* = \bar{c}A^*$
(d) $(AB)^* = B^*A^*$

7. Show that for vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{C}^n$, the complex dot product satisfies $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^*\mathbf{v}$. (This result is why we defined the complex dot product as we did. It gives us the analogue of the formula $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v}$ for vectors in $\mathbb{R}^n$.)

For real matrices, we have seen the importance of symmetric matrices, especially in our study of diagonalization. Recall that a real matrix A is symmetric if $A^T = A$. For complex matrices, the following definition is the correct generalization.

Definition A square complex matrix A is called ***Hermitian*** if $A^* = A$—that is, if it is equal to its own conjugate transpose.

Hermitian matrices are named after the French mathematician Charles Hermite (1822–1901). Hermite is best known for his proof that the number e is transcendental, but he also was the first to use the term *orthogonal matrices*, and he proved that symmetric (and Hermitian) matrices have real eigenvalues.

8. Prove that the diagonal entries of a Hermitian matrix must be real.

9. Which of the following matrices are Hermitian?

(a) $A = \begin{bmatrix} 2 & 1+i \\ 1-i & i \end{bmatrix}$

(b) $A = \begin{bmatrix} -1 & 2-3i \\ 2-3i & 5 \end{bmatrix}$

(c) $A = \begin{bmatrix} -3 & -1+5i \\ 1-5i & 3 \end{bmatrix}$

(d) $A = \begin{bmatrix} 1 & 1+4i & 3-i \\ 1-4i & 2 & i \\ 3+i & -i & 0 \end{bmatrix}$

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

(f) $A = \begin{bmatrix} 3 & 0 & -2 \\ 0 & 2 & 1 \\ -2 & 1 & 5 \end{bmatrix}$

10. Prove that the eigenvalues of a Hermitian matrix are real numbers. [*Hint:* The proof of Theorem 5.18 can be adapted by making use of the conjugate transpose operation.]

11. Prove that if A is a Hermitian matrix, then eigenvectors corresponding to distinct eigenvalues of A are orthogonal. [*Hint:* Adapt the proof of Theorem 5.19 using $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^*\mathbf{v}$ instead of $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v}$.]

Recall that a square real matrix Q is orthogonal if $Q^{-1} = Q^T$. The next definition provides the complex analogue.

Definition A square complex matrix U is called ***unitary*** if $U^{-1} = U^*$.

Just as for orthogonal matrices, in practice it is not necessary to compute U^{-1} directly. You need only show that $U^*U = I$ to verify that U is unitary.

12. Which of the following matrices are unitary? For those that are unitary, give their inverses.

(a) $\begin{bmatrix} i/\sqrt{2} & -i/\sqrt{2} \\ i/\sqrt{2} & i/\sqrt{2} \end{bmatrix}$

(b) $\begin{bmatrix} 1+i & 1+i \\ 1-i & -1+i \end{bmatrix}$

(c) $\begin{bmatrix} 3/5 & -4/5 \\ 4i/5 & 3i/5 \end{bmatrix}$

(d) $\begin{bmatrix} (1+i)/\sqrt{6} & 0 & 2/\sqrt{6} \\ 0 & 1 & 0 \\ (-1-i)/\sqrt{3} & 0 & 1/\sqrt{3} \end{bmatrix}$

Unitary matrices behave in most respects like orthogonal matrices. The following problem gives some alternative characterizations of unitary matrices.

13. Prove that the following statements are equivalent for a square complex matrix U:

(a) U is unitary.
(b) The columns of U form an orthonormal set in $\mathbb{C}^n$ with respect to the complex dot product.
(c) The rows of U form an orthonormal set in $\mathbb{C}^n$ with respect to the complex dot product.
(d) $\|U\mathbf{x}\| = \|\mathbf{x}\|$ for every $\mathbf{x}$ in $\mathbb{C}^n$.
(e) $U\mathbf{x} \cdot U\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$ for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{C}^n$.

[*Hint:* Adapt the proofs of Theorems 5.4–5.7.]

14. Repeat Problem 12, this time by applying the criterion in part (b) or part (c) of Problem 13.

The next definition is the natural generalization of orthogonal diagonalizability to complex matrices.

Definition A square complex matrix A is called ***unitarily diagonalizable*** if there exists a unitary matrix U and a diagonal matrix D such that

$$U^*AU = D$$

The process for diagonalizing a unitarily diagonalizable $n \times n$ matrix A mimics the real case. The columns of U must form an orthonormal basis for $\mathbb{C}^n$ consisting of eigenvectors of A. Therefore, we (1) compute the eigenvalues of A, (2) find a basis for each eigenspace, (3) ensure that each eigenspace basis consists of orthonormal vectors (using the Gram-Schmidt Process, with the complex dot product, if necessary), (4) form the matrix U whose columns are the orthonormal eigenvectors just found. Then U^*AU will be a diagonal matrix D whose diagonal entries are the eigenvalues of A, arranged in the same order as the corresponding eigenvectors in the columns of U.

15. In each of the following, find a unitary matrix U and a diagonal matrix D such that $U^*AU = D$.

(a) $A = \begin{bmatrix} 2 & i \\ -i & 2 \end{bmatrix}$

(b) $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

(c) $A = \begin{bmatrix} -1 & 1+i \\ 1-i & 0 \end{bmatrix}$

(d) $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1-i \\ 0 & 1+i & 3 \end{bmatrix}$

The matrices in (a), (c), and (d) of the preceding problem are all Hermitian. It turns out that every Hermitian matrix is unitarily diagonalizable. (This is the *Complex Spectral Theorem,* which can be proved by adapting the proof of Theorem 5.20.) At this point you probably suspect that the converse of this result must also be true—namely, that every unitarily diagonalizable matrix must be Hermitian. But unfortunately this is *false*! (Can you see where the complex analogue of the proof of Theorem 5.17 breaks down?)

For a specific counterexample, take the matrix in part (b) of Problem 15. It is not Hermitian, but it *is* unitarily diagonalizable.

It turns out that the correct characterization of unitary diagonalizability is the following theorem, the proof of which can be found in more advanced textbooks.

See *Linear Algebra with Applications* by S. J. Leon (Upper Saddle River, NJ: Prentice-Hall, 2002).

A square complex matrix A is ***unitarily diagonalizable*** if and only if

$$A^*A = AA^*$$

A matrix A for which $A^*A = AA^*$ is called ***normal.***

16. Show that every Hermitian matrix, every unitary matrix, and every *skew-Hermitian* matrix ($A^* = -A$) is normal. (Note that in the real case, this result refers to symmetric, orthogonal, and skew-symmetric matrices, respectively.)

17. Prove that if a square complex matrix is unitarily diagonalizable, then it must be normal.

Geometric Inequalities and Optimization Problems

This exploration will introduce some powerful (and perhaps surprising) applications of various inequalities, such as the Cauchy-Schwarz Inequality. As you will see, certain maximization/minimization problems (*optimization problems*) that typically arise in a calculus course can be solved without using calculus at all!

Recall that the Cauchy-Schwarz Inequality in $\mathbb{R}^n$ states that for all vectors $\mathbf{u}$ and $\mathbf{v}$,

$$|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\|$$

with equality if and only if $\mathbf{u}$ and $\mathbf{v}$ are scalar multiples of each other. If $\mathbf{u} = [x_1 \ \cdots \ x_n]^T$ and $\mathbf{v} = [y_1 \ \cdots \ y_n]^T$, the above inequality is equivalent to

$$|x_1y_1 + \cdots + x_ny_n| \leq \sqrt{x_1^2 + \cdots + x_n^2}\sqrt{y_1^2 + \cdots + y_n^2}$$

Squaring both sides and using summation notation, we have

$$\left(\sum_{i=1}^{n} x_iy_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)$$

Equality holds if and only if there is some scalar k such that $y_i = kx_i$ for $i = 1, \ldots, n$.

Let's begin by using Cauchy-Schwarz to derive a special case of one of the most useful of all inequalities.

1. Let x and y be nonnegative real numbers. Apply the Cauchy-Schwarz Inequality to $\mathbf{u} = \begin{bmatrix} \sqrt{x} \\ \sqrt{y} \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} \sqrt{y} \\ \sqrt{x} \end{bmatrix}$ to show that

$$\sqrt{xy} \leq \frac{x + y}{2} \tag{1}$$

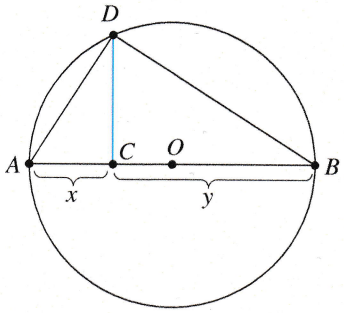


Figure 7.4

with equality if and only if $x = y$.

2. (a) Prove inequality (1) directly. [*Hint:* Square both sides.] (b) Figure 7.4 shows a circle with center O and diameter $AB = AC + CB = x + y$. The segment $\overline{CD}$ is perpendicular to $\overline{AB}$. Prove that $CD = \sqrt{xy}$ and use this result to deduce inequality (1). [*Hint:* Use similar triangles.]

The right-hand side of inequality (1) is the familiar ***arithmetic mean*** (or *average*) of the numbers x and y. The left-hand side shows the less familiar ***geometric mean*** of x and y. Accordingly, inequality (1) is known as the ***Arithmetic Mean–Geometric Mean Inequality (AMGM)***. It holds more generally; for n nonnegative variables $x_1, \ldots, x_n$, it states

$$\sqrt[n]{x_1 x_2 \cdots x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n}$$

with equality if and only if $x_1 = x_2 = \cdots = x_n$.

In words, the AMGM Inequality says that the geometric mean of a set of nonnegative numbers is always less than or equal to their arithmetic mean, and the two are the same precisely when all of the numbers are the same. (For the general proof, see Appendix B.)

We now explore how such an inequality can be applied to optimization problems. Here is a typical calculus problem.

Example 7.9

Prove that among all rectangles whose perimeter is 100 units, the square has the largest area.

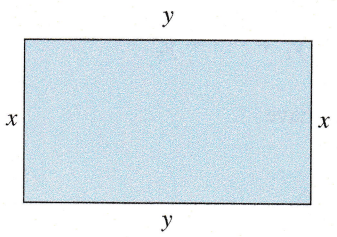


Figure 7.5

Solution If we let x and y be the dimensions of the rectangle (see Figure 7.5), then the area we want to maximize is given by

$$A = xy$$

We are given that the perimeter satisfies

$$2x + 2y = 100$$

which is the same as $x + y = 50$. We can relate xy and $x + y$ using the AMGM Inequality:

$$\sqrt{xy} \le \frac{x+y}{2} \quad \text{or, equivalently,} \quad xy \le \tfrac{1}{4}(x+y)^2$$

Since $x + y = 50$ is a *constant* (and this is the key), we see that the maximum value of $A = xy$ is $50^2/4 = 625$ and it occurs when $x = y = 25$.

Not a derivative in sight! Isn't that impressive? Notice that in this maximization problem, the crucial step was showing that the right-hand side of the AMGM Inequality was *constant*. In a similar fashion, we may be able to apply the inequality to a *minimization* problem if we can arrange for the left-hand side to be constant.

Example 7.10

Prove that among all rectangular prisms with volume 8 m^3, the cube has the minimum surface area.

Solution As shown in Figure 7.6, if the dimensions of such a prism are x, y, and z, then its volume is given by

$$V = xyz$$

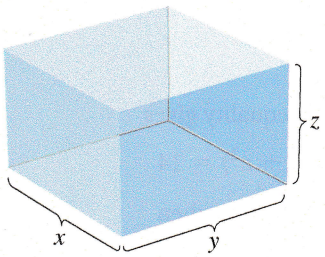


Figure 7.6

Thus, we are given that $xyz = 8$. The surface area to be minimized is

$$S = 2xy + 2yz + 2zx$$

Since this is a three-variable problem, the obvious thing to try is the version of the AMGM Inequality for $n = 3$—namely,

$$\sqrt[3]{xyz} \le \frac{x+y+z}{3}$$

Unfortunately, the expression for S does not appear here. However, the AMGM Inequality also implies that

$$\begin{aligned} \frac{S}{3} &= \frac{2xy + 2yz + 2zx}{3} \\ &\ge \sqrt[3]{(2xy)(2yz)(2zx)} \\ &= 2\sqrt[3]{(xyz)^2} \\ &= 2\sqrt[3]{64} = 8 \end{aligned}$$

which is equivalent to $S \ge 24$. Therefore, the minimum value of S is 24, and it occurs when

$$2xy = 2yz = 2zx$$

(Why?) This implies that $x = y = z = 2$ (i.e., the rectangular prism is a cube).

3. Prove that among all rectangles with area 100 square units, the square has the smallest perimeter.

4. What is the minimum value of $f(x) = x + \dfrac{1}{x}$ for $x > 0$?

5. A cardboard box with a square base and an open top is to be constructed from a square of cardboard 10 cm on a side by cutting out four squares at the corners and folding up the sides. What should the dimensions of the box be in order to make the enclosed volume as large as possible?

6. Find the minimum value of $f(x, y, z) = (x + y)(y + z)(z + x)$ if x, y, and z are positive real numbers such that $xyz = 1$.

7. For $x > y > 0$, find the minimum value of $x + \dfrac{8}{y(x - y)}$. [*Hint:* A substitution might help.]

The Cauchy-Schwarz Inequality itself can be applied to similar problems, as the next example illustrates.

Example 7.11 Find the maximum value of the function $f(x, y, z) = 3x + y + 2z$ subject to the constraint $x^2 + y^2 + z^2 = 1$. Where does the maximum value occur?

Solution This sort of problem is usually handled by techniques covered in a multivariable calculus course. Here's how to use the Cauchy-Schwarz Inequality. The function $3x + y + 2z$ has the form of a dot product, so we let

$$\mathbf{u} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

Then the componentwise form of the Cauchy-Schwarz Inequality gives

$$(3x + y + 2z)^2 \leq (3^2 + 1^2 + 2^2)(x^2 + y^2 + z^2) = 14$$

Thus, the maximum value of our function is $\sqrt{14}$, and it occurs when

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = k \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

Therefore, $x = 3k$, $y = k$, and $z = 2k$, so $3(3k) + k + 2(2k) = \sqrt{14}$. It follows that $k = 1/\sqrt{14}$, and hence

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3/\sqrt{14} \\ 1/\sqrt{14} \\ 2/\sqrt{14} \end{bmatrix}$$

8. Find the maximum value of $f(x, y, z) = x + 2y + 4z$ subject to $x^2 + 2y^2 + z^2 = 1$.

9. Find the minimum value of $f(x, y, z) = x^2 + y^2 + \dfrac{z^2}{2}$ subject to $x + y + z = 10$.

10. Find the maximum value of $\sin\theta + \cos\theta$.

11. Find the point on the line $x + 2y = 5$ that is closest to the origin.

There are many other inequalities that can be used to solve optimization problems. The ***quadratic mean*** of the numbers $x_1, \ldots, x_n$ is defined as

$$\sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}}$$

If $x_1, \ldots, x_n$ are nonzero, their ***harmonic mean*** is given by

$$\frac{n}{1/x_1 + 1/x_2 + \cdots + 1/x_n}$$

It turns out that the quadratic, arithmetic, geometric, and harmonic means are all related.

12. Let x and y be positive real numbers. Show that

$$\sqrt{\frac{x^2 + y^2}{2}} \geq \frac{x + y}{2} \geq \sqrt{xy} \geq \frac{2}{1/x + 1/y}$$

with equality if and only if $x = y$. (The middle inequality is just AMGM, so you need only establish the first and third inequalities.)

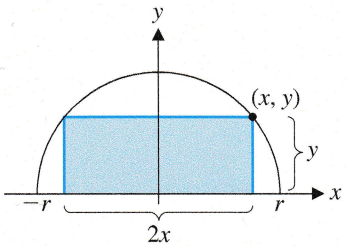


Figure 7.7

13. Find the area of the largest rectangle that can be inscribed in a semicircle of radius r (Figure 7.7).

14. Find the minimum value of the function

$$f(x, y) = \frac{(x + y)^2}{xy}$$

for $x, y > 0$. [*Hint:* $(x + y)^2/xy = (x + y)(1/x + 1/y)$.]

15. Let x and y be positive real numbers with $x + y = 1$. Show that the minimum value of

$$f(x, y) = \left(x + \frac{1}{x}\right)^2 + \left(y + \frac{1}{y}\right)^2$$

is $\frac{25}{2}$, and determine the values of x and y for which it occurs.



7.2 Norms and Distance Functions

In the last section, you saw that it is possible to define length and distance in an inner product space. As you will see shortly, there are also some versions of these two concepts that are not defined in terms of an inner product.

To begin, we need to specify the properties that we want a "length function" to have. The following definition does this, using as its basis Theorem 1.3 and the Triangle Inequality.

Definition A *norm* on a vector space V is a mapping that associates with each vector $\mathbf{v}$ a real number $\|\mathbf{v}\|$, called the *norm* of $\mathbf{v}$, such that the following properties are satisfied for all vectors $\mathbf{u}$ and $\mathbf{v}$ and all scalars c:

1. $\|\mathbf{v}\| \geq 0$, and $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
2. $\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$
3. $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$

A vector space with a norm is called a ***normed linear space.***

Example 7.12

Show that in an inner product space, $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$ defines a norm.

Solution Clearly, $\sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} \geq 0$. Moreover,

$$\sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = 0 \Leftrightarrow \langle \mathbf{v}, \mathbf{v} \rangle = 0 \Leftrightarrow \mathbf{v} = \mathbf{0}$$

by the definition of inner product. This proves property (1).

For property (2), we only need to note that

$$\|c\mathbf{v}\| = \sqrt{\langle c\mathbf{v}, c\mathbf{v} \rangle} = \sqrt{c^2 \langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{c^2}\sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = |c|\|\mathbf{v}\|$$

Property (3) is just the Triangle Inequality, which we verified in Theorem 7.4.

We now look at some examples of norms that are not defined in terms of an inner product. Example 7.13 is the mathematical generalization to $\mathbb{R}^n$ of the taxicab norm that we explored in the Introduction to this chapter.

Example 7.13

The *sum norm* $\|\mathbf{v}\|_s$ of a vector $\mathbf{v}$ in $\mathbb{R}^n$ is the sum of the absolute values of its components. That is, if $\mathbf{v} = [v_1 \cdots v_n]^T$, then

$$\|\mathbf{v}\|_s = |v_1| + \cdots + |v_n|$$

Show that the sum norm is a norm.

Solution Clearly, $\|\mathbf{v}\|_s = |v_1| + \cdots + |v_n| \geq 0$, and the only way to achieve equality is if $|v_1| = \cdots = |v_n| = 0$. But this is so if and only if $v_1 = \cdots = v_n = 0$ or, equivalently, $\mathbf{v} = \mathbf{0}$, proving property (1). For property (2), we see that $c\mathbf{v} = [cv_1 \quad \cdots \quad cv_n]^T$, so

$$\|c\mathbf{v}\|_s = |cv_1| + \cdots + |cv_n| = |c|(|v_1| + \cdots + |v_n|) = |c|\|\mathbf{v}\|_s$$

Finally, the Triangle Inequality holds, because if $\mathbf{u} = [u_1 \quad \cdots \quad u_n]^T$, then

$$\begin{aligned}\|\mathbf{u} + \mathbf{v}\|_s &= |u_1 + v_1| + \cdots + |u_n + v_n| \\ &\leq (|u_1| + |v_1|) + \cdots + (|u_n| + |v_n|) \\ &= (|u_1| + \cdots + |u_n|) + (|v_1| + \cdots + |v_n|) = \|\mathbf{u}\|_s + \|\mathbf{v}\|_s\end{aligned}$$

The sum norm is also known as the **1-*norm*** and is often denoted by $\|\mathbf{v}\|_1$. On $\mathbb{R}^2$, it is the same as the taxicab norm. As Example 7.13 shows, it is possible to have several norms on the same vector space. Example 7.14 illustrates another norm on $\mathbb{R}^n$.

Example 7.14

The ***max norm*** $\|\mathbf{v}\|_m$ of a vector $\mathbf{v}$ in $\mathbb{R}^n$ is the largest number among the absolute values of its components. That is, if $\mathbf{v} = [v_1 \quad \cdots \quad v_n]^T$, then

$$\|\mathbf{v}\|_m = \max\{|v_1|, \ldots, |v_n|\}$$

Show that the max norm is a norm.

Solution Again, it is clear that $\|\mathbf{v}\|_m \geq 0$. If $\|\mathbf{v}\|_m = 0$, then the largest of $|v_1|, \ldots, |v_n|$ is zero, and so they all are. Hence, $v_1 = \cdots = v_n = 0$, so $\mathbf{v} = \mathbf{0}$. This verifies property (1). Next, we observe that for any scalar c,

$$\|c\mathbf{v}\|_m = \max\{|cv_1|, \ldots, |cv_n|\} = |c|\max\{|v_1|, \ldots, |v_n|\} = |c|\|\mathbf{v}\|_m$$

Finally, for $\mathbf{u} = [u_1 \quad \cdots \quad u_n]^T$, we have

$$\begin{aligned}\|\mathbf{u} + \mathbf{v}\|_m &= \max\{|u_1 + v_1|, \ldots, |u_n + v_n|\} \\ &\leq \max\{|u_1| + |v_1|, \ldots, |u_n| + |v_n|\} \\ &\leq \max\{|u_1|, \ldots, |u_n|\} + \max\{|v_1|, \ldots, |v_n|\} = \|\mathbf{u}\|_m + \|\mathbf{v}\|_m\end{aligned}$$

(Why is the second inequality true?) This verifies the Triangle Inequality.

The max norm is also known as the **∞-*norm*** or ***uniform norm*** and is often denoted by $\|\mathbf{v}\|_\infty$. In general, it is possible to define a norm $\|\mathbf{v}\|_p$ on $\mathbb{R}^n$ by

$$\|\mathbf{v}\|_p = (|v_1|^p + \cdots + |v_n|^p)^{1/p}$$

for any real number $p \geq 1$. For $p = 1$, $\|\mathbf{v}\|_1 = \|\mathbf{v}\|_s$, justifying the term 1-norm. For $p = 2$,

$$\|\mathbf{v}\|_2 = (|v_1|^2 + \cdots + |v_n|^2)^{1/2} = \sqrt{v_1^2 + \cdots + v_n^2}$$

which is just the familiar norm on $\mathbb{R}^n$ obtained from the dot product. Called the **2-*norm*** or ***Euclidean norm,*** it is often denoted by $\|\mathbf{v}\|_E$. As p gets large, it can be shown using calculus that $\|\mathbf{v}\|_p$ approaches the max norm $\|\mathbf{v}\|_m$. This justifies the use of the alternative notation $\|\mathbf{v}\|_\infty$ for this norm.

Example 7.15

For a vector $\mathbf{v}$ in $\mathbb{Z}_2^n$, define $\|\mathbf{v}\|_H$ to be $w(\mathbf{v})$, the weight of $\mathbf{v}$. Show that it defines a norm.

Solution Certainly, $\|\mathbf{v}\|_H = w(\mathbf{v}) \geq 0$, and the only vector whose weight is zero is the zero vector. Therefore, property (1) is true. Since the only candidates for a scalar c are 0 and 1, property (2) is immediate.

To verify the Triangle Inequality, first observe that if $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{Z}_2^n$, then $w(\mathbf{u} + \mathbf{v})$ counts the number of places in which $\mathbf{u}$ and $\mathbf{v}$ differ. [For example, if

$$\mathbf{u} = [1 \quad 1 \quad 0 \quad 1 \quad 0]^T \text{ and } \mathbf{v} = [0 \quad 1 \quad 1 \quad 1 \quad 1]^T$$

then $\mathbf{u} + \mathbf{v} = [1 \quad 0 \quad 1 \quad 0 \quad 1]^T$, so $w(\mathbf{u} + \mathbf{v}) = 3$, in agreement with the fact that $\mathbf{u}$ and $\mathbf{v}$ differ in exactly three positions.] Suppose that both $\mathbf{u}$ and $\mathbf{v}$ have zeros in n_0 positions and 1s in n_1 positions, $\mathbf{u}$ has a 0 and $\mathbf{v}$ has a 1 in n_{01} positions, and $\mathbf{u}$ has a 1 and $\mathbf{v}$ has a 0 in n_{10} positions. (In the example above, $n_0 = 0$, $n_1 = 2$, $n_{01} = 2$, and $n_{10} = 1$.) Now

$$w(\mathbf{u}) = n_1 + n_{10}, \; w(\mathbf{v}) = n_1 + n_{01}, \text{ and } w(\mathbf{u} + \mathbf{v}) = n_{10} + n_{01}$$

Therefore,

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|_H = w(\mathbf{u} + \mathbf{v}) &= n_{10} + n_{01} \\ &= (n_1 + n_{10}) + (n_1 + n_{01}) - 2n_1 \\ &\leq (n_1 + n_{10}) + (n_1 + n_{01}) \\ &= w(\mathbf{u}) + w(\mathbf{v}) = \|\mathbf{u}\|_H + \|\mathbf{v}\|_H \end{aligned}$$

The norm $\|\mathbf{v}\|_H$ is called the ***Hamming norm***.

Distance Functions

For any norm, we can define a distance function just as we did in the last section—namely,

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$$

Example 7.16

Let $\mathbf{u} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. Compute $d(\mathbf{u}, \mathbf{v})$ relative to (a) the Euclidean norm, (b) the sum norm, and (c) the max norm.

Solution Each calculation requires knowing that $\mathbf{u} - \mathbf{v} = \begin{bmatrix} 4 \\ -3 \end{bmatrix}$.

(a) As is by now quite familiar,

$$d_E(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_E = \sqrt{4^2 + (-3)^2} = \sqrt{25} = 5$$

(b) $d_s(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_s = |4| + |-3| = 7$

(c) $d_m(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_m = \max\{|4|, |-3|\} = 4$

The distance function on $\mathbb{Z}_2^n$ determined by the Hamming norm is called the ***Hamming distance.*** We will explore its use in error-correcting codes in Section 8.5. Example 7.17 provides an illustration of the Hamming distance.

Example 7.17 Find the Hamming distance between

$$\mathbf{u} = [1 \quad 1 \quad 0 \quad 1 \quad 0]^T \text{ and } \mathbf{v} = [0 \quad 1 \quad 1 \quad 1 \quad 1]^T$$

Solution Since we are working over $\mathbb{Z}_2$, $\mathbf{u} - \mathbf{v} = \mathbf{u} + \mathbf{v}$. But

$$d_H(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} + \mathbf{v}\|_H = w(\mathbf{u} + \mathbf{v})$$

As we noted in Example 7.15, this is just the number of positions in which $\mathbf{u}$ and $\mathbf{v}$ differ. The given vectors are the same ones used in that example; the calculation is therefore exactly the same. Hence, $d_H(\mathbf{u}, \mathbf{v}) = 3$.

Theorem 7.5 summarizes the most important properties of a distance function.

Theorem 7.5 Let d be a distance function defined on a normed linear space V. The following properties hold for all vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ in V:

a. $d(\mathbf{u}, \mathbf{v}) \geq 0$, and $d(\mathbf{u}, \mathbf{v}) = 0$ if and only if $\mathbf{u} = \mathbf{v}$.
b. $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$
c. $d(\mathbf{u}, \mathbf{w}) \leq d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w})$

Proof (a) Using property (1) from the definition of a norm, it is easy to check that $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| \geq 0$, with equality holding if and only if $\mathbf{u} - \mathbf{v} = \mathbf{0}$ or, equivalently, $\mathbf{u} = \mathbf{v}$.

(b) You are asked to prove property (b) in Exercise 19.

(c) We apply the Triangle Inequality to obtain

$$\begin{aligned} d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w}) &= \|\mathbf{u} - \mathbf{v}\| + \|\mathbf{v} - \mathbf{w}\| \\ &\geq \|(\mathbf{u} - \mathbf{v}) + (\mathbf{v} - \mathbf{w})\| \\ &= \|\mathbf{u} - \mathbf{w}\| = d(\mathbf{u}, \mathbf{w}) \end{aligned}$$

A function d satisfying the three properties of Theorem 7.5 is also called a ***metric,*** and a vector space that possesses such a function is called a ***metric space.*** These are very important in many branches of mathematics and are studied in detail in more advanced courses.

Matrix Norms

We can define norms for matrices exactly as we defined norms for vectors in $\mathbb{R}^n$. After all, the vector space M_{mn} of all $m \times n$ matrices is isomorphic to $\mathbb{R}^{mn}$, so this is not difficult to do. Of course, properties (1), (2), and (3) of a norm will also hold in the setting of matrices. It turns out that, for matrices, the norms that are most useful satisfy an additional property. (We will restrict our attention to square matrices, but it is possible to generalize everything to arbitrary matrices.)

Definition A ***matrix norm*** on M_{nn} is a mapping that associates with each $n \times n$ matrix A a real number $\|A\|$, called the ***norm*** of A, such that the following properties are satisfied for all $n \times n$ matrices A and B and all scalars c.

1. $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = O$.
2. $\|cA\| = |c|\|A\|$
3. $\|A + B\| \leq \|A\| + \|B\|$
4. $\|AB\| \leq \|A\|\|B\|$

A matrix norm on M_{nn} is said to be ***compatible*** with a vector norm $\|\mathbf{x}\|$ on $\mathbb{R}^n$ if, for all $n \times n$ matrices A and all vectors $\mathbf{x}$ in $\mathbb{R}^n$, we have

$$\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$$

Example 7.18

The ***Frobenius norm*** $\|A\|_F$ of a matrix A is obtained by stringing out the entries of the matrix into a vector and then taking the Euclidean norm. In other words, $\|A\|_F$ is just the square root of the sum of the squares of the entries of A. So, if $A = [a_{ij}]$, then

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$$

(a) Find the Frobenius norm of

$$A = \begin{bmatrix} 3 & -1 \\ 2 & 4 \end{bmatrix}$$

(b) Show that the Frobenius norm is compatible with the Euclidean norm.

(c) Show that the Frobenius norm is a matrix norm.

Solution (a) $\|A\|_F = \sqrt{3^2 + (-1)^2 + 2^2 + 4^2} = \sqrt{30}$

Before we continue, observe that if $\mathbf{A}_1 = [3 \quad -1]$ and $\mathbf{A}_2 = [2 \quad 4]$ are the row vectors of A, then $\|\mathbf{A}_1\|_E = \sqrt{3^2 + (-1)^2}$ and $\|\mathbf{A}_2\|_E = \sqrt{2^2 + 4^2}$. Thus,

$$\|A\|_F = \sqrt{\|\mathbf{A}_1\|_E^2 + \|\mathbf{A}_2\|_E^2}$$

Similarly, if $\mathbf{a}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and $\mathbf{a}_2 = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$ are the column vectors of A, then

$$\|A\|_F = \sqrt{\|\mathbf{a}_1\|_E^2 + \|\mathbf{a}_2\|_E^2}$$

It is easy to see that these facts extend to $n \times n$ matrices in general. We will use these observations to solve parts (b) and (c).

(b) Write $$A = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_n \end{bmatrix}$$

Then

$$\begin{aligned}\|A\mathbf{x}\|_E &= \left\| \begin{bmatrix} A_1\mathbf{x} \\ \vdots \\ A_n\mathbf{x} \end{bmatrix} \right\|_E \\ &= \sqrt{\|\mathbf{A}_1\mathbf{x}\|_E^2 + \cdots + \|\mathbf{A}_n\mathbf{x}\|_E^2} \\ &\le \sqrt{\|\mathbf{A}_1\|_E^2\|\mathbf{x}\|_E^2 + \cdots + \mathbf{A}_n\|_E^2\|\mathbf{x}\|_E^2} \\ &= (\sqrt{\|\mathbf{A}_1\|_\mathbf{E}^2 + \cdots + \|\mathbf{A}_\mathbf{n}\|_\mathbf{E}^2})\|\mathbf{x}\|_\mathbf{E} \\ &= \|\mathbf{A}\|_F\|\mathbf{x}\|_E\end{aligned}$$

where the inequality arises from the Cauchy-Schwarz Inequality applied to the dot products of the row vectors $\mathbf{A}_i$ with the column vector $\mathbf{x}$. (Do you see how Cauchy-Schwarz has been applied?) Hence, the Frobenius norm is compatible with the Euclidean norm.

(c) Let $\mathbf{b}_i$ denote the ith column of B. Using the matrix-column representation of the product AB, we have

$$\begin{aligned}\|AB\|_F &= \|\,[A\mathbf{b}_1 \cdots A\mathbf{b}_n]\,\|_F \\ &= \sqrt{\|A\mathbf{b}_1\|_E^2 + \cdots + \|A\mathbf{b}_n\|_E^2} \\ &\le \sqrt{\|A\|_F^2\|\mathbf{b}_1\|_\mathbf{E}^2 + \cdots + \|A\|_F^2\|\mathbf{b}_n\|_E^2} \qquad \text{by part (b)} \\ &= \|A\|_F\sqrt{\|\mathbf{b}_1\|_E^2 + \cdots + \|\mathbf{b}_n\|_E^2} \\ &= \|A\|_F\|B\|_F\end{aligned}$$

which proves property (4) of the definition of a matrix norm. Properties (1) through (3) are true, since the Frobenius norm is derived from the Euclidean norm, which satisfies these properties. Therefore, the Frobenius norm is a matrix norm.

For many applications, the Frobenius matrix norm is not the best (or the easiest) one to use. The most useful types of matrix norms arise from considering the effect of the matrix transformation corresponding to the square matrix A. This transformation maps a vector $\mathbf{x}$ into $A\mathbf{x}$. One way to measure the "size" of A is to compare $\|\mathbf{x}\|$ and $\|A\mathbf{x}\|$ using any convenient (vector) norm. Let's think ahead. Whatever definition of $\|A\|$ we arrive at, we know we are going to want it to be compatible with the vector norm we are using; that is, we will need

$$\|A\mathbf{x}\| \le \|A\|\,\|\mathbf{x}\| \quad \text{or} \quad \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \le \|A\| \quad \text{for } \mathbf{x} \neq \mathbf{0}$$

The expression $\dfrac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$ measures the "stretching capability" of A. If we normalize each nonzero vector $\mathbf{x}$ by dividing it by its norm, we get unit vectors $\hat{\mathbf{x}} = \dfrac{1}{\|\mathbf{x}\|}\mathbf{x}$ and thus

$$\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{1}{\|\mathbf{x}\|}\|A\mathbf{x}\| = \left\|\frac{1}{\|\mathbf{x}\|}(A\mathbf{x})\right\| = \left\|A\left(\frac{1}{\|\mathbf{x}\|}\mathbf{x}\right)\right\| = \|A\hat{\mathbf{x}}\|$$

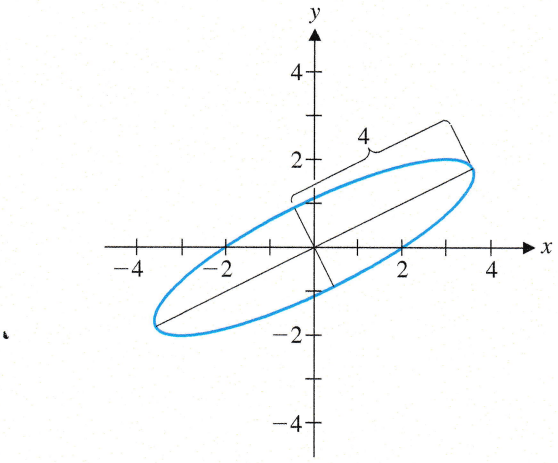


Figure 7.8

If $\mathbf{x}$ ranges over all nonzero vectors in $\mathbb{R}^n$, then $\hat{\mathbf{x}}$ ranges over all *unit* vectors (i.e., the unit sphere) and the set of all vectors $A\hat{\mathbf{x}}$ determines some curve in $\mathbb{R}^n$. For example, Figure 7.8 shows how the matrix $A = \begin{bmatrix} 3 & 2 \\ 2 & 0 \end{bmatrix}$ affects the unit circle in $\mathbb{R}^2$—it maps it into an ellipse. With the Euclidean norm, the maximum value of $\|A\hat{\mathbf{x}}\|$ is clearly just half the length of the principal axis—in this case, 4 units. We express this by writing $\max_{\|\hat{\mathbf{x}}\|=1} \|A\hat{\mathbf{x}}\| = 4$.

In Section 7.4, we will see that this is not an isolated phenomenon. That is,

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\hat{\mathbf{x}}\|=1} \|A\hat{\mathbf{x}}\|$$

always exists, and there is a particular unit vector $\mathbf{y}$ for which $\|A\mathbf{y}\|$ *is* maximum. Now we prove that $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ defines a matrix norm.

Theorem 7.6 If $\|\mathbf{x}\|$ is a vector norm on $\mathbb{R}^n$, then $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ defines a matrix norm on M_{nn} that is compatible with the vector norm that induces it.

Proof (1) Certainly, $\|A\mathbf{x}\| \geq 0$ for all vectors $\mathbf{x}$, so, in particular, this inequality is true if $\|\mathbf{x}\| = 1$. Hence, $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| \geq 0$ also. If $\|A\| = 0$, then we must have $\|A\mathbf{x}\| = 0$—and, hence, $A\mathbf{x} = \mathbf{0}$—for all $\mathbf{x}$ with $\|\mathbf{x}\| = 1$. In particular, $A\mathbf{e}_i = \mathbf{0}$ for each of the standard basis vectors $\mathbf{e}_i$ in $\mathbb{R}^n$. But $A\mathbf{e}_i$ is just the ith column of A, so we must have $A = O$. Conversely, if $A = O$, it is clear that $\|A\| = 0$. (Why?)

(2) Let c be a scalar. Then

$$\|cA\| = \max_{\|\mathbf{x}\|=1} \|cA\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} |c|\,\|A\mathbf{x}\| = |c| \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = |c|\,\|A\|$$

(3) Let B be an $n \times n$ matrix and let $\mathbf{y}$ be a unit vector for which

$$\|A + B\| = \max_{\|\mathbf{x}\|=1} \|(A + B)\mathbf{x}\| = \|(A + B)\mathbf{y}\|$$

Then

$$\begin{aligned} \|A + B\| &= \|(A + B)\mathbf{y}\| \\ &= \|A\mathbf{y} + B\mathbf{y}\| \\ &\leq \|A\mathbf{y}\| + \|B\mathbf{y}\| \\ &\leq \|A\| + \|B\| \end{aligned}$$

(Where does the second inequality come from?) Next, we show that our definition is compatible with the vector norm [property (5)] and then use this fact to complete the proof that we have a matrix norm.

(5) If $\mathbf{x} = \mathbf{0}$, then the inequality $\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$ is true, since both sides are zero. If $\mathbf{x} \neq \mathbf{0}$, then from the comments preceding this theorem,

$$\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \leq \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\|$$

Hence, $\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$.

(4) Let $\mathbf{z}$ be a unit vector such that $\|AB\| = \max_{\|\mathbf{x}\|=1} \|(AB)\mathbf{x}\| = \|AB\mathbf{z}\|$. Then

$$\begin{aligned} \|AB\| &= \|AB\mathbf{z}\| \\ &= \|A(B\mathbf{z})\| \\ &\leq \|A\| \|B\mathbf{z}\| && \text{by property (5)} \\ &\leq \|A\| \|B\| \|\mathbf{z}\| && \text{by property (5)} \\ &= \|A\| \|B\| \end{aligned}$$

This completes the proof that $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ defines a matrix norm on M_{nn} that is compatible with the vector norm that induces it.

Definition The matrix norm $\|A\|$ in Theorem 7.6 is called the ***operator norm*** induced by the vector norm $\|\mathbf{x}\|$.

The term *operator norm* reflects the fact that a matrix transformation arising from a square matrix is also called a *linear operator*. This norm is therefore a measure of the stretching capability of a linear operator.

The three most commonly used operator norms are those induced by the sum norm, the Euclidean norm, and the max norm—namely,

$$\|A\|_1 = \max_{\|\mathbf{x}\|_S=1} \|A\mathbf{x}\|_S, \quad \|A\|_2 = \max_{\|\mathbf{x}\|_E=1} \|A\mathbf{x}\|_E, \quad \|A\|_\infty = \max_{\|\mathbf{x}\|_M=1} \|A\mathbf{x}\|_M$$

respectively. The first and last of these turn out to have especially nice formulas that make them very easy to compute.

Theorem 7.7 Let A be an $n \times n$ matrix with column vectors $\mathbf{a}_i$ and row vectors $\mathbf{A}_i$ for $i = 1, \ldots, n$.

a. $\|A\|_1 = \max_{j=1,\ldots,n} \{\|\mathbf{a}_j\|_s\} = \max_{j=1,\ldots,n} \left\{ \sum_{i=1}^{n} |a_{ij}| \right\}$

b. $\|A\|_\infty = \max_{i=1,\ldots,n} \{\|\mathbf{A}_i\|_s\} = \max_{i=1,\ldots,n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}$

In other words, $\|A\|_1$ is the largest absolute column sum, and $\|A\|_\infty$ is the largest absolute row sum. Before we prove the theorem, let's look at an example to see how easy it is to use.

Example 7.19 Let

$$A = \begin{bmatrix} 1 & -3 & 2 \\ 4 & -1 & -2 \\ -5 & 1 & 3 \end{bmatrix}$$

Find $\|A\|_1$ and $\|A\|_\infty$.

Solution Clearly, the largest absolute column sum is in the first column, so

$$\|A\|_1 = \|\mathbf{a}_1\|_s = |1| + |4| + |-5| = 10$$

The third row has the largest absolute row sum, so

$$\|A\|_\infty = \|\mathbf{A}_3\|_s = |-5| + |1| + |3| = 9$$

With reference to the definition $\|A\|_1 = \max_{\|\mathbf{x}\|_s = 1} \|A\mathbf{x}\|_s$, we see that the maximum value of 10 is actually achieved when we take $\mathbf{x} = \mathbf{e}_1$, for then

$$\|A\mathbf{e}_1\|_s = \|\mathbf{a}_1\|_s = 10 = \|A\|_1$$

For $\|A\|_\infty = \max_{\|\mathbf{x}\|_m = 1} \|A\mathbf{x}\|_m$, if we take

$$\mathbf{x} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

we obtain

$$\|A\mathbf{x}\|_m = \left\| \begin{bmatrix} 1 & -3 & 2 \\ 4 & -1 & -2 \\ -5 & 1 & 3 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \right\|_m = \left\| \begin{bmatrix} -2 \\ -7 \\ 9 \end{bmatrix} \right\|_m$$
$$= \max\{|-2|, |-7|, |9|\} = 9 = \|A\|_\infty$$



We will use these observations in proving Theorem 7.7.

Proof of Theorem 7.7 The strategy is the same in the case of both the column sum and the row sum. If M represents the maximum value, we show that $\|A\mathbf{x}\| \leq M$ for all unit vectors $\mathbf{x}$. Then we find a specific unit vector $\mathbf{x}$ for which equality occurs. It is important to remember that for property (a) the vector norm is the sum norm whereas for property (b) it is the max norm.

(a) To prove (a), let $M = \max_{j=1,\ldots,n} \{\|\mathbf{a}_j\|_s\}$, the maximum absolute column sum, and let $\|\mathbf{x}\|_s = 1$. Then $|x_1| + \cdots + |x_n| = 1$, so

$$\begin{aligned}\|A\mathbf{x}\|_s &= \|x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n\|_s \\ &\leq |x_1|\|\mathbf{a}_1\|_s + \cdots + |x_n|\|\mathbf{a}_n\|_s \\ &\leq |x_1|M + \cdots + |x_n|M \\ &= (|x_1| + \cdots + |x_n|)M = 1 \cdot M = M\end{aligned}$$

If the maximum absolute column sum occurs in column k, then with $\mathbf{x} = \mathbf{e}_k$ we obtain

$$\|A\mathbf{e}_k\|_s = \|\mathbf{a}_k\|_s = M$$

Therefore, $\|A\|_1 = \max_{\|\mathbf{x}\|_s=1} \|A\mathbf{x}\|_s = M = \max_{j=1,\ldots,n} \{\|\mathbf{a}_j\|_s\}$, as required.

(b) The proof of property (b) is left as Exercise 32. ∎

In Section 7.4, we will discover a formula for the operator norm $\|A\|_2$, although it is not as computationally feasible as the formula for $\|A\|_1$ or $\|A\|_\infty$.

The Condition Number of a Matrix

In Exploration: Lies My Computer Told Me in Chapter 2, we encountered the notion of an *ill-conditioned* system of linear equations. Here is the definition as it applies to matrices.

Definition A matrix A is ***ill-conditioned*** if small changes in its entries can produce large changes in the solutions to $A\mathbf{x} = \mathbf{b}$. If small changes in the entries of A produce only small changes in the solutions to $A\mathbf{x} = \mathbf{b}$, then A is called ***well-conditioned.***

Although the definition applies to arbitrary matrices, we will restrict our attention to square matrices.

Example 7.20

Show that $A = \begin{bmatrix} 1 & 1 \\ 1 & 1.0005 \end{bmatrix}$ is ill-conditioned.

Solution If we take $\mathbf{b} = \begin{bmatrix} 3 \\ 3.0010 \end{bmatrix}$, then the solution to $A\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. However, if A changes to

$$A' = \begin{bmatrix} 1 & 1 \\ 1 & 1.0010 \end{bmatrix}$$

then the solution changes to $\mathbf{x}' = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. (Check these assertions.) Therefore, a relative change of $0.0005/1.0005 \approx 0.0005$, or about 0.05%, causes a change of $(2-1)/1 = 1$, or 100%, in x_1 and $(1-2)/2 = -0.5$, or -50%, in x_2. Hence, A is ill-conditioned.

We can use matrix norms to give a more precise way of determining when a matrix is ill-conditioned. Think of the change from A to A' as an error ΔA that, in turn, introduces an error $\Delta\mathbf{x}$ in the solution $\mathbf{x}$ to $A\mathbf{x} = \mathbf{b}$. Then $A' = A + \Delta A$ and $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$. In Example 7.20,

$$\Delta A = \begin{bmatrix} 0 & 0 \\ 0 & 0.0005 \end{bmatrix} \quad \text{and} \quad \Delta\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Then, since $A\mathbf{x} = \mathbf{b}$ and $A'\mathbf{x}' = \mathbf{b}$, we have $(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}$. Expanding and canceling off $A\mathbf{x} = \mathbf{b}$, we obtain

$$A(\Delta\mathbf{x}) + (\Delta A)\mathbf{x} + (\Delta A)(\Delta\mathbf{x}) = 0 \quad \text{or} \quad A(\Delta\mathbf{x}) = -\Delta A(\mathbf{x} + \Delta\mathbf{x})$$

Since we are assuming that $A\mathbf{x} = \mathbf{b}$ has a solution, A must be invertible. Therefore, we can rewrite the last equation as

$$\Delta\mathbf{x} = -A^{-1}(\Delta A)(\mathbf{x} + \Delta\mathbf{x}) = -A^{-1}(\Delta A)\mathbf{x}'$$

Taking norms of both sides (using a matrix norm that is compatible with a vector norm), we have

$$\begin{aligned} \|\Delta\mathbf{x}\| &= \|-A^{-1}(\Delta A)\mathbf{x}'\| = \|A^{-1}(\Delta A)\mathbf{x}'\| \\ &\leq \|A^{-1}(\Delta A)\|\,\|\mathbf{x}'\| \\ &\leq \|A^{-1}\|\,\|\Delta A\|\,\|\mathbf{x}'\| \end{aligned}$$

(What is the justification for each step?) Therefore,

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}'\|} \leq \|A^{-1}\|\,\|\Delta A\| = (\|A^{-1}\|\,\|A\|)\frac{\|\Delta A\|}{\|A\|}$$

The expression $\|A^{-1}\|\,\|A\|$ is called the ***condition number*** of A and is denoted by $\text{cond}(A)$. If A is not invertible, we define $\text{cond}(A) = \infty$.

What are we to make of the inequality just above? The ratio $\|\Delta A\|/\|A\|$ is a measure of the *relative change* in the matrix A, which we are assuming to be small. Similarly, $\|\Delta\mathbf{x}\|/\|\mathbf{x}'\|$ is a measure of the relative error created in the solution to $A\mathbf{x} = \mathbf{b}$ (although, in this case, the error is measured relative to the *new* solution, $\mathbf{x}'$, not the original one, $\mathbf{x}$). Thus, the inequality

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}'\|} \leq \text{cond}(A)\frac{\|\Delta A\|}{\|A\|} \tag{1}$$

gives an upper bound on how large the relative error in the solution can be in terms of the relative error in the coefficient matrix. The larger the condition number, the more ill-conditioned the matrix, since there is more "room" for the error to be large relative to the solution.

Remarks

- The condition number of a matrix depends on the choice of norm. The most commonly used norms are the operator norms $\|A\|_1$ and $\|A\|_\infty$.
- For any norm, $\text{cond}(A) \geq 1$. (See Exercise 45.)

Example 7.21

Find the condition number of $A = \begin{bmatrix} 1 & 1 \\ 1 & 1.0005 \end{bmatrix}$ relative to the ∞-norm.

Solution We first compute

$$A^{-1} = \begin{bmatrix} 2001 & -2000 \\ -2000 & 2000 \end{bmatrix}$$

Therefore, in the ∞-norm (maximum absolute row sum),

$$\|A\|_\infty = 1 + 1.0005 = 2.0005 \quad \text{and} \quad \|A^{-1}\|_\infty = 2001 + |-2000| = 4001$$

so $\text{cond}_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty = 4001(2.0005) \approx 8004.$

It turns out that if the condition number is large relative to one compatible matrix norm, it will be large relative to *any* compatible matrix norm. For example, it can be shown that for matrix A in Examples 7.20 and 7.21, $\text{cond}_1(A) \approx 8004$, $\text{cond}_2(A) \approx 8002$ (relative to the 2-norm), and $\text{cond}_F(A) \approx 8002$ (relative to the Frobenius norm).

The Convergence of Iterative Methods

In Section 2.5, we explored two iterative methods for solving a system of linear equations: Jacobi's method and the Gauss-Seidel method. In Theorem 2.9, we stated without proof that if A is a strictly diagonally dominant $n \times n$ matrix, then both of these methods converge to the solution of $A\mathbf{x} = \mathbf{b}$. We are now in a position to prove this theorem. Indeed, one of the important uses of matrix norms is to establish the convergence properties of various iterative methods.

We will deal only with Jacobi's method here. (The Gauss-Seidel method can be handled using similar techniques, but it requires a bit more care.) The key is to rewrite the iterative process in terms of matrices. Let's revisit Example 2.37 with this in mind. The system of equations is

$$\begin{aligned} 7x_1 - x_2 &= 5 \\ 3x_1 - 5x_2 &= -7 \end{aligned} \tag{2}$$

so $$A = \begin{bmatrix} 7 & -1 \\ 3 & -5 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 5 \\ -7 \end{bmatrix}$$

We rewrote Equation (2) as

$$\begin{aligned} x_1 &= \frac{5 + x_2}{7} \\ x_2 &= \frac{7 + 3x_1}{5} \end{aligned} \tag{3}$$

which is equivalent to

$$\begin{aligned} 7x_1 &= x_2 + 5 \\ -5x_2 &= -3x_1 - 7 \end{aligned} \tag{4}$$

or, in terms of matrices,

$$\begin{bmatrix} 7 & 0 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 5 \\ -7 \end{bmatrix} \tag{5}$$

Study Equation (5) carefully: The matrix on the left-hand side contains the diagonal entries of A, while on the right-hand side we see the *negative* of the off-diagonal entries of A and the vector $\mathbf{b}$. So, if we decompose A as

$$A = \begin{bmatrix} 7 & -1 \\ 3 & -5 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 3 & 0 \end{bmatrix} + \begin{bmatrix} 7 & 0 \\ 0 & -5 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} = L + D + U$$

then Equation (5) can be written as

$$D\mathbf{x} = -(L + U)\mathbf{x} + \mathbf{b}$$

or, equivalently,

$$\mathbf{x} = -D^{-1}(L + U)\mathbf{x} + D^{-1}\mathbf{b} \tag{6}$$

since the matrix D is invertible. Equation (6) is the matrix version of Equation (3). It is easy to see that we can do this in general: An $n \times n$ matrix A can be written as $A = L + D + U$, where D is the diagonal part of A and L and U are, respectively, the portions of A below and above the diagonal. The system $A\mathbf{x} = \mathbf{b}$ can then be written in the form of Equation (6), provided D is invertible—which it is if A is strictly diagonally dominant. (Why?) To simplify the notation, let's let $M = -D^{-1}(L + U)$ and $\mathbf{c} = D^{-1}\mathbf{b}$ so that Equation (6) becomes

$$\mathbf{x} = M\mathbf{x} + \mathbf{c} \tag{7}$$

Recall how we use this equation in Jacobi's method. We start with an initial vector $\mathbf{x}_0$ and plug it into the right-hand side of Equation (7) to get the first iterate $\mathbf{x}_1$—that is, $\mathbf{x}_1 = M\mathbf{x}_0 + \mathbf{c}$. Then we plug $\mathbf{x}_1$ into the right-hand side of Equation (7) to get the second iterate $\mathbf{x}_2 = M\mathbf{x}_1 + \mathbf{c}$. In general, we have

$$\mathbf{x}_{k+1} = M\mathbf{x}_k + \mathbf{c} \tag{8}$$

for $k \geq 0$. For Example 2.37, we have

$$M = -D^{-1}(L + U) = -\begin{bmatrix} 7 & 0 \\ 0 & -5 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -1 \\ 3 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{7} \\ \frac{3}{5} & 0 \end{bmatrix}$$

and

$$\mathbf{c} = D^{-1}\mathbf{b} = \begin{bmatrix} 7 & 0 \\ 0 & -5 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ -7 \end{bmatrix} = \begin{bmatrix} \frac{5}{7} \\ \frac{7}{5} \end{bmatrix}$$

so
$$\mathbf{x}_1 = \begin{bmatrix} 0 & \frac{1}{7} \\ \frac{3}{5} & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{5}{7} \\ \frac{7}{5} \end{bmatrix} = \begin{bmatrix} \frac{5}{7} \\ \frac{7}{5} \end{bmatrix} \approx \begin{bmatrix} 0.714 \\ 1.400 \end{bmatrix}$$
$$\mathbf{x}_2 = \begin{bmatrix} 0 & \frac{1}{7} \\ \frac{3}{5} & 0 \end{bmatrix}\begin{bmatrix} 0.714 \\ 1.400 \end{bmatrix} + \begin{bmatrix} \frac{5}{7} \\ \frac{7}{5} \end{bmatrix} \approx \begin{bmatrix} 0.914 \\ 1.829 \end{bmatrix}$$

and so on. (These are exactly the same calculations we did in Example 2.37, but written in matrix form.)

To show that Jacobi's method will converge, we need to show that the iterates $\mathbf{x}_k$ approach the actual solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$. It is enough to show that the ***error vectors*** $\mathbf{x}_k - \mathbf{x}$ approach the zero vector. From our calculations above, $A\mathbf{x} = \mathbf{b}$ is equivalent to $\mathbf{x} = M\mathbf{x} + \mathbf{c}$. Using Equation (8), we then have

$$\begin{aligned} \mathbf{x}_{k+1} - \mathbf{x} &= M\mathbf{x}_k + \mathbf{c} - (M\mathbf{x} + \mathbf{c}) \\ &= M(\mathbf{x}_k - \mathbf{x}) \end{aligned}$$

Now we take the norm of both sides of this equation. (At this point, it is not important which norm we use as long as we choose a matrix norm that is compatible with a vector norm.) We have

$$\|\mathbf{x}_{k+1} - \mathbf{x}\| = \|M(\mathbf{x}_k - \mathbf{x})\| \le \|M\|\,\|\mathbf{x}_k - \mathbf{x}\| \tag{9}$$

If we can show that $\|M\| < 1$, then we will have $\|\mathbf{x}_{k+1} - \mathbf{x}\| < \|\mathbf{x}_k - \mathbf{x}\|$ for all $k \ge 0$, and it follows that $\|\mathbf{x}_k - \mathbf{x}\|$ approaches zero, so the error vectors $\mathbf{x}_k - \mathbf{x}$ approach the zero vector.

The fact that strict diagonal dominance is defined in terms of the absolute values of the entries in the *rows* of a matrix suggests that the ∞-norm of a matrix (the operator norm induced by the max norm) is the one to choose. If $A = [a_{ij}]$, then

$$M = \begin{bmatrix} 0 & -a_{12}/a_{11} & \cdots & -a_{1n}/a_{11} \\ -a_{21}/a_{22} & 0 & \cdots & -a_{2n}/a_{22} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1}/a_{nn} & -a_{n2}/a_{nn} & \cdots & 0 \end{bmatrix}$$

(verify this), so, by Theorem 7.7, $\|M\|_\infty$ is the maximum absolute row sum of M. Suppose it occurs in the kth row. Then

$$\begin{aligned} \|M\|_\infty &= \left|\frac{-a_{k1}}{a_{kk}}\right| + \cdots + \left|\frac{-a_{k,k-1}}{a_{kk}}\right| + \left|\frac{-a_{k,k+1}}{a_{kk}}\right| + \cdots + \left|\frac{-a_{kn}}{a_{kk}}\right| \\ &= \frac{|a_{k1}| + \cdots + |a_{k,k-1}| + |a_{k,k+1}| + \cdots + |a_{kn}|}{|a_{kk}|} < 1 \end{aligned}$$

since A is strictly diagonally dominant. Thus, $\|M\|_\infty < 1$, so $\|\mathbf{x}_k - \mathbf{x}\| \to 0$, as we wished to show.

Example 7.22

Compute $\|M\|_\infty$ in Example 2.37 and use this value to find the number of iterations required to approximate the solution to three-decimal-place accuracy (after rounding) if the initial vector is $\mathbf{x}_0 = \mathbf{0}$.

Solution We have already computed $M = \begin{bmatrix} 0 & \frac{1}{7} \\ \frac{3}{5} & 0 \end{bmatrix}$, so $\|M\|_\infty = \frac{3}{5} = 0.6 < 1$ (implying that Jacobi's method converges in Example 2.37, as we saw). The approximate solution $\mathbf{x}_k$ will be accurate to three decimal places if the error vector $\mathbf{x}_k - \mathbf{x}$ has the property that each of its components is less than 0.0005 in absolute value. (Why?) Thus, we need only guarantee that the *maximum* absolute component of $\mathbf{x}_k - \mathbf{x}$ is less than 0.0005. In other words, we need to find the smallest value of k such that

$$\|\mathbf{x}_k - \mathbf{x}\|_m < 0.0005$$

Using Equation (9) above, we see that

$$\|\mathbf{x}_k - \mathbf{x}\|_m \leq \|M\|_\infty \|\mathbf{x}_{k-1} - \mathbf{x}\|_m \leq \|M\|_\infty^2 \|\mathbf{x}_{k-2} - \mathbf{x}\|_m \leq \cdots \leq \|M\|_\infty^k \|\mathbf{x}_0 - \mathbf{x}\|_m$$

Now $\|M\|_\infty = 0.6$ and $\|\mathbf{x}_0 - \mathbf{x}\|_m \approx \|\mathbf{x}_0 - \mathbf{x}_1\|_m = \|\mathbf{x}_1\|_m = \left\| \begin{bmatrix} 0.714 \\ 1.400 \end{bmatrix} \right\|_m = 1.4$, so

$$\|M\|_\infty^k \|\mathbf{x}_0 - \mathbf{x}\|_m \approx (0.6)^k (1.4)$$

(If we knew the exact solution in advance, we could use it instead of $\mathbf{x}_1$. In practice, this is not the case, so we use an approximation to the solution, as we have done here.) Therefore, we need to find k such that

$$(0.6)^k (1.4) < 0.0005$$

We can solve this inequality by taking logarithms (base 10) of both sides. We have

$$\begin{aligned} \log_{10}((0.6)^k(1.4)) < \log_{10}(5 \times 10^{-4}) &\Rightarrow k\log_{10}(0.6) + \log_{10}(1.4) < \log_{10}5 - 4 \\ &\Rightarrow -0.222k + 0.146 < -3.301 \\ &\Rightarrow k > 15.5 \end{aligned}$$

Since k must be an integer, we can therefore conclude that $k = 16$ will work and that 16 iterations of Jacobi's method will give us three-decimal-place accuracy in this example. (In fact, it appears from our calculations in Example 2.37 that we get this degree of accuracy sooner, but our goal here was only to come up with an estimate.)

Exercises 7.2

In Exercises 1–3, let $\mathbf{u} = \begin{bmatrix} -1 \\ 4 \\ -5 \end{bmatrix}$ *and* $\mathbf{v} = \begin{bmatrix} 2 \\ -2 \\ 0 \end{bmatrix}$.

1. Compute the Euclidean norm, the sum norm, and the max norm of $\mathbf{u}$.
2. Compute the Euclidean norm, the sum norm, and the max norm of $\mathbf{v}$.
3. Compute $d(\mathbf{u}, \mathbf{v})$ relative to the Euclidean norm, the sum norm, and the max norm.
4. **(a)** What does $d_s(\mathbf{u}, \mathbf{v})$ measure?
   (b) What does $d_m(\mathbf{u}, \mathbf{v})$ measure?

In Exercises 5 and 6, let $\mathbf{u} = [1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1]^T$ *and* $\mathbf{v} = [0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1]^T$.

5. Compute the Hamming norms of $\mathbf{u}$ and $\mathbf{v}$.
6. Compute the Hamming distance between $\mathbf{u}$ and $\mathbf{v}$.
7. **(a)** For which vectors $\mathbf{v}$ is $\|\mathbf{v}\|_E = \|\mathbf{v}\|_m$? Explain your answer.

(b) For which vectors $\mathbf{v}$ is $\|\mathbf{v}\|_s = \|\mathbf{v}\|_m$? Explain your answer.

(c) For which vectors $\mathbf{v}$ is $\|\mathbf{v}\|_s = \|\mathbf{v}\|_m = \|\mathbf{v}\|_E$? Explain your answer.

8. (a) Under what conditions on $\mathbf{u}$ and $\mathbf{v}$ is $\|\mathbf{u} + \mathbf{v}\|_E = \|\mathbf{u}\|_E + \|\mathbf{v}\|_E$? Explain your answer.

(b) Under what conditions on $\mathbf{u}$ and $\mathbf{v}$ is $\|\mathbf{u} + \mathbf{v}\|_s = \|\mathbf{u}\|_s + \|\mathbf{v}\|_s$? Explain your answer.

(c) Under what conditions on $\mathbf{u}$ and $\mathbf{v}$ is $\|\mathbf{u} + \mathbf{v}\|_m = \|\mathbf{u}\|_m + \|\mathbf{v}\|_m$? Explain your answer.

9. Show that for all $\mathbf{v}$ in $\mathbb{R}^n$, $\|\mathbf{v}\|_m \leq \|\mathbf{v}\|_E$.

10. Show that for all $\mathbf{v}$ in $\mathbb{R}^n$, $\|\mathbf{v}\|_E \leq \|\mathbf{v}\|_s$.

11. Show that for all $\mathbf{v}$ in $\mathbb{R}^n$, $\|\mathbf{v}\|_s \leq n\|\mathbf{v}\|_m$.

12. Show that for all $\mathbf{v}$ in $\mathbb{R}^n$, $\|\mathbf{v}\|_E \leq \sqrt{n}\|\mathbf{v}\|_m$.

13. Draw the unit circles in $\mathbb{R}^2$ relative to the sum norm and the max norm.

14. By showing that the identity of Exercise 33 in Section 7.1 fails, show that the sum norm does not arise from any inner product.

In Exercises 15–18, prove that $\| \ \|$ defines a norm on the vector space V.

15. $V = \mathbb{R}^2$, $\left\| \begin{bmatrix} a \\ b \end{bmatrix} \right\| = \max\{|2a|, |3b|\}$

16. $V = M_{mn}$, $\|A\| = \max_{i,j}\{|a_{ij}|\}$

17. $V = \mathscr{C}[0, 1]$, $\|f\| = \int_0^1 |f(x)|\,dx$

18. $\|f\| = \max_{0 \leq x \leq 1} |f(x)|$

19. Prove Theorem 7.5(b).

In Exercises 20–25, compute $\|A\|_F$, $\|A\|_1$, and $\|A\|_\infty$.

20. $A = \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix}$

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

22. $A = \begin{bmatrix} 1 & 5 \\ -2 & -1 \end{bmatrix}$

23. $A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 1 & 3 \end{bmatrix}$

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

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

In Exercises 26–31, find vectors $\mathbf{x}$ and $\mathbf{y}$ with $\|\mathbf{x}\|_s = 1$ and $\|\mathbf{y}\|_m = 1$ such that $\|A\|_1 = \|A\mathbf{x}\|_s$ and $\|A\|_\infty = \|A\mathbf{y}\|_m$, where A is the matrix in the given exercise.

26. Exercise 20 **27.** Exercise 21 **28.** Exercise 22

29. Exercise 23 **30.** Exercise 24 **31.** Exercise 25

32. Prove Theorem 7.7(b).

33. (a) If $\|A\|$ is an operator norm, prove that $\|I\| = 1$, where I is an identity matrix.

(b) Is there a vector norm that induces the Frobenius norm as an operator norm? Why or why not?

34. Let $\|A\|$ be a matrix norm that is compatible with a vector norm $\|\mathbf{x}\|$. Prove that $\|A\| \geq |\lambda|$ for every eigenvalue λ of A.

In Exercises 35–40, find $\text{cond}_1(A)$ and $\text{cond}_\infty(A)$. State whether the given matrix is ill-conditioned.

35. $A = \begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix}$

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

37. $A = \begin{bmatrix} 1 & 0.99 \\ 1 & 1 \end{bmatrix}$

38. $A = \begin{bmatrix} 150 & 200 \\ 3001 & 4002 \end{bmatrix}$

39. $A = \begin{bmatrix} 1 & 1 & 1 \\ 5 & 5 & 6 \\ 1 & 0 & 0 \end{bmatrix}$

40. $A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$

41. Let $A = \begin{bmatrix} 1 & k \\ 1 & 1 \end{bmatrix}$.

(a) Find a formula for $\text{cond}_\infty(A)$ in terms of k.

(b) What happens to $\text{cond}_\infty(A)$ as k approaches 1?

42. Consider the linear system $A\mathbf{x} = \mathbf{b}$, where A is invertible. Suppose an error $\Delta\mathbf{b}$ changes $\mathbf{b}$ to $\mathbf{b}' = \mathbf{b} + \Delta\mathbf{b}$. Let $\mathbf{x}'$ be the solution to the new system; that is, $A\mathbf{x}' = \mathbf{b}'$. Let $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$ so that $\Delta\mathbf{x}$ represents the resulting error in the solution of the system. Show that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \text{cond}(A)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

for any compatible matrix norm.

43. Let $A = \begin{bmatrix} 10 & 10 \\ 10 & 9 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 100 \\ 99 \end{bmatrix}$.

(a) Compute $\text{cond}_\infty(A)$.

(b) Suppose A is changed to $A' = \begin{bmatrix} 10 & 10 \\ 10 & 11 \end{bmatrix}$. How large a relative change can this change produce in the solution to $A\mathbf{x} = \mathbf{b}$? [*Hint:* Use inequality (1) from this section.]

(c) Solve the systems using A and A' and determine the actual relative error.

(d) Suppose $\mathbf{b}$ is changed to $\mathbf{b}' = \begin{bmatrix} 100 \\ 101 \end{bmatrix}$. How large a relative change can this change produce in the solution to $A\mathbf{x} = \mathbf{b}$? [*Hint:* Use Exercise 42.]

(e) Solve the systems using $\mathbf{b}$ and $\mathbf{b}'$ and determine the actual relative error.

44. Let $A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 5 & 0 \\ 1 & -1 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.

(a) Compute $\text{cond}_1(A)$.

(b) Suppose A is changed to $A' = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 0 \\ 1 & -1 & 2 \end{bmatrix}$. How large a relative change can this change produce in the solution to $A\mathbf{x} = \mathbf{b}$? [*Hint:* Use inequality (1) from this section.]

(c) Solve the systems using A and A' and determine the actual relative error.

(d) Suppose $\mathbf{b}$ is changed to $\mathbf{b}' = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$. How large a relative change can this change produce in the solution to $A\mathbf{x} = \mathbf{b}$? [*Hint:* Use Exercise 42.]

(e) Solve the systems using $\mathbf{b}$ and $\mathbf{b}'$ and determine the actual relative error.

45. Show that if A is an invertible matrix, then $\text{cond}(A) \geq 1$ with respect to any matrix norm.

46. Show that if A and B are invertible matrices, then $\text{cond}(AB) \leq \text{cond}(A)\text{cond}(B)$ with respect to any matrix norm.

47. Let A be an invertible matrix and let λ_1 and λ_n be the eigenvalues with the largest and smallest absolute values, respectively. Show that

$$\text{cond}(A) \geq \frac{|\lambda_1|}{|\lambda_n|}$$

[*Hint:* See Exercise 34 and Theorem 4.18(b) in Section 4.3.]

CAS *In Exercises 48–51, write the given system in the form of Equation (7). Then use the method of Example 7.22 to estimate the number of iterations of Jacobi's method that will be needed to approximate the solution to three-decimal-place accuracy. (Use $\mathbf{x}_0 = \mathbf{0}$.) Compare your answer with the solution computed in the given exercise from Section 2.5.*

48. Exercise 1, Section 2.5 **49.** Exercise 3, Section 2.5

50. Exercise 4, Section 2.5 **51.** Exercise 5, Section 2.5

Exercise 52(c) refers to the Leontief model of an open economy, as discussed in Sections 2.4 and 3.7.

52. Let A be an $n \times n$ matrix such that $\|A\| < 1$, where the norm is either the sum norm or the max norm.

(a) Prove that $A^n \to O$ as $n \to \infty$.

(b) Deduce from (a) that $I - A$ is invertible and

$$(I - A)^{-1} = I + A + A^2 + A^3 + \cdots$$

[*Hint:* See the proof of Theorem 3.34.]

(c) Show that (b) can be used to prove Corollaries 3.35 and 3.36.



7.3 Least Squares Approximation

In many branches of science, experimental data are used to infer a mathematical relationship among the variables being measured. For example, we might measure the height of a tree at various points in time and try to deduce a function that expresses the tree's height h in terms of time t. Or, we might measure the size p of a population over time and try to find a rule that relates p to t. Relationships between variables are also of interest in business; for example, a company producing widgets may be interested in knowing the relationship between its total costs c and the number n of widgets produced.

In each of these examples, the data come in the form of two measurements: one for the independent variable and one for the (supposedly) dependent variable. Thus, we have a set of *data points* (x_i, y_i), and we are looking for a function that best approximates the relationship between the independent variable x and the dependent variable y. Figure 7.9 shows examples in which experimental data points are plotted, along with a curve that approximately "fits" the data.

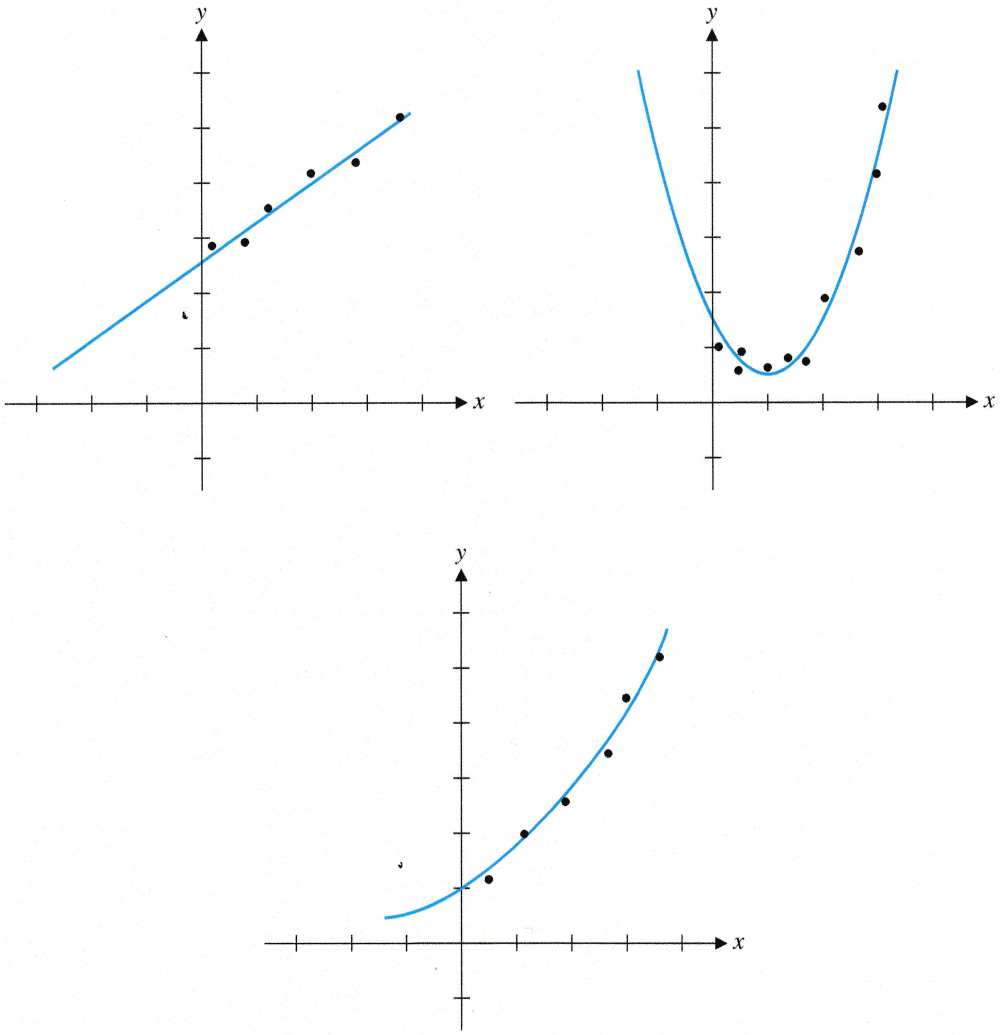


Figure 7.9
Curves of "best fit"

Roger Cotes (1682–1716) was an English mathematician who, while a fellow at Cambridge, edited the second edition of Newton's *Principia*. Although he published little, he made important discoveries in the theory of logarithms, integral calculus, and numerical methods.

The method of least squares, which we are about to consider, is attributed to Gauss. A new asteroid, Ceres, was discovered on New Year's Day, 1801, but it disappeared behind the sun shortly after it was observed. Astronomers predicted when and where Ceres would reappear, but their calculations differed greatly from those done, independently, by Gauss. Ceres reappeared on December 7, 1801, almost exactly where Gauss had predicted it would be. Although he did not disclose his methods at the time, Gauss had used his least squares approximation method, which he described in a paper in 1809. The same method was actually known earlier; Cotes anticipated the method in the early 18th century, and Legendre published a paper on it in 1806. Nevertheless, Gauss is generally given credit for the method of least squares approximation.

We begin our exploration of approximation with a more general result.

The Best Approximation Theorem

In the sciences, there are many problems that can be phrased generally as "What is the best approximation to X of type Y?" X might be a set of data points, a function, a vector, or many other things, while Y might be a particular type of function, a vector belonging to a certain vector space, etc. A typical example of such a problem is finding the vector $\mathbf{w}$ in a subspace W of a vector space V that best approximates (i.e., is closest to) a given vector $\mathbf{v}$ in V. This problem gives rise to the following definition.

Definition If W is a subspace of a normed linear space V and if $\mathbf{v}$ is a vector in V, then the ***best approximation to*** $\mathbf{v}$ ***in*** W is the vector $\bar{\mathbf{v}}$ in W such that

$$\|\mathbf{v} - \bar{\mathbf{v}}\| < \|\mathbf{v} - \mathbf{w}\|$$

for every vector $\mathbf{w}$ in W different from $\bar{\mathbf{v}}$.

In $\mathbb{R}^2$ or $\mathbb{R}^3$, we are used to thinking of "shortest distance" as corresponding to "perpendicular distance." In algebraic terminology, "shortest distance" relates to the notion of orthogonal projection: If W is a subspace of $\mathbb{R}^n$ and $\mathbf{v}$ is a vector in $\mathbb{R}^n$, then we expect $\text{proj}_W(\mathbf{v})$ to be the vector in W that is closest to $\mathbf{v}$ (Figure 7.10).

Since orthogonal projection can be defined in any inner product space, we have the following theorem.

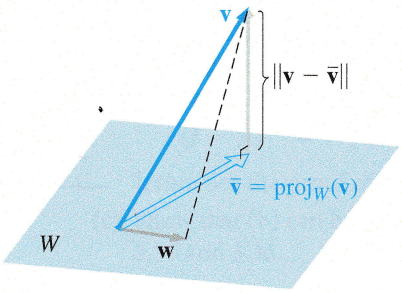


Figure 7.10
If $\bar{\mathbf{v}} = \text{proj}_W(\mathbf{v})$, then
$\|\mathbf{v} - \bar{\mathbf{v}}\| < \|\mathbf{v} - \mathbf{w}\|$ for all $\mathbf{w} \neq \bar{\mathbf{v}}$

Theorem 7.8 **The Best Approximation Theorem**

If W is a finite-dimensional subspace of an inner product space V and if $\mathbf{v}$ is a vector in V, then $\text{proj}_W(\mathbf{v})$ is the best approximation to $\mathbf{v}$ in W.

Proof Let $\mathbf{w}$ be a vector in W different from $\text{proj}_W(\mathbf{v})$. Then $\text{proj}_W(\mathbf{v}) - \mathbf{w}$ is also in W, so $\mathbf{v} - \text{proj}_W(\mathbf{v}) = \text{perp}_W(\mathbf{v})$ is orthogonal to $\text{proj}_W(\mathbf{v}) - \mathbf{w}$, by Exercise 43 in Section 7.1. Pythagoras' Theorem now implies that

$$\begin{aligned}\|\mathbf{v} - \text{proj}_W(\mathbf{v})\|^2 + \|\text{proj}_W(\mathbf{v}) - \mathbf{w}\|^2 &= \|(\mathbf{v} - \text{proj}_W(\mathbf{v})) + (\text{proj}_W(\mathbf{v}) - \mathbf{w})\|^2 \\ &= \|\mathbf{v} - \mathbf{w}\|^2\end{aligned}$$

as Figure 7.10 illustrates. However, $\|\text{proj}_W(\mathbf{v}) - \mathbf{w}\|^2 > 0$, since $\mathbf{w} \neq \text{proj}_W(\mathbf{v})$, so

$$\|\mathbf{v} - \text{proj}_W(\mathbf{v})\|^2 < \|\mathbf{v} - \text{proj}_W(\mathbf{v})\|^2 + \|\text{proj}_W(\mathbf{v}) - \mathbf{w}\|^2 = \|\mathbf{v} - \mathbf{w}\|^2$$

or, equivalently,

$$\|\mathbf{v} - \text{proj}_W(\mathbf{v})\| < \|\mathbf{v} - \mathbf{w}\|$$

Example 7.23

Let $\mathbf{u}_1 = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$, $\mathbf{u}_2 = \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix}$, and $\mathbf{v} = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}$. Find the best approximation to $\mathbf{v}$ in the plane $W = \text{span}(\mathbf{u}_1, \mathbf{u}_2)$ and find the Euclidean distance from $\mathbf{v}$ to W.

Solution The vector in W that best approximates $\mathbf{v}$ is $\text{proj}_W(\mathbf{v})$. Since $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal,

$$\begin{aligned}\text{proj}_W(\mathbf{v}) &= \left(\frac{\mathbf{u}_1 \cdot \mathbf{v}}{\mathbf{u}_1 \cdot \mathbf{u}_1}\right)\mathbf{u}_1 + \left(\frac{\mathbf{u}_2 \cdot \mathbf{v}}{\mathbf{u}_2 \cdot \mathbf{u}_2}\right)\mathbf{u}_2 \\ &= \tfrac{2}{6}\begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} + \tfrac{16}{30}\begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ -\frac{2}{5} \\ \frac{1}{5} \end{bmatrix}\end{aligned}$$

The distance from $\mathbf{v}$ to W is the distance from $\mathbf{v}$ to the point in W closest to $\mathbf{v}$. But this distance is just $\|\text{perp}_W(\mathbf{v})\| = \|\mathbf{v} - \text{proj}_W(\mathbf{v})\|$. We compute

$$\mathbf{v} - \text{proj}_W(\mathbf{v}) = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ -\frac{2}{5} \\ \frac{1}{5} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{12}{5} \\ \frac{24}{5} \end{bmatrix}$$

so $$\|\mathbf{v} - \text{proj}_W(\mathbf{v})\| = \sqrt{0^2 + (\tfrac{12}{5})^2 + (\tfrac{24}{5})^2} = \sqrt{\tfrac{720}{25}} = 12\sqrt{5}/5$$

which is the distance from $\mathbf{v}$ to W.

In Section 7.5, we will look at other examples of the Best Approximation Theorem when we explore the problem of approximating functions.

Remark The orthogonal projection of a vector $\mathbf{v}$ onto a subspace W is defined in terms of an orthogonal basis for W. The Best Approximation Theorem gives us an alternative proof that $\text{proj}_W(\mathbf{v})$ does not depend on the choice of this basis, since there can be only one vector in W that is closest to $\mathbf{v}$—namely, $\text{proj}_W(\mathbf{v})$.

Least Squares Approximation

We now turn to the problem of finding a curve that "best fits" a set of data points. Before we can proceed, however, we need to define what we mean by "best fit." Suppose the data points (1, 2), (2, 2), and (3, 4) have arisen from measurements taken during some experiment. Also suppose we have reason to believe that the x and y values are related by a linear function; that is, we expect the points to lie on some line with equation $y = a + bx$. If our measurements were accurate, all three points would satisfy this equation and we would have

$$2 = a + b \cdot 1 \qquad 2 = a + b \cdot 2 \qquad 4 = a + b \cdot 3$$

This is a system of three linear equations in two variables:

$$\begin{aligned} a + b &= 2 \\ a + 2b &= 2 \\ a + 3b &= 4 \end{aligned} \quad \text{or} \quad \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}$$

Unfortunately, this system is inconsistent (since the three points do not lie on a straight line). So we will settle for a line that comes "as close as possible" to passing through our points. For any line, we will measure the vertical distance from each data point to the line (representing the *errors* in the y-direction), and then we will try to choose the line that minimizes the *total error*. Figure 7.11 illustrates the situation.

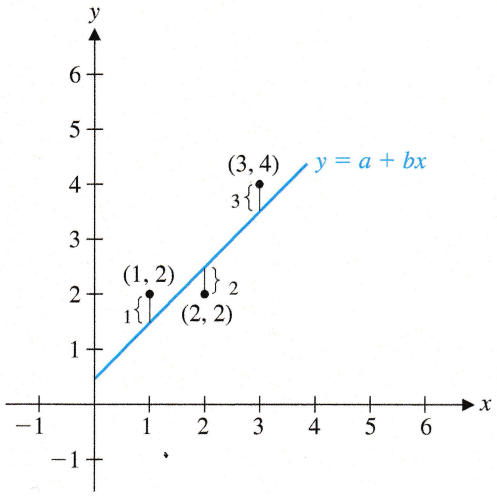


Figure 7.11
Finding the line that minimizes $\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2$

If the errors are denoted by ε_1, ε_2, and ε_3, then we can form the ***error vector***

$$\mathbf{e} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

We want $\mathbf{e}$ to be as small as possible, so $\|\mathbf{e}\|$ must be as close to zero as possible. Which norm should we use? It turns out that the familiar Euclidean norm is the best choice. (The sum norm would also be a sensible choice, since $\|\mathbf{e}\|_s = |\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3|$ is the actual sum of the errors in Figure 7.11. However, the absolute value signs are hard to work with, and, as you will soon see, the choice of the Euclidean norm leads to some very nice formulas.) So we are going to minimize

$$\|\mathbf{e}\| = \sqrt{\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2} \quad \text{or, equivalently,} \quad \|\mathbf{e}\|^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2$$

This is where the term "least squares" comes from: We need to find the smallest sum of squares, in the sense of the foregoing equation. The number $\|\mathbf{e}\|$ is called the ***least squares error*** of the approximation.

From Figure 7.11, we also obtain the following formulas for ε_1, ε_2, and ε_3 in our example:

$$\varepsilon_1 = 2 - (a + b \cdot 1) \quad \varepsilon_2 = 2 - (a + b \cdot 2) \quad \varepsilon_3 = 4 - (a + b \cdot 3)$$

Example 7.24

Which of the following lines gives the smallest least squares error for the data points (1, 2), (2, 2), and (3, 4)?

(a) $y = 1 + x$

(b) $y = -2 + 2x$

(c) $y = \frac{2}{3} + x$

Solution Table 7.1 shows the necessary calculations.

Table 7.1

	$y = 1 + x$	$y = -2 + 2x$	$y = \frac{2}{3} + x$
ε_1	$2 - (1 + 1) = 0$	$2 - (-2 + 2) = 2$	$2 - (\frac{2}{3} + 1) = \frac{1}{3}$
ε_2	$2 - (1 + 2) = -1$	$2 - (-2 + 4) = 0$	$2 - (\frac{2}{3} + 2) = -\frac{2}{3}$
ε_3	$4 - (1 + 3) = 0$	$4 - (-2 + 6) = 0$	$4 - (\frac{2}{3} + 3) = \frac{1}{3}$
$\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2$	$0^2 + (-1)^2 + 0^2 = 1$	$2^2 + 0^2 + 0^2 = 4$	$(\frac{1}{3})^2 + (-\frac{2}{3})^2 + (\frac{1}{3})^2 = \frac{2}{3}$
$\|\mathbf{e}\|$	1	2	$\sqrt{\frac{2}{3}} \approx 0.816$

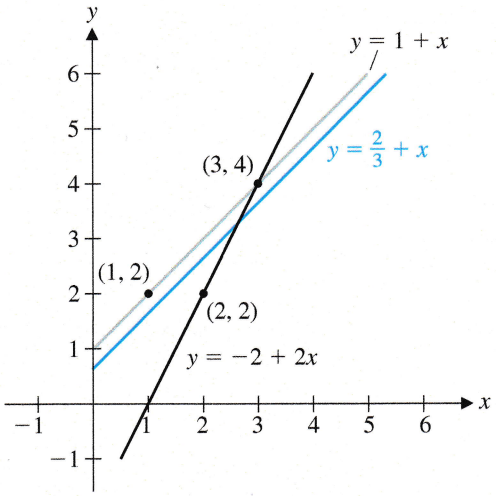


Figure 7.12

We see that the line $y = \frac{2}{3} + x$ produces the smallest least squares error among these three lines. Figure 7.12 shows the data points and all three lines.

It turns out that the line $y = \frac{2}{3} + x$ in Example 7.24 gives the smallest least squares error of *any* line, even though it passes through *none* of the given points. The rest of this section is devoted to illustrating why this is so.

In general, suppose we have n data points $(x_1, y_1), \ldots, (x_n, y_n)$ and a line $y = a + bx$. Our error vector is

$$\mathbf{e} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where $\varepsilon_i = y_i - (a + bx_i)$. The line $y = a + bx$ that minimizes $\varepsilon_1^2 + \cdots + \varepsilon_n^2$ is called the ***least squares approximating line*** (or the ***line of best fit***) for the points $(x_1, y_1), \ldots, (x_n, y_n)$. As noted prior to Example 7.24, we can express this problem in matrix form. If the given points were actually on the line $y = a + bx$, then the n linear equations

$$\begin{aligned} a + bx_1 &= y_1 \\ &\vdots \\ a + bx_n &= y_n \end{aligned}$$

would all be true (i.e., the system would be consistent). Our interest is in the case where the points are *not* collinear, in which case the system is *inconsistent*. In matrix form, we have

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

which is of the form $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The error vector $\mathbf{e}$ is just $\mathbf{b} - A\mathbf{x}$ (check this), and we want to minimize $\|\mathbf{e}\|^2$ or, equivalently, $\|\mathbf{e}\|$. We can therefore rephrase our problem in terms of matrices as follows.

Definition If A is an $m \times n$ matrix and $\mathbf{b}$ is in $\mathbb{R}^m$, a ***least squares solution*** of $A\mathbf{x} = \mathbf{b}$ is a vector $\bar{\mathbf{x}}$ in $\mathbb{R}^n$ such that

$$\|\mathbf{b} - A\bar{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$$

for all $\mathbf{x}$ in $\mathbb{R}^n$.

Solution of the Least Squares Problem

Any vector of the form $A\mathbf{x}$ is in the column space of A, and as $\mathbf{x}$ varies over all vectors in $\mathbb{R}^n$, $A\mathbf{x}$ varies over all vectors in $\text{col}(A)$. A least squares solution of $A\mathbf{x} = \mathbf{b}$ is therefore equivalent to a vector $\bar{\mathbf{y}}$ in $\text{col}(A)$ such that

$$\|\mathbf{b} - \bar{\mathbf{y}}\| \leq \|\mathbf{b} - \mathbf{y}\|$$

for all $\mathbf{y}$ in $\text{col}(A)$. In other words, we need the closest vector in $\text{col}(A)$ to $\mathbf{b}$. By the Best Approximation Theorem, the vector we want is the orthogonal projection of $\mathbf{b}$ onto $\text{col}(A)$. Thus, if $\bar{\mathbf{x}}$ is a least squares solution of $A\mathbf{x} = \mathbf{b}$, we have

$$A\bar{\mathbf{x}} = \text{proj}_{\text{col}(A)}(\mathbf{b}) \tag{1}$$

In order to find $\bar{\mathbf{x}}$, it would appear that we need to first compute $\text{proj}_{\text{col}(A)}(\mathbf{b})$ and then solve the system (1). However, there is a better way to proceed.

We know that

$$\mathbf{b} - A\bar{\mathbf{x}} = \mathbf{b} - \text{proj}_{\text{col}(A)}(\mathbf{b}) = \text{perp}_{\text{col}(A)}(\mathbf{b})$$

is orthogonal to $\text{col}(A)$. So $\mathbf{b} - A\bar{\mathbf{x}}$ is in $(\text{col}(A))^{\perp} = \text{null}(A^T)$. Therefore $A^T(\mathbf{b} - A\bar{\mathbf{x}}) = \mathbf{0}$, which, in turn, is equivalent to $A^T\mathbf{b} - A^TA\bar{\mathbf{x}} = \mathbf{0}$ or

$$A^TA\bar{\mathbf{x}} = A^T\mathbf{b}$$

This represents a system of equations known as the ***normal equations*** for $\bar{\mathbf{x}}$.

We have just established that the solutions of the normal equations for $\bar{\mathbf{x}}$ are precisely the least squares solutions of $A\mathbf{x} = \mathbf{b}$. This proves the first part of the following theorem.

Theorem 7.9 **The Least Squares Theorem**

Let A be an $m \times n$ matrix and let $\mathbf{b}$ be in $\mathbb{R}^m$. Then $A\mathbf{x} = \mathbf{b}$ always has at least one least squares solution $\bar{\mathbf{x}}$. Moreover:

a. $\bar{\mathbf{x}}$ is a least squares solution of $A\mathbf{x} = \mathbf{b}$ if and only if $\bar{\mathbf{x}}$ is a solution of the normal equations $A^TA\bar{\mathbf{x}} = A^T\mathbf{b}$.

b. A has linearly independent columns if and only if A^TA is invertible. In this case, the least squares solution of $A\mathbf{x} = \mathbf{b}$ is unique and is given by

$$\bar{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$$

Proof We have already established property (a). For property (b), we note that the n columns of A are linearly independent if and only if $\text{rank}(A) = n$. But this is true if and only if A^TA is invertible, by Theorem 3.28. If A^TA is invertible, then the unique solution of $A^TA\bar{\mathbf{x}} = A^T\mathbf{b}$ is clearly $\bar{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$.

Example 7.25 Find a least squares solution to the inconsistent system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & 5 \\ 2 & -2 \\ -1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}$$

Solution We compute

$$A^TA = \begin{bmatrix} 1 & 2 & -1 \\ 5 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 2 & -2 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 0 \\ 0 & 30 \end{bmatrix}$$

and

$$A^T\mathbf{b} = \begin{bmatrix} 1 & 2 & -1 \\ 5 & -2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 2 \\ 16 \end{bmatrix}$$

The normal equations $A^TA\bar{\mathbf{x}} = A^T\mathbf{b}$ are just

$$\begin{bmatrix} 6 & 0 \\ 0 & 30 \end{bmatrix} \bar{\mathbf{x}} = \begin{bmatrix} 2 \\ 16 \end{bmatrix}$$

which yield $\bar{\mathbf{x}} = \begin{bmatrix} \frac{1}{3} \\ \frac{8}{15} \end{bmatrix}$. The fact that this solution is unique was guaranteed by Theorem 7.9(b), since the columns of A are clearly linearly independent.

Remark We could have phrased Example 7.25 as follows: Find the best approximation to $\mathbf{b}$ in the column space of A. The resulting equations give the system $A\mathbf{x} = \mathbf{b}$ whose least squares solution we just found. (Verify this.) In this case, the components of $\bar{\mathbf{x}}$ are the *coefficients* of that linear combination of the columns of A that produces the best approximation to $\mathbf{b}$—namely,

$$\tfrac{1}{3}\begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} + \tfrac{8}{15}\begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ -\frac{2}{5} \\ \frac{1}{5} \end{bmatrix}$$

This is exactly the result of Example 7.23. Compare the two approaches.

Example 7.26 Find the least squares approximating line for the data points (1, 2), (2, 2), and (3, 4) from Example 7.24.

Solution We have already seen that the corresponding system $A\mathbf{x} = \mathbf{b}$ is

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}$$

where $y = a + bx$ is the line we seek. Since the columns of A are clearly linearly independent, there will be a unique least squares solution, by part (b) of the Least Squares Theorem. We compute

$$A^TA = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \quad \text{and} \quad A^T\mathbf{b} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 8 \\ 18 \end{bmatrix}$$

Hence, we can solve the normal equations $A^TA\bar{\mathbf{x}} = A^T\mathbf{b}$, using Gaussian elimination to obtain

$$[A^TA \mid A^T\mathbf{b}] = \left[\begin{array}{cc|c} 3 & 6 & 8 \\ 6 & 14 & 18 \end{array}\right] \longrightarrow \left[\begin{array}{cc|c} 1 & 0 & \frac{2}{3} \\ 0 & 1 & 1 \end{array}\right]$$

So $\bar{\mathbf{x}} = \begin{bmatrix} \frac{2}{3} \\ 1 \end{bmatrix}$, from which we see that $a = \frac{2}{3}$, $b = 1$ are the coefficients of the least squares approximating line: $y = \frac{2}{3} + x$.



The line we just found is the line in Example 7.24(c), so we have justified our claim that this line produces the smallest least squares error for the data points (1, 2), (2, 2), and (3, 4). Notice that if $\bar{\mathbf{x}}$ is a least squares solution of $A\mathbf{x} = \mathbf{b}$, we may compute the least squares error as

$$\|\mathbf{e}\| = \|\mathbf{b} - A\bar{\mathbf{x}}\|$$

Since $A\bar{\mathbf{x}} = \text{proj}_{\text{col}(A)}(\mathbf{b})$, this is just the length of $\text{perp}_{\text{col}(A)}(\mathbf{b})$—that is, the distance from $\mathbf{b}$ to the column space of A. In Example 7.26, we had

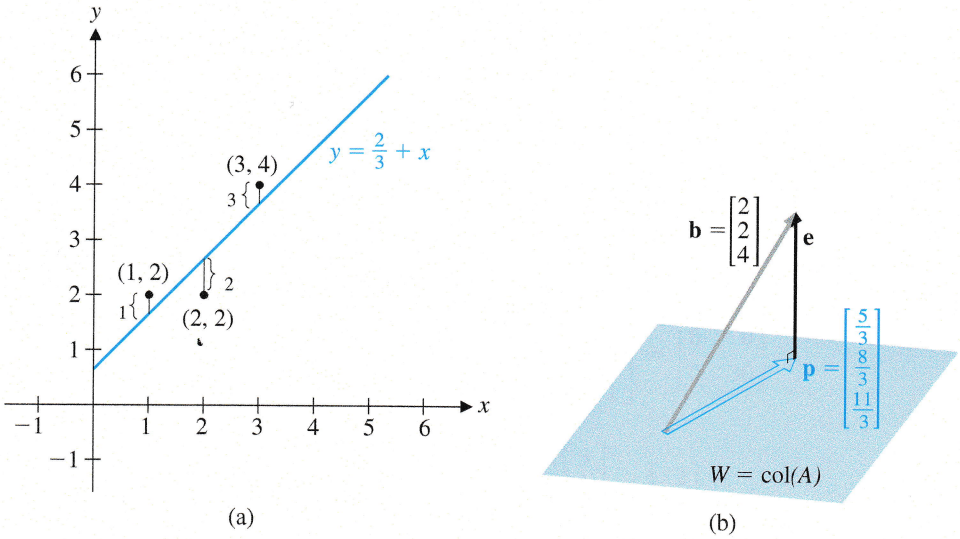
$$\mathbf{e} = \mathbf{b} - A\bar{\mathbf{x}} = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} \frac{2}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{bmatrix}$$

so, as in Example 7.24(c), we have a least squares error of $\|\mathbf{e}\| = \sqrt{\frac{2}{3}} \approx 0.816$.

Remark Note that the columns of A in Example 7.26 are linearly independent, so $(A^TA)^{-1}$ exists, and we could calculate $\bar{\mathbf{x}}$ as $\bar{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$. However, it is almost always easier to solve the normal equations using Gaussian elimination (or to let your CAS do it for you!).

It is interesting to look at Example 7.26 from two different geometric points of view. On the one hand, we have the least squares approximating line $y = \frac{2}{3} + x$, with corresponding errors $\varepsilon_1 = \frac{1}{3}$, $\varepsilon_2 = -\frac{2}{3}$, and $\varepsilon_3 = \frac{1}{3}$, as shown in Figure 7.13(a). Equivalently, we have the projection of $\mathbf{b}$ onto the column space of A, as shown in Figure 7.13(b). Here,

$$\mathbf{p} = \text{proj}_{\text{col}(A)}(\mathbf{b}) = A\bar{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} \frac{2}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ \frac{8}{3} \\ \frac{11}{3} \end{bmatrix}$$



(a) (b)

Figure 7.13

and the least squares error vector is $\mathbf{e} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$. [What would Figure 7.13(b) look like if the data points *were* collinear?]

Example 7.27

Find the least squares approximating line and the least squares error for the points (1, 1), (2, 2), (3, 2), and (4, 3).

Solution Let $y = a + bx$ be the equation of the line we seek. Then, substituting the four points into this equation, we obtain

$$\begin{aligned} a + b &= 1 \\ a + 2b &= 2 \\ a + 3b &= 2 \\ a + 4b &= 3 \end{aligned} \quad \text{or} \quad \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$

So we want the least squares solution of $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \end{bmatrix}$$

Since the columns of A are linearly independent, the solution we want is

$$\bar{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b} = \begin{bmatrix} \frac{1}{2} \\ \frac{3}{5} \end{bmatrix}$$

(Check this calculation.) Therefore, we take $a = \frac{1}{2}$ and $b = \frac{3}{5}$, producing the least squares approximating line $y = \frac{1}{2} + \frac{3}{5}x$, as shown in Figure 7.14.

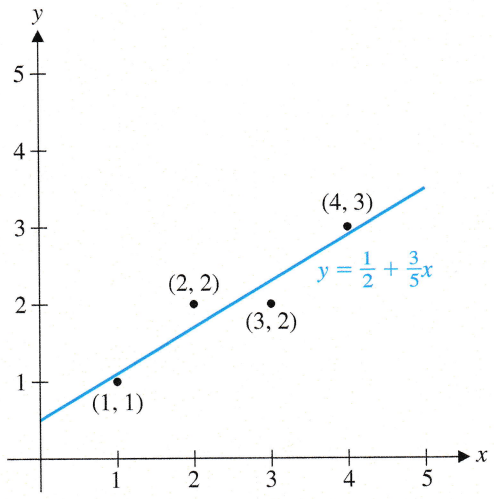


Figure 7.14

Since

$$\mathbf{e} = \mathbf{b} - A\bar{\mathbf{x}} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{3}{5} \end{bmatrix} = \begin{bmatrix} -\frac{1}{10} \\ \frac{3}{10} \\ -\frac{3}{10} \\ \frac{1}{10} \end{bmatrix}$$

the least squares error is $\|\mathbf{e}\| = \sqrt{5}/5 \approx 0.447$.

We can use the method of least squares to approximate data points by curves other than straight lines.

Example 7.28

Find the parabola that gives the best least squares approximation to the points $(-1, 1)$, $(0, -1)$, $(1, 0)$, and $(2, 2)$.

Solution The equation of a parabola is a quadratic $y = a + bx + cx^2$. Substituting the given points into this quadratic, we obtain the linear system

$$\begin{array}{rcl} a - b + c &=& 1 \\ a &=& -1 \\ a + b + c &=& 0 \\ a + 2b + 4c &=& 2 \end{array} \quad \text{or} \quad \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 2 \end{bmatrix}$$

Thus, we want the least squares approximation of $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 2 \end{bmatrix}$$

We compute

$$A^TA = \begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{bmatrix} \quad \text{and} \quad A^T\mathbf{b} = \begin{bmatrix} 2 \\ 3 \\ 9 \end{bmatrix}$$

so the normal equations are given by

$$\begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{bmatrix} \bar{\mathbf{x}} = \begin{bmatrix} 2 \\ 3 \\ 9 \end{bmatrix}$$

whose solution is

$$\bar{\mathbf{x}} = \begin{bmatrix} -\frac{7}{10} \\ -\frac{3}{5} \\ 1 \end{bmatrix}$$

Thus, the least squares approximating parabola has the equation

$$y = -\tfrac{7}{10} - \tfrac{3}{5}x + x^2$$

as shown in Figure 7.15.

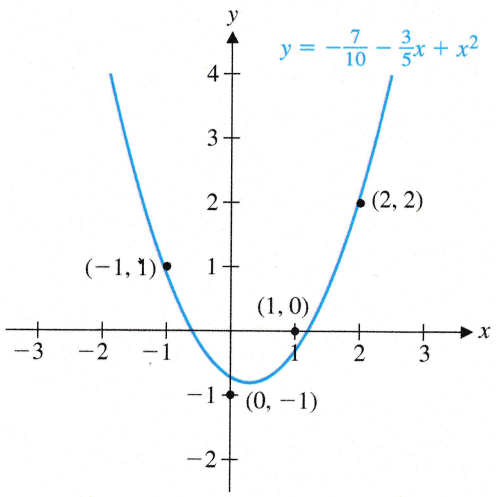


Figure 7.15
A least squares approximating parabola

One of the important uses of least squares approximation is to estimate constants associated with various processes. The next example illustrates this application in the context of population growth. Recall from Section 6.7 that a population that is growing (or decaying) exponentially satisfies an equation of the form $p(t) = ce^{kt}$, where $p(t)$ is the size of the population at time t and c and k are constants. Clearly, $c = p(0)$, but k is not so easy to determine. It is easy to see that

$$k = \frac{p'(t)}{p(t)}$$

which explains why k is sometimes referred to as the *relative growth rate* of the population: It is the ratio of the growth rate $p'(t)$ to the size of the population $p(t)$.

CAS **Example 7.29**

Table 7.2 gives the population of the world at 10-year intervals for the second half of the 20th century. Assuming an exponential growth model, find the relative growth rate and predict the world's population in 2010.

Table 7.2

Year	Population (in billions)
1950	2.56
1960	3.04
1970	3.71
1980	4.46
1990	5.28
2000	6.08

Source: U.S. Bureau of the Census, International Data Base

Solution Let's agree to measure time t in 10-year intervals so that $t = 0$ is 1950, $t = 1$ is 1960, and so on. Since $c = p(0) = 2.56$, the equation for the growth rate of the population is

$$p = 2.56e^{kt}$$

How can we use the method of least squares on this equation? If we take the natural logarithm of both sides, we convert the equation into a linear one:

$$\begin{aligned} \ln p &= \ln(2.56e^{kt}) \\ &= \ln 2.56 + \ln(e^{kt}) \\ &\approx 0.94 + kt \end{aligned}$$

Plugging in the values of t and p from Table 7.2 yields the following system (where we have rounded calculations to three decimal places):

$$\begin{aligned} 0.94 &= 0.94 \\ k &= 0.172 \\ 2k &= 0.371 \\ 3k &= 0.555 \\ 4k &= 0.724 \\ 5k &= 0.865 \end{aligned}$$

We can ignore the first equation (it just corresponds to the initial condition $c = p(0) = 2.56$). The remaining equations correspond to a system $A\mathbf{x} = \mathbf{b}$, with

$$A = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 0.172 \\ 0.371 \\ 0.555 \\ 0.724 \\ 0.865 \end{bmatrix}$$

Since $A^TA = 55$ and $A^T\mathbf{b} = 9.80$, the corresponding normal equations are just the single equation

$$55\bar{\mathbf{x}} = 9.80$$

Therefore, $k = \bar{\mathbf{x}} = 9.80/55 \approx 0.178$. Consequently, the least squares solution has the form $p = 2.56e^{0.178t}$ (see Figure 7.16).

The world's population in 2010 corresponds to $t = 6$, from which we obtain

$$p(6) = 2.56e^{0.178(6)} \approx 7.448$$

Thus, if our model is accurate, there will be approximately 7.45 billion people on Earth in the year 2010. (The U.S. Census Bureau estimates that the global population will be "only" 6.82 billion in 2010. Why do you think our estimate is higher?)

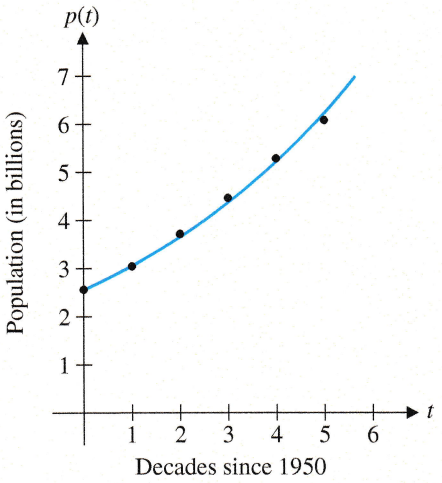


Figure 7.16

Least Squares via the *QR* Factorization

It is often the case that the normal equations for a least squares problem are ill-conditioned. Therefore, a small numerical error in performing Gaussian elimination will result in a large error in the least squares solution. Consequently, in practice, other methods are usually used to compute least squares approximations.

It turns out that the *QR* factorization of A yields a more reliable way of computing the least squares approximation of $A\mathbf{x} = \mathbf{b}$.

Theorem 7.10

Let A be an $m \times n$ matrix with linearly independent columns and let $\mathbf{b}$ be in $\mathbb{R}^m$. If $A = QR$ is a *QR* factorization of A, then the unique least squares solution $\bar{\mathbf{x}}$ of $A\mathbf{x} = \mathbf{b}$ is

$$\bar{\mathbf{x}} = R^{-1}Q^T\mathbf{b}$$

Proof Recall from Theorem 5.16 that the *QR* factorization $A = QR$ involves an $m \times n$ matrix Q with orthonormal columns and an invertible upper triangular matrix R. From the Least Squares Theorem, we have

$$\begin{aligned} A^TA\bar{\mathbf{x}} &= A^T\mathbf{b} \\ \Rightarrow (QR)^TQR\bar{\mathbf{x}} &= (QR)^T\mathbf{b} \\ \Rightarrow R^TQ^TQR\bar{\mathbf{x}} &= R^TQ^T\mathbf{b} \\ \Rightarrow R^TR\bar{\mathbf{x}} &= R^TQ^T\mathbf{b} \end{aligned}$$

since $Q^TQ = I$. (Why?)

Since R is invertible, so is R^T, and hence we have

$$R\bar{\mathbf{x}} = Q^T\mathbf{b} \quad \text{or, equivalently,} \quad \bar{\mathbf{x}} = R^{-1}Q^T\mathbf{b}$$

Remark Since R is upper triangular, in practice it is easier to solve $R\bar{\mathbf{x}} = Q^T\mathbf{b}$ directly than to invert R and compute $R^{-1}Q^T\mathbf{b}$.

Example 7.30 Use the QR factorization to find a least squares solution of $A\mathbf{x} = \mathbf{b}$, where

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

Solution From Example 5.15,

$$A = QR = \begin{bmatrix} 1/2 & 3\sqrt{5}/10 & -\sqrt{6}/6 \\ -1/2 & 3\sqrt{5}/10 & 0 \\ -1/2 & \sqrt{5}/10 & \sqrt{6}/6 \\ 1/2 & \sqrt{5}/10 & \sqrt{6}/3 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1/2 \\ 0 & \sqrt{5} & 3\sqrt{5}/2 \\ 0 & 0 & \sqrt{6}/2 \end{bmatrix}$$

We have

$$Q^T\mathbf{b} = \begin{bmatrix} 1/2 & -1/2 & -1/2 & 1/2 \\ 3\sqrt{5}/10 & 3\sqrt{5}/10 & \sqrt{5}/10 & \sqrt{5}/10 \\ -\sqrt{6}/6 & 0 & \sqrt{6}/6 & \sqrt{6}/3 \end{bmatrix} \begin{bmatrix} 2 \\ -3 \\ -2 \\ 0 \end{bmatrix} = \begin{bmatrix} 7/2 \\ -\sqrt{5}/2 \\ -2\sqrt{6}/3 \end{bmatrix}$$

so we require the solution to $R\bar{\mathbf{x}} = Q^T\mathbf{b}$, or

$$\begin{bmatrix} 2 & 1 & 1/2 \\ 0 & \sqrt{5} & 3\sqrt{5}/2 \\ 0 & 0 & \sqrt{6}/2 \end{bmatrix} \bar{\mathbf{x}} = \begin{bmatrix} 7/2 \\ -\sqrt{5}/2 \\ -2\sqrt{6}/3 \end{bmatrix}$$

Back substitution quickly yields

$$\bar{\mathbf{x}} = \begin{bmatrix} 4/3 \\ 3/2 \\ -4/3 \end{bmatrix}$$

Orthogonal Projection Revisited

One of the nice byproducts of the least squares method is a new formula for the orthogonal projection of a vector onto a subspace of $\mathbb{R}^m$.

Theorem 7.11 Let W be a subspace of $\mathbb{R}^m$ and let A be an $m \times n$ matrix whose columns form a basis for W. If $\mathbf{v}$ is any vector in $\mathbb{R}^m$, then the orthogonal projection of $\mathbf{v}$ onto W is the vector

$$\text{proj}_W(\mathbf{v}) = A(A^TA)^{-1}A^T\mathbf{v}$$

The linear transformation $P : \mathbb{R}^m \to \mathbb{R}^m$ that projects $\mathbb{R}^m$ onto W has $A(A^TA)^{-1}A^T$ as its standard matrix.

Proof Given the way we have constructed A, its column space is W. Since the columns of A are linearly independent, the Least Squares Theorem guarantees that there is a unique least squares solution to $A\mathbf{x} = \mathbf{v}$ given by

$$\bar{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{v}$$

By Equation (1),

$$A\bar{\mathbf{x}} = \text{proj}_{\text{col}(A)}(\mathbf{v}) = \text{proj}_W(\mathbf{v})$$

Therefore, $\text{proj}_W(\mathbf{v}) = A((A^TA)^{-1}A^T\mathbf{v}) = (A(A^TA)^{-1}A^T)\mathbf{v}$

as required. Since this equation holds for all $\mathbf{v}$ in $\mathbb{R}^m$, the last statement of the theorem follows immediately.

We will illustrate Theorem 7.11 by revisiting Example 5.11.

Example 7.31

Find the orthogonal projection of $\mathbf{v} = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix}$ onto the plane W in $\mathbb{R}^3$ with equation $x - y + 2z = 0$, and give the standard matrix of the orthogonal projection transformation onto W.

Solution As in Example 5.11, we will take as a basis for W the set

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

We form the matrix

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

with these basis vectors as its columns. Then

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

so $$(A^TA)^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

By Theorem 7.11, the standard matrix of the orthogonal projection transformation onto W is

$$A(A^TA)^{-1}A^T = A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & \frac{1}{6} & -\frac{1}{3} \\ \frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

so the orthogonal projection of $\mathbf{v}$ onto W is

$$\text{proj}_W(\mathbf{v}) = A(A^TA)^{-1}A^T\mathbf{v} = \begin{bmatrix} \frac{5}{6} & \frac{1}{6} & -\frac{1}{3} \\ \frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ \frac{1}{3} \\ -\frac{2}{3} \end{bmatrix}$$

which agrees with our solution to Example 5.11.

Remark Since the projection of a vector onto a subspace W is unique, the standard matrix of this linear transformation (as given by Theorem 7.11) cannot depend on the choice of basis for W. In other words, with a different basis for W, we have a different matrix A, but the matrix $A(A^TA)^{-1}A^T$ will be the same! (You are asked to verify this in Exercise 43.)

The Pseudoinverse of a Matrix

If A is an $n \times n$ matrix with linearly independent columns, then it is invertible, and the unique solution to $A\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = A^{-1}\mathbf{b}$. If $m > n$ and A is $m \times n$ with linearly independent columns, then $A\mathbf{x} = \mathbf{b}$ has no exact solution, but the best approximation is given by the unique least squares solution $\bar{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$. The matrix $(A^TA)^{-1}A^T$ therefore plays the role of an "inverse of A" in this situation.

Definition If A is a matrix with linearly independent columns, then the ***pseudoinverse*** of A is the matrix A^+ defined by

$$A^+ = (A^TA)^{-1}A^T$$

Observe that if A is $m \times n$, then A^+ is $n \times m$.

Example 7.32

Find the pseudoinverse of $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$.

Solution We have already done most of the calculations in Example 7.26. Using our previous work, we have

$$A^+ = (A^TA)^{-1}A^T = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} \frac{4}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

The pseudoinverse is a convenient shorthand notation for some of the concepts we have been exploring. For example, if A is $m \times n$ with linearly independent columns, the least squares solution of $A\mathbf{x} = \mathbf{b}$ is given by

$$\bar{\mathbf{x}} = A^+\mathbf{b}$$

and the standard matrix of the orthogonal projection P from $\mathbb{R}^m$ onto $\text{col}(A)$ is

$$[P] = AA^+$$

If A is actually a square matrix, then it is easy to show that $A^+ = A^{-1}$ (see Exercise 53). In this case, the least squares solution of $A\mathbf{x} = \mathbf{b}$ is the *exact* solution, since

$$\bar{\mathbf{x}} = A^+\mathbf{b} = A^{-1}\mathbf{b} = \mathbf{x}$$

The projection matrix becomes $[P] = AA^+ = AA^{-1} = I$. (What is the geometric interpretation of this equality?)

Theorem 7.12 summarizes the key properties of the pseudoinverse of a matrix. (Before reading the proof of this theorem, verify these properties for the matrix in Example 7.32.)

Theorem 7.12 Let A be a matrix with linearly independent columns. Then the pseudoinverse A^+ of A satisfies the following properties, called the ***Penrose conditions*** for A:

a. $AA^+A = A$
b. $A^+AA^+ = A^+$
c. AA^+ and A^+A are symmetric.

Proof We prove condition (a) and half of condition (c) and leave the proofs of the remaining conditions as Exercises 54 and 55.

(a) We compute

$$\begin{aligned} AA^+A &= A((A^TA)^{-1}A^T)A \\ &= A(A^TA)^{-1}(A^TA) \\ &= AI = A \end{aligned}$$

(c) By Theorem 3.4, A^TA is symmetric. Therefore, $(A^TA)^{-1}$ is also symmetric, by Exercise 46 in Section 3.3. Taking the transpose of AA^+, we have

$$\begin{aligned} (AA^+)^T &= (A(A^TA)^{-1}A^T)^T \\ &= (A^T)^T((A^TA)^{-1})^TA^T \\ &= A(A^TA)^{-1}A^T \\ &= AA^+ \end{aligned}$$

Exercise 56 explores further properties of the pseudoinverse. In the next section, we will see how to extend the definition of A^+ to handle *all* matrices, whether or not the columns of A are linearly independent.

Exercises 7.3

CAS

In Exercises 1–3, consider the data points (1, 0), (2, 1), *and* (3, 5). *Compute the least squares error for the given line. In each case, plot the points and the line.*

1. $y = -2 + 2x$ **2.** $y = x$ **3.** $y = -3 + \frac{5}{2}x$

In Exercises 4–6, consider the data points (−5, 3), (0, 3), (5, 2), *and* (10, 0). *Compute the least squares error for the given line. In each case, plot the points and the line.*

4. $y = 3 - \frac{1}{3}x$ **5.** $y = \frac{5}{2}$ **6.** $y = 2 - \frac{1}{5}x$

In Exercises 7–14, find the least squares approximating line for the given points and compute the corresponding least squares error.

7. (1, 0), (2, 1), (3, 5)
8. (1, 6), (2, 3), (3, 1)
9. (0, 4), (1, 1), (2, 0)
10. (0, 3), (1, 3), (2, 5)
11. (−5, −1), (0, 1), (5, 2), (10, 4)

12. $(-5, 3), (0, 3), (5, 2), (10, 0)$

13. $(1, 1), (2, 3), (3, 4), (4, 5), (5, 7)$

14. $(1, 10), (2, 8), (3, 5), (4, 3), (5, 0)$

In Exercises 15–18, find the least squares approximating parabola for the given points.

15. $(1, 1), (2, -2), (3, 3), (4, 4)$

16. $(1, 6), (2, 0), (3, 0), (4, 2)$

17. $(-2, 4), (-1, 7), (0, 3), (1, 0), (2, -1)$

18. $(-2, 0), (-1, -11), (0, -10), (1, -9), (2, 8)$

In Exercises 19–22, find a least squares solution of $A\mathbf{x} = \mathbf{b}$ by constructing and solving the normal equations.

19. $A = \begin{bmatrix} 3 & 1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

20. $A = \begin{bmatrix} 1 & -2 \\ 3 & -2 \\ 2 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

21. $A = \begin{bmatrix} 1 & -2 \\ 0 & -3 \\ 2 & 5 \\ 3 & 0 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 4 \\ 1 \\ -2 \\ 4 \end{bmatrix}$

22. $A = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 5 \\ -1 \\ 2 \end{bmatrix}$

In Exercises 23 and 24, show that the least squares solution of $A\mathbf{x} = \mathbf{b}$ is not unique and solve the normal equations to find all the least squares solutions.

23. $A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & -1 & 1 & 1 \\ 1 & -1 & 1 & 0 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ -3 \\ 2 \\ 4 \end{bmatrix}$

24. $A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 5 \\ 3 \\ -1 \\ 1 \end{bmatrix}$

In Exercises 25 and 26, find the best approximation to a solution of the given system of equations.

25.
$$\begin{aligned} x + y - z &= 2 \\ -y + 2z &= 6 \\ 3x + 2y - z &= 11 \\ -x + z &= 0 \end{aligned}$$

26.
$$\begin{aligned} 2x + 3y + z &= 21 \\ x + y + z &= 7 \\ -x + y - z &= 14 \\ 2y + z &= 0 \end{aligned}$$

In Exercises 27 and 28, a QR factorization of A is given. Use it to find a least squares solution of $A\mathbf{x} = \mathbf{b}$.

27. $A = \begin{bmatrix} 2 & 1 \\ 2 & 0 \\ 1 & 1 \end{bmatrix}, Q = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}, R = \begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$

28. $A = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ -1 & 1 \end{bmatrix}, Q = \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{2} \\ 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \end{bmatrix},$

$R = \begin{bmatrix} \sqrt{6} & -\sqrt{6}/2 \\ 0 & 1/\sqrt{2} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

29. A tennis ball is dropped from various heights, and the height of the ball on the first bounce is measured. Use the data in Table 7.3 to find the least squares approximating line for bounce height b as a linear function of initial height h.

Table 7.3

h (cm)	20	40	48	60	80	100
b (cm)	14.5	31	36	45.5	59	73.5

30. Hooke's Law states that the length L of a spring is a linear function of the force F applied to it. (See Figure 7.17 and Example 6.92.) Accordingly, there are constants a and b such that

$$L = a + bF$$

Table 7.4 shows the results of attaching various weights to a spring.

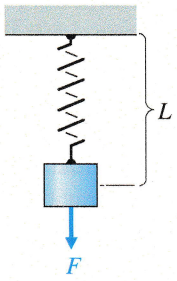


Figure 7.17

Table 7.4

F (oz)	2	4	6	8
L (in.)	7.4	9.6	11.5	13.6

Table 7.5

Year of Birth	1920	1930	1940	1950	1960	1970	1980	1990
Life Expectancy (years)	54.1	59.7	62.9	68.2	69.7	70.8	73.7	75.4

Source: World Almanac and Book of Facts. New York: World Almanac Books, 1999

(a) Determine the constants a and b by finding the least squares approximating line for these data. What does a represent?
(b) Estimate the length of the spring when a weight of 5 ounces is attached.

31. Table 7.5 gives life expectancies for people born in the United States in the given years.
(a) Determine the least squares approximating line for these data and use it to predict the life expectancy of someone born in 2000.
(b) How good is this model? Explain.

32. When an object is thrown straight up into the air, Newton's Second Law of Motion states that its height $s(t)$ at time t is given by

$$s(t) = s_0 + v_0 t + \tfrac{1}{2}gt^2$$

where v_0 is its initial velocity and g is the constant of acceleration due to gravity. Suppose we take the measurements shown in Table 7.6.

Table 7.6

Time (s)	0.5	1	1.5	2	3
Height (m)	11	17	21	23	18

(a) Find the least squares approximating quadratic for these data.
(b) Estimate the height at which the object was released (in m), its initial velocity (in m/s), and its acceleration due to gravity (in m/s^2).
(c) Approximately when will the object hit the ground?

33. Table 7.7 gives the population of the United States at 10-year intervals for the years 1950–2000.
(a) Assuming an exponential growth model of the form $p(t) = ce^{kt}$, where $p(t)$ is the population at time t, use least squares to find the equation for the growth rate of the population. [*Hint:* Let $t = 0$ be 1950.]
(b) Use the equation to estimate the U.S. population in 2010.

Table 7.7

Year	Population (in millions)
1950	150
1960	179
1970	203
1980	227
1990	250
2000	281

Source: U.S. Bureau of the Census

34. Table 7.8 shows average major league baseball salaries for the years 1970–2005.
(a) Find the least squares approximating quadratic for these data.
(b) Find the least squares approximating exponential for these data.
(c) Which equation gives the better approximation? Why?
(d) What do you estimate the average major league baseball salary will be in 2010 and 2015?

Table 7.8

Year	Average Salary (thousands of dollars)
1970	29.3
1975	44.7
1980	143.8
1985	371.6
1990	597.5
1995	1110.8
2000	1895.6
2005	2476.6

Source: Major League Baseball Players Association

35. A 200 mg sample of radioactive polonium-210 is observed as it decays. Table 7.9 shows the mass remaining at various times.

Assuming an exponential decay model, use least squares to find the half-life of polonium-210. (See Section 6.7.)

Table 7.9

Time (days)	0	30	60	90
Mass (mg)	200	172	148	128

36. Find the plane $z = a + bx + cy$ that best fits the data points $(0, -4, 0)$, $(5, 0, 0)$, $(4, -1, 1)$, $(1, -3, 1)$, and $(-1, -5, -2)$.

In Exercises 37–42, find the standard matrix of the orthogonal projection onto the subspace W. Then use this matrix to find the orthogonal projection of $\mathbf{v}$ onto W.

37. $W = \text{span}\left(\begin{bmatrix}1\\1\end{bmatrix}\right), \mathbf{v} = \begin{bmatrix}3\\4\end{bmatrix}$

38. $W = \text{span}\left(\begin{bmatrix}1\\-2\end{bmatrix}\right), \mathbf{v} = \begin{bmatrix}1\\1\end{bmatrix}$

39. $W = \text{span}\left(\begin{bmatrix}1\\1\\1\end{bmatrix}\right), \mathbf{v} = \begin{bmatrix}1\\2\\3\end{bmatrix}$

40. $W = \text{span}\left(\begin{bmatrix}2\\2\\-1\end{bmatrix}\right), \mathbf{v} = \begin{bmatrix}1\\0\\0\end{bmatrix}$

41. $W = \text{span}\left(\begin{bmatrix}1\\0\\-1\end{bmatrix}, \begin{bmatrix}1\\1\\1\end{bmatrix}\right), \mathbf{v} = \begin{bmatrix}1\\0\\0\end{bmatrix}$

42. $W = \text{span}\left(\begin{bmatrix}1\\-2\\1\end{bmatrix}, \begin{bmatrix}1\\0\\-1\end{bmatrix}\right), \mathbf{v} = \begin{bmatrix}1\\2\\3\end{bmatrix}$

43. Verify that the standard matrix of the projection onto W in Example 7.31 (as constructed by Theorem 7.11) does not depend on the choice of basis. Take

$$\left\{\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\3\\1\end{bmatrix}\right\}$$

as a basis for W and repeat the calculations to show that the resulting projection matrix is the same.

44. Let A be a matrix with linearly independent columns and let $P = A(A^TA)^{-1}A^T$ be the matrix of orthogonal projection onto col(A).

(a) Show that P is symmetric.

(b) Show that P is idempotent.

In Exercises 45–52, compute the pseudoinverse of A.

45. $A = \begin{bmatrix}1\\2\end{bmatrix}$

46. $A = \begin{bmatrix}1\\-1\\2\end{bmatrix}$

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

48. $A = \begin{bmatrix}1 & 3\\3 & 1\\2 & 2\end{bmatrix}$

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

50. $A = \begin{bmatrix}1 & 2\\3 & 4\end{bmatrix}$

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

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

53. **(a)** Show that if A is a square matrix with linearly independent columns, then $A^+ = A^{-1}$.

(b) If A is an $m \times n$ matrix with orthonormal columns, what is A^+?

54. Prove Theorem 7.12(b).

55. Prove the remaining part of Theorem 7.12(c).

56. Let A be a matrix with linearly independent columns. Prove the following:

(a) $(cA)^+ = (1/c)A^+$ for all scalars $c \neq 0$.

(b) $(A^+)^+ = A$ if A is a square matrix.

(c) $(A^T)^+ = (A^+)^T$ if A is a square matrix.

57. Let n data points $(x_1, y_1), \ldots, (x_n, y_n)$ be given. Show that if the points do not all lie on the same vertical line, then they have a unique least squares approximating line.

58. Let n data points $(x_1, y_1), \ldots, (x_n, y_n)$ be given. Generalize Exercise 57 to show that if at least $k + 1$ of $x_1, \ldots, x_n$ are distinct, then the given points have a unique least squares approximating polynomial of degree at most k.



7.4 The Singular Value Decomposition

In Chapter 5, we saw that every symmetric matrix A can be factored as $A = PDP^T$, where P is an orthogonal matrix and D is a diagonal matrix displaying the eigenvalues of A. If A is not symmetric, such a factorization is not possible, but as we learned in Chapter 4, we may still be able to factor a square matrix A as $A = PDP^{-1}$, where D is as before but P is now simply an invertible matrix. However, not every matrix is diagonalizable, so it may surprise you that we will now show that *every* matrix (symmetric or not, square or not) has a factorization of the form $A = PDQ^T$, where P and Q are orthogonal and D is a diagonal matrix! This remarkable result is the *singular value decomposition* (SVD), and it is one of the most important of all matrix factorizations.

In this section, we will show how to compute the SVD of a matrix and then consider some of its many applications. Along the way, we will tie up some loose ends by answering a few questions that were left open in previous sections.

The Singular Values of a Matrix

For any $m \times n$ matrix A, the $n \times n$ matrix A^TA is symmetric and hence can be orthogonally diagonalized, by the Spectral Theorem. Not only are the eigenvalues of A^TA all real (Theorem 5.18), they are all *nonnegative*. To show this, let λ be an eigenvalue of A^TA with corresponding unit eigenvector $\mathbf{v}$. Then

$$\begin{aligned} 0 \leq \|A\mathbf{v}\|^2 &= (A\mathbf{v}) \cdot (A\mathbf{v}) = (A\mathbf{v})^T A\mathbf{v} = \mathbf{v}^T A^T A\mathbf{v} \\ &= \mathbf{v}^T \lambda \mathbf{v} = \lambda(\mathbf{v} \cdot \mathbf{v}) = \lambda \|\mathbf{v}\|^2 = \lambda \end{aligned}$$

It therefore makes sense to take (positive) square roots of these eigenvalues.

Definition If A is an $m \times n$ matrix, the ***singular values*** of A are the square roots of the eigenvalues of A^TA and are denoted by $\sigma_1, \ldots, \sigma_n$. It is conventional to arrange the singular values so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

Example 7.33 Find the singular values of

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

Solution The matrix

$$A^TA = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

has eigenvalues $\lambda_1 = 3$ and $\lambda_2 = 1$. Consequently, the singular values of A are $\sigma_1 = \sqrt{\lambda_1} = \sqrt{3}$ and $\sigma_2 = \sqrt{\lambda_2} = 1$.

To understand the significance of the singular values of an $m \times n$ matrix A, consider the eigenvectors of A^TA. Since A^TA is symmetric, we know that there is an *orthonormal* basis for $\mathbb{R}^n$ that consists of eigenvectors of A^TA. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be such a basis corresponding to the eigenvalues of A^TA, ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. From our calculations just before the definition,

$$\lambda_i = \|A\mathbf{v}_i\|^2$$

Therefore,

$$\sigma_i = \sqrt{\lambda_i} = \|A\mathbf{v}_i\|$$

In other words, the singular values of A are the lengths of the vectors $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$.

Geometrically, this result has a nice interpretation. Consider Example 7.33 again. If $\mathbf{x}$ lies on the unit circle in $\mathbb{R}^2$ (i.e., $\|\mathbf{x}\| = 1$), then

$$\|A\mathbf{x}\|^2 = (A\mathbf{x}) \cdot (A\mathbf{x}) = (A\mathbf{x})^T(A\mathbf{x}) = \mathbf{x}^TA^TA\mathbf{x}$$

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 + 2x_1x_2 + 2x_2^2$$

which we recognize is a quadratic form. By Theorem 5.23, the maximum and minimum values of this quadratic form, subject to the constraint $\|\mathbf{x}\| = 1$, are $\lambda_1 = 3$ and $\lambda_2 = 1$, respectively, and they occur at the corresponding eigenvectors of A^TA—that is, when $\mathbf{x} = \mathbf{v}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ and $\mathbf{x} = \mathbf{v}_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$, respectively. Since

$$\|A\mathbf{v}_i\|^2 = \mathbf{v}_i^TA^TA\mathbf{v}_i = \lambda_i$$

for $i = 1, 2$, we see that $\sigma_1 = \|A\mathbf{v}_1\| = \sqrt{3}$ and $\sigma_2 = \|A\mathbf{v}_2\| = 1$ are the maximum and minimum values of the lengths $\|A\mathbf{x}\|$ as $\mathbf{x}$ traverses the unit circle in $\mathbb{R}^2$.

Now, the linear transformation corresponding to A maps $\mathbb{R}^2$ onto the plane in $\mathbb{R}^3$ with equation $x - y - z = 0$ (verify this), and the image of the unit circle under this transformation is an ellipse that lies in this plane. (We will verify this fact in general shortly; see Figure 7.18.) So σ_1 and σ_2 are the lengths of half of the major and minor axes of this ellipse, as shown in Figure 7.19.

We can now describe the singular value decomposition of a matrix.

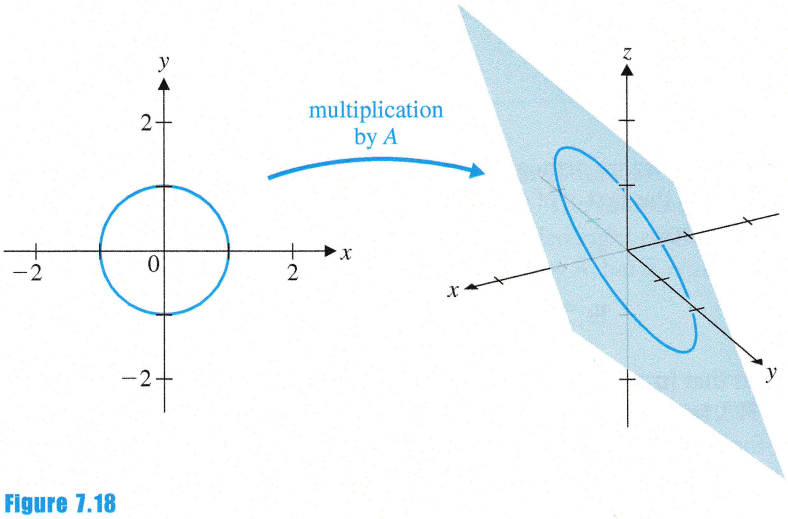


Figure 7.18
The matrix A transforms the unit circle in $\mathbb{R}^2$ into an ellipse in $\mathbb{R}^3$

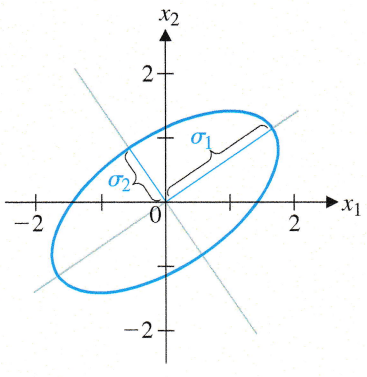


Figure 7.19

The Singular Value Decomposition

We want to show that an $m \times n$ matrix A can be factored as

$$A = U\Sigma V^T$$

where U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, and Σ is an $m \times n$ "diagonal" matrix. If the *nonzero* singular values of A are

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

and $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$, then Σ will have the block form

$$\Sigma = \begin{bmatrix} \overbrace{D}^{r} & \overbrace{O}^{n-r} \\ O & O \end{bmatrix} \begin{matrix} \}\, r \\ \}\, m-r \end{matrix}, \quad \text{where} \quad D = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \tag{1}$$

and each matrix O is a zero matrix of the appropriate size. (If $r = m$ or $r = n$, some of these will not appear.) Some examples of such a matrix Σ with $r = 2$ are

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

(What is D in each case?)

To construct the orthogonal matrix V, we first find an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbb{R}^n$ consisting of eigenvectors of the $n \times n$ symmetric matrix A^TA. Then

$$V = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_n]$$

is an orthogonal $n \times n$ matrix.

For the orthogonal matrix U, we first note that $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_n\}$ is an orthogonal set of vectors in $\mathbb{R}^m$. To see this, suppose that $\mathbf{v}_i$ is the eigenvector of A^TA corresponding to the eigenvalue λ_i. Then, for $i \neq j$, we have

$$\begin{aligned} (A\mathbf{v}_i) \cdot (A\mathbf{v}_j) &= (A\mathbf{v}_i)^T A\mathbf{v}_j \\ &= \mathbf{v}_i^T A^T A\mathbf{v}_j \\ &= \mathbf{v}_i^T \lambda_j \mathbf{v}_j \\ &= \lambda_j(\mathbf{v}_i \cdot \mathbf{v}_j) = 0 \end{aligned}$$

since the eigenvectors $\mathbf{v}_i$ are orthogonal. Now recall that the singular values satisfy $\sigma_i = \|A\mathbf{v}_i\|$ and that the first r of these are nonzero. Therefore, we can normalize $A\mathbf{v}_1, \ldots, A\mathbf{v}_r$ by setting

$$\mathbf{u}_i = \frac{1}{\sigma_i} A\mathbf{v}_i \quad \text{for } i = 1, \ldots, r$$

This guarantees that $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal set in $\mathbb{R}^m$, but if $r < m$ it will not be a basis for $\mathbb{R}^m$. In this case, we extend the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ to an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ for $\mathbb{R}^m$. (This is the only tricky part of the construction; we will describe techniques for carrying it out in the examples below and in the exercises.) Then we set

$$U = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_m]$$

All that remains to be shown is that this works; that is, we need to verify that with U, V, and Σ as described, we have $A = U\Sigma V^T$. Since $V^T = V^{-1}$, this is equivalent to showing that

$$AV = U\Sigma$$

We know that $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$ for $i = 1, \ldots, r$

and $\|A\mathbf{v}_i\| = \sigma_i = 0$ for $i = r + 1, \ldots, n$. Hence,

$$A\mathbf{v}_i = \mathbf{0} \quad \text{for } i = r + 1, \ldots, n$$

Therefore,

$$\begin{aligned} AV &= A[\mathbf{v}_1 \ \cdots \ \mathbf{v}_n] \\ &= [A\mathbf{v}_1 \ \cdots \ A\mathbf{v}_n] \\ &= [A\mathbf{v}_1 \ \cdots \ A\mathbf{v}_r \ \mathbf{0} \ \cdots \ \mathbf{0}] \\ &= [\sigma_1\mathbf{u}_1 \ \cdots \ \sigma_r\mathbf{u}_r \ \mathbf{0} \ \cdots \ \mathbf{0}] \\ &= [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m] \left[\begin{array}{ccc|c} \sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & O \\ 0 & \cdots & \sigma_r & \\ \hline & O & & O \end{array}\right] \\ &= U\Sigma \end{aligned}$$

as required.

We have just proved the following extremely important theorem.

Theorem 7.13 **The Singular Value Decomposition**

Let A be an $m \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$. Then there exist an $m \times m$ orthogonal matrix U, an $n \times n$ orthogonal matrix V, and an $m \times n$ matrix Σ of the form shown in Equation (1) such that

$$A = U\Sigma V^T$$

A factorization of A as in Theorem 7.13 is called a ***singular value decomposition (SVD)*** of A. The columns of U are called ***left singular vectors*** of A, and the columns of V are called ***right singular vectors*** of A. The matrices U and V are not uniquely determined by A, but Σ *must* contain the singular values of A, as in Equation (1). (See Exercise 25.)

Example 7.34

Find a singular value decomposition for the following matrices:

(a) $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$

Solution (a) We compute

$$A^TA = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and find that its eigenvalues are $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = 0$, with corresponding eigenvectors

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

(Verify this.) These vectors are orthogonal, so we normalize them to obtain

$$\mathbf{v}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}$$

The singular values of A are $\sigma_1 = \sqrt{2}$, $\sigma_2 = \sqrt{1} = 1$, and $\sigma_3 = \sqrt{0} = 0$. Thus,

$$V = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

To find U, we compute

$$\mathbf{u}_1 = \frac{1}{\sigma_1}A\mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\mathbf{u}_2 = \frac{1}{\sigma_2}A\mathbf{v}_2 = \frac{1}{1}\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

These vectors already form an orthonormal basis (the standard basis) for $\mathbb{R}^2$, so we have

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This yields the SVD

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix} = U\Sigma V^T$$

which can be easily checked. (Note that V had to be transposed. Also note that the singular value σ_3 does not appear in Σ.)

(b) This is the matrix in Example 7.33, so we already know that the singular values are $\sigma_1 = \sqrt{3}$ and $\sigma_2 = 1$, corresponding to $\mathbf{v}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$. So

$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

For U, we compute

$$\mathbf{u}_1 = \frac{1}{\sigma_1} A\mathbf{v}_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$$

and

$$\mathbf{u}_2 = \frac{1}{\sigma_2} A\mathbf{v}_2 = \frac{1}{1} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

This time, we need to extend $\{\mathbf{u}_1, \mathbf{u}_2\}$ to an orthonormal basis for $\mathbb{R}^3$. There are several ways to proceed; one method is to use the Gram-Schmidt Process, as in Example 5.14. We first need to find a linearly independent set of three vectors that contains $\mathbf{u}_1$ and $\mathbf{u}_2$. If $\mathbf{e}_3$ is the third standard basis vector in $\mathbb{R}^3$, it is clear that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{e}_3\}$ is linearly independent. (Here, you should be able to determine this by inspection, but a reliable method to use in general is to row reduce the matrix with these vectors as its columns and use the Fundamental Theorem.) Applying Gram-Schmidt (with normalization) to $\{\mathbf{u}_1, \mathbf{u}_2, \hat{\mathbf{e}}_3\}$ (only the last step is needed), we find

$$\mathbf{u}_3 = \begin{bmatrix} -1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}$$

so

$$U = \begin{bmatrix} 2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix}$$

and we have the SVD

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = U\Sigma V^T$$



There is another form of the singular value decomposition, analogous to the spectral decomposition of a symmetric matrix. It is obtained from the SVD by an outer product expansion and is very useful in applications. We can obtain this version of the SVD by imitating what we did to obtain the spectral decomposition.

Accordingly, we have

$$A = U\Sigma V^T = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m]\left[\begin{array}{ccc|c} \sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & O \\ 0 & \cdots & \sigma_r & \\ \hline & O & & O \end{array}\right]\begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix}$$

$$= [\mathbf{u}_1 \ \cdots \ \mathbf{u}_r \,|\, \mathbf{u}_{r+1} \ \cdots \ \mathbf{u}_m]\left[\begin{array}{ccc|c} \sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & O \\ 0 & \cdots & \sigma_r & \\ \hline & O & & O \end{array}\right]\left[\begin{array}{c} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \hline \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{array}\right]$$

$$= [\mathbf{u}_1 \ \cdots \ \mathbf{u}_r]\begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}\begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} + [\mathbf{u}_{r+1} \ \cdots \ \mathbf{u}_m][O]\begin{bmatrix} \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix}$$

$$= [\mathbf{u}_1 \ \cdots \ \mathbf{u}_r]\begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}\begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix}$$

$$= [\sigma_1\mathbf{u}_1 \ \cdots \ \sigma_r\mathbf{u}_r]\begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix}$$

$$= \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T$$

using block multiplication and the column-row representation of the product. The following theorem summarizes the process for obtaining this ***outer product form of the SVD.***

Theorem 7.14 **The Outer Product Form of the SVD**

Let A be an $m \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$. Let $\mathbf{u}_1, \ldots, \mathbf{u}_r$ be left singular vectors and let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be right singular vectors of A corresponding to these singular values. Then

$$A = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T$$

Remark If A is a positive definite, symmetric matrix, then Theorems 7.13 and 7.14 both reduce to results that we already know. In this case, it is not hard to show that the SVD generalizes the Spectral Theorem and that Theorem 7.14 generalizes the spectral decomposition. (See Exercise 27.)

The SVD of a matrix A contains much important information about A, as outlined in the crucial Theorem 7.15.

Theorem 7.15 Let $A = U\Sigma V^T$ be a singular value decomposition of an $m \times n$ matrix A. Let $\sigma_1, \ldots, \sigma_r$ be all the nonzero singular values of A. Then:

a. The rank of A is r.

b. $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal basis for col(A).

c. $\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$ is an orthonormal basis for null(A^T).

d. $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ is an orthonormal basis for row(A).

e. $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is an orthonormal basis for null(A).

Proof (a) By Exercise 61 in Section 3.5, we have

$$\begin{aligned} \text{rank}(A) &= \text{rank}(U\Sigma V^T) \\ &= \text{rank}(\Sigma V^T) \\ &= \text{rank}(\Sigma) = r \end{aligned}$$

(b) We already know that $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal set. Therefore, it is linearly independent, by Theorem 5.1. Since $\mathbf{u}_i = (1/\sigma_i)A\mathbf{v}_i$ for $i = 1, \ldots, r$, each $\mathbf{u}_i$ is in the column space of A. (Why?) Furthermore,

$$r = \text{rank}(A) = \dim(\text{col}(A))$$

Therefore, $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal basis for col(A), by Theorem 6.10(c).

(c) Since $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ is an orthonormal basis for $\mathbb{R}^m$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is a basis for col(A), by property (b), it follows that $\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$ is an orthonormal basis for the orthogonal complement of col(A). But $(\text{col}(A)) = \text{null}(A^T)$, by Theorem 5.10.

(e) Since

$$A\mathbf{v}_{r+1} = \cdots = A\mathbf{v}_n = \mathbf{0}$$

the set $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is an orthonormal set contained in the null space of A. Therefore, $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is a linearly independent set of $n - r$ vectors in null(A). But

$$\dim(\text{null}(A)) = n - r$$

by the Rank Theorem, so $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is an orthonormal basis for null(A), by Theorem 6.10(c).

(d) Property (d) follows from property (e) and Theorem 5.10. (You are asked to prove this in Exercise 32.)

The SVD provides new geometric insight into the effect of matrix transformations. We have noted several times (without proof) that an $m \times n$ matrix transforms the unit sphere in $\mathbb{R}^n$ into an ellipsoid in $\mathbb{R}^m$. This point arose, for example, in our discussions of Perron's Theorem and of operator norms, as well as in the introduction to singular values in this section. We now prove this result.

Theorem 7.16 Let A be an $m \times n$ matrix with rank r. Then the image of the unit sphere in $\mathbb{R}^n$ under the matrix transformation that maps $\mathbf{x}$ to $A\mathbf{x}$ is

a. the surface of an ellipsoid in $\mathbb{R}^m$ if $r = n$.
b. a solid ellipsoid in $\mathbb{R}^m$ if $r < n$.

Proof Let $A = U\Sigma V^T$ be a singular value decomposition of the $m \times n$ matrix A. Let the left and right singular vectors of A be $\mathbf{u}_1, \ldots, \mathbf{u}_m$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$, respectively. Since $\text{rank}(A) = r$, the singular values of A satisfy

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \quad \text{and} \quad \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$$

by Theorem 7.15(a). Let $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ be a unit vector in $\mathbb{R}^n$. Now, since V is an orthogonal matrix, so is V^T, and hence $V^T\mathbf{x}$ is a unit vector, by Theorem 5.6. Now

$$V^T\mathbf{x} = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{v}_1^T\mathbf{x} \\ \vdots \\ \mathbf{v}_n^T\mathbf{x} \end{bmatrix}$$

so $(\mathbf{v}_1^T\mathbf{x})^2 + \cdots + (\mathbf{v}_n^T\mathbf{x})^2 = 1$.

By the outer product form of the SVD, we have $A = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T$. Therefore,

$$\begin{aligned} A\mathbf{x} &= \sigma_1\mathbf{u}_1\mathbf{v}_1^T\mathbf{x} + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^T\mathbf{x} \\ &= (\sigma_1\mathbf{v}_1^T\mathbf{x})\mathbf{u}_1 + \cdots + (\sigma_r\mathbf{v}_r^T\mathbf{x})\mathbf{u}_r \\ &= y_1\mathbf{u}_1 + \cdots + y_r\mathbf{u}_r \end{aligned}$$

where we are letting y_i denote the scalar $\sigma_i\mathbf{v}_i^T\mathbf{x}$.

(a) If $r = n$, then we must have $n \leq m$ and

$$\begin{aligned} A\mathbf{x} &= y_1\mathbf{u}_1 + \cdots + y_n\mathbf{u}_n \\ &= U\mathbf{y} \end{aligned}$$

where $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$. Therefore, again by Theorem 5.6, $\|A\mathbf{x}\| = \|U\mathbf{y}\| = \|\mathbf{y}\|$, since U is orthogonal. But

$$\left(\frac{y_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{y_n}{\sigma_n}\right)^2 = (\mathbf{v}_1^T\mathbf{x})^2 + \cdots + (\mathbf{v}_n^T\mathbf{x})^2 = 1$$

which shows that the vectors $A\mathbf{x}$ form the surface of an ellipsoid in $\mathbb{R}^m$. (Why?)

(b) If $r < n$, the only difference in the above steps is that the equation becomes

$$\left(\frac{y_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{y_r}{\sigma_r}\right)^2 \leq 1$$

since we are missing some terms. This inequality corresponds to a solid ellipsoid in $\mathbb{R}^m$.

Example 7.35 Describe the image of the unit sphere in $\mathbb{R}^3$ under the action of the matrix

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

Solution In Example 7.34(a), we found the following SVD of A:

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix}$$

Since $r = \text{rank}(A) = 2 < 3 = n$, the second part of Theorem 7.16 applies. The image of the unit sphere will satisfy the inequality

$$\left(\frac{y_1}{\sqrt{2}}\right)^2 + \left(\frac{y_2}{1}\right)^2 \leq 1 \quad \text{or} \quad \frac{y_1^2}{2} + y_2^2 \leq 1$$

relative to y_1y_2 coordinate axes in $\mathbb{R}^2$ (corresponding to the left singular vectors $\mathbf{u}_1$ and $\mathbf{u}_2$). Since $\mathbf{u}_1 = \mathbf{e}_1$ and $\mathbf{u}_2 = \mathbf{e}_2$, the image is as shown in Figure 7.20.

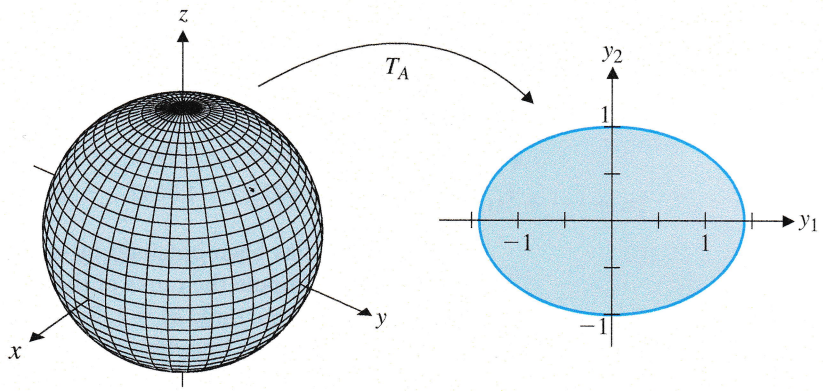


Figure 7.20

In general, we can describe the effect of an $m \times n$ matrix A on the unit sphere in $\mathbb{R}^n$ in terms of the effect of each factor in its SVD, $A = U\Sigma V^T$, from right to left. Since V^T is an orthogonal matrix, it maps the unit sphere to itself. The $m \times n$ matrix Σ does two things: The diagonal entries $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$ collapse $n - r$ of the dimensions of the unit sphere, leaving an r-dimensional unit sphere, which the nonzero diagonal entries $\sigma_1, \ldots, \sigma_r$ then distort into an ellipsoid. The orthogonal matrix U then aligns the axes of this ellipsoid with the orthonormal basis vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ in $\mathbb{R}^m$. (See Figure 7.21.)

Applications of the SVD

The singular value decomposition is an extremely useful tool, both practically and theoretically. We will look at just a few of its many applications.

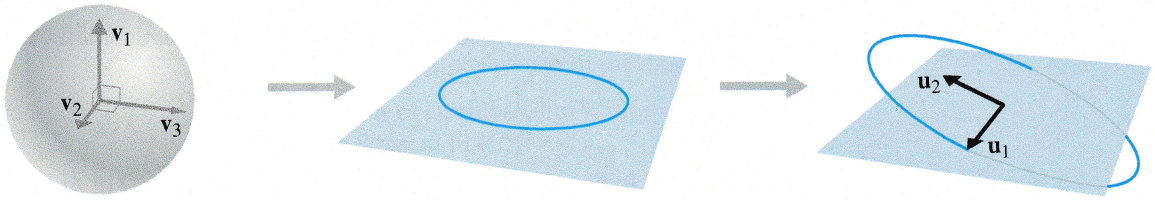


Figure 7.21

Rank Until now, we have not worried about calculating the rank of a matrix from a computational point of view. We compute the rank of a matrix by row reducing it to echelon form and counting the number of nonzero rows. However, as we have seen, roundoff errors can affect this process, especially if the matrix is ill-conditioned. Entries that should be zero may end up as very small nonzero numbers, affecting our ability to accurately determine the rank and other quantities associated with the matrix. In practice, the SVD is often used to find the rank of a matrix, since it is much more reliable when roundoff errors are present. The basic idea behind this approach is that the orthogonal matrices U and V in the SVD preserve lengths and thus do not introduce additional errors; any errors that occur will tend to show up in the matrix Σ.

CAS **Example 7.36**

Let

$$A = \begin{bmatrix} 8.1650 & -0.0041 & -0.0041 \\ 4.0825 & -3.9960 & 4.0042 \\ 4.0825 & 4.0042 & -3.9960 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 8.17 & 0 & 0 \\ 4.08 & -4 & 4 \\ 4.08 & 4 & -4 \end{bmatrix}$$

The matrix B has been obtained by rounding off the entries in A to two decimal places. If we compute the ranks of these two approximately equal matrices, we find that $\text{rank}(A) = 3$ but $\text{rank}(B) = 2$. By the Fundamental Theorem, this implies, among other things, that A is invertible but B is not.

The explanation for this critical difference between two matrices that are approximately equal lies in their SVDs. The singular values of A are 10, 8, and 0.01, so A has rank 3. The singular values of B are 10, 8, and 0, so B has rank 2.

In practical applications, it is often assumed that if a singular value is computed to be close to zero, then roundoff error has crept in and the actual value should be zero. In this way, "noise" can be filtered out. In this example, if we compute $A = U\Sigma V^T$ and replace

$$\Sigma = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 0.01 \end{bmatrix} \quad \text{by} \quad \Sigma' = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

then $U\Sigma' V^T = B$. (Try it!)

Matrix Norms and the Condition Number The SVD can provide simple formulas for certain expressions involving matrix norms. Consider, for example, the Frobenius norm of a matrix. The following theorem shows that it is completely determined by the singular values of the matrix.

Theorem 7.17 Let A be an $m \times n$ matrix and let $\sigma_1, \ldots, \sigma_r$ be all the nonzero singular values of A. Then

$$\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$$

The proof of this result depends on the following analogue of Theorem 5.6:

If A is an $m \times n$ matrix and Q is an $m \times m$ orthogonal matrix, then

$$\|QA\|_F = \|A\|_F \tag{2}$$

To show that this is true, we compute

$$\begin{aligned}\|QA\|_F^2 &= \|[Q\mathbf{a}_1 \quad \cdots \quad Q\mathbf{a}_n]\|_F^2 \\ &= \|Q\mathbf{a}_1\|_E^2 + \cdots + \|Q\mathbf{a}_n\|_E^2 \\ &= \|\mathbf{a}_1\|_E^2 + \cdots + \|\mathbf{a}_n\|_E^2 \\ &= \|A\|_F^2\end{aligned}$$

Proof of Theorem 7.17 Let $A = U\Sigma V^T$ be a singular value decomposition of A. Then, using Equation (2) twice, we have

$$\begin{aligned}\|A\|_F^2 &= \|U\Sigma V^T\|_F^2 \\ &= \|\Sigma V^T\|_F^2 = \|(\Sigma V^T)^T\|_F^2 \\ &= \|V\Sigma^T\|_F^2 = \|\Sigma^T\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2\end{aligned}$$

which establishes the result. ∎

CAS **Example 7.37** Verify Theorem 7.17 for the matrix A in Example 7.18.

Solution The matrix $A = \begin{bmatrix} 3 & -1 \\ 2 & 4 \end{bmatrix}$ has singular values 4.5150 and 3.1008. We check that

$$\sqrt{4.5150^2 + 3.1008^2} = \sqrt{30} = \|A\|_F$$

which agrees with Example 7.18. ◀

In Section 7.2, we commented that there is no easy formula for the operator 2-norm of a matrix A. Although that is true, the SVD of A provides us with a very nice expression for $\|A\|_2$. Recall that

$$\|A\|_2 = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

where the vector norm is the ordinary Euclidean norm. By Theorem 7.16, for $\|\mathbf{x}\| = 1$, the set of vectors $\|A\mathbf{x}\|$ lies on or inside an ellipsoid whose semi-axes have lengths

equal to the nonzero singular values of A. It follows immediately that the largest of these is σ_1, so

$$\|A\|_2 = \sigma_1$$

This provides us with a neat way to express the condition number of a (square) matrix with respect to the operator 2-norm. Recall that the condition number (with respect to the operator 2-norm) of an invertible matrix A is defined as

$$\text{cond}_2(A) = \|A^{-1}\|_2\|A\|_2$$

As you will be asked to show in Exercise 28, if $A = U\Sigma V^T$, then $A^{-1} = V\Sigma^{-1}U^T$. Therefore, the singular values of A^{-1} are $1/\sigma_1, \ldots, 1/\sigma_n$ (why?), and

$$1/\sigma_n \geq \cdots \geq 1/\sigma_1$$

It follows that $\|A^{-1}\|_2 = 1/\sigma_n$, so

$$\text{cond}_2(A) = \frac{\sigma_1}{\sigma_n}$$

Example 7.38

Find the 2-condition number of the matrix A in Example 7.36.

Solution Since $\sigma_1 = 10$ and $\sigma_3 = 0.01$,

$$\text{cond}_2(A) = \frac{\sigma_1}{\sigma_3} = \frac{10}{0.01} = 1000$$

This value is large enough to suggest that A may be ill-conditioned and we should be wary of the effect of roundoff errors.

The Pseudoinverse and Least Squares Approximation In Section 7.3, we produced the formula $A^+ = (A^TA)^{-1}A^T$ for the pseudoinverse of a matrix A. Clearly, this formula is valid only if A^TA is invertible, as we noted at the time. Equipped with the SVD, we can now define the pseudoinverse of *any* matrix, generalizing our previous formula.

E. H. Moore (1862–1932) was an American mathematician who worked in group theory, number theory, and geometry. He was the first head of the mathematics department at the University of Chicago when it opened in 1892. In 1920, he introduced a generalized matrix inverse that included rectangular matrices. His work did not receive much attention because of his obscure writing style.

Definition Let $A = U\Sigma V^T$ be an SVD for an $m \times n$ matrix A, where $\Sigma = \begin{bmatrix} D & O \\ O & O \end{bmatrix}$ and D is an $r \times r$ diagonal matrix containing the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ of A. The ***pseudoinverse*** (or ***Moore-Penrose inverse***) of A is the $n \times m$ matrix A^+ defined by

$$A^+ = V\Sigma^+U^T$$

where Σ^+ is the $n \times m$ matrix

$$\Sigma^+ = \begin{bmatrix} D^{-1} & O \\ O & O \end{bmatrix}$$

Example 7.39 Find the pseudoinverses of the matrices in Example 7.34.

Solution (a) From the SVD

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix} = U\Sigma V^T$$

we form

$$\Sigma^+ = \begin{bmatrix} 1/\sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Then

$$A^+ = V\Sigma^+ U^T = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1 \end{bmatrix}$$



Jerry Bauer

One of those who was unaware of Moore's work on matrix inverses was Roger Penrose (b.1931), who introduced his own notion of a generalized matrix inverse in 1955. Penrose has made many contributions to geometry and theoretical physics. He is also the inventor of a type of *nonperiodic tiling* that covers the plane with only two different shapes of tile, yet has no repeating pattern. He has received many awards, including the 1988 Wolf Prize in Physics, which he shared with Stephen Hawking. In 1994, he was knighted for services to science. Sir Roger Penrose is currently the Emeritus Rouse Ball Professor of Mathematics at the University of Oxford.

(b) We have the SVD

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = U\Sigma V^T$$

so

$$\Sigma^+ = \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

and

$$A^+ = V\Sigma^+ U^T = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix}$$

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



It is straightforward to check that this new definition of the pseudoinverse generalizes the old one, for if the $m \times n$ matrix $A = U\Sigma V^T$ has linearly independent columns, then direct substitution shows that $(A^TA)^{-1}A^T = V\Sigma^+ U^T$. (You are asked to verify this in Exercise 50.) Other properties of the pseudoinverse are explored in the exercises.

We have seen that when A has linearly independent columns, there is a unique least squares solution $\bar{\mathbf{x}}$ to $A\mathbf{x} = \mathbf{b}$; that is, the normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$ have the unique solution

$$\bar{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b} = A^+\mathbf{b}$$

When the columns of A are linearly dependent, then A^TA is not invertible, so the normal equations have infinitely many solutions. In this case, we will ask for the solution $\bar{\mathbf{x}}$ of *minimum length* (i.e., the one closest to the origin). It turns out that this time we simply use the general version of the pseudoinverse.

Theorem 7.18 The least squares problem $A\mathbf{x} = \mathbf{b}$ has a unique least squares solution $\bar{\mathbf{x}}$ of minimal length that is given by

$$\bar{\mathbf{x}} = A^+\mathbf{b}$$

Proof Let A be an $m \times n$ matrix of rank r with SVD $A = U\Sigma V^T$ (so that $A^+ = V\Sigma^+U^T$). Let $\mathbf{y} = V^T\mathbf{x}$ and let $\mathbf{c} = U^T\mathbf{b}$. Write $\mathbf{y}$ and $\mathbf{c}$ in block form as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}$$

where $\mathbf{y}_1$ and $\mathbf{c}_1$ are in $\mathbb{R}^r$.

We wish to minimize $\|\mathbf{b} - A\mathbf{x}\|$ or, equivalently, $\|\mathbf{b} - A\mathbf{x}\|^2$. Using Theorem 5.6 and the fact that U^T is orthogonal (because U is), we have

$$\begin{aligned} \|\mathbf{b} - A\mathbf{x}\|^2 &= \|U^T(\mathbf{b} - A\mathbf{x})\|^2 = \|U^T(\mathbf{b} - U\Sigma V^T\mathbf{x})\|^2 = \|U^T\mathbf{b} - U^TU\Sigma V^T\mathbf{x}\|^2 \\ &= \|\mathbf{c} - \Sigma\mathbf{y}\|^2 = \left\| \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} - \begin{bmatrix} D & O \\ O & O \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} \mathbf{c}_1 - D\mathbf{y}_1 \\ \mathbf{c}_2 \end{bmatrix} \right\|^2 \end{aligned}$$

The only part of this expression that we have any control over is $\mathbf{y}_1$, so the minimum value occurs when $\mathbf{c}_1 - D\mathbf{y}_1 = \mathbf{0}$ or, equivalently, when $\mathbf{y}_1 = D^{-1}\mathbf{c}_1$. So all least squares solutions $\mathbf{x}$ are of the form

$$\mathbf{x} = V\mathbf{y} = V \begin{bmatrix} D^{-1}\mathbf{c}_1 \\ \mathbf{y}_2 \end{bmatrix}$$

Set

$$\bar{\mathbf{x}} = V\bar{\mathbf{y}} = V \begin{bmatrix} D^{-1}\mathbf{c}_1 \\ \mathbf{0} \end{bmatrix}$$

We claim that this $\bar{\mathbf{x}}$ is the least squares solution of minimal length. To show this, let's suppose that

$$\mathbf{x}' = V\mathbf{y}' = V \begin{bmatrix} D^{-1}\mathbf{c}_1 \\ \mathbf{y}_2 \end{bmatrix}$$

is a different least squares solution (hence, $\mathbf{y}_2 \neq \mathbf{0}$). Then

$$\|\bar{\mathbf{x}}\| = \|V\bar{\mathbf{y}}\| = \|\bar{\mathbf{y}}\| < \|\mathbf{y}'\| = \|V\mathbf{y}'\| = \|\mathbf{x}'\|$$

as claimed.

We still must show that $\bar{\mathbf{x}}$ is equal to $A^+\mathbf{b}$. To do so, we simply compute

$$\begin{aligned} \bar{\mathbf{x}} = V\bar{\mathbf{y}} &= V \begin{bmatrix} D^{-1}\mathbf{c}_1 \\ \mathbf{0} \end{bmatrix} = V \begin{bmatrix} D^{-1} & O \\ O & O \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} \\ &= V\Sigma^+\mathbf{c} = V\Sigma^+U^T\mathbf{b} = A^+\mathbf{b} \end{aligned}$$

Example 7.40 Find the minimum length least squares solution of $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Solution The corresponding equations

$$x + y = 0$$
$$x + y = 1$$

are clearly inconsistent, so a least squares solution is our only hope. Moreover, the columns of A are linearly dependent, so there will be infinitely many least squares solutions—among which we want the one with minimal length.

An SVD of A is given by

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} = U\Sigma V^T$$

(Verify this.) It follows that

$$A^+ = V\Sigma^+ U^T = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}$$

so

$$\bar{\mathbf{x}} = A^+\mathbf{b} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}$$

You can see that the minimum least squares solution in Example 7.40 satisfies $x + y = \frac{1}{2}$. In a sense, this is a compromise between the two equations we started with. In Exercise 49, you are asked to solve the normal equations for this problem directly and to verify that this solution really is the one closest to the origin.

The Fundamental Theorem of Invertible Matrices It is appropriate to conclude by revisiting the Fundamental Theorem of Invertible Matrices one more time. Not surprisingly, the singular values of a square matrix tell us when the matrix is invertible.

Theorem 7.19 The Fundamental Theorem of Invertible Matrices: Final Version

Let A be an $n \times n$ matrix and let $T : V \to W$ be a linear transformation whose matrix $[T]_{\mathcal{C}\leftarrow\mathcal{B}}$ with respect to bases $\mathcal{B}$ and $\mathcal{C}$ of V and W, respectively, is A. The following statements are equivalent:

a. A is invertible.
b. $A\mathbf{x} = \mathbf{b}$ has a unique solution for every $\mathbf{b}$ in $\mathbb{R}^n$.
c. $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
d. The reduced row echelon form of A is I_n.
e. A is a product of elementary matrices.
f. $\text{rank}(A) = n$
g. $\text{nullity}(A) = 0$
h. The column vectors of A are linearly independent.
i. The column vectors of A span $\mathbb{R}^n$.
j. The column vectors of A form a basis for $\mathbb{R}^n$.
k. The row vectors of A are linearly independent.
l. The row vectors of A span $\mathbb{R}^n$.

m. The row vectors of A form a basis for $\mathbb{R}^n$.
n. $\det A \neq 0$
o. 0 is not an eigenvalue of A.
p. T is invertible.
q. T is one-to-one.
r. T is onto.
s. $\ker(T) = \{\mathbf{0}\}$
t. $\text{range}(T) = W$
u. 0 is not a singular value of A.

Proof First note that, by the definition of singular values, 0 is a singular value of A if and only if 0 is an eigenvalue of A^TA.

(a) $\Rightarrow$ (u) If A is invertible, so is A^T, and hence A^TA is as well. Therefore, property (o) implies that 0 is not an eigenvalue of A^TA, so 0 is not a singular value of A.

(u) $\Rightarrow$ (a) If 0 is not a singular value of A, then 0 is not an eigenvalue of A^TA. Therefore, A^TA is invertible, by the equivalence of properties (a) and (o). But then $\text{rank}(A) = n$, by Theorem 3.28, so A is invertible, by the equivalence of properties (a) and (f).

Vignette

Digital Image Compression

Among the many applications of the SVD, one of the most impressive is its use in compressing digital images so that they can be efficiently transmitted electronically (by satellite, fax, Internet, or the like). We have already discussed the problem of detecting and correcting errors in such transmissions. The problem we now wish to consider has to do with reducing the amount of information that has to be transmitted, without losing any essential information.

In the case of digital images, let's suppose we have a grayscale picture that is 340×280 pixels in size. Each pixel is one of 256 shades of gray, which we can represent by a number between 0 and 255. We can store this information in a 340×280 matrix A, but transmitting and manipulating these 95,200 numbers is very expensive. The idea behind image compression is that some parts of the picture are less interesting than others. For example, in a photograph of someone standing outside, there may be a lot of sky in the background, while the person's face contains a lot of detail. We can probably get away with transmitting every second or third pixel in the background, but we would like to keep all the pixels in the region of the face.

It turns out that the small singular values in the SVD of the matrix A come from the "boring" parts of the image, and we can ignore many of them. Suppose, then, that we have the SVD of A in outer product form

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T$$

Let $k \leq r$ and define $$A_k = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

Then A_k is an approximation to A that corresponds to keeping only the first k singular values and the corresponding singular vectors. For our 340×280 example, we may discover that it is enough to transmit only the data corresponding to the first 20 singular values. Then, instead of transmitting 95,200 numbers, we need only send 20 singular values plus the 20 vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{20}$ in $\mathbb{R}^{340}$ and the 20 vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{20}$ in $\mathbb{R}^{280}$, for a total of

$$20 + 20 \cdot 340 + 20 \cdot 280 = 12{,}420$$

numbers. This represents a substantial saving!



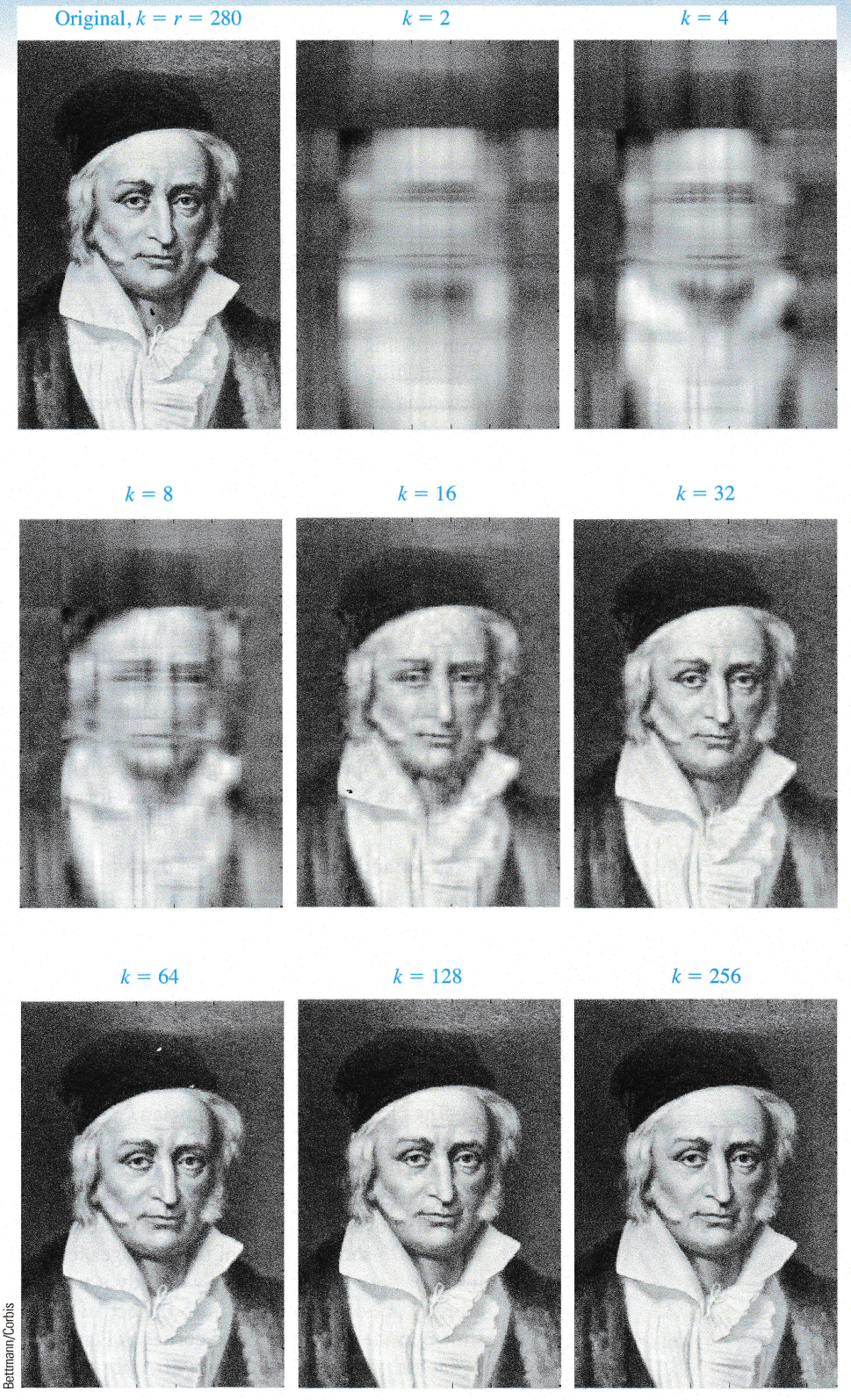
Bettmann/Corbis

Figure 7.22

The picture of the mathematician Gauss in Figure 7.22 is a 340×280 pixel image. It has 256 shades of gray, so the corresponding matrix A is 340×280, with entries between 0 and 255.

It turns out that the matrix A has rank 280. If we approximate A by A_k, as described above, we get an image that corresponds to the first k singular values of A. Figure 7.23 shows several of these images for values of k from 2 to 256. At first, the image is very blurry, but fairly quickly it takes shape. Notice that A_{32} already gives a pretty good approximation to the actual image (which comes from $A = A_{280}$, as shown in the upper left-hand corner of Figure 7.23).

Some of the singular values of A are $\sigma_1 = 49{,}096$, $\sigma_{16} = 22{,}589$, $\sigma_{32} = 10{,}187$, $\sigma_{64} = 484$, $\sigma_{128} = 182$, $\sigma_{256} = 5$, and $\sigma_{280} = 0.5$. The smaller singular values contribute very little to the image, which is why the approximations quickly look so close to the original.



Bettmann/Corbis

Figure 7.23

Exercises 7.4

In Exercises 1–10, find the singular values of the given matrix.

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

3. $A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ **4.** $A = \begin{bmatrix} \sqrt{2} & 1 \\ 0 & \sqrt{2} \end{bmatrix}$

5. $A = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ **6.** $A = \begin{bmatrix} 3 & 4 \end{bmatrix}$

7. $A = \begin{bmatrix} 0 & 0 \\ 0 & 3 \\ -2 & 0 \end{bmatrix}$ **8.** $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & -2 \end{bmatrix}$

9. $A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \end{bmatrix}$ **10.** $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & -3 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

In Exercises 11–20, find an SVD of the indicated matrix.

11. A in Exercise 3 **12.** $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

13. $A = \begin{bmatrix} 0 & -2 \\ -3 & 0 \end{bmatrix}$ **14.** $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

15. A in Exercise 5 **16.** A in Exercise 6

17. A in Exercise 7 **18.** A in Exercise 8

19. A in Exercise 9 **20.** $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

In Exercises 21–24, find the outer product form of the SVD for the matrix in the given exercises.

21. Exercises 3 and 11 **22.** Exercise 14

23. Exercises 7 and 17 **24.** Exercises 9 and 19

25. Show that the matrices U and V in the SVD are not uniquely determined. [*Hint:* Find an example in which it would be possible to make different choices in the construction of these matrices.]

26. Let A be a symmetric matrix. Show that the singular values of A are:
(a) the absolute values of the eigenvalues of A.
(b) the eigenvalues of A if A is positive definite.

27. **(a)** Show that, for a positive definite, symmetric matrix A, Theorem 7.13 gives the orthogonal diagonalization of A, as guaranteed by the Spectral Theorem.
(b) Show that, for a positive definite, symmetric matrix A, Theorem 7.14 gives the spectral decomposition of A.

28. If A is an invertible matrix with SVD $A = U\Sigma V^T$, show that Σ is invertible and that $A^{-1} = V\Sigma^{-1}U^T$ is an SVD of A^{-1}.

29. Show that if $A = U\Sigma V^T$ is an SVD of A, then the left singular vectors are eigenvectors of AA^T.

30. Show that A and A^T have the same singular values.

31. Let Q be an orthogonal matrix such that QA makes sense. Show that A and QA have the same singular values.

32. Prove Theorem 7.15(d).

33. What is the image of the unit circle in $\mathbb{R}^2$ under the action of the matrix in Exercise 3?

34. What is the image of the unit circle in $\mathbb{R}^2$ under the action of the matrix in Exercise 7?

35. What is the image of the unit sphere in $\mathbb{R}^3$ under the action of the matrix in Exercise 9?

36. What is the image of the unit sphere in $\mathbb{R}^3$ under the action of the matrix in Exercise 10?

In Exercises 37–40, compute (a) $\|A\|_2$ and (b) $\text{cond}_2(A)$ for the indicated matrix.

37. A in Exercise 3 **38.** A in Exercise 8

39. $A = \begin{bmatrix} 1 & 0.9 \\ 1 & 1 \end{bmatrix}$ **40.** $A = \begin{bmatrix} 10 & 10 & 0 \\ 100 & 100 & 1 \end{bmatrix}$

In Exercises 41–44, compute the pseudoinverse A^+ of A in the given exercise.

41. Exercise 3 **42.** Exercise 8

43. Exercise 9 **44.** Exercise 10

In Exercises 45–48, find A^+ and use it to compute the minimal length least squares solution to $A\mathbf{x} = \mathbf{b}$.

45. $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$

46. $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$

47. $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

48. $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

49. **(a)** Set up and solve the normal equations for the system of equations in Example 7.40.
(b) Find a parametric expression for the length of a solution vector in part (a).
(c) Find the solution vector of minimal length and verify that it is the one produced by the method of Example 7.40. [*Hint:* Recall how to find the coordinates of the vertex of a parabola.]

50. Verify that when A has linearly independent columns, the definitions of pseudoinverse in this section and in Section 7.3 are the same.

51. Verify that the pseudoinverse (as defined in this section) satisfies the Penrose conditions for A (Theorem 7.12 in Section 7.3).

52. Show that A^+ is the *only* matrix that satisfies the Penrose conditions for A. To do this, assume that A' is a matrix satisfying the Penrose conditions: (a) $AA'A = A$, (b) $A'AA' = A'$, and (c) AA' and $A'A$ are symmetric. Prove that $A' = A^+$. [*Hint:* Use the Penrose conditions for A^+ and A' to show that $A^+ = A'AA^+$ and $A' = A'AA^+$. It is helpful to note that condition (c) can be written as $AA' = (A')^TA^T$ and $A'A = A^T(A')^T$, with similar versions for A^+.]

53. Show that $(A^+)^+ = A$. [*Hint:* Show that A satisfies the Penrose conditions for A^+. By Exercise 52, A must therefore be $(A^+)^+$.]

54. Show that $(A^+)^T = (A^T)^+$. [*Hint:* Show that $(A^+)^T$ satisfies the Penrose conditions for A^T. By Exercise 52, $(A^+)^T$ must therefore be $(A^T)^+$.]

55. Show that if A is a symmetric, idempotent matrix, then $A^+ = A$.

56. Let Q be an orthogonal matrix such that QA makes sense. Show that $(QA)^+ = A^+Q^T$.

57. Prove that if A is a positive definite matrix with SVD $A = U\Sigma V^T$, then $U = V$.

58. Prove that for a diagonal matrix, the 1-, 2-, and ∞-norms are the same.

59. Prove that for any square matrix A, $\|A\|_2^2 \leq \|A\|_1\|A\|_\infty$. [*Hint:* $\|A\|_2^2$ is the square of the largest singular value of A and hence is equal to the largest eigenvalue of A^TA. Now use Exercise 34 in Section 7.2.]

a + bi *Every complex number can be written in polar form as $z = re^{i\theta}$, where $r = |z|$ is a nonnegative real number and θ is its argument, with $|e^{i\theta}| = 1$. (See Appendix C.) Thus, z has been decomposed into a stretching factor r and a rotation factor $e^{i\theta}$. There is an analogous decomposition $A = RQ$ for square matrices, called the* ***polar decomposition.***

60. Show that every square matrix A can be factored as $A = RQ$, where R is symmetric, positive semidefinite and Q is orthogonal. [*Hint:* Show that the SVD can be rewritten to give

$$A = U\Sigma V^T = U\Sigma(U^TU)V^T = (U\Sigma U^T)(UV^T)$$

Then show that $R = U\Sigma U^T$ and $Q = UV^T$ have the right properties.]

Find a polar decomposition of the matrices in Exercises 61–64.

61. A in Exercise 3

62. A in Exercise 14

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

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



7.5 Applications

Approximation of Functions

dy/dx

In many applications, it is necessary to approximate a given function by a "nicer" function. For example, we might want to approximate $f(x) = e^x$ by a linear function $g(x) = c + dx$ on some interval $[a, b]$. In this case, we have a continuous function f, and we want to approximate it as closely as possible on the interval $[a, b]$

by a function g in the subspace $\mathcal{P}_1$. The general problem can be phrased as follows:

Given a continuous function f on an interval $[a, b]$ and a subspace W of $\mathcal{C}[a, b]$, find the function "closest" to f in W.

The problem is analogous to the least squares fitting of data points, except now we have infinitely many data points—namely, the points on the graph of the function f. What should "approximate" mean in this context? Once again, the Best Approximation Theorem holds the answer.

The given function f lives in the vector space $\mathcal{C}[a, b]$ of continuous functions on the interval $[a, b]$. This is an inner product space, with inner product

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,dx$$

If W is a finite-dimensional subspace of $\mathcal{C}[a, b]$, then the best approximation to f in W is given by the projection of f onto W, by Theorem 7.8. Furthermore, if $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is an orthogonal basis for W, then

$$\text{proj}_W(f) = \frac{\langle \mathbf{u}_1, f \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\mathbf{u}_1 + \cdots + \frac{\langle \mathbf{u}_k, f \rangle}{\langle \mathbf{u}_k, \mathbf{u}_k \rangle}\mathbf{u}_k$$

Example 7.41 Find the best linear approximation to $f(x) = e^x$ on the interval $[-1, 1]$.

Solution Linear functions are polynomials of degree 1, so we use the subspace $W = \mathcal{P}_1[-1, 1]$ of $\mathcal{C}[-1, 1]$ with the inner product

$$\langle f, g \rangle = \int_{-1}^1 f(x)g(x)\,dx$$

A basis for $\mathcal{P}_1[-1, 1]$ is given by $\{1, x\}$. Since

$$\langle 1, x \rangle = \int_{-1}^1 x\,dx = 0$$

this is an orthogonal basis, so the best approximation to f in W is

$$\begin{aligned} g(x) = \text{proj}_W(e^x) &= \frac{\langle 1, e^x \rangle}{\langle 1, 1 \rangle}1 + \frac{\langle x, e^x \rangle}{\langle x, x \rangle}x \\ &= \frac{\displaystyle\int_{-1}^1 (1 \cdot e^x)\,dx}{\displaystyle\int_{-1}^1 (1 \cdot 1)\,dx} + \frac{\displaystyle\int_{-1}^1 xe^x\,dx}{\displaystyle\int_{-1}^1 x^2\,dx}x \\ &= \frac{e - e^{-1}}{2} + \frac{2e^{-1}}{\frac{2}{3}}x \\ &= \tfrac{1}{2}(e - e^{-1}) + 3e^{-1}x \approx 1.18 + 1.10x \end{aligned}$$

where we have used integration by parts to evaluate $\int_{-1}^{1} xe^x\,dx$. (Check these calculations.) See Figure 7.24.

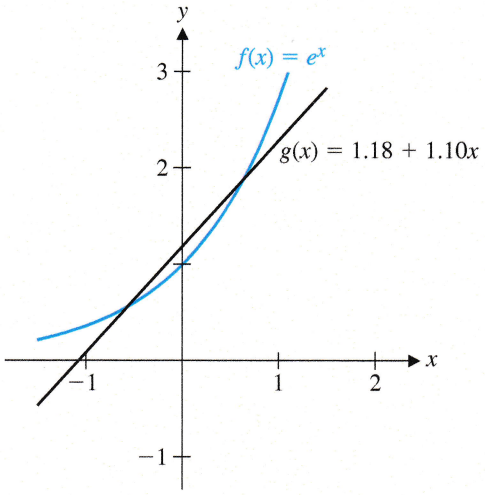


Figure 7.24

The error in approximating f by g is the one specified by the Best Approximation Theorem: the distance $\|f - g\|$ between f and g relative to the inner product on $\mathscr{C}[-1, 1]$. This error is just

$$\|f - g\| = \sqrt{\int_{-1}^{1} (f(x) - g(x))^2\,dx}$$

and is often called the ***root mean square error.*** With the aid of a CAS, we find that the root mean square error in Example 7.41 is

$$\|e^x - (\tfrac{1}{2}(e - e^{-1}) + 3e^{-1}x)\| = \sqrt{\int_{-1}^{1} (e^x - \tfrac{1}{2}(e - e^{-1}) - 3e^{-1}x)^2\,dx} \approx 0.23$$

Remark The root mean square error can be thought of as analogous to the area between the graphs of f and g on the specified interval. Recall that the area between the graphs of f and g on the interval $[a, b]$ is given by

$$\int_a^b |f(x) - g(x)|\,dx$$

(See Figure 7.25.)

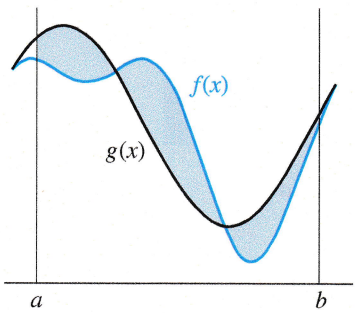


Figure 7.25

Although the equation in the above Remark is a sensible measure of the "error" between f and g, the absolute value sign makes it hard to work with. The root mean square error is easier to use and therefore preferable. The square root is necessary to "compensate" for the squaring and to keep the unit of measurement the same as it would be for the area between the curves. For comparison purposes, the area between the graphs of f and g in Example 7.41 is

$$\int_{-1}^{1} |e^x - \tfrac{1}{2}(e - e^{-1}) - 3e^{-1}x|\,dx \approx 0.28$$

Example 7.42 Find the best quadratic approximation to $f(x) = e^x$ on the interval $[-1, 1]$.

Solution A quadratic function is a polynomial of the form $g(x) = a + bx + cx^2$ in $W = \mathcal{P}_2[-1, 1]$. This time, the standard basis $\{1, x, x^2\}$ is not orthogonal. However, we can construct an orthogonal basis using the Gram-Schmidt Process, as we did in Example 7.8. The result is the set of Legendre polynomials

$$\{1, x, x^2 - \tfrac{1}{3}\}$$

Using this set as our basis, we compute the best approximation to f in W as $g(x) = \text{proj}_W(e^x)$. The linear terms in this calculation are exactly as in Example 7.41, so we only require the additional calculations

$$\langle x^2 - \tfrac{1}{3}, e^x \rangle = \int_{-1}^{1} (x^2 - \tfrac{1}{3})e^x \, dx = \int_{-1}^{1} x^2 e^x \, dx - \tfrac{1}{3}\int_{-1}^{1} e^x \, dx = \tfrac{2}{3}(e - 7e^{-1})$$

and $$\langle x^2 - \tfrac{1}{3}, x^2 - \tfrac{1}{3} \rangle = \int_{-1}^{1} (x^2 - \tfrac{1}{3})^2 \, dx = \int_{-1}^{1} (x^4 - \tfrac{2}{3}x^2 + \tfrac{1}{9}) \, dx = \tfrac{8}{45}$$

Then the best quadratic approximation to $f(x) = e^x$ on the interval $[-1, 1]$ is

$$\begin{aligned} g(x) = \text{proj}_W(e^x) &= \frac{\langle 1, e^x \rangle}{\langle 1, 1 \rangle} 1 + \frac{\langle x, e^x \rangle}{\langle x, x \rangle} x + \frac{\langle x^2 - \frac{1}{3}, e^x \rangle}{\langle x^2 - \frac{1}{3}, x^2 - \frac{1}{3} \rangle} (x^2 - \tfrac{1}{3}) \\ &= \tfrac{1}{2}(e - e^{-1}) + 3e^{-1}x + \frac{\frac{2}{3}(e - 7e^{-1})}{\frac{8}{45}} (x^2 - \tfrac{1}{3}) \\ &= \frac{3(11e^{-1} - e)}{4} + 3e^{-1}x + \frac{15(e - 7e^{-1})}{4} x^2 \approx 1.00 + 1.10x + 0.54x^2 \end{aligned}$$

(See Figure 7.26.)

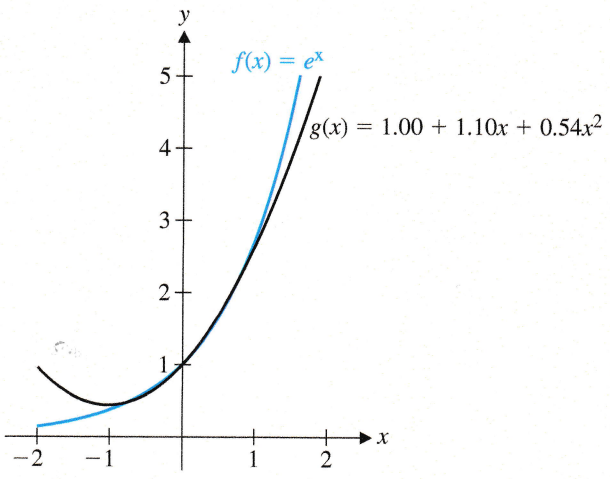


Figure 7.26

Notice how much better the quadratic approximation in Example 7.42 is than the linear approximation in Example 7.41. It turns out that, in the quadratic case, the root mean square error is

$$\|e^x - g(x)\| = \sqrt{\int_{-1}^{1} (e^x - g(x))^2 \, dx} \approx 0.04$$

In general, the higher the degree of the approximating polynomial, the smaller the error and the better the approximation.

In many applications, functions are approximated by combinations of sine and cosine functions. This method is particularly useful if the function being approximated displays periodic or almost periodic behavior (such as that of a sound wave, an electrical impulse, or the motion of a vibrating system). A function of the form

$$\begin{aligned} p(x) = a_0 &+ a_1 \cos x + a_2 \cos 2x + \cdots + a_n \cos nx + b_1 \sin x \\ &+ b_2 \sin 2x + \cdots + b_n \sin nx \end{aligned} \tag{1}$$

is called a ***trigonometric polynomial;*** if a_n and b_n are not both zero, then $p(x)$ is said to have ***order n.*** For example,

$$p(x) = 3 - \cos x + \sin 2x + 4 \sin 3x$$

is a trigonometric polynomial of order 3.

Let's restrict our attention to the vector space $\mathcal{C}[-\pi, \pi]$ with the inner product

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)g(x) \, dx$$

The trigonometric polynomials of the form in Equation (1) are linear combinations of the set

$$\mathcal{B} = \{1, \cos x, \ldots, \cos nx, \sin x, \ldots, \sin nx\}$$

The best approximation to a function f in $\mathcal{C}[-\pi, \pi]$ by a trigonometric polynomial of order n will therefore be $\text{proj}_W(f)$, where $W = \text{span}(\mathcal{B})$. It turns out that $\mathcal{B}$ is an orthogonal set and, hence, a basis for W. Verification of this fact involves showing that any two distinct functions in $\mathcal{B}$ are orthogonal with respect to the given inner product. Example 7.43 presents some of the necessary calculations; you are asked to provide the remaining ones in Exercises 17–19.

Example 7.43 Show that $\sin jx$ is orthogonal to $\cos kx$ in $\mathcal{C}[-\pi, \pi]$ for $j, k \geq 1$.

Solution Using a trigonometric identity, we compute as follows: If $j \neq k$, then

$$\begin{aligned} \int_{-\pi}^{\pi} \sin jx \cos kx \, dx &= \tfrac{1}{2} \int_{-\pi}^{\pi} [\sin(j + k)x + \sin(j - k)x] \, dx \\ &= -\tfrac{1}{2} \left[\frac{\cos(j + k)x}{j + k} + \frac{\cos(j - k)x}{j - k} \right]_{-\pi}^{\pi} \\ &= 0 \end{aligned}$$

since the cosine function is periodic with period 2π.

If $j = k$, then

$$\int_{-\pi}^{\pi} \sin kx \cos kx \, dx = \frac{1}{2k}[\sin^2 kx]_{-\pi}^{\pi} = 0$$

since $\sin k\pi = 0$ for any integer k.



In order to find the orthogonal projection of a function f in $\mathcal{C}[-\pi, \pi]$ onto the subspace W spanned by the orthogonal basis $\mathcal{B}$, we need to know the squares of the norms of the basis vectors. For example, using a half-angle formula, we have

$$\begin{aligned}\langle \sin kx, \sin kx \rangle &= \int_{-\pi}^{\pi} \sin^2 kx \, dx \\ &= \tfrac{1}{2}\int_{-\pi}^{\pi} (1 - \cos 2kx)\, dx \\ &= \tfrac{1}{2}\left[x - \frac{\sin 2kx}{2k}\right]_{-\pi}^{\pi} \\ &= \pi\end{aligned}$$

In Exercise 20, you are asked to show that $\langle \cos kx, \cos kx \rangle = \pi$ and $\langle 1, 1 \rangle = 2\pi$.

We now have

$$\text{proj}_W(f) = a_0 + a_1 \cos x + \cdots + a_n \cos nx + b_1 \sin x + \cdots + b_n \sin nx \qquad (2)$$

where

$$\begin{aligned} a_0 &= \frac{\langle 1, f \rangle}{\langle 1, 1 \rangle} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\, dx \\ a_k &= \frac{\langle \cos kx, f \rangle}{\langle \cos kx, \cos kx \rangle} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \cos kx \, dx \qquad (3) \\ b_k &= \frac{\langle \sin kx, f \rangle}{\langle \sin kx, \sin kx \rangle} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \sin kx \, dx \end{aligned}$$

for $k \geq 1$. The approximation to f given by Equations (2) and (3) is called the ***nth-order Fourier approximation*** to f on $[-\pi, \pi]$. The coefficients $a_0, a_1, \ldots, a_n, b_1, \ldots, b_n$ are called the ***Fourier coefficients*** of f.

Example 7.44

Find the fourth-order Fourier approximation to $f(x) = x$ on $[-\pi, \pi]$.

Solution Using formulas (3), we obtain

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} x \, dx = \frac{1}{2\pi}\left[\frac{x^2}{2}\right]_{-\pi}^{\pi} = 0$$

and for $k \geq 1$, integration by parts yields

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x \cos kx \, dx = \frac{1}{\pi}\left[\frac{x}{k}\sin kx + \frac{1}{k^2}\cos kx\right]_{-\pi}^{\pi} = 0$$



Photo Researchers

Jean-Baptiste Joseph Fourier (1768–1830) was a French mathematician and physicist who gained prominence through his investigation into the theory of heat. In his landmark solution of the so-called heat equation, he introduced techniques related to what are now known as Fourier series, a tool widely used in many branches of mathematics, physics, and engineering. Fourier was a political activist during the French revolution and became a favorite of Napoleon, accompanying him on his Egyptian campaign in 1798. Later Napoleon appointed Fourier Prefect of Isère, where he oversaw many important engineering projects. In 1808, Fourier was made a baron. He is commemorated by a plaque on the Eiffel Tower.

and
$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x \sin kx\, dx = \frac{1}{\pi}\left[-\frac{x}{k}\cos kx + \frac{1}{k^2}\sin kx\right]_{-\pi}^{\pi}$$
$$= \frac{1}{\pi}\left[\frac{-\pi\cos k\pi - \pi\cos(-k\pi)}{k}\right]$$
$$= \begin{cases} -\dfrac{2}{k} & \text{if } k \text{ is even} \\ \dfrac{2}{k} & \text{if } k \text{ is odd} \end{cases}$$
$$= \frac{2(-1)^{k+1}}{k}$$

It follows that the fourth-order Fourier approximation to $f(x) = x$ on $[-\pi, \pi]$ is

$$2(\sin x - \tfrac{1}{2}\sin 2x + \tfrac{1}{3}\sin 3x - \tfrac{1}{4}\sin 4x)$$

Figure 7.27 shows the first four Fourier approximations to $f(x) = x$ on $[-\pi, \pi]$.

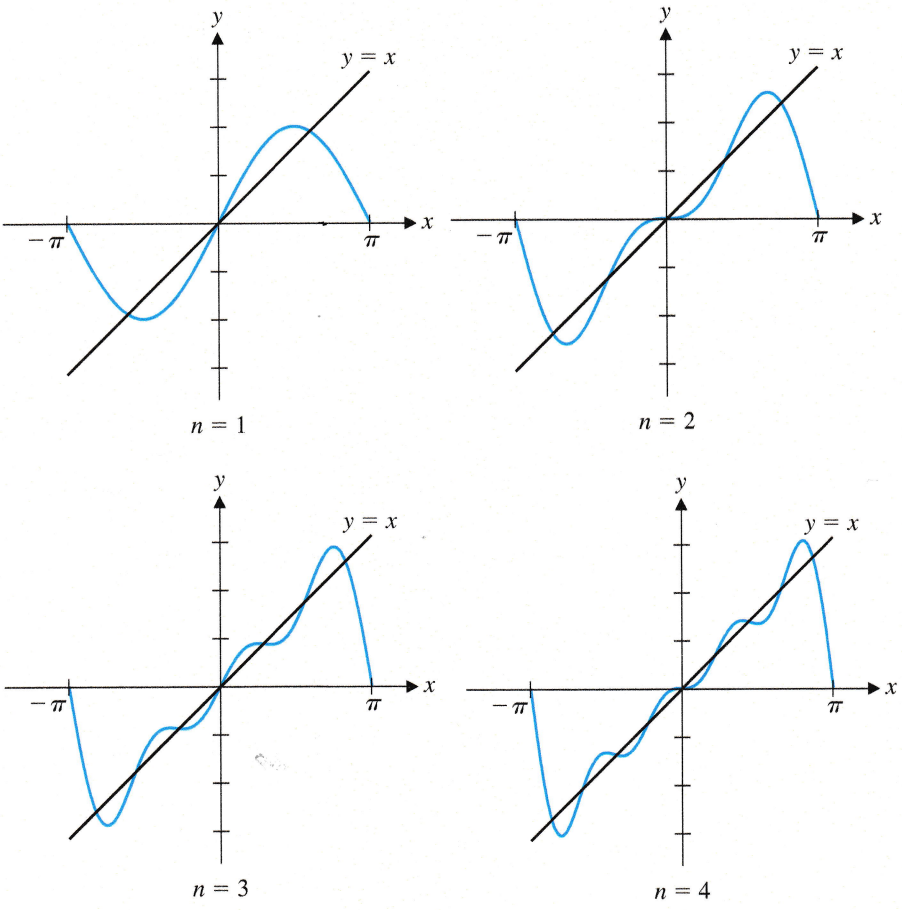


Figure 7.27

You can clearly see the approximations in Figure 7.27 improving, a fact that can be confirmed by computing the root mean square error in each case. As the order of the Fourier approximation increases, it can be shown that this error approaches zero. The trigonometric polynomial then becomes an *infinite series*, and we write

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

This is called the ***Fourier series*** of f on $[-\pi, \pi]$.

Mariner 9 used the Reed-Muller code R_5, whose minimum distance is $2^4 = 16$. By Theorem 1, this code can correct k errors, where $2k + 1 \leq 16$. The largest value of k for which this inequality is true is $k = 7$. Thus, R_5 not only contains exactly the right number of code vectors for transmitting 64 shades of gray but also is capable of correcting up to 7 errors, making it quite reliable. This explains why the images transmitted by *Mariner 9* were so sharp!

Exercises 7.5

Approximation of Functions

In Exercises 1–4, find the best linear approximation to f on the interval $[-1, 1]$.

1. $f(x) = x^2$ **2.** $f(x) = x^2 + 2x$

3. $f(x) = x^3$ **4.** $f(x) = \sin(\pi x/2)$

In Exercises 5 and 6, find the best quadratic approximation to f on the interval $[-1, 1]$.

5. $f(x) = |x|$ **6.** $f(x) = \cos(\pi x/2)$

7. Apply the Gram-Schmidt Process to the basis $\{1, x\}$ to construct an orthogonal basis for $\mathcal{P}_1[0, 1]$.

8. Apply the Gram-Schmidt Process to the basis $\{1, x, x^2\}$ to construct an orthogonal basis for $\mathcal{P}_2[0, 1]$.

In Exercises 9–12, find the best linear approximation to f on the interval $[0, 1]$.

9. $f(x) = x^2$ **10.** $f(x) = \sqrt{x}$

11. $f(x) = e^x$ **12.** $f(x) = \sin(\pi x/2)$

In Exercises 13–16, find the best quadratic approximation to f on the interval $[0, 1]$.

13. $f(x) = x^3$ **14.** $f(x) = \sqrt{x}$

15. $f(x) = e^x$ **16.** $f(x) = \sin(\pi x/2)$

17. Show that 1 is orthogonal to $\cos kx$ and $\sin kx$ in $\mathcal{C}[-\pi, \pi]$ for $k \geq 1$.

18. Show that $\cos jx$ is orthogonal to $\cos kx$ in $\mathcal{C}[-\pi, \pi]$ for $j \neq k, j, k \geq 1$.

19. Show that $\sin jx$ is orthogonal to $\sin kx$ in $\mathcal{C}[-\pi, \pi]$ for $j \neq k, j, k \geq 1$.

20. Show that $\|1\|^2 = 2\pi$ and $\|\cos kx\|^2 = \pi$ in $\mathcal{C}[-\pi, \pi]$.

In Exercises 21 and 22, find the third-order Fourier approximation to f on $[-\pi, \pi]$.

21. $f(x) = |x|$ **22.** $f(x) = x^2$

In Exercises 23–26, find the Fourier coefficients a_0, a_k, and b_k of f on $[-\pi, \pi]$.

23. $f(x) = \begin{cases} 0 & \text{if } -\pi \leq x < 0 \\ 1 & \text{if } 0 \leq x \leq \pi \end{cases}$

24. $f(x) = \begin{cases} -1 & \text{if } -\pi \leq x < 0 \\ 1 & \text{if } 0 \leq x \leq \pi \end{cases}$

25. $f(x) = \pi - x$ **26.** $f(x) = |x|$

*Recall that a function f is an **even function** if $f(-x) = f(x)$ for all x; f is called an **odd function** if $f(-x) = -f(x)$ for all x.*

27. (a) Prove that $\displaystyle\int_{-\pi}^{\pi} f(x)\,dx = 0$ if f is an odd function.

(b) Prove that the Fourier coefficients a_k are all zero if f is odd.

28. (a) Prove that $\displaystyle\int_{-\pi}^{\pi} f(x)\,dx = 2\int_{0}^{\pi} f(x)\,dx$ if f is an even function.

(b) Prove that the Fourier coefficients b_k are all zero if f is even.

Chapter Review



Key Definitions and Concepts

Best Approximation Theorem, 570
Cauchy-Schwarz Inequality, 539
condition number of a matrix, 562
distance, 535
Euclidean norm (2-norm), 553
Frobenius norm, 556
Fundamental Theorem of Invertible Matrices, 605
Hamming distance, 554
Hamming norm, 554
ill-conditioned matrix, 561
inner product, 531
inner product space, 531
least squares error, 572
least squares solution, 574, 604
Least Squares Theorem, 575
matrix norm, 556
max norm (∞-norm, uniform norm), 553
norm, 535, 552
normed linear space, 552
operator norm, 559
orthogonal basis, 537
orthogonal projection, 538, 583
orthogonal (set of) vectors, 537
orthonormal basis, 537
orthonormal set of vectors, 537
pseudoinverse of a matrix, 585, 602
singular value decomposition (SVD), 593
singular values, 590
singular vectors, 593
sum norm (1-norm), 552
Triangle Inequality, 540
unit sphere, 535
unit vector, 535
well-conditioned matrix, 561

Review Questions

1. Mark each of the following statements true or false:

(a) If $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, then $\langle \mathbf{u}, \mathbf{v} \rangle = u_1v_1 + \pi u_2v_2$ defines an inner product on $\mathbb{R}^2$.

(b) If $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, then $\langle \mathbf{u}, \mathbf{v} \rangle = 4u_1v_1 - 2u_1v_2 - 2u_2v_1 + 4u_2v_2$ defines an inner product on $\mathbb{R}^2$.

(c) $\langle A, B \rangle = \text{tr}(A) + \text{tr}(B)$ defines an inner product on M_{22}.

(d) If $\mathbf{u}$ and $\mathbf{v}$ are vectors in an inner product space with $\|\mathbf{u}\| = 4$, $\|\mathbf{v}\| = \sqrt{5}$, and $\langle \mathbf{u}, \mathbf{v} \rangle = 2$, then $\|\mathbf{u} + \mathbf{v}\| = 5$.

(e) The sum norm, max norm, and Euclidean norm on $\mathbb{R}^n$ are all equal to the absolute value function when $n = 1$.

(f) If a matrix A is well-conditioned, then cond(A) is small.

(g) If cond(A) is small, then the matrix A is well-conditioned.

(h) Every linear system has a unique least squares solution.

(i) If A is a matrix with orthonormal columns, then the standard matrix of an orthogonal projection onto the column space of A is $P = AA^T$.

(j) If A is a symmetric matrix, then the singular values of A are the same as the eigenvalues of A.

In Questions 2–4, determine whether the definition gives an inner product.

2. $\langle p(x), q(x) \rangle = p(0)q(1) + p(1)q(0)$ for $p(x), q(x)$ in $\mathcal{P}_1$

3. $\langle A, B \rangle = \text{tr}(A^TB)$ for A, B in M_{22}

4. $\langle f, g \rangle = (\max_{0 \le x \le 1} f(x))(\max_{0 \le x \le 1} g(x))$ for f, g in $\mathcal{C}[0, 1]$

In Questions 5 and 6, compute the indicated quantity using the specified inner product.

5. $\|1 + x + x^2\|$ if $\langle a_0 + a_1x + a_2x^2, b_0 + b_1x + b_2x^2 \rangle = a_0b_0 + a_1b_1 + a_2b_2$

6. $d(x, x^2)$ if $\langle p(x), q(x) \rangle = \int_0^1 p(x)q(x)\,dx$

In Questions 7 and 8, construct an orthogonal set of vectors by applying the Gram-Schmidt Process to the given set of vectors using the specified inner product.

7. $\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$ if $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T A\mathbf{v}$, where $A = \begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix}$

8. $\{1, x, x^2\}$ if $\langle p(x), q(x) \rangle = \int_0^1 p(x)q(x)\,dx$

In Questions 9 and 10, determine whether the definition gives a norm.

9. $\|\mathbf{v}\| = \mathbf{v}^T\mathbf{v}$ for $\mathbf{v}$ in $\mathbb{R}^n$

10. $\|p(x)\| = |p(0)| + |p(1) - p(0)|$ for $p(x)$ in $\mathcal{P}_1$

11. Show that the matrix $A = \begin{bmatrix} 1 & 0.1 & 0.11 \\ 0.1 & 0.11 & 0.111 \\ 0.11 & 0.111 & 0.1111 \end{bmatrix}$ is ill-conditioned.

12. Prove that if Q is an orthogonal $n \times n$ matrix, then its Frobenius norm is $\|Q\|_F = \sqrt{n}$.

13. Find the line of best fit through the points (1, 2), (2, 3), (3, 5), and (4, 7).

14. Find the least squares solution of

$$\begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 2 & -1 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 3 \end{bmatrix}.$$

15. Find the orthogonal projection of $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ onto the column space of $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$.

16. If $\mathbf{u}$ and $\mathbf{v}$ are orthonormal vectors, show that $P = \mathbf{u}\mathbf{u}^T + \mathbf{v}\mathbf{v}^T$ is the standard matrix of an orthogonal projection onto span $(\mathbf{u}, \mathbf{v})$. [*Hint:* Show that $P = A(A^TA)^{-1}A^T$ for some matrix A.]

In Questions 17 and 18, find (a) the singular values, (b) a singular value decomposition, and (c) the pseudoinverse of the matrix A.

17. $A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}$ **18.** $A = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \end{bmatrix}$

19. If P and Q are orthogonal matrices for which PAQ is defined, prove that PAQ has the same singular values as A.

20. If A is a square matrix for which $A^2 = O$, prove that $(A^+)^2 = O$.

Appendix A*

Mathematical Notation and Methods of Proof

Please, sir, I want some more.
—Oliver
Charles Dickens, *Oliver Twist*

Anyone who understands algebraic notation reads at a glance in an equation results reached arithmetically only with great labour and pains.
—Augustin Cournot
Researches into the Mathematical Principles of the Theory of Wealth
Translated by Nathaniel T. Bacon
Macmillan, 1897, p. 4

In this book, an effort has been made to use "mathematical English" as much as possible, keeping mathematical notation to a minimum. However, mathematical notation is a convenient shorthand that can greatly simplify the amount of writing we have to do. Moreover, it is commonly used in every branch of mathematics, so the ability to read and write mathematical notation is an essential ingredient of mathematical understanding. Finally, there are some theorems whose proofs become "obvious" if the right notation is used.

Proving theorems in mathematics is as much an art as a science. For the beginner, it is often hard to know what approach to use in proving a theorem; there are many approaches, any one of which might turn out to be the best. To become proficient at proofs, it is important to study as many examples as possible and to get plenty of practice.

This appendix summarizes basic mathematical notation applied to sets. Summation notation, a useful shorthand for dealing with sums, is also discussed. Finally, some approaches to proofs are illustrated with generic examples.

Set Notation

A ***set*** is a collection of objects, called the ***elements*** (or ***members***) of the set. Examples of sets include the set of all words in this text, the set of all books in your college library, the set of positive integers, and the set of all vectors in the plane whose equation is $2x + 3y - z = 0$.

It is often possible to list the elements of a set, in which case it is conventional to enclose the list within braces. For example, we have

$$\{1, 2, 3\}, \quad \{a, t, x, z\}, \quad \{2, 4, 6, \ldots, 100\}, \quad \left\{\frac{\pi}{4}, \frac{2\pi}{5}, \frac{\pi}{2}, \frac{4\pi}{7}, \ldots, \frac{5\pi}{6}\right\}$$

Note that ellipses (. . .) denote elements omitted when a pattern is present. (What is the pattern in the last two examples?) Infinite sets are often expressed using ellipses. For example, the set of positive integers is usually denoted by $\mathbb{N}$ or $\mathbb{Z}^+$, so

$$\mathbb{N} = \mathbb{Z}^+ = \{1, 2, 3, \ldots\}$$

The set of all integers is denoted by $\mathbb{Z}$, so

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

Two sets are considered to be ***equal*** if they contain exactly the same elements. The *order* in which elements are listed does not matter, and repetitions are not counted. Thus,

*Exercises and selected odd-numbered answers for this appendix can be found on the student companion website.

$$\{1, 2, 3\} = \{2, 1, 3\} = \{1, 3, 2, 1\}$$

The symbol $\in$ means "is an element of" or "is in," and the symbol $\notin$ denotes the negation—that is, "is not an element of" or "is not in." For example,

$$5 \in \mathbb{Z}^+ \quad \text{but} \quad 0 \notin \mathbb{Z}^+$$

It is often more convenient to describe a set in terms of a rule satisfied by all of its elements. In such cases, ***set builder notation*** is appropriate. The format is

$$\{x : x \text{ satisfies } P\}$$

where P represents a property or a collection of properties that the element x must satisfy. The colon is pronounced "such that." For example,

$$\{n : n \in \mathbb{Z}, n > 0\}$$

is read as "the set of all n such that n is an integer and n is greater than zero." This is just another way of describing the positive integers $\mathbb{Z}^+$. (We could also write $\mathbb{Z}^+ = \{n \in \mathbb{Z} : n > 0\}$.)

The ***empty set*** is the set with no elements. It is denoted by either $\varnothing$ or $\{\ \}$.

Example A.1

Describe in words the following sets:

(a) $A = \{n : n = 2k, k \in \mathbb{Z}\}$ (b) $B = \{m/n : m, n \in \mathbb{Z}, n \neq 0\}$
(c) $C = \{x \in \mathbb{R} : 4x^2 - 4x - 3 = 0\}$ (d) $D = \{x \in \mathbb{Z} : 4x^2 - 4x - 3 = 0\}$

Solution (a) A is the set of numbers n that are integer multiples of 2. Therefore, A is the set of all even integers.

(b) B is the set of all expressions of the form m/n, where m and n are integers and n is nonzero. This is the set of *rational numbers,* usually denoted by $\mathbb{Q}$. (Note that this way of describing $\mathbb{Q}$ produces many repetitions; however, our convention, as noted above, is that we include only one occurrence of each element. Thus, this expression precisely describes the set of all rational numbers.)

(c) C is the set of all real solutions of the equation $4x^2 - 4x - 3 = 0$. By factoring or using the quadratic formula, we find that the roots of this equation are $-\frac{1}{2}$ and $\frac{3}{2}$. (Verify this.) Therefore,

$$C = \{-\tfrac{1}{2}, \tfrac{3}{2}\}$$

(d) From the solution to (c) we see that there are *no* solutions to $4x^2 - 4x - 3 = 0$ in $\mathbb{R}$ that are integers. Therefore, D is the empty set, which we can express by writing $D = \varnothing$.

John Venn (1834–1923) was an English mathematician who studied at Cambridge University and later lectured there. He worked primarily in mathematical logic and is best known for inventing Venn diagrams.

If every element of a set A is also an element of a set B, then A is called a ***subset*** of B, denoted $A \subseteq B$. We can represent this situation schematically using a ***Venn diagram,*** as shown in Figure A.1. (The rectangle represents the *universal set,* a set large enough to contain all of the other sets in question—in this case, A and B.)

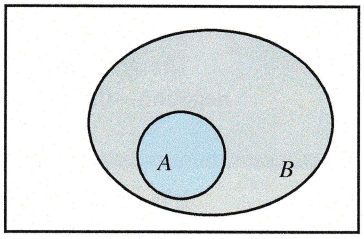


Figure A.1
$A \subseteq B$

Example A.2

(a) $\{1, 2, 3\} \subseteq \{1, 2, 3, 4, 5\}$

(b) $\mathbb{Z}^+ \subseteq \mathbb{Z} \subseteq \mathbb{R}$

(c) Let A be the set of all positive integers whose last two digits are 24 and let B be the set of all positive integers that are evenly divisible by 4. Then if n is in A, it is of the form

$$n = 100k + 24$$

for some integer k. (For example, $36{,}524 = 100 \cdot 365 + 24$.) But then

$$n = 100k + 24 = 4(25k + 6)$$

so $n/4 = 25k + 6$, which is an integer. Hence, n is evenly divisible by 4, so it is in B. Therefore, $A \subseteq B$.

We can show that two sets A and B are equal by showing that each is a subset of the other. This strategy is particularly useful if the sets are defined abstractly or if it is not easy to list and compare their elements.

Example A.3

Let A be the set of all positive integers whose last two digits form a number that is evenly divisible by 4. In the case of a one-digit number, we take its tens digit to be 0. Let B be the set of all positive integers that are evenly divisible by 4. Show that $A = B$.

Solution As in Example A.2(c), it is easy to see that $A \subseteq B$. If n is in A, then we can split off the number m formed by its last two digits by writing

$$n = 100k + m$$

for some integer k. But, since m is divisible by 4, we have $m = 4r$ for some integer r. Therefore,

$$n = 100k + m = 100k + 4r = 4(25k + r)$$

so n is also evenly divisible by 4. Hence, $A \subseteq B$.

To show that $B \subseteq A$, let n be in B. That is, n is evenly divisible by 4. Let's say that $n = 4s$, where s is an integer. If m is the number formed by the last two digits of n, then, as above, $n = 100k + m$ for some integer k. But now

$$m = n - 100k = 4s - 100k = 4(s - 25k)$$

which implies that m is evenly divisible by 4, since $s - 25k$ is an integer. Therefore, n is in A, and we have shown that $B \subseteq A$.

Since $A \subseteq B$ and $B \subseteq A$, we must have $A = B$.

The ***intersection*** of sets A and B is denoted by $A \cap B$ and consists of the elements that A and B have in common. That is,

$$A \cap B = \{x : x \in A \quad \text{and} \quad x \in B\}$$

Figure A.2 shows a Venn diagram of this case. The ***union*** of A and B is denoted by $A \cup B$ and consists of the elements that are in either A or B (or both). That is,

$$A \cup B = \{x : x \in A \quad \text{or} \quad x \in B\}$$

See Figure A.3.

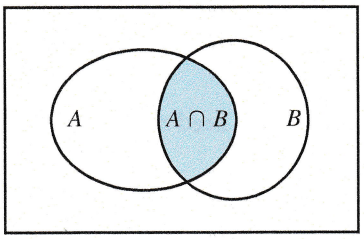


Figure A.2
$A \cap B$

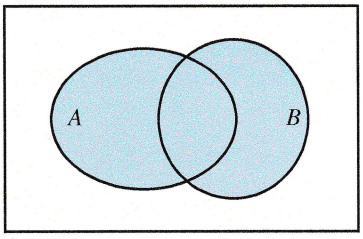


Figure A.3
$A \cup B$

Example A.4

Let $A = \{n^2 : n \in \mathbb{Z}^+, 1 \leq n \leq 4\}$ and let $B = \{n \in \mathbb{Z}^+ : n \leq 10 \text{ and } n \text{ is odd}\}$. Find $A \cap B$ and $A \cup B$.

Solution We see that

$$A = \{1^2, 2^2, 3^2, 4^2\} = \{1, 4, 9, 16\} \quad \text{and} \quad B = \{1, 3, 5, 7, 9\}$$

Therefore, $A \cap B = \{1, 9\}$ and $A \cup B = \{1, 3, 4, 5, 7, 9, 16\}$.

If $A \cap B = \varnothing$, then A and B are called ***disjoint sets.*** (See Figure A.4.) For example, the set of even integers and the set of odd integers are disjoint.

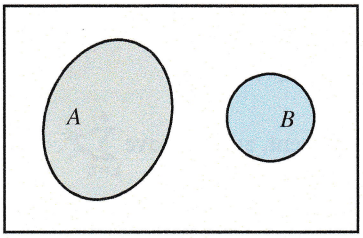


Figure A.4
Disjoint sets

Summation Notation

Summation notation is a convenient shorthand to use to write out a sum such as

$$1 + 2 + 3 + \cdots + 100$$

where we want to leave out all but a few terms. As in set notation, ellipses (. . .) convey that we have established a pattern and have simply left out some intermediate terms. In the above example, readers are expected to recognize that we are summing all of the positive integers from 1 to 100. However, ellipses can be ambiguous. For example, what would one make of the following sum?

$$1 + 2 + \cdots + 64$$

Is this the sum of all positive integers from 1 to 64 or just the powers of two, $1 + 2 + 4 + 8 + 16 + 32 + 64$? It is often clearer (and shorter) to use ***summation notation*** (or ***sigma notation***).

Σ is the capital Greek letter *sigma*, corresponding to S (for "sum"). Summation notation was introduced by Fourier in 1820 and was quickly adopted by the mathematical community.

We can abbreviate a sum of the form

$$a_1 + a_2 + \cdots + a_n \tag{1}$$

as

$$\sum_{k=1}^{n} a_k \tag{2}$$

which tells us to sum the terms a_k over all integers k ranging from 1 to n. An alternative version of this expression is

$$\sum_{1 \le k \le n} a_k$$

The subscript k is called the ***index of summation.*** It is a "dummy variable" in the sense that it does not appear in the actual sum in expression (1). Therefore, we can use any letter we like as the index of summation (as long as it doesn't already appear somewhere else in the expressions we are summing). Thus, expression (2) can also be written as

$$\sum_{i=1}^{n} a_i$$

The index of summation need not start at 1. The sum $a_3 + a_4 + \cdots + a_{99}$ becomes

$$\sum_{k=3}^{99} a_k$$

although we can arrange for the index to begin at 1 by rewriting the expression as

$$\sum_{k=1}^{97} a_{k+2}.$$

The key to using summation notation effectively is being able to recognize patterns.

Example A.5 Write the following sums using summation notation.

(a) $1 + 2 + 4 + \cdots + 64$ (b) $1 + 3 + 5 + \cdots + 99$ (c) $3 + 8 + 15 + \cdots + 99$

Solution (a) We recognize this expression as a sum of powers of 2:

$$1 + 2 + 4 + \cdots + 64 = 2^0 + 2^1 + 2^2 + \cdots + 2^6$$

Therefore, the index of summation appears as the exponent, and we have $\sum_{k=0}^{6} 2^k$.

(b) This expression is the sum of all the odd integers from 1 to 99. Every odd integer is of the form $2k + 1$, so the sum is

$$\begin{aligned} &1 + 3 + 5 + \cdots + 99 \\ &= (2 \cdot 0 + 1) + (2 \cdot 1 + 1) + (2 \cdot 2 + 1) + \cdots + (2 \cdot 49 + 1) \\ &= \sum_{k=0}^{49} (2k + 1) \end{aligned}$$

(c) The pattern here is less clear, but a little reflection reveals that each term is 1 less than a perfect square:

$$\begin{aligned} &3 + 8 + 15 + \cdots + 99 \\ &= (2^2 - 1) + (3^2 - 1) + (4^2 - 1) + \cdots + (10^2 - 1) \\ &= \sum_{k=2}^{10} (k^2 - 1) \end{aligned}$$



Example A.6 Rewrite each of the sums in Example A.5 so that the index of summation starts at 1.

Solution (a) If we use the change of variable $i = k + 1$, then, as k goes from 0 to 6, i goes from 1 to 7. Since $k = i - 1$, we obtain

$$\sum_{k=0}^{6} 2^k = \sum_{i=1}^{7} 2^{i-1}$$

(b) Using the same substitution as in part (a), we get

$$\sum_{k=0}^{49} (2k + 1) = \sum_{i=1}^{50} (2(i - 1) + 1) = \sum_{i=1}^{50} (2i - 1)$$

(c) The substitution $i = k - 2$ will work (try it), but it is easier just to add a term corresponding to $k = 1$, since $1^2 - 1 = 0$. Therefore,

$$\sum_{k=2}^{10} (k^2 - 1) = \sum_{k=1}^{10} (k^2 - 1)$$

Multiple summations arise when there is more than one index of summation, as there is with a matrix. The notation

$$\sum_{i,j=1}^{n} a_{ij} \tag{3}$$

means to sum the terms a_{ij} as i and j each range independently from 1 to n. The sum in expression (3) is equivalent to either

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$$

where we sum first over j and then over i (we always work from the inside out), or

$$\sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij}$$

where the order of summation is reversed.

Example A.7 Write out $\sum_{i,j=1}^{3} i^j$ using both possible orders of summation.

Solution

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

$$= (1^1 + 1^2 + 1^3) + (2^1 + 2^2 + 2^3) + (3^1 + 3^2 + 3^3)$$

$$= (1 + 1 + 1) + (2 + 4 + 8) + (3 + 9 + 27) = 56$$

and

$$\sum_{j=1}^{3}\sum_{i=1}^{3} i^j = \sum_{j=1}^{n}(1^j + 2^j + 3^j)$$
$$= (1^1 + 2^1 + 3^1) + (1^2 + 2^2 + 3^2) + (1^3 + 2^3 + 3^3)$$
$$= (1 + 2 + 3) + (1 + 4 + 9) + (1 + 8 + 27) = 56$$

Remark Of course, the value of the sum in Example A.7 is the same no matter which order of summation we choose, because the sum is *finite*. It is also possible to consider *infinite sums* (known as *infinite series* in calculus), but such sums do not always have a value and great care must be taken when rearranging or manipulating their terms. For example, suppose we let

$$S = \sum_{k=0}^{\infty} 2^k$$

Then
$$S = 1 + 2 + 4 + 8 + \cdots$$
$$= 1 + 2(1 + 2 + 4 + \cdots)$$
$$= 1 + 2S$$

from which it follows that $S = -1$. This is clearly nonsense, since S is a sum of *non-negative* terms! (Where is the error?)

How to Solve It is the title of a book by the mathematician George Pólya (1887–1985). Since its publication in 1945, *How to Solve It* has sold over a million copies and has been translated into 17 languages. Pólya was born in Hungary, but because of the political situation in Europe, he moved to the United States in 1940. He subsequently taught at Brown and Stanford Universities, where he did mathematical research and developed a well-deserved reputation as an outstanding teacher. The Pólya Prize is awarded annually by the Society for Industrial and Applied Mathematics for major contributions to areas of mathematics close to those on which Pólya worked. The Mathematical Association of America annually awards Pólya Lectureships to mathematicians demonstrating the high-quality exposition for which Pólya was known.

Methods of Proof

The notion of proof is at the very heart of mathematics. It is one thing to know *what* is true; it is quite another to know *why* it is true and to be able to demonstrate its truth by means of a logically connected sequence of statements. The intention here is not to try to teach you how to do proofs; you will become better at doing proofs by studying examples and by practicing—something you should do often as you work through this text. The intention of this brief section is simply to provide a few elementary examples of some types of proofs. The proofs of theorems in the text will provide further illustrations of "how to solve it."

Roughly speaking, mathematical proofs fall into two categories: ***direct proofs*** and ***indirect proofs.*** Many theorems have the structure "if P, then Q," where P (the *hypothesis,* or *premise*) and Q (the *conclusion*) are statements that are either true or false. We denote such an implication by $P \Rightarrow Q$. A direct proof proceeds by establishing a chain of implications

$$P \Rightarrow P_1 \Rightarrow P_2 \Rightarrow \cdots \Rightarrow P_n \Rightarrow Q$$

leading directly from P to Q.

Example A.8

Prove that any two consecutive perfect squares differ by an odd number. This instruction can be rephrased as "Prove that if a and b are consecutive perfect squares, then $a - b$ is odd." Hence, it has the form $P \Rightarrow Q$, with P being "a and b are consecutive perfect squares" and Q being "$a - b$ is odd."

Solution Assume that a and b are consecutive perfect squares, with $a > b$. Then

$$a = (n + 1)^2 \quad \text{and} \quad b = n^2$$

for some integer n. But now

$$\begin{aligned} a - b &= (n + 1)^2 - n^2 \\ &= n^2 + 2n + 1 - n^2 \\ &= 2n + 1 \end{aligned}$$

so $a - b$ is odd.

There are two types of indirect proofs that can be used to establish a conditional statement of the form $P \Rightarrow Q$. A ***proof by contradiction*** assumes that the hypothesis P is true, just as in a direct proof, but then supposes that the conclusion Q is *false*. The strategy then is to show that this is not possible (i.e., to rule out the possibility that the conclusion is false) by finding a contradiction to the truth of P. It then follows that Q must be true.

Example A.9

Let n be a positive integer. Prove that if n^2 is even, so is n. (Take a few minutes to try to find a direct proof of this assertion; it will help you to appreciate the indirect proof that follows.)

Solution Assume that n is a positive integer such that n^2 is even. Now suppose that n is not even. Then n is odd, so

$$n = 2k + 1$$

for some integer k. But if so, we have

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1$$

so n^2 is odd, since it is 1 more than the even number $4k^2 + 4k$. This contradicts our hypothesis that n^2 is even. We conclude that our supposition that n was *not* even must have been false; in other words, n must be even.

Closely related to the method of proof by contradiction is ***proof by contrapositive***. The *negative* of a statement P is the statement "it is not the case that P," abbreviated symbolically as $\neg P$ and pronounced "not P." For example, if P is "n is even," then $\neg P$ is "it is not the case that n is even"—in other words, "n is odd."

The *contrapositive* of the statement $P \Rightarrow Q$ is the statement $\neg Q \Rightarrow \neg P$. A conditional statement $P \Rightarrow Q$ and its contrapositive $\neg Q \Rightarrow \neg P$ are logically equivalent in the sense that they are either both true or both false. (For example, if $P \Rightarrow Q$ is a theorem, then so is $\neg Q \Rightarrow \neg P$. To see this, note that if the hypothesis $\neg Q$ is true, then Q is false. The conclusion $\neg P$ cannot be false, for if it were, then P would be true and our known theorem $P \Rightarrow Q$ would imply the truth of Q, giving us a contradiction. It follows that $\neg P$ is true and we have proved $\neg Q \Rightarrow \neg P$.) Here is a contrapositive proof of the assertion in Example A.9.

Example A.10

Let n be a positive integer. Prove that if n^2 is even, so is n.

Solution The contrapositive of the given statement is

"If n is not even, then n^2 is not even" or "If n is odd, so is n^2"

To prove this contrapositive, assume that n is odd. Then $n = 2k + 1$ for some integer k. As before, this means that $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1$ is odd, which completes the proof of the contrapositive. Since the contrapositive is true, so is the original statement.

Although we do not require a new method of proof to handle it, we will briefly consider how to prove an "if and only if" theorem. A statement of the form "P if and only if Q" signals a *double implication,* which we denote by $P \Leftrightarrow Q$. To prove such a statement, we must prove $P \Rightarrow Q$ and $Q \Rightarrow P$. To do so, we can use the techniques described earlier, where appropriate. It is important to notice that the "*if*" part of $P \Leftrightarrow Q$ is "P if Q," which is $Q \Rightarrow P$; the "only if" part of $P \Leftrightarrow Q$ is "P only if Q," meaning $P \Rightarrow Q$. The implication $P \Rightarrow Q$ is sometimes read as "P is sufficient for Q" or "Q is necessary for P"; $Q \Rightarrow P$ is read "Q is sufficient for P" or "P is necessary for Q." Taken together, they are $P \Leftrightarrow Q$, or "P is necessary and sufficient for Q" and vice versa.

Example A.11

A pawn is placed on a chessboard and is allowed to move one square at a time, either horizontally or vertically. A *pawn's tour* of a chessboard is a path taken by a pawn, moving as described, that visits each square exactly once, starting and ending on the same square. Prove that there is a pawn's tour of an $n \times n$ chessboard if and only if n is even.

Solution [$\Leftarrow$] ("if") Assume that n is even. It is easy to see that the strategy illustrated in Figure A.5 for a 6×6 chessboard will always give a pawn's tour.

[$\Rightarrow$] ("only if") Suppose that there is a pawn's tour of an $n \times n$ chessboard. We will give a proof by contradiction that n must be even. To this end, let's assume that n is odd. At each move, the pawn moves to a square of a different color. The total number of moves in its tour is n^2, which is also an odd number, according to the proof in Example A.10. Therefore, the pawn must end up on a square of the opposite color from that of the square on which it started. (Why?) This is impossible, since the pawn ends where it started, so we have a contradiction. It follows that n cannot be odd; hence, n is even and the proof is complete.

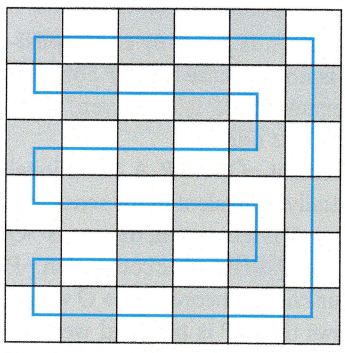


Figure A.5

Some theorems assert that several statements are *equivalent.* This means that each is true if and only if all of the others are true. Showing that n statements are equivalent requires $\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n^2 - n}{2}$ "if and only if" proofs. In practice, however, it is often easier to establish a "ring" of n implications that links all of the statements. The proof of the Fundamental Theorem of Invertible Matrices provides an excellent example of this approach.

Appendix B*

Mathematical Induction

The ability to spot patterns is one of the keys to success in mathematical problem solving. Consider the following pattern:

Great fleas have little fleas
upon their backs to bite 'em,
And little fleas have lesser fleas,
and so ad infinitum.
—Augustus De Morgan
A Budget of Paradoxes
Longmans, Green, and Company,
1872, p. 377

$$\begin{aligned} 1 &= 1 \\ 1 + 3 &= 4 \\ 1 + 3 + 5 &= 9 \\ 1 + 3 + 5 + 7 &= 16 \\ 1 + 3 + 5 + 7 + 9 &= 25 \end{aligned}$$

The sums are all perfect squares: $1^2, 2^2, 3^2, 4^2, 5^2$. It seems reasonable to conjecture that this pattern will continue to hold; that is, the sum of consecutive odd numbers, starting at 1, will always be a perfect square. Let's try to be more precise. If the sum is n^2, then the last odd number in the sum is $2n - 1$. (Check this in the five cases above.) In symbols, our conjecture becomes

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2 \quad \text{for all } n \geq 1 \tag{1}$$

Notice that Equation (1) is really an *infinite* collection of statements, one for each value of $n \geq 1$. Although our conjecture seems reasonable, we cannot assume that the pattern continues—we need to prove it. This is where ***mathematical induction*** comes in.

First Principle of Mathematical Induction

Let $S(n)$ be a statement about the positive integer n. If

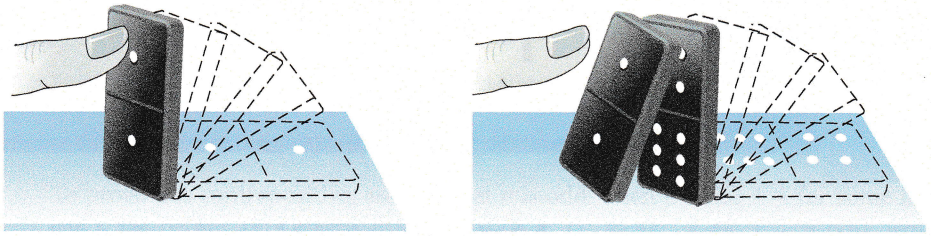
1. $S(1)$ is true and
2. for all $k \geq 1$, the truth of $S(k)$ implies the truth of $S(k + 1)$

then $S(n)$ is true for all $n \geq 1$.

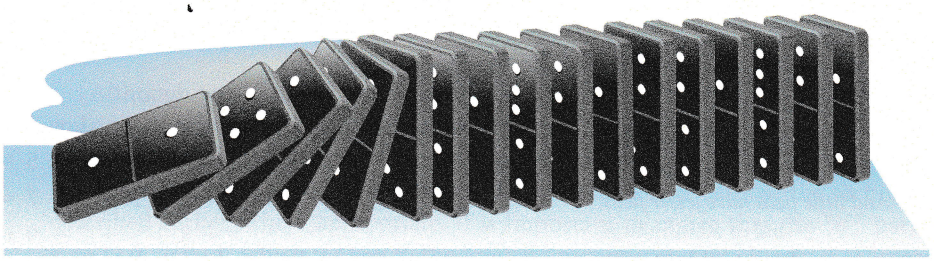
Verifying that $S(1)$ is true is called the ***basis step.*** The assumption that $S(k)$ is true for some $k \geq 1$ is called the ***induction hypothesis.*** Using the induction hypothesis to prove that $S(k + 1)$ is then true is called the ***induction step.*** Mathematical induction has been referred to as the *domino principle* because it is analogous to showing that a line of dominoes will fall down if (1) the first domino can be knocked down (the basis step) and (2) knocking down any domino (the induction hypothesis) will knock over the next domino (the induction step). See Figure B.1.

We now use the principle of mathematical induction to prove Equation (1).

*Exercises and selected odd-numbered answers for this appendix can be found on the student companion website.



If the first domino falls, and . . . each domino that falls knocks down the next one, . . .



then all the dominoes can be made to fall by pushing over the first one.

Figure B.1

Example B.1 Use mathematical induction to prove that

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2$$

for all $n \geq 1$.

Solution For $n = 1$, the sum on the left-hand side is just 1, while the right-hand side is 1^2. Since $1 = 1^2$, this completes the basis step.

Now assume that the formula is true for some integer $k \geq 1$. That is, assume that

$$1 + 3 + 5 + \cdots + (2k - 1) = k^2$$

(This is the induction hypothesis.) The induction step consists of proving that the formula is true when $n = k + 1$. We see that when $n = k + 1$, the left-hand side of formula (1) is

$$\begin{aligned} 1 + 3 + 5 + \cdots + (2(k+1) - 1) &= 1 + 3 + 5 + \cdots + (2k + 1) \\ &= \underbrace{1 + 3 + 5 + \cdots + (2k - 1)}_{k^2} + (2k + 1) \\ &= (k + 1)^2 \end{aligned}$$

$+ 2k + 1$ ← by the induction hypothesis

which is the right-hand side of Equation (1) when $n = k + 1$.

This completes the induction step, and we conclude that Equation (1) is true for all $n \geq 1$, by the principle of mathematical induction.

The next example gives a proof of a useful formula for the sum of the first n positive integers. The formula appears several times in the text; for example, see the solution to Exercise 51 in Section 2.4.

Example B.2 Prove that

$$1 + 2 + \cdots + n = \frac{n(n + 1)}{2}$$

for all $n \geq 1$.

Solution The formula is true for $n = 1$, since

$$1 = \frac{1(1 + 1)}{2}$$

Assume that the formula is true for $n = k$; that is,

$$1 + 2 + \cdots + k = \frac{k(k + 1)}{2}$$

We need to show that the formula is true when $n = k + 1$; that is, we must prove that

$$1 + 2 + \cdots + (k + 1) = \frac{(k + 1)[(k + 1) + 1]}{2}$$

But we see that

$$\begin{aligned}
1 + 2 + \cdots + (k + 1) &= (1 + 2 + \cdots + k) + (k + 1) \\
&= \frac{k(k + 1)}{2} + (k + 1) && \text{by the induction hypothesis} \\
&= \frac{k(k + 1) + 2(k + 1)}{2} \\
&= \frac{k^2 + 3k + 2}{2} \\
&= \frac{(k + 1)(k + 2)}{2} \\
&= \frac{(k + 1)[(k + 1) + 1]}{2}
\end{aligned}$$

which is what we needed to show.

This completes the induction step, and we conclude that the formula is true for all $n \geq 1$, by the principle of mathematical induction.

In a similar vein, we can prove that the sum of the squares of the first n positive integers satisfies the formula

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n + 1)(2n + 1)}{6}$$

for all $n \geq 1$. (Verify this for yourself.)

The basis step need not be for $n = 1$, as the next two examples illustrate.

Example B.3

Prove that $n! > 2^n$ for all integers $n \geq 4$.

Solution The basis step here is when $n = 4$. The inequality is clearly true in this case, since

$$4! = 24 > 16 = 2^4$$

Assume that $k! > 2^k$ for some integer $k \geq 4$. Then

$$\begin{aligned}(k+1)! &= (k+1)k! \\ &> (k+1)2^k && \text{by the induction hypothesis} \\ &\geq 5 \cdot 2^k && \text{since } k \geq 4 \\ &> 2 \cdot 2^k = 2^{k+1}\end{aligned}$$

which verifies the inequality for $n = k + 1$ and completes the induction step.

We conclude that $n! > 2^n$ for all integers $n \geq 4$, by the principle of mathematical induction.

If a is a nonzero real number and $n \geq 0$ is an integer, we can give a recursive definition of the power a^n that is compatible with mathematical induction. We define $a^0 = 1$ and, for $n \geq 0$,

$$a^{n+1} = a^n a$$

(This form avoids the ellipses used in the version $a^n = \overbrace{aa \cdots a}^{n \text{ times}}$.) We can now use mathematical induction to verify a familiar property of exponents.

Example B.4

Let a be a nonzero real number. Prove that $a^m a^n = a^{m+n}$ for all integers $m, n \geq 0$.

Solution At first glance, it is not clear how to proceed, since there are *two* variables, m and n. But we simply need to keep one of them fixed and perform our induction using the other. So, let $m \geq 0$ be a fixed integer. When $n = 0$, we have

$$a^m a^0 = a^m \cdot 1 = a^m = a^{m+0}$$

using the definition $a^0 = 1$. Hence, the basis step is true.

Now assume that the formula holds when $n = k$, where $k \geq 0$. Then $a^m a^k = a^{m+k}$. For $n = k + 1$, using our recursive definition and the fact that addition and multiplication are associative, we see that

$$\begin{aligned}a^m a^{k+1} &= a^m(a^k a) && \text{by definition} \\ &= (a^m a^k)a \\ &= a^{m+k}a && \text{by the induction hypothesis} \\ &= a^{(m+k)+1} && \text{by definition} \\ &= a^{m+(k+1)}\end{aligned}$$

Therefore, the formula is true for $n = k + 1$, and the induction step is complete.

We conclude that $a^m a^n = a^{m+n}$ for all integers m, $n \geq 0$, by the principle of mathematical induction.

In Examples B.1 through B.4, the use of the induction hypothesis during the induction step is relatively straightforward. However, this is not always the case. An alternative version of the principle of mathematical induction is often more useful.

Second Principle of Mathematical Induction

Let $S(n)$ be a statement about the positive integer n. If

1. $S(1)$ is true and
2. the truth of $S(1), S(2), \ldots, S(k)$ implies the truth of $S(k + 1)$

then $S(n)$ is true for all $n \geq 1$.

The only difference between the two principles of mathematical induction is in the induction hypothesis: The first version assumes that $S(k)$ is true, whereas the second version assumes that all of $S(1), S(2), \ldots, S(k)$ are true. This makes the second principle seem weaker than the first, since we need to assume more in order to prove $S(k + 1)$ (although, paradoxically, the second principle is sometimes called *strong* induction). In fact, however, the two principles are logically equivalent: Each one implies the other. (Can you see why?)

The next example presents an instance in which the second principle of mathematical induction is easier to use than the first. Recall that a prime number is a positive integer whose only positive integer factors are 1 and itself.

Example B.5

Prove that every positive integer $n \geq 2$ either is prime or can be factored into a product of primes.

Solution The result is clearly true when $n = 2$, since 2 is prime. Now assume that for all integers n between 2 and k, n either is prime or can be factored into a product of primes. Let $n = k + 1$. If $k + 1$ is prime, we are done. Otherwise, it must factor into a product of two smaller integers—say,

$$k + 1 = ab$$

Since $2 \leq a, b \leq k$ (why?), the induction hypothesis applies to a and b. Therefore,

$$a = p_1 \cdots p_r \quad \text{and} \quad b = q_1 \cdots q_s$$

where the p's and q's are all prime. Then

$$ab = p_1 \cdots p_r q_1 \cdots q_s$$

gives a factorization of ab into primes, completing the induction step.

We conclude that the result is true for all integers $n \geq 2$, by the second principle of mathematical induction.

Do you see why the first principle of mathematical induction would have been difficult to use here?

We conclude with a highly nontrivial example that involves a combination of induction and *backward* induction. The result is the Arithmetic Mean–Geometric Mean Inequality, discussed in Chapter 7 in Exploration: Geometric Inequalities and Optimization Problems. The clever proof in Example B.6 is due to Cauchy.

Example B.6

Let $x_1, \ldots, x_n$ be nonnegative real numbers. Prove that

$$\sqrt[n]{x_1 x_2 \cdots x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n}$$

for all integers $n \geq 2$.

Solution For $n = 2$, the inequality becomes $\sqrt{xy} \leq (x + y)/2$. You are asked to verify this in Problems 1 and 2 of the Exploration mentioned above.

If $S(n)$ is the stated inequality, we will prove that $S(k)$ implies $S(2k)$. Assume that $S(k)$ is true; that is,

$$\sqrt[k]{x_1 x_2 \cdots x_k} \leq \frac{x_1 + x_2 + \cdots + x_k}{k}$$

for all nonnegative real numbers $x_1, \ldots, x_k$. Let

$$x_1 = \frac{y_1 + y_2}{2}, \quad x_2 = \frac{y_3 + y_4}{2}, \quad \ldots, \quad x_k = \frac{y_{2k-1} + y_{2k}}{2}$$

Then

$$\begin{aligned}
\sqrt[2k]{y_1 \cdots y_{2k}} &= \sqrt[k]{\sqrt{y_1 \cdots y_{2k}}} = \sqrt[k]{\sqrt{y_1 y_2} \cdots \sqrt{y_{2k-1} y_{2k}}} \\
&\leq \sqrt[k]{\left(\frac{y_1 + y_2}{2}\right) \cdots \left(\frac{y_{2k-1} + y_{2k}}{2}\right)} && \text{by } S(2) \\
&= \sqrt[k]{x_1 \cdots x_k} \\
&\leq \frac{x_1 + x_2 + \cdots + x_k}{k} && \text{by } S(k) \\
&= \frac{\left(\frac{y_1 + y_2}{2}\right) + \cdots + \left(\frac{y_{2k-1} + y_{2k}}{2}\right)}{k} \\
&= \frac{y_1 + \cdots + y_{2k}}{2k}
\end{aligned}$$

which verifies $S(2k)$.

Thus, the Arithmetic Mean–Geometric Mean Inequality is true for $n = 2, 4, 8, \ldots$—the powers of 2. In order to complete the proof, we need to "fill in the gaps." We will use backward induction to prove that $S(k)$ implies $S(k - 1)$. Assuming $S(k)$ is true, let

$$x_k = \frac{x_1 + x_2 + \cdots + x_{k-1}}{k - 1}.$$

Then

$$\sqrt[k]{x_1x_2\cdots x_{k-1}\left(\frac{x_1+x_2+\cdots+x_{k-1}}{k-1}\right)} \le \frac{x_1+x_2+\cdots+\left(\frac{x_1+x_2+\cdots+x_{k-1}}{k-1}\right)}{k}$$

$$= \frac{kx_1+kx_2+\cdots+kx_{k-1}}{k(k-1)}$$

$$= \frac{x_1+x_2+\cdots+x_{k-1}}{k-1}$$

Equivalently,

$$x_1x_2\cdots x_{k-1}\left(\frac{x_1+x_2+\cdots+x_{k-1}}{k-1}\right) \le \left(\frac{x_1+x_2+\cdots+x_{k-1}}{k-1}\right)^k$$

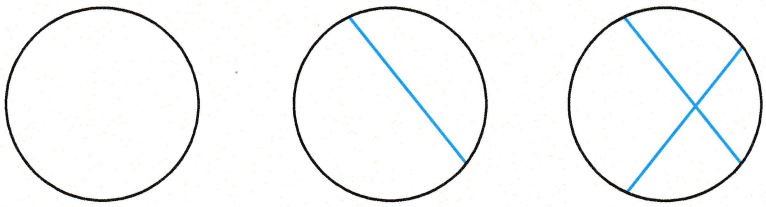
or

$$x_1x_2\cdots x_{k-1} \le \left(\frac{x_1+x_2+\cdots+x_{k-1}}{k-1}\right)^{k-1}$$

Taking the $(k-1)$th root of both sides yields $S(k-1)$.

The two inductions, taken together, show that the Arithmetic Mean–Geometric Mean Inequality is true for all $n \ge 2$.

Remark Although mathematical induction is a powerful and indispensable tool, it cannot work miracles. That is, it cannot prove that a pattern or formula holds if it does not. Consider the diagrams in Figure B.2, which show the maximum number of regions $R(n)$ into which a circle can be subdivided by n straight lines.



$R(0) = 1 = 2^0$ $R(1) = 2 = 2^1$ $R(2) = 4 = 2^2$

Figure B.2

Based on the evidence in Figure B.2, we might conjecture that $R(n) = 2^n$ for $n \ge 0$ and try to prove this conjecture using mathematical induction. We would not succeed, since this formula is not correct! If we had considered one more case, we would have discovered that $R(3) = 7 \ne 8 = 2^3$, thereby demolishing our conjecture. In fact, the correct formula turns out to be

$$R(n) = \frac{n^2+n+2}{2}$$

which *can* be verified by induction. (Can you do it?)

For other examples in which a pattern appears to be true, only to disappear when enough cases are considered, see Richard K. Guy's delightful article "The Strong Law of Small Numbers" in the *American Mathematical Monthly*, Vol. 95 (1988), pp. 697–712.

Appendix C*

Complex Numbers

A ***complex number*** is a number of the form $a + bi$, where a and b are real numbers and i is a symbol with the property that $i^2 = -1$. The real number a is considered to be a special type of complex number, since $a = a + 0i$. If $z = a + bi$ is a complex number, then the ***real part*** of z, denoted by Re z, is a, and the ***imaginary part*** of z, denoted by Im z, is b. Two complex numbers $a + bi$ and $c + di$ are ***equal*** if their real parts are equal and their imaginary parts are equal—that is, if $a = c$ and $b = d$. A complex number $a + bi$ can be identified with the point (a, b) and plotted in the plane (called the ***complex plane,*** or the ***Argand plane***), as shown in Figure C.1. In the complex plane, the horizontal axis is called the ***real axis*** and the vertical axis is called the ***imaginary axis.***

[The] extension of the number concept to include the irrational, and we will at once add, the imaginary, is the greatest forward step which pure mathematics has ever taken.
—Hermann Hankel
Theorie der Complexen Zahlensysteme
Leipzig, 1867, p. 60

There is nothing "imaginary" about complex numbers—they are just as "real" as the real numbers. The term *imaginary* arose from the study of polynomial equations such as $x^2 + 1 = 0$, whose solutions are not "real" (i.e., real numbers). It is worth remembering that at one time negative numbers were thought of as "imaginary" too.

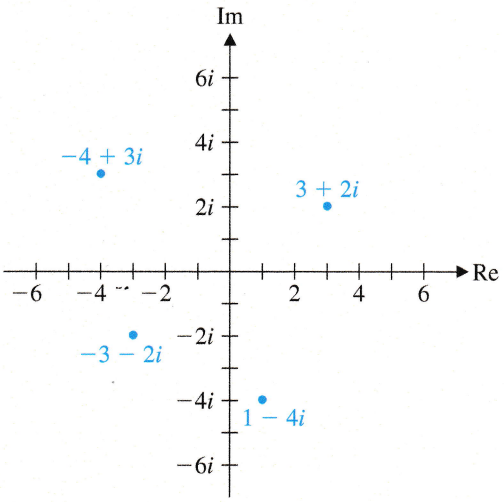


Figure C.1
The complex plane

Operations on Complex Numbers

The ***sum*** of the complex numbers $a + bi$ and $c + di$ is defined as

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

Notice that, with the identification of $a + bi$ with (a, b), $c + di$ with (c, d), and $(a + c) + (b + d)i$ with $(a + c, b + d)$, addition of complex numbers is the same as vector addition. The ***product*** of $a + bi$ and $c + di$ is

$$\begin{aligned}(a + bi)(c + di) &= a(c + di) + bi(c + di)\\ &= ac + adi + bci + bdi^2\end{aligned}$$

Jean-Robert Argand (1768–1822) was a French accountant and amateur mathematician. His geometric interpretation of complex numbers appeared in 1806 in a book that he published privately. He was not, however, the first to give such an interpretation. The Norwegian-Danish surveyor Caspar Wessel (1745–1818) gave the same version of the complex plane in 1787, but his paper was not noticed by the mathematical community until after his death.

*Exercises and selected odd-numbered answers for this appendix can be found on the student companion website.

Since $i^2 = -1$, this expression simplifies to $(ac - bd) + (ad + bc)i$. Thus, we have

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i$$

Observe that, as a special case, $a(c + di) = ac + adi$, so the ***negative*** of $c + di$ is $-(c + di) = (-1)(c + di) = -c - di$. This fact allows us to compute the ***difference*** of $a + bi$ and $c + di$ as

$$\begin{aligned}(a + bi) - (c + di) &= (a + bi) + (-1)(c + di)\\ &= (a + (-c)) + (b + (-d))i\\ &= (a - c) + (b - d)i\end{aligned}$$

Example C.1

Find the sum, difference, and product of $3 - 4i$ and $-1 + 2i$.

Solution The sum is

$$(3 - 4i) + (-1 + 2i) = (3 - 1) + (-4 + 2)i = 2 - 2i$$

The difference is

$$(3 - 4i) - (-1 + 2i) = (3 - (-1)) + (-4 - 2)i = 4 - 6i$$

The product is

$$\begin{aligned}(3 - 4i)(-1 + 2i) &= -3 + 6i + 4i - 8i^2\\ &= -3 + 10i - 8(-1) = 5 + 10i\end{aligned}$$

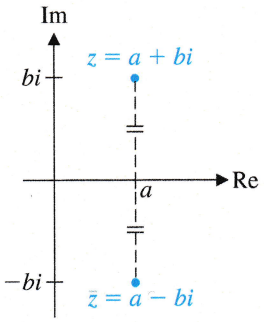


Figure C.2
Complex conjugates

The ***conjugate*** of $z = a + bi$ is the complex number

$$\bar{z} = a - bi$$

($\bar{z}$ is pronounced "z bar.") Figure C.2 gives the geometric interpretation of the conjugate.

To find the quotient of two complex numbers, we multiply the numerator and the denominator by the conjugate of the denominator.

Example C.2

Express $\dfrac{-1 + 2i}{3 + 4i}$ in the form $a + bi$.

Solution We multiply the numerator and denominator by $\overline{3 + 4i} = 3 - 4i$. Using Example C.1, we obtain

$$\frac{-1 + 2i}{3 + 4i} = \frac{-1 + 2i}{3 + 4i} \cdot \frac{3 - 4i}{3 - 4i} = \frac{5 + 10i}{3^2 + 4^2} = \frac{5 + 10i}{25} = \frac{1}{5} + \frac{2}{5}i$$

On the following page is a summary of some of the properties of conjugates. The proofs follow from the definition of conjugate; you should verify them for yourself.

1. $\overline{\overline{z}} = z$
2. $\overline{z + w} = \overline{z} + \overline{w}$
3. $\overline{zw} = \overline{z}\,\overline{w}$
4. If $z \neq 0$, then $\overline{(w/z)} = \overline{w}/\overline{z}$.
5. z is real if and only if $\overline{z} = z$.

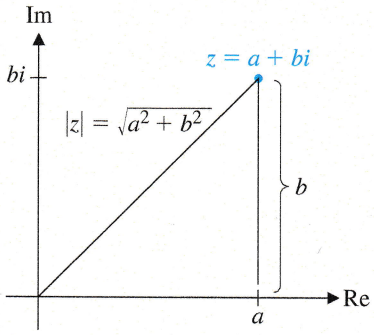


Figure C.3

The ***absolute value*** (or ***modulus***) $|z|$ of a complex number $z = a + bi$ is its distance from the origin. As Figure C.3 shows, Pythagoras' Theorem gives

$$|z| = |a + bi| = \sqrt{a^2 + b^2}$$

Observe that

$$z\overline{z} = (a + bi)(a - bi) = a^2 - abi + bai - b^2i^2 = a^2 + b^2$$

Hence,

$$z\overline{z} = |z|^2$$

This gives us an alternative way of describing the division process for the quotient of two complex numbers. If w and $z \neq 0$ are two complex numbers, then

$$\frac{w}{z} = \frac{w}{z} \cdot \frac{\overline{z}}{\overline{z}} = \frac{w\overline{z}}{z\overline{z}} = \frac{w\overline{z}}{|z|^2}$$

Below is a summary of some of the properties of absolute value. (You should try to prove these using the definition of absolute value and other properties of complex numbers.)

1. $|z| = 0$ if and only if $z = 0$.
2. $|z| = |\overline{z}|$
3. $|zw| = |z||w|$
4. If $z \neq 0$, then $\left|\frac{1}{z}\right| = \frac{1}{|z|}$.
5. $|z + w| \leq |z| + |w|$

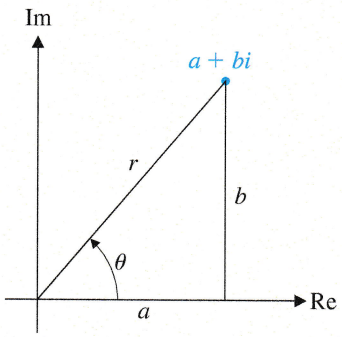


Figure C.4

Polar Form

As you have seen, the complex number $z = a + bi$ can be represented geometrically by the point (a, b). This point can also be expressed in terms of ***polar coordinates*** (r, θ), where $r \geq 0$, as shown in Figure C.4. We have

$$a = r\cos\theta \quad \text{and} \quad b = r\sin\theta$$

so

$$z = a + bi = r\cos\theta + (r\sin\theta)i$$

Thus, any complex number can be written in the ***polar form***

$$z = r(\cos\theta + i\sin\theta)$$

where $r = |z| = \sqrt{a^2 + b^2}$ and $\tan\theta = b/a$. The angle θ is called an ***argument*** of z and is denoted by $\arg z$. Observe that $\arg z$ is not unique: Adding or subtracting any integer multiple of 2π gives another argument of z. However, there is only one argument θ that satisfies

$$-\pi < \theta \leq \pi$$

This is called the ***principal argument*** of z and is denoted by $\text{Arg}\, z$.

Example C.3

Write the following complex numbers in polar form using their principal arguments:

(a) $z = 1 + i$ (b) $w = 1 - \sqrt{3}i$

Solution (a) We compute

$$r = |z| = \sqrt{1^2 + 1^2} = \sqrt{2} \quad \text{and} \quad \tan\theta = \frac{1}{1} = 1$$

Therefore, $\text{Arg}\, z = \theta = \dfrac{\pi}{4}$ $(= 45°)$, and we have

$$z = \sqrt{2}\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right)$$

as shown in Figure C.5.

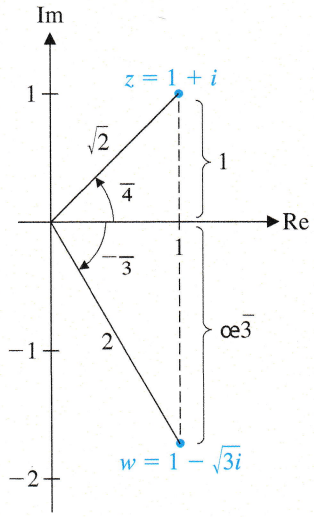


Figure C.5

(b) We have

$$r = |w| = \sqrt{1^2 + (-\sqrt{3})^2} = \sqrt{4} = 2 \quad \text{and} \quad \tan\theta = \frac{-\sqrt{3}}{1} = -\sqrt{3}$$

Since w lies in the fourth quadrant, we must have $\text{Arg}\, z = \theta = -\dfrac{\pi}{3}$ $(= -60°)$. Therefore,

$$w = 2\left(\cos\left(-\frac{\pi}{3}\right) + i\sin\left(-\frac{\pi}{3}\right)\right)$$

See Figure C.5.

The polar form of complex numbers can be used to give geometric interpretations of multiplication and division. Let

$$z_1 = r_1(\cos\theta_1 + i\sin\theta_1) \quad \text{and} \quad z_2 = r_2(\cos\theta_2 + i\sin\theta_2)$$

Multiplying, we obtain

$$\begin{aligned} z_1 z_2 &= r_1 r_2(\cos\theta_1 + i\sin\theta_1)(\cos\theta_2 + i\sin\theta_2) \\ &= r_1 r_2[(\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2)] \end{aligned}$$

Using the trigonometric identities

$$\begin{aligned} \cos(\theta_1 + \theta_2) &= \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 \\ \sin(\theta_1 + \theta_2) &= \sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2 \end{aligned}$$

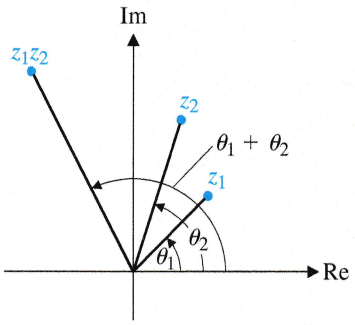


Figure C.6

we obtain

$$z_1z_2 = r_1r_2[\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)] \tag{1}$$

which is the polar form of a complex number with absolute value r_1r_2 and argument $\theta_1 + \theta_2$. This shows that

$$|z_1z_2| = |z_1||z_2| \quad \text{and} \quad \arg(z_1z_2) = \arg z_1 + \arg z_2$$

Equation (1) says that *to multiply two complex numbers, we multiply their absolute values and add their arguments.* See Figure C.6.

Similarly, using the subtraction identities for sine and cosine, we can show that

$$\frac{z_1}{z_2} = \frac{r_1}{r_2}[\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2)] \quad \text{if } z \neq 0$$

(Verify this.) Therefore,

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|} \quad \text{and} \quad \arg\left(\frac{z_1}{z_2}\right) = \arg z_1 - \arg z_2$$

and we see that *to divide two complex numbers, we divide their absolute values and subtract their arguments.*

As a special case of the last result, we obtain a formula for the reciprocal of a complex number in polar form. Setting $z_1 = 1$ (and therefore $\theta_1 = 0$) and $z_2 = z$ (and therefore $\theta_2 = \theta$), we obtain the following:

If $z = r(\cos\theta + i\sin\theta)$ is nonzero, then

$$\frac{1}{z} = \frac{1}{r}(\cos\theta - i\sin\theta)$$

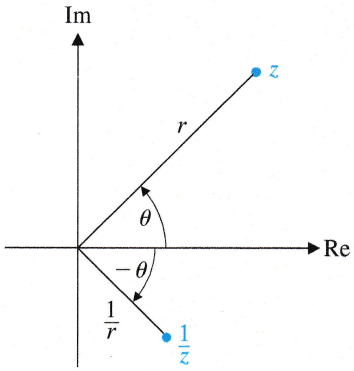


Figure C.7

See Figure C.7.

Example C.4

Find the product of $1 + i$ and $1 - \sqrt{3}i$ in polar form.

Solution From Example C.3, we have

$$1 + i = \sqrt{2}\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right) \quad \text{and} \quad 1 - \sqrt{3}i = 2\left(\cos\left(-\frac{\pi}{3}\right) + i\sin\left(-\frac{\pi}{3}\right)\right)$$

Therefore,

$$\begin{aligned}(1 + i)(1 - \sqrt{3}i) &= 2\sqrt{2}\left[\cos\left(\frac{\pi}{4} - \frac{\pi}{3}\right) + i\sin\left(\frac{\pi}{4} - \frac{\pi}{3}\right)\right] \\ &= 2\sqrt{2}\left[\cos\left(-\frac{\pi}{12}\right) + i\sin\left(-\frac{\pi}{12}\right)\right]\end{aligned}$$

See Figure C.8.

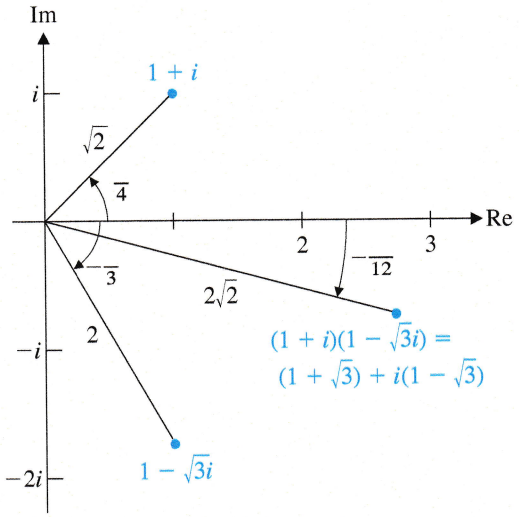


Figure C.8

Remark Since $(1 + i)(1 - \sqrt{3}i) = (1 + \sqrt{3}) + i(1 - \sqrt{3})$ (check this), we must have

$$1 + \sqrt{3} = 2\sqrt{2}\cos\left(-\frac{\pi}{12}\right) = -2\sqrt{2}\cos\left(\frac{\pi}{12}\right)$$

and

$$1 - \sqrt{3} = 2\sqrt{2}\sin\left(-\frac{\pi}{12}\right) = -2\sqrt{2}\sin\left(\frac{\pi}{12}\right)$$

(Why?) This implies that

$$\cos\left(\frac{\pi}{12}\right) = \frac{1 + \sqrt{3}}{2\sqrt{2}} \quad \text{and} \quad \sin\left(\frac{\pi}{12}\right) = \frac{\sqrt{3} - 1}{2\sqrt{2}}$$

We therefore have a method for finding the sine and cosine of an angle such as $\pi/12$ that is not a special angle but that can be obtained as a sum or difference of special angles.

De Moivre's Theorem

If n is a positive integer and $z = r(\cos\theta + i\sin\theta)$, then repeated use of Equation (1) yields formulas for the powers of z:

$$z^2 = r^2(\cos 2\theta + i\sin 2\theta)$$
$$z^3 = zz^2 = r^3(\cos 3\theta + i\sin 3\theta)$$
$$z^4 = zz^3 = r^4(\cos 4\theta + i\sin 4\theta)$$
$$\vdots$$

In general, we have the following result, known as ***De Moivre's Theorem.***

Abraham De Moivre (1667–1754) was a French mathematician who made important contributions to trigonometry, analytic geometry, probability, and statistics.

Theorem C.1 **De Moivre's Theorem**

If $z = r(\cos\theta + i\sin\theta)$ and n is a positive integer, then

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

Stated differently, we have

$$|z^n| = |z|^n \text{ and } \arg(z^n) = n\arg z$$

In words, De Moivre's Theorem says that *to take the nth power of a complex number, we take the nth power of its absolute value and multiply its argument by n.*

Example C.5 Find $(1 + i)^6$.

Solution From Example C.3(a), we have

$$1 + i = \sqrt{2}\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right)$$

Hence, De Moivre's Theorem gives

$$\begin{aligned}(1 + i)^6 &= (\sqrt{2})^6\left(\cos\frac{6\pi}{4} + i\sin\frac{6\pi}{4}\right)\\ &= 8\left(\cos\frac{3\pi}{2} + i\sin\frac{3\pi}{2}\right)\\ &= 8(0 + i(-1)) = -8i\end{aligned}$$

See Figure C.9, which shows $1 + i, (1 + i)^2, (1 + i)^3, \ldots, (1 + i)^6$.

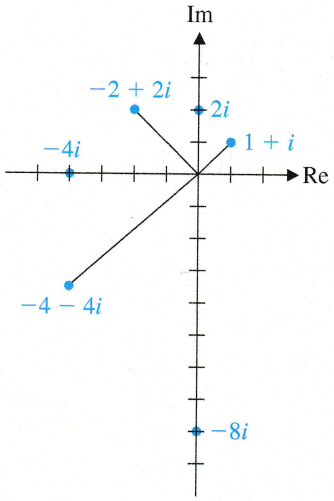


Figure C.9
Powers of $1 + i$

We can also use De Moivre's Theorem to find nth roots of complex numbers. An nth root of the complex number z is any complex number w such that

$$w^n = z$$

In polar form, we have

$$w = s(\cos\varphi + i\sin\varphi) \quad \text{and} \quad z = r(\cos\theta + i\sin\theta)$$

so, by De Moivre's Theorem,

$$s^n(\cos n\varphi + i\sin n\varphi) = r(\cos\theta + i\sin\theta)$$

Equating the absolute values, we see that

$$s^n = r \quad \text{or} \quad s = r^{1/n} = \sqrt[n]{r}$$

We must also have

$$\cos n\varphi = \cos\theta \quad \text{and} \quad \sin n\varphi = \sin\theta$$

(Why?) Since the sine and cosine functions each have period 2π, these equations imply that $n\varphi$ and θ differ by an integer multiple of 2π; that is,

$$n\varphi = \theta + 2k\pi \quad \text{or} \quad \varphi = \frac{\theta + 2k\pi}{n}.$$

where k is an integer. Therefore,

$$w = r^{1/n}\left[\cos\left(\frac{\theta + 2k\pi}{n}\right) + i\sin\left(\frac{\theta + 2k\pi}{n}\right)\right]$$

describes the possible nth roots of z as k ranges over the integers. It is not hard to show that $k = 0, 1, 2, \ldots, n - 1$ produce distinct values of w, so there are exactly n different nth roots of $z = r(\cos\theta + i\sin\theta)$. We summarize this result as follows:

Let $z = r(\cos\theta + i\sin\theta)$ and let n be a positive integer. Then z has exactly n distinct nth roots given by

$$r^{1/n}\left[\cos\left(\frac{\theta + 2k\pi}{n}\right) + i\sin\left(\frac{\theta + 2k\pi}{n}\right)\right] \qquad (2)$$

for $k = 0, 1, 2, \ldots, n - 1$.

Example C.6 Find the three cube roots of -27.

Solution In polar form, $-27 = 27(\cos\pi + i\sin\pi)$. It follows that the cube roots of -27 are given by

$$(-27)^{1/3} = 27^{1/3}\left[\cos\left(\frac{\pi + 2k\pi}{3}\right) + i\sin\left(\frac{\pi + 2k\pi}{3}\right)\right] \quad \text{for } k = 0, 1, 2$$

Using formula (2) with $n = 3$, we obtain

$$27^{1/3}\left[\cos\frac{\pi}{3} + i\sin\frac{\pi}{3}\right] = 3\left(\frac{1}{2} + \frac{\sqrt{3}}{2}i\right) = \frac{3}{2} + \frac{3\sqrt{3}}{2}i$$

$$27^{1/3}\left[\cos\left(\frac{\pi + 2\pi}{3}\right) + i\sin\left(\frac{\pi + 2\pi}{3}\right)\right] = 3(\cos\pi + i\sin\pi) = -3$$

$$27^{1/3}\left[\cos\left(\frac{\pi + 4\pi}{3}\right) + i\sin\left(\frac{\pi + 4\pi}{3}\right)\right] = 3\left(\cos\frac{5\pi}{3} + i\sin\frac{5\pi}{3}\right)$$

$$= 3\left(\frac{1}{2} - \frac{\sqrt{3}}{2}i\right) = \frac{3}{2} - \frac{3\sqrt{3}}{2}i$$

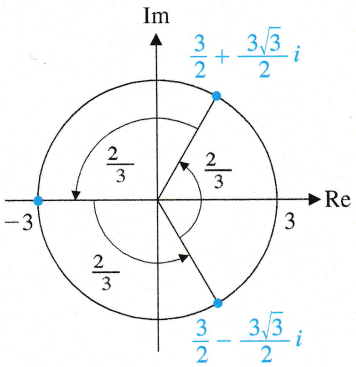
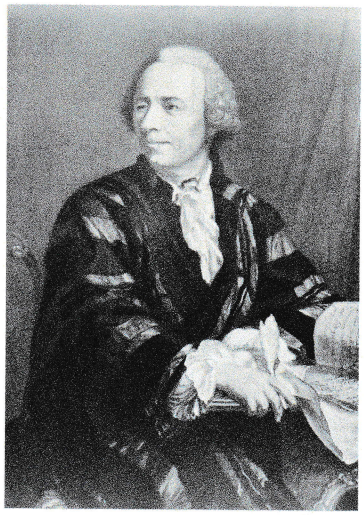


Figure C.10
The cube roots of -27

As Figure C.10 shows, the three cube roots of -27 are equally spaced $2\pi/3$ radians (120°) apart around a circle of radius 3 centered at the origin.

In general, formula (2) implies that the nth roots of $z = r(\cos\theta + i\sin\theta)$ will lie on a circle of radius $r^{1/n}$ centered at the origin. Moreover, they will be equally spaced $2\pi/n$ radians ($360/n$°) apart. (Verify this.) Thus, if we can find one nth root of z, the remaining nth roots of z can be obtained by rotating the first root through successive increments of $2\pi/n$ radians. Had we known this in Example C.6, we could have used the fact that the real cube root of -27 is -3 and then rotated it twice through an angle of $2\pi/3$ radians (120°) to get the other two cube roots.



© Bettmann/CORBIS

Leonhard Euler (1707–1783) was the most prolific mathematician of all time. He has over 900 publications to his name, and his collected works fill over 70 volumes. There are so many results attributed to him that "Euler's formula" or "Euler's Theorem" can mean many different things, depending on the context.

Euler worked in so many areas of mathematics, it is difficult to list them all. His contributions to calculus and analysis, differential equations, number theory, geometry, topology, mechanics, and other areas of applied mathematics continue to be influential. He also introduced much of the notation we currently use, including π, e, i, Σ for summation, Δ for difference, and $f(x)$ for a function, and was the first to treat sine and cosine as functions.

Euler was born in Switzerland but spent most of his mathematical life in Russia and Germany. In 1727, he joined the St. Petersburg Academy of Sciences, which had been founded by Catherine I, the wife of Peter the Great. He went to Berlin in 1741 at the invitation of Frederick the Great, but returned in 1766 to St. Petersburg, where he remained until his death. When he was young, he lost the vision in one eye as the result of an illness, and by 1776 he had lost the vision in the other eye and was totally blind. Remarkably, his mathematical output did not diminish, and he continued to be productive until the day he died.

Euler's Formula

In calculus, you learn that the function e^z has a power series expansion

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

that converges for every real number z. It can be shown that this expansion also works when z is a complex number and that the complex exponential function e^z obeys the usual rules for exponents. The sine and cosine functions also have power series expansions:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + - \cdots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + - \cdots$$

If we let $z = ix$, where x is a real number, then we have

$$e^z = e^{ix} = 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \cdots$$

Using the fact that $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, $i^5 = i$, and so on, repeating in a cycle of length 4, we see that

$$\begin{aligned} e^{ix} &= 1 + ix - \frac{x^2}{2!} - \frac{ix^3}{3!} + \frac{x^4}{4!} + \frac{ix^5}{5!} - \frac{x^6}{6!} - \frac{ix^7}{7!} + + - - \cdots \\ &= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + - \cdots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + - \cdots\right) \\ &= \cos x + i \sin x \end{aligned}$$

This remarkable result is known as ***Euler's formula.***

Theorem C.2 **Euler's Formula**

For any real number x,

$$e^{ix} = \cos x + i \sin x$$

Using Euler's formula, we see that the polar form of a complex number can be written more compactly as

$$z = r(\cos\theta + i\sin\theta) = re^{i\theta}$$

For example, from Example C.3(a), we have

$$1 + i = \sqrt{2}\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right) = \sqrt{2}e^{i\pi/4}$$

We can also go in the other direction and convert a complex exponential back into polar or standard form.

Example C.7 Write the following in the form $a + bi$:

(a) $e^{i\pi}$ (b) $e^{2+i\pi/4}$

Solution (a) Using Euler's formula, we have

$$e^{i\pi} = \cos\pi + i\sin\pi = -1 + i \cdot 0 = -1$$

(If we write this equation as $e^{i\pi} + 1 = 0$, we obtain what is surely one of the most remarkable equations in mathematics. It contains the fundamental operations of addition, multiplication, and exponentiation; the additive identity 0 and the multiplicative identity 1; the two most important transcendental numbers, π and e; and the complex unit i—all in one equation!)

(b) Using rules for exponents together with Euler's formula, we obtain

$$e^{2+i\pi/4} = e^2e^{i\pi/4} = e^2\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right) = e^2\left(\frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}\right)$$

$$= \frac{e^2\sqrt{2}}{2} + \frac{e^2\sqrt{2}}{2}i$$

If $z = re^{i\theta} = r(\cos\theta + i\sin\theta)$, then

$$\bar{z} = r(\cos\theta - i\sin\theta) \tag{3}$$

The trigonometric identities

$$\cos(-\theta) = \cos\theta \text{ and } \sin(-\theta) = -\sin\theta$$

allow us to rewrite Equation (3) as

$$\bar{z} = r(\cos(-\theta) + i\sin(-\theta)) = re^{i(-\theta)}$$

This gives the following useful formula for the conjugate:

If $z = re^{i\theta}$, then

$$\bar{z} = re^{-i\theta}$$

Note Euler's formula gives a quick, one-line proof of De Moivre's Theorem:

$$[r(\cos\theta + i\sin\theta)]^n = (re^{i\theta})^n = r^n e^{in\theta} = r^n(\cos n\theta + i\sin n\theta)$$

Appendix D*

Polynomials

A ***polynomial*** is a function p of a single variable x that can be written in the form

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \tag{1}$$

where $a_0, a_1, \ldots, a_n$ are constants ($a_n \neq 0$), called the ***coefficients*** of p. With the convention that $x^0 = 1$, we can use summation notation to write p as

$$p(x) = \sum_{k=0}^{n} a_k x^k$$

The integer n is called the ***degree*** of p, which is denoted by writing $\deg p = n$. A polynomial of degree zero is called a ***constant polynomial.***

Euler gave the most algebraic of the proofs of the existence of the roots of [a polynomial] equation. . . . I regard it as unjust to ascribe this proof exclusively to Gauss, who merely added the finishing touches.
—Georg Frobenius, 1907
Quoted on the MacTutor History of Mathematics archive, http://www-history.mcs.st-and.ac.uk/history/

Example D.1 Which of the following are polynomials?

(a) $2 - \frac{1}{3}x + \sqrt{2}x^2$ (b) $2 - \dfrac{1}{3x^2}$ (c) $\sqrt{2x^2}$

(d) $\ln\left(\dfrac{2e^{5x^3}}{e^{3x}}\right)$ (e) $\dfrac{x^2 - 5x + 6}{x - 2}$ (f) $\sqrt{x}$

(g) $\cos(2\cos^{-1}x)$ (h) e^x

Solution (a) This is the only one that is obviously a polynomial.

(b) A polynomial of the form shown in Equation (1) cannot become infinite as x approaches a finite value [$\lim_{x\to c} p(x) \neq \pm\infty$], whereas $2 - 1/3x^2$ approaches $-\infty$ as x approaches zero. Hence, it is not a polynomial.

(c) We have

$$\sqrt{2x^2} = \sqrt{2}\sqrt{x^2} = \sqrt{2}|x|$$

which is equal to $\sqrt{2}x$ when $x \geq 0$ and to $-\sqrt{2}x$ when $x < 0$. Therefore, this expression is formed by "splicing together" two polynomials (a *piecewise polynomial*), but it is not a polynomial itself.

*Exercises and selected odd-numbered answers for this appendix can be found on the student companion website.

(d) Using properties of exponents and logarithms, we have

$$\ln\left(\frac{2e^{5x^3}}{e^{3x}}\right) = \ln(2e^{5x^3-3x}) = \ln 2 + \ln(e^{5x^3-3x})$$
$$= \ln 2 + 5x^3 - 3x = \ln 2 - 3x + 5x^3$$

so this expression is a polynomial.

(e) The domain of this function consists of all real numbers $x \neq 2$. For these values of x, the function simplifies to

$$\frac{x^2 - 5x + 6}{x - 2} = \frac{(x-2)(x-3)}{x-2} = x - 3$$

so we can say that it is a polynomial *on its domain.*

(f) We see that this function cannot be a polynomial (even on its domain $x \geq 0$), since repeated differentiation of a polynomial of the form shown in Equation (1) eventually results in zero and $\sqrt{x}$ does not have this property. (Verify this.)

(g) The domain of this expression is $-1 \leq x \leq 1$. Let $\theta = \cos^{-1} x$ so that $\cos\theta = x$. Using a trigonometric identity, we see that

$$\cos(2\cos^{-1} x) = \cos 2\theta = 2\cos^2\theta - 1 = 2x^2 - 1$$

so this expression is a polynomial on its domain.

(h) Analyzing this expression as we did the one in (f), we conclude that it is not a polynomial.

Two polynomials are ***equal*** if the coefficients of corresponding powers of x are all equal. In particular, equal polynomials must have the same degree. The ***sum*** of two polynomials is obtained by adding together the coefficients of corresponding powers of x.

Example D.2 Find the sum of $2 - 4x + x^2$ and $1 + 2x - x^2 + 3x^3$.

Solution We compute

$$(2 - 4x + x^2) + (1 + 2x - x^2 + 3x^3) = (2 + 1) + (-4 + 2)x$$
$$+ (1 + (-1))x^2 + (0 + 3)x^3$$
$$= 3 - 2x + 3x^3$$

where we have "padded" the first polynomial by giving it an x^3 coefficient of zero.

We define the ***difference*** of two polynomials analogously, subtracting coefficients instead of adding them. The ***product*** of two polynomials is obtained by repeatedly using the distributive law and then gathering together corresponding powers of x.

Example D.3 Find the product of $2 - 4x + x^2$ and $1 + 2x - x^2 + 3x^3$.

Solution We obtain

$$\begin{aligned}
&(2 - 4x + x^2)(1 + 2x - x^2 + 3x^3) \\
&\quad = 2(1 + 2x - x^2 + 3x^3) - 4x(1 + 2x - x^2 + 3x^3) \\
&\qquad + x^2(1 + 2x - x^2 + 3x^3) \\
&\quad = (2 + 4x - 2x^2 + 6x^3) + (-4x - 8x^2 + 4x^3 - 12x^4) \\
&\qquad + (x^2 + 2x^3 - x^4 + 3x^5) \\
&\quad = 2 + (4x - 4x) + (-2x^2 - 8x^2 + x^2) + (6x^3 + 4x^3 + 2x^3) \\
&\qquad + (-12x^4 - x^4) + 3x^5 \\
&\quad = 2 - 9x^2 + 12x^3 - 13x^4 + 3x^5
\end{aligned}$$



Observe that for two polynomials p and q, we have

$$\deg(pq) = \deg p + \deg q$$

If p and q are polynomials with $\deg q \leq \deg p$, we can divide q into p, using long division to obtain the quotient p/q. The next example illustrates the procedure, which is the same as for long division of one integer into another. Just as the quotient of two integers is not, in general, an integer, the quotient of two polynomials is not, in general, another polynomial.

Example D.4 Compute $\dfrac{1 + 2x - x^2 + 3x^3}{2 - 4x + x^2}$.

Solution We will perform long division. It is helpful to write each polynomial with *decreasing* powers of x. Accordingly, we have

$$x^2 - 4x + 2\overline{)3x^3 - x^2 + 2x + 1}$$

We begin by dividing x^2 into $3x^3$ to obtain the partial quotient $3x$. We then multiply $3x$ by the divisor $x^2 - 4x + 2$, subtract the result, and bring down the next term from the dividend $(3x^3 - x^2 + 2x + 1)$:

$$\begin{array}{r}
3x \\
x^2 - 4x + 2\overline{)3x^3 - x^2 + 2x + 1} \\
\underline{3x^3 - 12x^2 + 6x} \\
11x^2 - 4x + 1
\end{array}$$

Then we repeat the process with $11x^2$, multiplying 11 by $x^2 - 4x + 2$ and subtracting the result from $11x^2 - 4x + 1$. We obtain

$$\begin{array}{r}
3x + 11 \\
x^2 - 4x + 2\overline{)3x^3 - x^2 + 2x + 1} \\
\underline{3x^3 - 12x^2 + 6x} \\
11x^2 - 4x + 1 \\
\underline{11x^2 - 44x + 22} \\
40x - 21
\end{array}$$

We now have a remainder $40x - 21$. Its degree is less than that of the divisor $x^2 - 4x + 2$, so the process stops, and we have found that

$$3x^3 - x^2 + 2x + 1 = (x^2 - 4x + 2)(3x + 11) + (40x - 21)$$

or

$$\frac{3x^3 - x^2 + 2x + 1}{x^2 - 4x + 2} = 3x + 11 + \frac{40x - 21}{x^2 - 4x + 2}$$



Example D.4 can be generalized to give the following result, known as the ***division algorithm.***

Theorem D.1 **The Division Algorithm**

If f and g are polynomials with $\deg g \leq \deg f$, then there are polynomials q and r such that

$$f(x) = g(x)q(x) + r(x)$$

where either $r = 0$ or $\deg r < \deg g$.

In Example D.4,

$$f(x) = 3x^3 - x^2 + 2x + 1, \quad g(x) = x^2 - 4x + 2, \quad q(x) = 3x + 11,$$

$$\text{and} \quad r(x) = 40x - 21$$

In the division algorithm, if the remainder is zero, then

$$f(x) = g(x)q(x)$$

and we say that g is a ***factor*** of f. (Notice that q is also a factor of f.) There is a close connection between the factors of a polynomial and its zeros. A ***zero*** of a polynomial f is a number a such that $f(a) = 0$. [The number a is also called a ***root*** of the polynomial equation $f(x) = 0$.] The following result, known as the ***Factor Theorem,*** establishes the connection between factors of a polynomial and its zeros.

Theorem D.2 **The Factor Theorem**

Let f be a polynomial and let a be a constant. Then a is a zero of f if and only if $x - a$ is a factor of $f(x)$.

Proof By the division algorithm,

$$f(x) = (x - a)q(x) + r(x)$$

where either $r(x) = 0$ or $\deg r < \deg(x - a) = 1$. Thus, in either case, $r(x) = r$ is a constant. Now,

$$f(a) = (a - a)q(a) + r = r$$

so $f(a) = 0$ if and only if $r = 0$, which is equivalent to

$$f(x) = (x - a)q(x)$$

as we needed to prove.

There is no method that is guaranteed to find the zeros of a given polynomial. However, there are some guidelines that are useful in special cases. The case of a polynomial with *integer* coefficients is particularly interesting. The following result, known as the ***Rational Roots Theorem,*** gives criteria for a zero of such a polynomial to be a *rational* number.

Theorem D.3 **The Rational Roots Theorem**

Let

$$f(x) = a_0 + a_1x + \cdots + a_nx^n$$

be a polynomial with integer coefficients and let a/b be a rational number written in lowest terms. If a/b is a zero of f, then a_0 is a multiple of a and a_n is a multiple of b.

Proof If a/b is a zero of f, then

$$a_0 + a_1\left(\frac{a}{b}\right) + \cdots + a_{n-1}\left(\frac{a}{b}\right)^{n-1} + a_n\left(\frac{a}{b}\right)^n = 0$$

Multiplying through by b^n, we have

$$a_0b^n + a_1ab^{n-1} + \cdots + a_{n-1}a^{n-1}b + a_na = 0 \tag{1}$$

which implies that

$$a_0b^n + a_1ab^{n-1} + \cdots + a_{n-1}a^{n-1}b = -a_na^n \tag{2}$$

The left-hand side of Equation (2) is a multiple of b, so a_na^n must be a multiple of b also. Since a/b is in lowest terms, a and b have no common factors greater than 1. Therefore, a_n must be a multiple of b.

We can also write Equation (1) as

$$-a_0b^n = a_1ab^{n-1} + \cdots + a_{n-1}a^{n-1}b + a_na^n$$

and a similar argument shows that a_0 must be a multiple of a. (Show this.)

Example D.5 Find all the rational roots of the equation

$$6x^3 + 13x^2 - 4 = 0 \tag{3}$$

Solution If a/b is a root of this equation, then 6 is a multiple of b and -4 is a multiple of a, by the Rational Roots Theorem. Therefore,

$$a \in \{\pm 1, \pm 2, \pm 4\} \quad \text{and} \quad b \in \{\pm 1, \pm 2, \pm 3, \pm 6\}$$

Forming all possible rational numbers a/b with these choices of a and b, we see that the only possible rational roots of the given equation are

$$\pm 1, \pm 2, \pm 4, \pm \tfrac{1}{2}, \pm \tfrac{1}{3}, \pm \tfrac{2}{3}, \pm \tfrac{4}{3}, \pm \tfrac{1}{6}$$

Substituting these values into Equation (3) one at a time, we find that -2, $-\frac{2}{3}$, and $\frac{1}{2}$ are the only values from this list that are actually roots. (Check these.) As we will see shortly, a polynomial equation of degree 3 cannot have more than three roots, so these are not only all the *rational* roots of Equation (3) but also its *only* roots.

We can improve on the trial-and-error method of Example D.5 in various ways. For example, once we find one root a of a given polynomial equation $f(x) = 0$, we know that $x - a$ is a factor of $f(x)$—say, $f(x) = (x - a)g(x)$. We can therefore divide $f(x)$ by $x - a$ (using long division) to find $g(x)$. Since $\deg g < \deg f$, the roots of $g(x) = 0$ [which are also roots of $f(x) = 0$] may be easier to find. In particular, if $g(x)$ is a quadratic polynomial, we have access to the ***quadratic formula.***

Suppose

$$ax^2 + bx + c = 0$$

(We may assume that a is positive, since multiplying both sides by -1 would produce an equivalent equation otherwise.) Then, completing the square, we have

$$a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}\right) = \frac{b^2}{4a} - c$$

(Verify this.) Equivalently,

$$a\left(x + \frac{b}{2a}\right)^2 = \frac{b^2}{4a} - c \quad \text{or} \quad \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}$$

Therefore,

$$x + \frac{b}{2a} = \pm\sqrt{\frac{b^2 - 4ac}{4a^2}} = \frac{\pm\sqrt{b^2 - 4ac}}{2a}$$

or

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

Let's revisit the equation from Example D.5 with the quadratic formula in mind.

Example D.6

Find the roots of $6x^3 + 13x^2 - 4 = 0$.

Solution Let's suppose we use the Rational Roots Theorem to discover that $x = -2$ is a rational root of $6x^3 + 13x^2 - 4 = 0$. Then $x + 2$ is a factor of $6x^3 + 13x^2 - 4$, and long division gives

$$6x^3 + 13x^2 - 4 = (x + 2)(6x^2 + x - 2)$$

(Check this.) We can now apply the quadratic formula to the second factor to find that its zeros are

$$x = \frac{-1 \pm \sqrt{1^2 - 4(6)(-2)}}{2 \cdot 6}$$
$$= \frac{-1 \pm \sqrt{49}}{12} = \frac{-1 \pm 7}{12}$$
$$= \tfrac{6}{12}, -\tfrac{8}{12}$$

or, in lowest terms, $\frac{1}{2}$ and $-\frac{2}{3}$. Thus, the three roots of Equation (3) are -2, $\frac{1}{2}$, and $-\frac{2}{3}$, as we determined in Example D.5.

Remark The Factor Theorem establishes a connection between the zeros of a polynomial and its *linear* factors. However, a polynomial without linear factors may still have factors of higher degree. Furthermore, when asked to factor a polynomial, we need to know the number system to which the coefficients of the factors are supposed to belong.

For example, consider the polynomial

$$p(x) = x^4 + 1$$

Over the *rational numbers* $\mathbb{Q}$, the only possible zeros of p are 1 and -1, by the Rational Roots Theorem. A quick check shows that neither of these actually works, so $p(x)$ has no *linear* factors with rational coefficients, by the Factor Theorem. However, $p(x)$ may still factor into a product of two *quadratics*. We will check for quadratic factors using the method of ***undetermined coefficients.***

Suppose that

$$x^4 + 1 = (x^2 + ax + b)(x^2 + cx + d)$$

Expanding the right-hand side and comparing coefficients, we obtain the equations

$$\begin{aligned} a + c &= 0 \\ b + ac + d &= 0 \\ bc + ad &= 0 \\ bd &= 1 \end{aligned}$$

If $a = 0$, then $c = 0$ and $d = -b$. This gives $-b^2 = 1$, which has no solutions in $\mathbb{Q}$. Hence, we may assume that $a \neq 0$. Then $c = -a$, and we obtain $d = b$. It now follows that $b^2 = 1$, so $b = 1$ or $b = -1$. This implies that $a^2 = 2$ or $a^2 = -2$, respectively, neither of which has solutions in $\mathbb{Q}$. It follows that $x^4 + 1$ cannot be factored over $\mathbb{Q}$. We say that it is ***irreducible*** over $\mathbb{Q}$.

However, over the *real numbers* $\mathbb{R}$, $x^4 + 1$ does factor. The calculations we have just done show that

$$x^4 + 1 = (x^2 + \sqrt{2}x + 1)(x^2 - \sqrt{2}x + 1)$$

(Why?) To see whether we can factor further, we apply the quadratic formula. We see that the first factor has zeros

$$x = \frac{-\sqrt{2} \pm \sqrt{(\sqrt{2})^2 - 4}}{2} = \frac{-\sqrt{2} \pm \sqrt{-2}}{2} = \frac{\sqrt{2}}{2}(-1 \pm i) = -\frac{1}{\sqrt{2}} \pm \frac{1}{\sqrt{2}}i$$

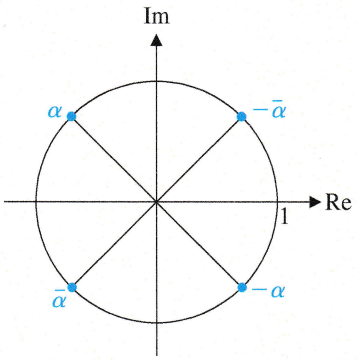


Figure D.1

which are in $\mathbb{C}$ but not in $\mathbb{R}$. Hence, $x^2 + \sqrt{2}x + 1$ cannot be factored into linear factors over $\mathbb{R}$. Similarly, $x^2 - \sqrt{2}x + 1$ cannot be factored into linear factors over $\mathbb{R}$.

Our calculations show that a complete factorization of $x^4 + 1$ is possible over the *complex numbers* $\mathbb{C}$. The four zeros of $x^4 + 1$ are

$$\alpha = -\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i, \quad \bar{\alpha} = -\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}i, \quad -\bar{\alpha} = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i,$$
$$-\alpha = \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}i$$

which, as Figure D.1 shows, all lie on the unit circle in the complex plane. Thus, the factorization of $x^4 + 1$ is

$$x^4 + 1 = (x - \alpha)(x - \bar{\alpha})(x + \bar{\alpha})(x + \alpha)$$

The preceding Remark illustrates several important properties of polynomials. Notice that the polynomial $p(x) = x^4 + 1$ satisfies $\deg p = 4$ and has exactly four zeros in $\mathbb{C}$. Furthermore, its complex zeros occur in *conjugate pairs*; that is, its complex zeros can be paired up as

$$\{\alpha, \bar{\alpha}\} \quad \text{and} \quad \{-\alpha, -\bar{\alpha}\}$$

These last two facts are true in general. The first is an instance of the ***Fundamental Theorem of Algebra (FTA)***, a result that was first proved by Gauss in 1797.

Theorem D.4 **The Fundamental Theorem of Algebra**

Every polynomial of degree n with real or complex coefficients has exactly n zeros (counting multiplicities) in $\mathbb{C}$.

This important theorem is sometimes stated as

"Every polynomial with real or complex coefficients has a zero in $\mathbb{C}$."

Let's call this statement FTA′. Certainly, FTA implies FTA′. Conversely, if FTA′ is true, then if we have a polynomial p of degree n, it has a zero α in $\mathbb{C}$. The Factor Theorem then tells us that $x - \alpha$ is a factor of $p(x)$, so

$$p(x) = (x - \alpha)q(x)$$

where q is a polynomial of degree $n - 1$ (also with real or complex coefficients). We can now apply FTA′ to q to get another zero, and so on, making FTA true. This argument can be made into a nice induction proof. (Try it.)

It is not possible to give a formula (along the lines of the quadratic formula) for the zeros of polynomials of degree 5 or more. (The work of Abel and Galois confirmed this; see page 311.) Consequently, other methods must be used to prove FTA. The proof that Gauss gave uses topological methods and can be found in more advanced mathematics courses.

Now suppose that

$$p(x) = a_0 + a_1x + \cdots + a_nx^n$$

is a polynomial with real coefficients. Let α be a complex zero of p so that

$$a_0 + a_1\alpha + \cdots + a_n\alpha^n = p(\alpha) = 0$$

Then, using properties of conjugates, we have

$$\begin{aligned} p(\overline{\alpha}) &= a_0 + a_1\overline{\alpha} + \cdots + a_n\overline{\alpha}^n = \overline{a}_0 + \overline{a}_1\overline{\alpha} + \cdots + \overline{a}_n\overline{\alpha}^n \\ &= \overline{a_0 + a_1\alpha + \cdots + a_n\alpha^n} \\ &= \overline{p(\alpha)} = \overline{0} = 0 \end{aligned}$$

Thus, $\overline{\alpha}$ is also a zero of p. This proves the following result:

The complex zeros of a polynomial with real coefficients occur in conjugate pairs.

In some situations, we do not need to know *what* the zeros of a polynomial are—we only need to know *where* they are located. For example, we might only need to know whether the zeros are positive or negative (as in Theorem 4.35). One theorem that is useful in this regard is ***Descartes' Rule of Signs.*** It allows us to make certain predictions about the number of positive zeros of a polynomial with real coefficients based on the signs of these coefficients.

Descartes' stated this rule in his 1637 book *La Géometrie,* but did not give a proof. Several mathematicians later furnished a proof, and Gauss provided a somewhat sharper version of the theorem in 1828.

Given a polynomial $a_0 + a_1x + \cdots + a_nx^n$, write its nonzero coefficients in order. Replace each positive coefficient by a plus sign and each negative coefficient by a minus sign. We will say that the polynomial has k ***sign changes*** if there are k places where the coefficients change sign. For example, the polynomial $2 - 3x + 4x^3 + x^4 - 7x^5$ has the sign pattern

$$\underbrace{+ -}_{} \; \underbrace{+}_{} \; \underbrace{+ -}_{}$$

so it has three sign changes, as indicated.

Theorem D.5 **Descartes' Rule of Signs**

Let p be a polynomial with real coefficients that has k sign changes. Then the number of positive zeros of p (counting multiplicities) is at most k.

In words, Descartes' Rule of Signs says that a real polynomial cannot have more positive zeros than it has sign changes.

Example D.7

Show that the polynomial $p(x) = 4 + 2x^2 - 7x^4$ has exactly one positive zero.

Solution The coefficients of p have the sign pattern $+ + -$, which has only one sign change. So, by Descartes' Rule of Signs, p has at most one positive zero. But $p(0) = 4$ and $p(1) = -1$, so there is a zero somewhere in the interval $(0, 1)$. Hence, this is the only positive zero of p.



We can also use Descartes' Rule of Signs to give a bound on the number of *negative* zeros of a polynomial with real coefficients. Let

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

and let b be a negative zero of p. Then $b = -c$ for $c > 0$, and we have

$$\begin{aligned} 0 = p(b) &= a_0 + a_1b + a_2b^2 + \cdots + a_nb^n \\ &= a_0 - a_1c + a_2c^2 - + \cdots + (-1)^n a_nc^n \end{aligned}$$

But $$p(-x) = a_0 - a_1x + a_2x^2 - + \cdots + (-1)^n a_nx^n$$

so c is a positive zero of $p(-x)$. Therefore, $p(x)$ has exactly as many negative zeros as $p(-x)$ has positive zeros. Combined with Descartes' Rule of Signs, this observation yields the following:

Let p be a polynomial with real coefficients. Then the number of negative zeros of p is at most the number of sign changes of $p(-x)$.

Example D.8 Show that the zeros of $p(x) = 1 + 3x + 2x^2 + x^5$ cannot all be real.

Solution The coefficients of $p(x)$ have no sign changes, so p has no positive zeros. Since $p(-x) = 1 - 3x + 2x^2 - x^5$ has three sign changes among its coefficients, p has at most three negative zeros. We note that 0 is not a zero of p either, so p has at most three real zeros. Therefore, it has at least two complex (nonreal) zeros.

Answers to Selected Odd-Numbered Exercises

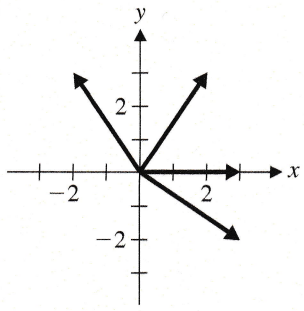
Answers are easy. It's asking the right questions [that's] hard.

—Doctor Who
"The Face of Evil,"
By Chris Boucher
BBC, 1977

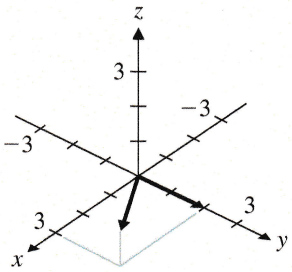
Chapter 1

Exercises 1.1

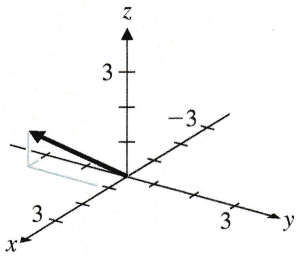
1.



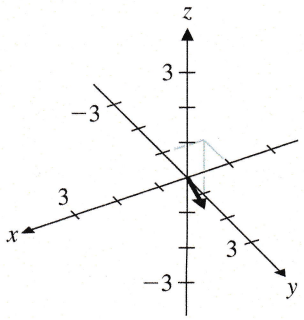
3. (a), (b)



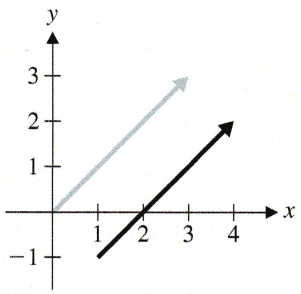
(c)



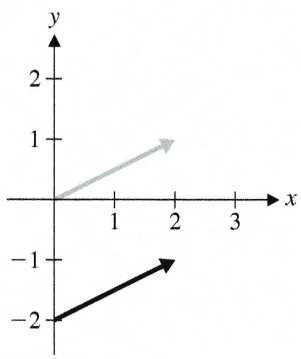
(d)



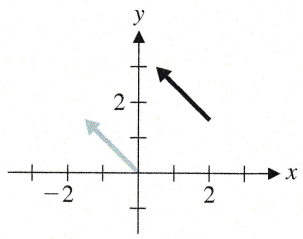
5. (a)



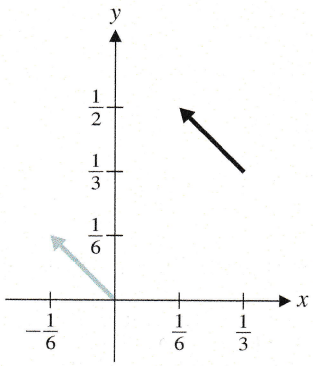
(b)



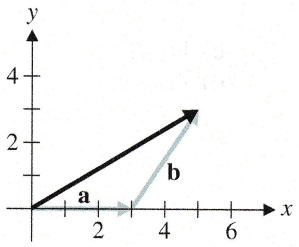
(c)



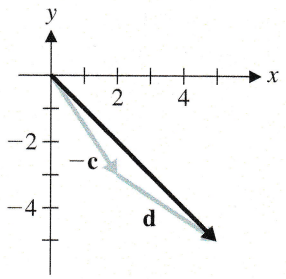
(d)



7. $\mathbf{a} + \mathbf{b} = [5, 3]$



9. $\mathbf{d} - \mathbf{c} = [5, -5]$



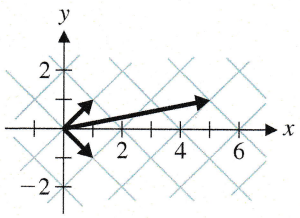
11. $[3, -2, 3]$

13. $\mathbf{u} = \begin{bmatrix} 1/2 \\ \sqrt{3}/2 \end{bmatrix}$, $\mathbf{v} = \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix}$, $\mathbf{u} + \mathbf{v} = \begin{bmatrix} (1 - \sqrt{3})/2 \\ (\sqrt{3} - 1)/2 \end{bmatrix}$, $\mathbf{u} - \mathbf{v} = \begin{bmatrix} (1 + \sqrt{3})/2 \\ (1 + \sqrt{3})/2 \end{bmatrix}$

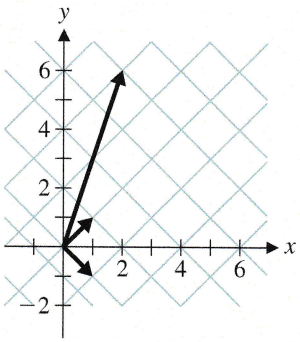
15. $5\mathbf{a}$

17. $\mathbf{x} = 3\mathbf{a}$

19.



21. $\mathbf{w} = -2\mathbf{u} + 4\mathbf{v}$



25. $\mathbf{u} + \mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

27. $\mathbf{u} + \mathbf{v} = [0, 1, 0, 0]$

29.

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

·	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

31. 0

33. 1

35. 0

37. 2, 0, 3

39. 5

41. [1, 1, 0]

43. [0, 0, 2, 2], [2, 3, 1, 1]

45. $x = 2$

47. No solution

49. $x = 3$

51. No solution

53. $x = 2$

55. $x = 1$, or $x = 5$

57. **(a)** All $a \neq 0$ **(b)** $a = 1, 5$

(c) a and m can have no common factors other than 1 [i.e., the greatest common divisor (gcd) of a and m is 1].

Exercises 1.2

1. -1

3. 11

5. 2

7. $\sqrt{5}, \begin{bmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$

9. $\sqrt{14}, \begin{bmatrix} 1/\sqrt{14} \\ 2/\sqrt{14} \\ 3/\sqrt{14} \end{bmatrix}$

11. $\sqrt{6}$, $[1/\sqrt{6}, 1/\sqrt{3}, 1/\sqrt{2}, 0]$
13. $\sqrt{17}$ **15.** $\sqrt{6}$
17. **(a)** $\mathbf{u} \cdot \mathbf{v}$ is a scalar, not a vector.
(c) $\mathbf{v} \cdot \mathbf{w}$ is a scalar and $\mathbf{u}$ is a vector.
19. Acute **21.** Acute **23.** Acute
25. 60° **27.** ≈88.10° **29.** ≈14.34°
31. Since $\overrightarrow{AB} \cdot \overrightarrow{AC} = \begin{bmatrix} -4 \\ 1 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ -3 \end{bmatrix} = 0$, $\angle BAC$ is a right angle.

33. If we take the cube to be a unit cube (as in Figure 1.34), the four diagonals are given by the vectors

$$\mathbf{d}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{d}_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \mathbf{d}_3 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{d}_4 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

Since $\mathbf{d}_i \cdot \mathbf{d}_j \neq 0$ for all $i \neq j$ (six possibilities), no two diagonals are perpendicular.
35. $D = (-2, 1, 1)$
37. 5 mi/h at an angle of ≈ 53.13° to the bank
39. 60°

41. $\begin{bmatrix} -\frac{3}{5} \\ \frac{4}{5} \end{bmatrix}$ **43.** $\begin{bmatrix} \frac{3}{2} \\ -\frac{3}{2} \\ \frac{3}{2} \\ -\frac{3}{2} \end{bmatrix}$ **45.** $\begin{bmatrix} -0.301 \\ 0.033 \\ -0.252 \end{bmatrix}$

47. $\mathcal{A} = \sqrt{45}/2$ **49.** $k = -2, 3$

51. $\mathbf{v}$ is of the form $k\begin{bmatrix} b \\ -a \end{bmatrix}$, where k is a scalar.

53. The Cauchy-Schwarz Inequality would be violated.

Exercises 1.3

1. **(a)** $\begin{bmatrix} 3 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 0$ **(b)** $3x + 2y = 0$

3. **(a)** $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + t\begin{bmatrix} -1 \\ 3 \end{bmatrix}$
(b) $x = 1 - t$
$y = \quad 3t$

5. **(a)** $\begin{bmatrix} x \\ y \\ z \end{bmatrix} = t\begin{bmatrix} 1 \\ -1 \\ 4 \end{bmatrix}$ **(b)** $x = \quad t$
$y = -t$
$z = \quad 4t$

7. **(a)** $\begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 2$ **(b)** $3x + 2y + z = 2$

9. **(a)** $\begin{bmatrix} x \\ y \\ z \end{bmatrix} = s\begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} + t\begin{bmatrix} -3 \\ 2 \\ 1 \end{bmatrix}$
(b) $x = 2s - 3t$
$y = \quad s + 2t$
$z = 2s + \quad t$

11. $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} + t\begin{bmatrix} 2 \\ 2 \end{bmatrix}$

13. $\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + s\begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} + t\begin{bmatrix} -1 \\ 0 \\ -2 \end{bmatrix}$

15. **(a)** $x = \quad t$
$y = -1 + 3t$
$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} + t\begin{bmatrix} 1 \\ 3 \end{bmatrix}$

17. Direction vectors for the two lines are given by $\mathbf{d}_1 = \begin{bmatrix} 1 \\ m_1 \end{bmatrix}$ and $\mathbf{d}_2 = \begin{bmatrix} 1 \\ m_2 \end{bmatrix}$. The lines are perpendicular if and only if $\mathbf{d}_1$ and $\mathbf{d}_2$ are orthogonal. But $\mathbf{d}_1 \cdot \mathbf{d}_2 = 0$ if and only if $1 + m_1 m_2 = 0$ or, equivalently, $m_1 m_2 = -1$.

19. **(a)** Perpendicular **(b)** Parallel
(c) Perpendicular **(d)** Perpendicular

21. $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} + t\begin{bmatrix} 3 \\ 2 \end{bmatrix}$

23. $\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 3 \end{bmatrix} + t\begin{bmatrix} -1 \\ 3 \\ -1 \end{bmatrix}$

25. **(a)** $x = 0, x = 1, y = 0, y = 1, z = 0, z = 1$
(b) $x - y = 0$ **(c)** $x + y - z = 0$

27. $3\sqrt{2}/2$ **29.** $2\sqrt{3}/3$ **31.** $(\frac{1}{2}, \frac{1}{2})$
33. $(\frac{4}{3}, \frac{4}{3}, \frac{8}{3})$ **35.** $18\sqrt{13}/13$ **37.** $\frac{5}{3}$
43. ≈78.9° **45.** ≈80.4°

Exercises 1.4

1. 13 N at approx N 67.38 E
3. $8\sqrt{3}$ N at an angle of 30° to $\mathbf{f}_1$
5. 4 N at an angle of 60° to $\mathbf{f}_2$
7. 5 N at an angle of 60° to the given force, $5\sqrt{3}$ N perpendicular to the 5 N force
9. $750\sqrt{2}$ N
11. 980 N
13. ≈ 117.6 N in the 15 cm wire, ≈ 88.2 N in the 20 cm wire

Review Questions

1. (a) T (c) F (e) T (g) F (i) T

3. $\mathbf{x} = \begin{bmatrix} 8 \\ 11 \end{bmatrix}$ **5.** 120° **7.** $\begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \\ 0 \end{bmatrix}$

9. $2x + 3y - z = 7$ **11.** $\sqrt{6}/2$

13. The Cauchy-Schwarz Inequality would be violated.

15. $2\sqrt{6}/3$ **17.** $x = 2$ **19.** 3

Chapter 2

Exercises 2.1

1. Linear

3. Not linear because of the x^{-1} term

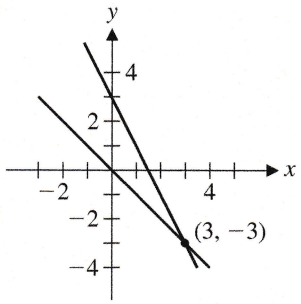
5. Not linear

7. $2x + 4y = 7$

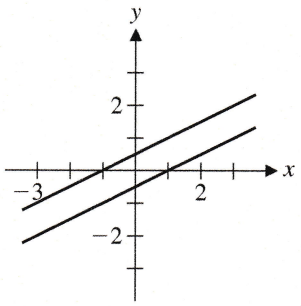
9. $x + y = 4 (x, y \neq 0)$

11. $\left\{ \begin{bmatrix} 2t \\ t \end{bmatrix} \right\}$ **13.** $\left\{ \begin{bmatrix} 4 - 2s - 3t \\ s \\ t \end{bmatrix} \right\}$

15. Unique solution, $x = 3, y = -3$



17. No solution



19. [7, 3]

21. $[\frac{2}{3}, \frac{1}{3}, -\frac{1}{3}]$

23. [5, −2, 1, 1]

25. [2, −7, −32]

27. $\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 2 & 1 & 3 \end{array}\right]$

29. $\left[\begin{array}{cc|c} 1 & 5 & -1 \\ -1 & 1 & -5 \\ 2 & 4 & 4 \end{array}\right]$

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

33. [1, 1]

35. [4, −1]

37. No solution

39. (a) $\begin{aligned} 2x + y &= 3 \\ 4x + 2y &= 6 \end{aligned}$ (b) $\begin{aligned} x &= \tfrac{3}{2} - \tfrac{1}{2}s \\ y &= s \end{aligned}$

41. Let $u = \dfrac{1}{x}$ and $v = \dfrac{1}{y}$. The solution is $x = \frac{1}{3}, y = -\frac{1}{2}$.

43. Let $u = \tan x$, $v = \sin y$, $w = \cos z$. One solution is $x = \pi/4, y = -\pi/6, z = \pi/3$. (There are infinitely many solutions.)

Exercises 2.2

1. No

3. Reduced row echelon form

5. No

7. No

9. (a) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ **11.** (b) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$

13. (b) $\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$

15. Perform elementary row operations in the order $R_4 + 29R_3$, $8R_3$, $R_4 - 3R_2$, $R_2 \leftrightarrow R_3$, $R_4 - R_1$, $R_3 + 2R_1$, and, finally, $R_2 + 2R_1$.

17. One possibility is to perform elementary row operations on A in the order $R_2 - 3R_1$, $\frac{1}{2}R_2$, $R_1 + 2R_2$, $R_2 + 3R_1$, $R_1 \leftrightarrow R_2$.

19. *Hint:* Pick a random 2×2 matrix and try this—carefully!

21. This is really two elementary row operations combined: $3R_2$ and $R_2 - 2R_1$.

23. Exercise 1: 3; Exercise 3: 2; Exercise 5: 2; Exercise 7: 3

25. $\begin{bmatrix} 2 \\ 5 \\ 1 \end{bmatrix}$ **27.** $t\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ **29.** $\begin{bmatrix} 2 \\ -1 \end{bmatrix}$

31. $\begin{bmatrix} 24 \\ -10 \\ 0 \\ 0 \end{bmatrix} + r\begin{bmatrix} 6 \\ -2 \\ 1 \\ 0 \end{bmatrix} + s\begin{bmatrix} 0 \\ 6 \\ 0 \\ 1 \end{bmatrix} + t\begin{bmatrix} 12 \\ -6 \\ 0 \\ 0 \end{bmatrix}$

33. No solution

35. Unique solution

37. Infinitely many solutions

39. *Hint:* Show that if $ad - bc \neq 0$, the rank of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is 2. (There are two cases: $a = 0$ and $a \neq 0$.) Use the Rank Theorem to deduce that the given system must have a unique solution.

41. **(a)** No solution if $k = -1$
(b) A unique solution if $k \neq \pm 1$
(c) Infinitely many solutions if $k = 1$

43. **(a)** No solution if $k = 1$
(b) A unique solution if $k \neq -2, 1$
(c) Infinitely many solutions if $k = -2$

45. $\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} + t \begin{bmatrix} 9 \\ -10 \\ -7 \end{bmatrix}$

49. No intersection

51. The required vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ are the solutions of the homogeneous system with augmented matrix

$$\begin{bmatrix} u_1 & u_2 & u_3 & 0 \\ v_1 & v_2 & v_3 & 0 \end{bmatrix}$$

By Theorem 3, there are infinitely many solutions. If $u_1 \neq 0$ and $u_1v_2 - u_2v_1 \neq 0$, the solutions are given by

$$t \begin{bmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{bmatrix}$$

But a direct check shows that these are still solutions even if $u_1 = 0$ and/or $u_1v_2 - u_2v_1 = 0$.

53. $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$ **55.** $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ **57.** $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$

Exercises 2.3

1. Yes **3.** No **5.** Yes **7.** Yes

9. We need to show that the vector equation $x \begin{bmatrix} 1 \\ 1 \end{bmatrix} + y \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$ has a solution for all values of a and b. This vector equation is equivalent to the linear system whose augmented matrix is $\left[\begin{array}{cc|c} 1 & 1 & a \\ 1 & -1 & b \end{array}\right]$. Row reduction yields $\left[\begin{array}{cc|c} 1 & 1 & a \\ 0 & -2 & b - a \end{array}\right]$, from which we can see that there is a (unique) solution. [Further row operations yield $x = (a + b)/2$, $y = (a - b)/2$.] Hence, $\mathbb{R}^2 = \text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)$.

11. We need to show that the vector equation $x \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + y \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + z \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ has a solution for all values of a, b, and c. This vector equation is equivalent to the linear system whose augmented matrix is $\left[\begin{array}{ccc|c} 1 & 1 & 0 & a \\ 0 & 1 & 1 & b \\ 1 & 0 & 1 & c \end{array}\right]$. Row reduction yields $\left[\begin{array}{ccc|c} 1 & 1 & 0 & a \\ 0 & 1 & 1 & b \\ 0 & 0 & 2 & b + c - a \end{array}\right]$, from which we can see that there is a (unique) solution. [Further row operations yield $x = (a - b + c)/2$, $y = (a + b - c)/2$, $z = (-a + b + c)/2$.] Hence, $\mathbb{R}^3 = \text{span}\left(\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\right)$.

13. **(a)** The line through the origin with direction vector $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$
(b) The line with general equation $2x + y = 0$

15. **(a)** The plane through the origin with direction vectors $\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ -1 \end{bmatrix}$
(b) The plane with general equation $2x - y + 4z = 0$

17. Substitution yields the linear system

$$\begin{aligned} a \quad\quad + 3c &= 0 \\ -a + b - 3c &= 0 \end{aligned}$$

whose solution is $t \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}$. It follows that there are infinitely many solutions, the simplest perhaps being $a = -3$, $b = 0$, $c = 1$.

19. $\mathbf{u} = \mathbf{u} + 0(\mathbf{u} + \mathbf{v}) + 0(\mathbf{u} + \mathbf{v} + \mathbf{w})$
$\mathbf{v} = (-1)\mathbf{u} + (\mathbf{u} + \mathbf{v}) + 0(\mathbf{u} + \mathbf{v} + \mathbf{w})$
$\mathbf{w} = 0\mathbf{u} + (-1)(\mathbf{u} + \mathbf{v}) + (\mathbf{u} + \mathbf{v} + \mathbf{w})$

21. (c) We must show that $\text{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = \text{span}(\mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)$. We know that $\text{span}(\mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3) \subseteq \mathbb{R}^3 = \text{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. From Exercise 19, $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ all belong to $\text{span}(\mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)$. Therefore, by Exercise 21(b), $\text{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = \text{span}(\mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)$.

23. Linearly independent

25. Linearly dependent, $-\begin{bmatrix}0\\1\\2\end{bmatrix} + \begin{bmatrix}2\\1\\3\end{bmatrix} = \begin{bmatrix}2\\0\\1\end{bmatrix}$

27. Linearly dependent, since the set contains the zero vector

29. Linearly independent

31. Linearly dependent, $\begin{bmatrix}3\\-1\\1\\-1\end{bmatrix} + \begin{bmatrix}-1\\3\\1\\-1\end{bmatrix} + \begin{bmatrix}-1\\-1\\1\\3\end{bmatrix} = \begin{bmatrix}1\\1\\3\\1\end{bmatrix}$

43. (a) Yes **(b)** No

Exercises 2.4

1. $x_1 = 160, x_2 = 120, x_3 = 160$

3. two small, three medium, four large

5. 65 bags of house blend, 30 bags of special blend, 45 bags of gourmet blend

7. $4FeS_2 + 11O_2 \longrightarrow 2Fe_2O_3 + 8SO_2$

9. $2C_4H_{10} + 13O_2 \longrightarrow 8CO_2 + 10H_2O$

11. $2C_5H_{11}OH + 15O_2 \longrightarrow 12H_2O + 10CO_2$

13. $Na_2CO_3 + 4C + N_2 \longrightarrow 2NaCN + 3CO$

15. (a) $f_1 = 30 - t$, $f_2 = -10 + t$, $f_3 = t$ **(b)** $f_1 = 15, f_3 = 15$

(c) $0 \le f_1 \le 20$, $0 \le f_2 \le 20$, $10 \le f_3 \le 30$

(d) Negative flow would mean that water was flowing backward, against the direction of the arrow.

17. (a) $f_1 = -200 + s + t$, $f_2 = 300 - s - t$, $f_3 = s$, $f_4 = 150 - t$, $f_5 = t$ **(b)** $200 \le f_3 \le 300$

(c) If $f_3 = s = 0$, then $f_5 = t \ge 200$ (from the f_1 equation), but $f_5 = t \le 150$ (from the f_4 equation). This is a contradiction.

(d) $50 \le f_3 \le 300$

19. $I_1 = 3$ amps, $I_2 = 5$ amps, $I_3 = 2$ amps

21. (a) $I = 10$ amps, $I_1 = I_5 = 6$ amps, $I_2 = I_4 = 4$ amps, $I_3 = 2$ amps

(b) $R_{eff} = \frac{7}{5}$ ohms

(c) Yes; change it to 4 ohms.

23. Farming : Manufacturing = 2 : 3

25. The painter charges \$39/hr, the plumber \$42/hr, the electrician \$54/hr.

27. (a) Coal should produce \$100 million and steel \$160 million.

(b) Coal should reduce production by $\approx$ \$4.2 million and steel should increase production by $\approx$ \$5.7 million.

29. (a) Yes; push switches 1, 2, and 3 or switches 3, 4, and 5.

(b) No

31. The states that can be obtained are represented by those vectors

$$\begin{bmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{bmatrix}$$

in $\mathbb{Z}_2^5$ for which $x_1 + x_2 + x_4 + x_5 = 0$. (There are 16 such possibilities.)

33. If 0 = off, 1 = light blue, and 2 = dark blue, then the linear system that arises has augmented matrix

$$\left[\begin{array}{ccccc|c}1&1&0&0&0&2\\1&1&1&0&0&1\\0&1&1&1&0&2\\0&0&1&1&1&1\\0&0&0&1&1&2\end{array}\right]$$

which reduces over $\mathbb{Z}_3$ to

$$\left[\begin{array}{ccccc|c}1&0&0&0&1&1\\0&1&0&0&2&1\\0&0&1&0&0&2\\0&0&0&1&1&2\\0&0&0&0&0&0\end{array}\right]$$

This yields the solutions

$$\begin{bmatrix}x_1\\x_2\\x_3\\x_4\\x_5\end{bmatrix} = \begin{bmatrix}1\\1\\2\\2\\0\end{bmatrix} + t\begin{bmatrix}2\\1\\0\\2\\1\end{bmatrix}$$

where t is in $\mathbb{Z}_3$. Hence, there are exactly three solutions:

$$\begin{bmatrix}1\\1\\2\\2\\0\end{bmatrix}, \begin{bmatrix}0\\2\\2\\1\\1\end{bmatrix}, \begin{bmatrix}2\\0\\2\\0\\2\end{bmatrix}$$

where each entry indicates the number of times the corresponding switch should be pushed.

35. **(a)** Push squares 3 and 7.
(b) The 9×9 coefficient matrix A is row equivalent to $\mathbb{Z}_2$, so for any $\mathbf{b}$ in $\mathbb{Z}_2^9$, $A\mathbf{x} = \mathbf{b}$ has a unique solution.

37. Grace is 15, and Hans is 5.
39. 1200 and 600 square yards
41. **(a)** $a = 4 - d, b = 5 - d, c = -2 + d, d$ is arbitrary
(b) No solution
43. **(a)** No solution
(b) $[a, b, c, d, e, f] = [4, 5, 6, -3, -1, 0] + f[-1, -1, -1, 1, 1, 1]$
45. **(a)** $y = x^2 - 2x + 1$ **(b)** $y = x^2 + 6x + 10$
47. $A = 1, B = 2$
49. $A = -\frac{1}{5}, B = \frac{1}{3}, C = 0, D = -\frac{2}{15}, E = -\frac{1}{5}$
51. $a = \frac{1}{2}, b = \frac{1}{2}, c = 0$

Exercises 2.5

1.

n	0	1	2	3	4	5
x_1	0	0.8571	0.9714	0.9959	0.9991	0.9998
x_2	0	0.8000	0.9714	0.9943	0.9992	0.9998

Exact solution: $x_1 = 1, x_2 = 1$

3.

n	0	1	2	3	4	5	6
x_1	0	0.2222	0.2539	0.2610	0.2620	0.2622	0.2623
x_2	0	0.2857	0.3492	0.3582	0.3603	0.3606	0.3606

Exact solution (to four decimal places): $x_1 = 0.2623$, $x_2 = 0.3606$

5.

n	0	1	2	3	4	5	6	7	8
x_1	0	0.3333	0.2500	0.3055	0.2916	0.3009	0.2986	0.3001	0.2997
x_2	0	0.2500	0.0834	0.1250	0.0972	0.1042	0.0996	0.1008	0.1000
x_3	0	0.3333	0.2500	0.3055	0.2916	0.3009	0.2986	0.3001	0.2997

Exact solution: $x_1 = 0.3, x_2 = 0.1, x_3 = 0.3$

7.

n	0	1	2	3	4
x_1	0	0.8571	0.9959	0.9998	1.0000
x_2	0	0.9714	0.9992	1.0000	1.0000

After three iterations, the Gauss-Seidel method is within 0.001 of the exact solution. Jacobi's method took four iterations to reach the same accuracy.

9.

n	0	1	2	3	4
x_1	0	0.2222	0.2610	0.2622	0.2623
x_2	0	0.3492	0.3603	0.3606	0.3606

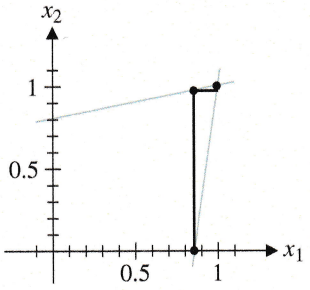
After three iterations, the Gauss-Seidel method is within 0.001 of the exact solution. Jacobi's method took four iterations to reach the same accuracy.

11.

n	0	1	2	3	4	5	6
x_1	0	0.3333	0.2777	0.2962	0.2993	0.2998	0.3000
x_2	0	0.1667	0.1112	0.1020	0.1004	0.1000	0.1000
x_3	0	0.2777	0.2962	0.2993	0.2998	0.3000	0.3000

After four iterations, the Gauss-Seidel method is within 0.001 of the exact solution. Jacobi's method took seven iterations to reach the same accuracy.

13.



15.

n	0	1	2	3	4
x_1	0	3	−5	19	−53
x_2	0	−4	8	−28	80

If the equations are interchanged and the Gauss-Seidel method is applied to the equivalent system

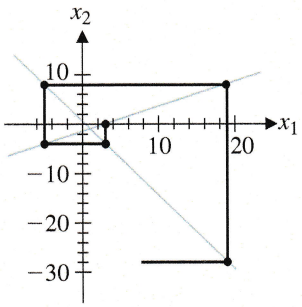
$$3x_1 + 2x_2 = 1$$
$$x_1 - 2x_2 = 3$$

we obtain

n	0	1	2	3	4	5	6	7	8
x_1	0	0.3333	1.2222	0.9260	1.0247	0.9918	1.0027	0.9991	1.0003
x_2	0	−1.3333	−0.8889	−1.0370	−0.9876	−1.0041	−0.9986	−1.0004	−0.9998

After seven iterations, the process has converged to within 0.001 of the exact solution $x_1 = 1, x_2 = -1$.

17.



19.

n	0	1	2	3	4	5	6
x_1	0	−1.6	14.97	8.550	10.740	9.839	10.120
x_2	0	25.9	11.408	14.051	11.615	11.718	11.249
x_3	0	−10.35	−9.311	−11.200	−11.322	−11.721	−11.816

n	7	8	9	10	11	12
x_1	9.989	10.022	10.002	10.005	10.001	10.001
x_2	11.187	11.082	11.052	11.026	11.015	11.008
x_3	−11.912	−11.948	−11.973	−11.985	−11.992	−11.996

After 12 iterations, the Gauss-Seidel method has converged to within 0.01 of the exact solution $x_1 = 10, x_2 = 11, x_3 = -12$.

21.

n	13	14	15	16
x_1	10.0004	10.0003	10.0001	10.0001
x_2	11.0043	11.0023	11.0014	11.0007
x_3	−11.9976	−11.9986	−11.9993	−11.9996

23. The Gauss-Seidel method produces

n	0	1	2	3	4	5	6	7	8	9
x_1	0	0	12.5	21.875	24.219	24.805	24.951	24.988	24.997	24.999
x_2	0	0	18.75	21.438	24.609	24.902	24.976	24.994	24.998	24.999
x_3	0	50	68.75	73.438	74.609	74.902	74.976	74.994	74.998	74.999
x_4	0	62.5	71.875	74.219	74.805	74.951	74.988	74.997	74.999	75.000

The exact solution is $x_1 = 25$, $x_2 = 25$, $x_3 = 75$, $x_4 = 75$.

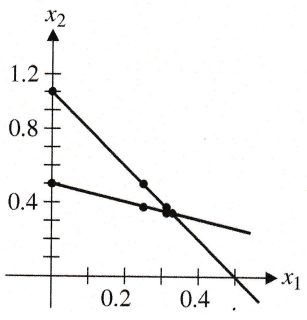
25. The Gauss-Seidel method produces the following iterates:

n	0	1	2	3	4	5	6
t_1	0	20	21.25	22.8125	23.3301	23.6596	23.7732
t_2	0	5	11.25	13.3203	14.6386	15.0926	15.2732
t_3	0	21.25	24.6094	26.9873	27.7303	27.9626	28.0352
t_4	0	2.5	5.8594	8.2373	8.9804	9.2126	9.2852
t_5	0	7.1875	14.6289	16.2829	16.7578	16.9036	16.9491
t_6	0	23.0469	24.9072	25.3207	25.4394	25.4759	25.4873

n	7	8	9	10	11	12
t_1	23.8093	23.8206	23.8242	23.8252	23.8256	23.8257
t_2	15.2824	15.2966	15.3010	15.3024	15.3029	15.3029
t_3	28.0579	28.0650	28.0671	28.0678	28.0681	28.0681
t_4	9.3079	9.3150	9.3172	9.3178	9.3181	9.3181
t_5	16.9633	16.9677	16.9690	16.9695	16.9696	16.9696
t_6	25.4908	25.4919	25.4922	25.4924	25.4924	25.4924

27. (a)

n	0	1	2	3	4	5	6
x_1	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{5}{16}$	$\frac{5}{16}$	$\frac{21}{64}$
x_2	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{11}{32}$	$\frac{11}{32}$



(b) $2x_1 + x_2 = 1$
$x_1 + 2x_2 = 1$

(c)

n	0	1	2	3	4	5	6	7
x_1	0	0	0.25	0.3125	0.3281	0.3320	0.3330	0.3332
x_2	1	0.5	0.375	0.3438	0.3360	0.3340	0.3335	0.3334

[Columns 1, 2, and 3 of this table are the *odd-numbered* columns 1, 3, and 5 from the table in part (a).] The iterates are converging to $x_1 = x_2 = 0.3333$.

(d) $x_1 = x_2 = \frac{1}{3}$

Review Questions

1. **(a)** F **(c)** F **(e)** T **(g)** T **(i)** F

3. $\begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}$ **5.** $\begin{bmatrix} 6 \\ 2 \end{bmatrix}$ **7.** $k = -1$ **9.** (0, 3, 1)

11. $x - 2y + z = 0$ **13.** **(a)** Yes **15.** 1 or 2

17. If $c_1(\mathbf{u} + \mathbf{v}) + c_2(\mathbf{u} - \mathbf{v}) = \mathbf{0}$, then $(c_1 + c_2)\mathbf{u} + (c_1 - c_2)\mathbf{v} = \mathbf{0}$. Linear independence of $\mathbf{u}$ and $\mathbf{v}$ implies $c_1 + c_2 = 0$ and $c_1 - c_2 = 0$. Solving this system, we get $c_1 = c_2 = 0$. Hence $\mathbf{u} + \mathbf{v}$ and $\mathbf{u} - \mathbf{v}$ are linearly independent.

19. Their ranks must be equal.

Chapter 3

Exercises 3.1

1. $\begin{bmatrix} 3 & -6 \\ -5 & 7 \end{bmatrix}$ **3.** Not possible

5. $\begin{bmatrix} 12 & -6 & 3 \\ -4 & 12 & 14 \end{bmatrix}$ **7.** $\begin{bmatrix} 3 & 3 \\ 19 & 27 \end{bmatrix}$

9. [10] **11.** $\begin{bmatrix} -4 & -2 \\ 8 & 4 \end{bmatrix}$

13. $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ **15.** $\begin{bmatrix} 27 & 0 \\ -49 & 125 \end{bmatrix}$

17. $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$

19. $B = \begin{bmatrix} 1.50 & 1.00 & 2.00 \\ 1.75 & 1.50 & 1.00 \end{bmatrix}$, $BA = \begin{bmatrix} 650.00 & 462.50 \\ 675.00 & 406.25 \end{bmatrix}$

Column i corresponds to warehouse i, row 1 contains the costs of shipping by truck, and row 2 contains the costs of shipping by train.

21. $\begin{bmatrix} 1 & -2 & 3 \\ 2 & 1 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}$

23. $AB = [2\mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_3 \quad 3\mathbf{a}_1 - \mathbf{a}_2 + 6\mathbf{a}_3 \quad \mathbf{a}_2 + 4\mathbf{a}_3]$ (where $\mathbf{a}_i$ is the ith column of A)

25. $\begin{bmatrix} 2 & 3 & 0 \\ -6 & -9 & 0 \\ 4 & 6 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 2 & -12 & -8 \\ -1 & 6 & 4 \\ 1 & -6 & -4 \end{bmatrix}$

27. $BA = \begin{bmatrix} 2\mathbf{A}_1 + 3\mathbf{A}_2 \\ \mathbf{A}_1 - \mathbf{A}_2 + \mathbf{A}_3 \\ -\mathbf{A}_1 + 6\mathbf{A}_2 + 4\mathbf{A}_3 \end{bmatrix}$ (where $\mathbf{A}_i$ is the ith row of A)

29. If $\mathbf{b}_i$ is the ith column of B, then $A\mathbf{b}_i$ is the ith column of AB. If the columns of B are linearly dependent, then there are scalars $c_1, \ldots, c_n$ (not all zero) such that $c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n = \mathbf{0}$. But then $c_1(A\mathbf{b}_1) + \cdots + c_n(A\mathbf{b}_n) = A(c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n) = A\mathbf{0} = \mathbf{0}$, so the columns of AB are linearly dependent.

31. $\left[\begin{array}{cc|c} 3 & 2 & 0 \\ -1 & 1 & 0 \\ \hline 0 & 0 & 5 \end{array}\right]$ **33.** $\left[\begin{array}{cc|c} 1 & 2 & 2 & 0 \\ 3 & 4 & 5 & 3 \\ \hline 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & -1 \end{array}\right]$

35. **(a)** $A^2 = \begin{bmatrix} -1 & 1 \\ -1 & 0 \end{bmatrix}$, $A^3 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$,
$A^4 = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}$, $A^5 = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$, $A^6 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,
$A^7 = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$

(b) $A^{2001} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$

37. $A^n = \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix}$

39. **(a)** $\begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}$ **(c)** $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 2 & 4 & 8 & 16 \\ 3 & 9 & 27 & 81 \end{bmatrix}$

Exercises 3.2

1. $X = \begin{bmatrix} 5 & 4 \\ 3 & 5 \end{bmatrix}$ **3.** $X = \begin{bmatrix} -\frac{2}{3} & \frac{4}{3} \\ \frac{10}{3} & 4 \end{bmatrix}$

5. $B = 2A_1 + A_2$ **7.** Not possible

9. $\text{span}(A_1, A_2) = \left\{ \begin{bmatrix} c_1 & 2c_1 + c_2 \\ -c_1 + 2c_2 & c_1 + c_2 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} w & x \\ 2x - 5w & x - w \end{bmatrix} \right\}$

11. $\text{span}(A_1, A_2, A_3) = \left\{ \begin{bmatrix} c_1 - c_2 + c_3 & 2c_2 + c_3 & -c_1 + c_3 \\ 0 & c_1 + c_2 & 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} -3b + 4c + 5e & b & c \\ 0 & e & 0 \end{bmatrix} \right\}$

13. Linearly independent **15.** Linearly independent

23. $a = d, c = 0$ **25.** $3b = 2c, a = d - c$

27. $a = d, b = c = 0$

29. Let $A = [a_{ij}]$ and $B = [b_{ij}]$ be upper triangular $n \times n$ matrices and let $i > j$. Then, by the definition of an upper triangular matrix,

$a_{i1} = a_{i2} = \cdots = a_{i,i-1} = 0$ and
$b_{ij} = b_{i+1,j} = \cdots = b_{nj} = 0$
Now let $C = AB$. Then

$$\begin{aligned} c_{ij} &= a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots a_{i,i-1}b_{i-1,j} + a_{ii}b_{ij} \\ &\quad + a_{i,i+1}b_{i+1,j} + \cdots + a_{in}b_{nj} \\ &= 0 \cdot b_{1j} + 0 \cdot b_{2j} + \cdots 0 \cdot b_{i-1,j} + a_{ii} \cdot 0 \\ &\quad + a_{i,i+1} \cdot 0 + \cdots + a_{in} \cdot 0 = 0 \end{aligned}$$

from which it follows that C is upper triangular.

35. (a) A, B symmetric $\Rightarrow (A + B)^T = A^T + B^T = A + B \Rightarrow A + B$ is symmetric

37. Matrices (b) and (c) are skew-symmetric.

41. Either A or B (or both) must be the zero matrix.

43. (b) $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 5 & 7 \\ 5 & 7 & 9 \end{bmatrix} + \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix}$

47. *Hint:* Use the trace.

Exercises 3.3

1. $\begin{bmatrix} 2 & -7 \\ -1 & 4 \end{bmatrix}$ **3.** Not invertible

5. Not invertible **7.** $\begin{bmatrix} -1.6 & -2.8 \\ 0.\overline{3} & 1 \end{bmatrix}$

9. $\begin{bmatrix} a/(a^2 + b^2) & b/(a^2 + b^2) \\ -b/(a^2 + b^2) & a/(a^2 + b^2) \end{bmatrix}$ **11.** $\begin{bmatrix} -5 \\ 9 \end{bmatrix}$

13. (a) $\mathbf{x}_1 = \begin{bmatrix} 4 \\ -\frac{1}{2} \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -5 \\ 2 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 6 \\ -2 \end{bmatrix}$

(c) The method in part (b) uses fewer multiplications.

17. (b) $(AB)^{-1} = A^{-1}B^{-1}$ if and only if $AB = BA$

21. $X = A^{-1}(BA)^2B^{-1}$

23. $X = (AB)^{-1}BA + A$

25. $E = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ **27.** $E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$

29. $E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$ **31.** $\begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix}$

33. $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ **35.** $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$ **37.** $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/c & 0 \\ 0 & 0 & 1 \end{bmatrix}$

39. $A = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$

43. (a) If A is invertible, then $BA = CA \Rightarrow (BA)A^{-1} = (CA)A^{-1} \Rightarrow B(AA^{-1}) = C(AA^{-1}) \Rightarrow BI = CI \Rightarrow B = C$.

45. *Hint:* Rewrite $A^2 - 2A + I = O$ as $A(2I - A) = I$.

47. If AB is invertible, then there exists a matrix X such that $(AB)X = I$. But then $A(BX) = I$ too, so A is invertible (with inverse BX).

49. $\begin{bmatrix} \frac{1}{10} & \frac{2}{5} \\ \frac{3}{10} & \frac{1}{5} \end{bmatrix}$ **51.** $\begin{bmatrix} 1/(a^2 + 1) & -a/(a^2 + 1) \\ a/(a^2 + 1) & 1/(a^2 + 1) \end{bmatrix}$

53. Not invertible

55. $\begin{bmatrix} 1/a & 0 & 0 \\ -1/a^2 & 1/a & 0 \\ 1/a^3 & -1/a^2 & 1/a \end{bmatrix}, a \neq 0$

57. $\begin{bmatrix} -11 & -2 & 5 & -4 \\ 4 & 1 & -2 & 2 \\ 5 & 1 & -2 & 2 \\ 9 & 2 & -4 & 3 \end{bmatrix}$

59. $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -a/d & -b/d & -c/d & 1/d \end{bmatrix}, d \neq 0$

61. Not invertible **63.** $\begin{bmatrix} 4 & 6 & 4 \\ 5 & 3 & 2 \\ 0 & 6 & 5 \end{bmatrix}$

69. $\left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline -2 & -3 & 1 & 0 \\ -1 & -2 & 0 & 1 \end{array}\right]$ **71.** $\left[\begin{array}{cc|cc} -1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{array}\right]$

Exercises 3.4

1. $\begin{bmatrix} -2 \\ 1 \end{bmatrix}$ **3.** $\begin{bmatrix} -3/2 \\ -2 \\ -1 \end{bmatrix}$ **5.** $\begin{bmatrix} -7 \\ -15 \\ -2 \\ 2 \end{bmatrix}$.

7. $\begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 0 & 5 \end{bmatrix}$

9. $\begin{bmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 8 & 3 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 3 \end{bmatrix}$

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

13. $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 1 & -2 \\ 0 & 3 & 3 & 1 \\ 0 & 0 & 0 & 5 \end{bmatrix}$

15. $L^{-1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, $U^{-1} = \begin{bmatrix} -\frac{1}{2} & \frac{1}{12} \\ 0 & \frac{1}{6} \end{bmatrix}$,

$A^{-1} = \begin{bmatrix} -5/12 & 1/12 \\ 1/6 & 1/6 \end{bmatrix}$

19. $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

21. $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

23. $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 5 & 1 \end{bmatrix}\begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 4 \\ 0 & 0 & -16 \end{bmatrix}$

25. $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 1 & 1 & 2 \\ 0 & -1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}$

27. $\begin{bmatrix} 4 \\ -1 \\ -2 \end{bmatrix}$ **31.** $\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} -2 & 0 \\ 0 & 6 \end{bmatrix}\begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & 1 \end{bmatrix}$

Exercises 3.5

1. Subspace **3.** Subspace

5. Subspace **7.** Not a subspace

11. **b** is in col(A), **w** is not in row(A).

15. No

17. $\{[1 \quad 0 \quad -1], [0 \quad 1 \quad 2]\}$ is a basis for row(A); $\left\{\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}$ is a basis for col(A); $\left\{\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}\right\}$ is a basis for null(A).

19. $\{[1 \quad 0 \quad 1 \quad 0], [0 \quad 1 \quad -1 \quad 0], [0 \quad 0 \quad 0 \quad 1]\}$ is a basis for row(A); $\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}\right\}$ is a basis for col(A); $\left\{\begin{bmatrix} -1 \\ 1 \\ 1 \\ 0 \end{bmatrix}\right\}$ is a basis for null(A).

21. $\{[1 \quad 0 \quad -1], [1 \quad 1 \quad 1]\}$ is a basis for row(A); $\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}$ is a basis for col(A)

23. $\{[1 \quad 1 \quad 0 \quad 1], [0 \quad 1 \quad -1 \quad 1], [0 \quad 1 \quad -1 \quad -1]\}$ is a basis for row(A); $\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right\}$ is a basis for col(A)

25. Both $\{[1 \quad 0 \quad -1], [0 \quad 1 \quad 2]\}$ and $\{[1 \quad 0 \quad -1], [1 \quad 1 \quad 1]\}$ are linearly independent spanning sets for row(A) $= \{[a \quad b \quad -a + 2b]\}$. Both $\left\{\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}$ and $\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}$ are linearly independent spanning sets for col(A) $= \mathbb{R}^2$.

27. $\left\{\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}\right\}$

29. $\{[1 \quad 0 \quad 0], [0 \quad 1 \quad 0], [0 \quad 0 \quad 1]\}$

31. $\{[2 \quad -3 \quad 1], [1 \quad -1 \quad 0], [4 \quad -4 \quad 1]\}$

35. rank(A) $= 2$, nullity(A) $= 1$

37. rank(A) $= 3$, nullity(A) $= 1$

39. If A is 3×5, then rank(A) ≤ 3, so there cannot be more than three linearly independent columns.

41. nullity(A) $= 2, 3, 4,$ or 5

43. If $a = -1$, then $\text{rank}(A) = 1$; if $a = 2$, then $\text{rank}(A) = 2$; for $a \neq -1, 2$, $\text{rank}(A) = 3$.

45. Yes **47.** Yes **49.** No

51. $\mathbf{w}$ is in $\text{span}(\mathcal{B})$ if and only if the linear system with augmented matrix $[\mathcal{B} \mid \mathbf{w}]$ is consistent, which is true in this case, since

$$[\mathcal{B} \mid \mathbf{w}] = \left[\begin{array}{ccc|c} 1 & 1 & 1 \\ 2 & 0 & 6 \\ 0 & -1 & 2 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 1 & 0 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{array}\right]$$

From this reduced row echelon form, it is also clear that $[\mathbf{w}]_{\mathcal{B}} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$.

53. $\text{rank}(A) = 2$, $\text{nullity}(A) = 1$

55. $\text{rank}(A) = 3$, $\text{nullity}(A) = 1$

57. Let $\mathbf{A}_1, \ldots, \mathbf{A}_m$ be the row vectors of A so that $\text{row}(A) = \text{span}(\mathbf{A}_1, \ldots, \mathbf{A}_m)$. If $\mathbf{x}$ is in $\text{null}(A)$, then, since $A\mathbf{x} = \mathbf{0}$, we also have $\mathbf{A}_i \cdot \mathbf{x} = 0$ for $i = 1, \ldots, m$, by the row-column definition of matrix multiplication. If $\mathbf{r}$ is in $\text{row}(A)$, then $\mathbf{r}$ is of the form $\mathbf{r} = c_1\mathbf{A}_1 + \cdots + c_m\mathbf{A}_m$. Therefore,

$$\begin{aligned} \mathbf{r} \cdot \mathbf{x} &= (c_1\mathbf{A}_1 + \cdots + c_m\mathbf{A}_m) \cdot \mathbf{x} \\ &= c_1(\mathbf{A}_1 \cdot \mathbf{x}) + \cdots + c_m(\mathbf{A}_m \cdot \mathbf{x}) = 0 \end{aligned}$$

59. (a) If a set of columns of AB is linearly independent, then the corresponding columns of B are linearly independent (by an argument similar to that needed to prove Exercise 29 in Section 3.1). It follows that the *maximum* number k of linearly independent columns of AB [i.e., $k = \text{rank}(AB)$] is not more than the *maximum* number r of linearly independent columns of B [i.e., $r = \text{rank}(B)$]. In other words, $\text{rank}(AB) \leq \text{rank}(B)$.

61. (a) From Exercise 59(a), $\text{rank}(UA) \leq \text{rank}(A)$ and $\text{rank}(A) = \text{rank}((U^{-1}U)A) = \text{rank}(U^{-1}(UA)) \leq \text{rank}(UA)$. Hence, $\text{rank}(UA) = \text{rank}(A)$.

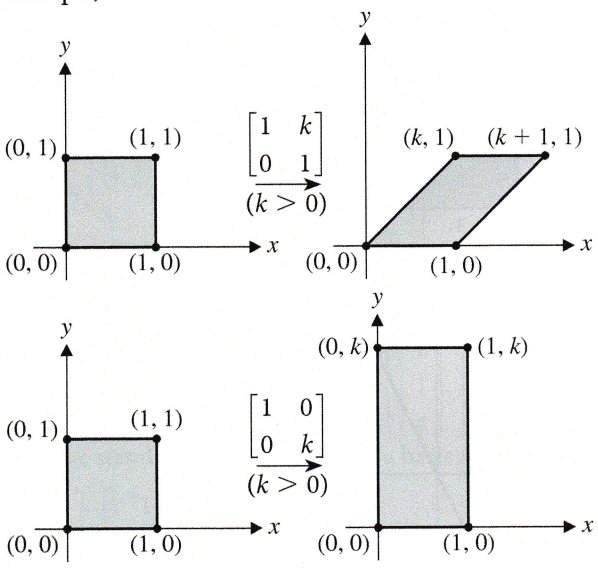
Exercises 3.6

1. $T(\mathbf{u}) = \begin{bmatrix} 0 \\ 11 \end{bmatrix}$, $T(\mathbf{v}) = \begin{bmatrix} 8 \\ 1 \end{bmatrix}$

11. $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ **13.** $\begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & -3 \end{bmatrix}$

15. $[F] = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ **17.** $[D] = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$

19. $\begin{bmatrix} k & 0 \\ 0 & 1 \end{bmatrix}$ stretches or contracts in the x-direction (combined with a reflection in the y-axis if $k < 0$); $\begin{bmatrix} 1 & 0 \\ 0 & k \end{bmatrix}$ stretches or contracts in the y-direction (combined with a reflection in the x-axis if $k < 0$); $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ is a reflection in the line $y = x$; $\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$ is a *shear* in the x-direction; $\begin{bmatrix} 1 & 0 \\ k & 1 \end{bmatrix}$ is a *shear* in the y-direction. For example,



21. $\begin{bmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{bmatrix}$ **23.** $\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$

25. $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ **27.** $\begin{bmatrix} -\frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix}$

31. $[S \circ T] = \begin{bmatrix} -8 & 5 \\ 4 & 1 \end{bmatrix}$

33. $[S \circ T] = \begin{bmatrix} 0 & 6 & -6 \\ 1 & -2 & 2 \end{bmatrix}$

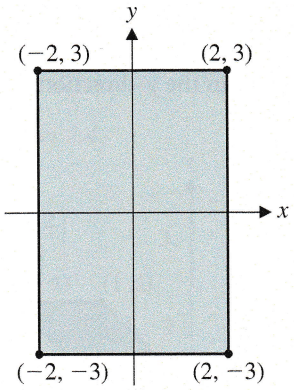
35. $[S \circ T] = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$

37. $\begin{bmatrix} -\sqrt{3}/2 & 1/2 \\ 1/2 & \sqrt{3}/2 \end{bmatrix}$ **39.** $\begin{bmatrix} -\sqrt{3}/2 & -1/2 \\ 1/2 & -\sqrt{3}/2 \end{bmatrix}$

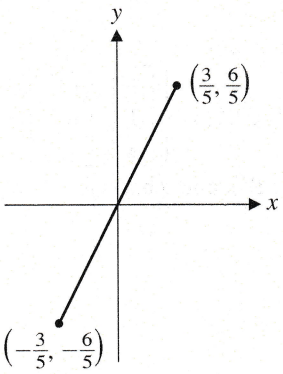
45. In vector form, let the parallel lines be given by $\mathbf{x} = \mathbf{p} + t\mathbf{d}$ and $\mathbf{x}' = \mathbf{p}' + t\mathbf{d}$. Their images are $T(\mathbf{x}) = T(\mathbf{p} + t\mathbf{d}) = T(\mathbf{p}) + tT(\mathbf{d})$ and $T(\mathbf{x}') = T(\mathbf{p}' + t\mathbf{d}) = T(\mathbf{p}') + tT(\mathbf{d})$. Suppose $T(\mathbf{d}) \neq \mathbf{0}$. If $T(\mathbf{p}') - T(\mathbf{p})$ is parallel to $T(\mathbf{d})$, then the images represent the same line; otherwise the images represent distinct parallel lines. On the other hand, if $T(\mathbf{d}) = \mathbf{0}$,

then the images represent two distinct points if $T(\mathbf{p}') \neq T(\mathbf{p})$ and single point otherwise.

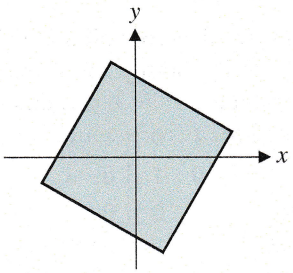
47.



49.



51.



Exercises 3.7

1. $\mathbf{x}_1 = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0.38 \\ 0.62 \end{bmatrix}$

3. 64%

5. $\mathbf{x}_1 = \begin{bmatrix} 150 \\ 120 \\ 120 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 155 \\ 120 \\ 115 \end{bmatrix}$

7. $\frac{5}{18}$

9. (a) $P = \begin{bmatrix} 0.662 & 0.250 \\ 0.338 & 0.750 \end{bmatrix}$

(b) 0.353

(c) 42.5% wet, 57.5% dry

11. (a) $P = \begin{bmatrix} 0.08 & 0.09 & 0.11 \\ 0.07 & 0.11 & 0.05 \\ 0.85 & 0.80 & 0.84 \end{bmatrix}$

(b) 0.08, 0.1062, 0.1057, 0.1057, 0.1057

(c) 10.6% good, 5.5% fair, 83.9% poor

13. The entries of the vector $\mathbf{j}P$ are just the column sums of the matrix P. So P is stochastic if and only if $\mathbf{j}P = \mathbf{j}$.

15. 4

17. 9.375

19. Yes, $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

21. No

23. No

25. Yes, $\mathbf{x} = \begin{bmatrix} 10 \\ 27 \\ 35 \end{bmatrix}$

27. Productive

29. Not productive

31. $\mathbf{x} = \begin{bmatrix} 10 \\ 16 \end{bmatrix}$

33. Yes, $\mathbf{x} = \begin{bmatrix} 10 \\ 6 \\ 8 \end{bmatrix}$

37. $\mathbf{x}_1 = \begin{bmatrix} 45 \\ 6 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 120 \\ 27 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 375 \\ 72 \end{bmatrix}$

39. $\mathbf{x}_1 = \begin{bmatrix} 500 \\ 70 \\ 50 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 720 \\ 350 \\ 35 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 1175 \\ 504 \\ 175 \end{bmatrix}$

41. (a) For L_1, we have $\mathbf{x}_1 = \begin{bmatrix} 50 \\ 8 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 40 \\ 40 \end{bmatrix}$,

$\mathbf{x}_3 = \begin{bmatrix} 200 \\ 32 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 160 \\ 160 \end{bmatrix}, \mathbf{x}_5 = \begin{bmatrix} 800 \\ 128 \end{bmatrix}, \mathbf{x}_6 = \begin{bmatrix} 640 \\ 640 \end{bmatrix}$,

$\mathbf{x}_7 = \begin{bmatrix} 3200 \\ 512 \end{bmatrix}, \mathbf{x}_8 = \begin{bmatrix} 2560 \\ 2560 \end{bmatrix}, \mathbf{x}_9 = \begin{bmatrix} 12800 \\ 2048 \end{bmatrix}$,

$\mathbf{x}_{10} = \begin{bmatrix} 10240 \\ 10240 \end{bmatrix}$.

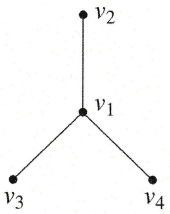
(b) The first population oscillates between two states,while the second approaches a steady state.

43. The population oscillates through a cycle of three states (for the relative population): If $0.1 < s \leq 1$, the actual population is growing; if $s = 0.1$, the actual population goes through a cycle of length 3; and if $0 \leq s < 0.1$, the actual population is declining (and will eventually die out).

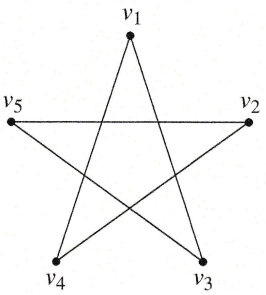
45. $A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$.

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

49.

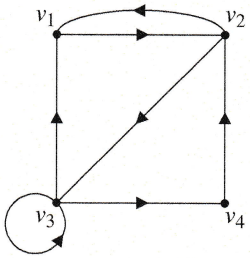


51.

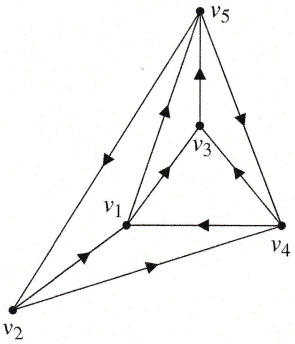


53. $A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$ **55.** $A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

57.



59.

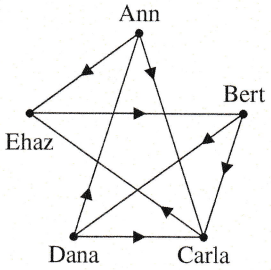


61. 2 **63.** 3 **65.** 0 **67.** 3

69. (a) Vertex i is not adjacent to any other vertices.

71. If we use direct wins only, P_2 is in first place; P_3, P_4, and P_6 tie for second place; and P_1 and P_5 tie for third place. If we combine direct and indirect wins, the players rank as follows: P_2 in first place, followed by P_6, P_4, P_3, P_5, and P_1.

73. (a)



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

(b) two steps; all of the off-diagonal entries of the second row of $A + A^2$ are nonzero.

(d) If the graph has n vertices, check the (i, j) entry of the powers A^k for $k = 1, \ldots, n-1$. Vertex i is connected to vertex j by a path of length k if and only if $(A^k)_{ij} \neq 0$.

75. $(AA^T)_{ij}$ counts the number of vertices adjacent to *both* vertex i and vertex j.

77. Bipartite **79.** Bipartite

Review Questions

1. (a) T **(c)** F **(e)** T **(g)** T **(i)** T

3. Impossible **5.** $\begin{bmatrix} \frac{17}{83} & -\frac{1}{83} \\ -\frac{1}{83} & \frac{5}{166} \end{bmatrix}$

7. $\begin{bmatrix} 1 & 3 \\ 3 & 9 \end{bmatrix} + \begin{bmatrix} 4 & 10 \\ 10 & 25 \end{bmatrix}$ **9.** $\begin{bmatrix} 0 & -9 \\ 2 & 4 \\ 1 & -6 \end{bmatrix}$

11. Because $(I - A)(I + A + A^2) = I - A^3 = I - O = I$, $(I - A)^{-1} = I + A + A^2$.

13. A basis for row(A) is $\{[1, -2, 0, -1, 0], [0, 0, 1, 2, 0], [0, 0, 0, 0, 1]\}$; a basis for col$(A)$ is $\left\{\begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 5 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 5 \\ 1 \\ 6 \end{bmatrix}\right\}$ (or the standard basis for $\mathbf{R}^3$); and a basis for null(A) is $\left\{\begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \\ 0 \end{bmatrix}\right\}$.

15. An invertible matrix has a trivial (zero) null space. If A is invertible, then so is A^T, and so both A and A^T have trivial null spaces. If A is not invertible, then A and A^T need not have the same null space. For example, take $A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$.

17. Because A has n linearly independent columns, rank$(A) = n$. Hence rank$(A^TA) = n$ by Theorem 3.28. Because A^TA is $n \times n$, this implies that A^TA is invertible, by the Fundamental Theorem of Invertible Matrices. AA^T need not be invertible. For example, take $A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

19. $\begin{bmatrix} -1/5\sqrt{2} & -3/5\sqrt{2} \\ 2/5\sqrt{2} & 6/5\sqrt{2} \end{bmatrix}$

Chapter 4

Exercises 4.1

1. $A\mathbf{v} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3\mathbf{v}$, $\lambda = 3$

3. $A\mathbf{v} = \begin{bmatrix} -3 \\ 6 \end{bmatrix} = -3\mathbf{v}, \lambda = -3$

5. $A\mathbf{v} = \begin{bmatrix} 6 \\ -3 \\ 3 \end{bmatrix} = 3\mathbf{v}, \lambda = 3$

7. $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ **9.** $\begin{bmatrix} 4 \\ 1 \end{bmatrix}$ **11.** $\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$

13. $\lambda = 1, E_1 = \text{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right); \lambda = -1, E_{-1} = \text{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$

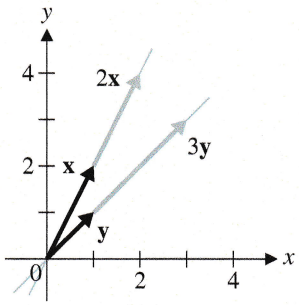
15. $\lambda = 0, E_0 = \text{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right); \lambda = 1, E_1 = \text{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$

17. $\lambda = 2, E_2 = \text{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right); \lambda = 3, E_3 = \text{span}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$

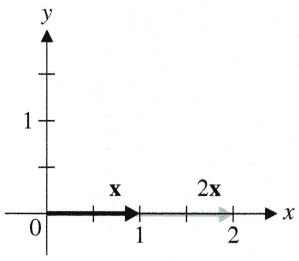
19. $\mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \lambda = 1; \mathbf{v} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \lambda = 2$

21. $\mathbf{v} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \lambda = 2; \mathbf{v} = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \lambda = 0$

23. $\lambda = 2, E_2 = \text{span}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right); \lambda = 3, E_3 = \text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)$



25. $\lambda = 2, E_2 = \text{span}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$



27. $\lambda = 1 + i, E_{1+i} = \text{span}\left(\begin{bmatrix} 1 \\ i \end{bmatrix}\right); \lambda = 1 - i, E_{1-i} = \text{span}\left(\begin{bmatrix} 1 \\ -i \end{bmatrix}\right)$

29. $\lambda = 1 + i, E_{1+i} = \text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right); \lambda = 1 - i, E_{1-i} = \text{span}\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)$

31. $\lambda = 1, 2$ **33.** $\lambda = 4$

Exercises 4.2

1. 16 **3.** 0 **5.** -18 **7.** 6
9. -12 **11.** $a^2b + ab^2$ **13.** 4 **15.** $abdg$
17. 0 **25.** 2 **27.** -24 **29.** 0
31. 0 **33.** -24 **35.** 8 **37.** -4
39. -8 **45.** $k \neq 0, 2$ **47.** -6 **49.** $-\frac{3}{2}$
51. $(-2)3^n$
53. $\det(AB) = (\det A)(\det B) = (\det B)(\det A) = \det(BA)$
55. 0, 1 **57.** $x = \frac{3}{2}, y = -\frac{1}{2}$
59. $x = -1, y = 0, z = 1$ **61.** $\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$

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

Exercises 4.3

1. (a) $\lambda^2 - 7\lambda + 12$ **(b)** $\lambda = 3, 4$
(c) $E_3 = \text{span}\left(\begin{bmatrix} 3 \\ 2 \end{bmatrix}\right); E_4 = \text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)$
(d) The algebraic and geometric multiplicities are all 1.

3. (a) $-\lambda^3 + 2\lambda^2 + 5\lambda - 6$
(b) $\lambda = -2, 1, 3$
(c) $E_{-2} = \text{span}\left(\begin{bmatrix} 1 \\ -3 \\ 0 \end{bmatrix}\right); E_1 = \text{span}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right);$
$E_3 = \text{span}\left(\begin{bmatrix} 1 \\ 2 \\ 10 \end{bmatrix}\right)$
(d) The algebraic and geometric multiplicities are all 1.

5. (a) $-\lambda^3 + \lambda^2$ **(b)** $\lambda = 0, 1$
(c) $E_0 = \text{span}\left(\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}\right); E_1 = \text{span}\left(\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right)$
(d) $\lambda = 0$ has algebraic multiplicity 2 and geometric multiplicity 1; $\lambda = 1$ has algebraic and geometric multiplicity 1.

7. (a) $-\lambda^3 + 9\lambda^2 - 27\lambda + 27$
(b) $\lambda = 3$

(c) $E_3 = \text{span}\left(\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}\right)$

(d) $\lambda = 3$ has algebraic multiplicity 3 and geometric multiplicity 2.

9. (a) $\lambda^4 - 6\lambda^3 + 9\lambda^2 + 4\lambda - 12$

(b) $\lambda = -1, 2, 3$

(c) $E_{-1} = \text{span}\left(\begin{bmatrix} 0 \\ 0 \\ -2 \\ 1 \end{bmatrix}\right); E_2 = \text{span}\left(\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}\right);$

$E_3 = \text{span}\left(\begin{bmatrix} 0 \\ 0 \\ 2 \\ 1 \end{bmatrix}\right)$

(d) $\lambda = -1$ and $\lambda = 3$ have algebraic and geometric multiplicity 1; $\lambda = 2$ has algebraic multiplicity 2 and geometric multiplicity 1.

11. (a) $\lambda^4 - 4\lambda^3 + 2\lambda^2 + 4\lambda - 3$

(b) $\lambda = -1, 1, 3$

(c) $E_{-1} = \text{span}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}\right);$

$E_1 = \text{span}\left(\begin{bmatrix} -2 \\ 0 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \\ 0 \\ 3 \end{bmatrix}\right);$

$E_3 = \text{span}\left(\begin{bmatrix} 0 \\ 0 \\ 2 \\ 1 \end{bmatrix}\right)$

(d) $\lambda = -1$ and $\lambda = 3$ have algebraic and geometric multiplicity 1; $\lambda = 1$ has algebraic and geometric multiplicity 2.

15. $\begin{bmatrix} 2^{-9} + 3 \cdot 2^{10} \\ -2^{-9} + 3 \cdot 2^{10} \end{bmatrix}$ **17.** $\begin{bmatrix} 2 \\ (2 \cdot 3^{20} - 1)/3^{20} \\ 2 \end{bmatrix}$

23. (a) $\lambda = -2, E_{-2} = \text{span}\left(\begin{bmatrix} 2 \\ -5 \end{bmatrix}\right); \lambda = 5, E_5 =$

$\text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)$

(b) (i) $\lambda = -\frac{1}{2}, E_{-1/2} = \text{span}\left(\begin{bmatrix} 2 \\ -5 \end{bmatrix}\right); \lambda = \frac{1}{5}, E_{1/5} =$

$\text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)$

(iii) $\lambda = 0, E_0 = \text{span}\left(\begin{bmatrix} 2 \\ -5 \end{bmatrix}\right); \lambda = 7,$

$E_7 = \text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)$

27. $\begin{bmatrix} -3 & 4 & -12 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, -\lambda^3 - 3\lambda^2 + 4\lambda - 12$

35. $A^2 = 4A - 5I, A^3 = 11A - 20I$

$A^4 = 24A - 55I$

37. $A^{-1} = -\frac{1}{5}A + \frac{4}{5}I, A^{-2} = -\frac{4}{25}A + \frac{11}{25}I$

Exercises 4.4

1. The characteristic polynomial of A is $\lambda^2 - 5\lambda + 1$, but that of B is $\lambda^2 - 2\lambda + 1$.

3. The eigenvalues of A are $\lambda = 2$ and $\lambda = 4$, but those of B are $\lambda = 1$ and $\lambda = 4$.

5. $\lambda_1 = 4, E_4 = \text{span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right); \lambda_2 = 3, E_3 = \text{span}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)$

7. $\lambda_1 = 6, E_6 = \text{span}\left(\begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix}\right); \lambda_2 = -2, E_{-2} =$

$\text{span}\left(\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}\right)$

9. Not diagonalizable

11. $P = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -2 & 0 \end{bmatrix}, D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

13. Not diagonalizable

15. $P = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}$

17. $\begin{bmatrix} 35839 & -69630 \\ -11605 & 24234 \end{bmatrix}$

19. $\begin{bmatrix} (3^k + 3(-1)^k)/4 & (3^{k+1} - 3(-1)^k)/4 \\ (3^k - (-1)^k)/4 & (3^{k+1} + (-1)^k)/4 \end{bmatrix}$

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

23.

$$\begin{bmatrix} (5 + 2^{k+2} + (-3)^k)/10 & (2^k - (-3)^k)/5 & (-5 + 2^{k+2} + (-3)^k)/10 \\ (2^{k+1} - 2(-3)^k)/5 & (2^k + 4(-3)^k)/5 & (2^{k+1} - 2(-3)^k)/5 \\ (-5 + 2^{k+2} + (-3)^k)/10 & (2^k - (-3)^k)/5 & (5 + 2^{k+2} + (-3)^k)/10 \end{bmatrix}$$

25. $k = 0$ **27.** $k = 0$

29. All real values of k

37. If $A \sim B$, then there is an invertible matrix P such that $B = P^{-1}AP$. Therefore, we have

$$\begin{aligned} \operatorname{tr}(B) &= \operatorname{tr}(P^{-1}AP) = \operatorname{tr}(P^{-1}(AP)) = \operatorname{tr}((AP)P^{-1}) \\ &= \operatorname{tr}(APP^{-1}) = \operatorname{tr}(AI) = \operatorname{tr}(A) \end{aligned}$$

using Exercise 45 in Section 3.2.

39. $P = \begin{bmatrix} 7 & -2 \\ 10 & -3 \end{bmatrix}$

41. $P = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{3}{2} & -\frac{3}{2} & 1 \\ -\frac{5}{2} & -\frac{3}{2} & 0 \end{bmatrix}$

51. (b) $\dim E_{-1} = 1$, $\dim E_1 = 2$, $\dim E_2 = 3$

Exercises 4.5

1. (a) $\begin{bmatrix} 1 \\ 2.5 \end{bmatrix}$, 6.000

(b) $\lambda_1 = 6$

3. (a) $\begin{bmatrix} 1 \\ 0.618 \end{bmatrix}$, 2.618

(b) $\lambda_1 = (3 + \sqrt{5})/2 \approx 2.618$

5. (a) $m_5 = 11.001$, $\mathbf{y}_5 = \begin{bmatrix} -0.333 \\ 1.000 \end{bmatrix}$

7. (a) $m_8 = 10.000$, $\mathbf{y}_8 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

9.

k	0	1	2	3	4	5
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 26 \\ 8 \end{bmatrix}$	$\begin{bmatrix} 17.692 \\ 5.923 \end{bmatrix}$	$\begin{bmatrix} 18.018 \\ 6.004 \end{bmatrix}$	$\begin{bmatrix} 17.999 \\ 6.000 \end{bmatrix}$	$\begin{bmatrix} 18.000 \\ 6.000 \end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.308 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.335 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.333 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.333 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.333 \end{bmatrix}$
m_k	1	26	17.692	18.018	17.999	18.000

Therefore, $\lambda_1 \approx 18$, $\mathbf{v}_1 \approx \begin{bmatrix} 1 \\ 0.333 \end{bmatrix}$.

11.

k	0	1	2	3	4	5	6
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 7 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 7.571 \\ 2.857 \end{bmatrix}$	$\begin{bmatrix} 7.755 \\ 3.132 \end{bmatrix}$	$\begin{bmatrix} 7.808 \\ 3.212 \end{bmatrix}$	$\begin{bmatrix} 7.823 \\ 3.234 \end{bmatrix}$	$\begin{bmatrix} 7.827 \\ 3.240 \end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.286 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.377 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.404 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.411 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.413 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.414 \end{bmatrix}$
m_k	1	7	7.571	7.755	7.808	7.823	7.827

Therefore, $\lambda_1 \approx 7.827$, $\mathbf{v}_1 \approx \begin{bmatrix} 1 \\ 0.414 \end{bmatrix}$.

13.

k	0	1	2	3	4	5
$\mathbf{x}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}21\\15\\13\end{bmatrix}$	$\begin{bmatrix}16.809\\12.238\\10.714\end{bmatrix}$	$\begin{bmatrix}17.011\\12.371\\10.824\end{bmatrix}$	$\begin{bmatrix}16.999\\12.363\\10.818\end{bmatrix}$	$\begin{bmatrix}17.000\\12.363\\10.818\end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0.714\\0.619\end{bmatrix}$	$\begin{bmatrix}1\\0.728\\0.637\end{bmatrix}$	$\begin{bmatrix}1\\0.727\\0.636\end{bmatrix}$	$\begin{bmatrix}1\\0.727\\0.636\end{bmatrix}$	$\begin{bmatrix}1\\0.727\\0.636\end{bmatrix}$
m_k	1	21	16.809	17.011	16.999	17.000

Therefore, $\lambda_1 \approx 17, \mathbf{v}_1 \approx \begin{bmatrix}1\\0.727\\0.636\end{bmatrix}$.

15. $\lambda_1 \approx 5, \mathbf{v}_1 \approx \begin{bmatrix}1\\0\\0.333\end{bmatrix}$

17.

k	0	1	2	3	4	5	6
$\mathbf{x}_k$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}7\\2\end{bmatrix}$	$\begin{bmatrix}7.571\\2.857\end{bmatrix}$	$\begin{bmatrix}7.755\\3.132\end{bmatrix}$	$\begin{bmatrix}7.808\\3.212\end{bmatrix}$	$\begin{bmatrix}7.823\\3.234\end{bmatrix}$	$\begin{bmatrix}7.827\\3.240\end{bmatrix}$
$R(\mathbf{x}_k)$	7	7.755	7.823	7.828	7.828	7.828	7.828
$\mathbf{y}_k$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0.286\end{bmatrix}$	$\begin{bmatrix}1\\0.377\end{bmatrix}$	$\begin{bmatrix}1\\0.404\end{bmatrix}$	$\begin{bmatrix}1\\0.411\end{bmatrix}$	$\begin{bmatrix}1\\0.413\end{bmatrix}$	$\begin{bmatrix}1\\0.414\end{bmatrix}$

19.

k	0	1	2	3	4	5
$\mathbf{x}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}21\\15\\13\end{bmatrix}$	$\begin{bmatrix}16.809\\12.238\\10.714\end{bmatrix}$	$\begin{bmatrix}17.011\\12.371\\10.824\end{bmatrix}$	$\begin{bmatrix}16.999\\12.363\\10.818\end{bmatrix}$	$\begin{bmatrix}17.000\\12.363\\10.818\end{bmatrix}$
$R(\mathbf{x}_k)$	16.333	16.998	17.000	17.000	17.000	17.000
$\mathbf{y}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0.714\\0.619\end{bmatrix}$	$\begin{bmatrix}1\\0.728\\0.637\end{bmatrix}$	$\begin{bmatrix}1\\0.727\\0.636\end{bmatrix}$	$\begin{bmatrix}1\\0.727\\0.636\end{bmatrix}$	$\begin{bmatrix}1\\0.727\\0.636\end{bmatrix}$

21.

k	0	1	2	3	4	5	6	7	8
$\mathbf{x}_k$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}5\\4\end{bmatrix}$	$\begin{bmatrix}4.8\\3.2\end{bmatrix}$	$\begin{bmatrix}4.667\\2.667\end{bmatrix}$	$\begin{bmatrix}4.571\\2.286\end{bmatrix}$	$\begin{bmatrix}4.500\\2.000\end{bmatrix}$	$\begin{bmatrix}4.444\\1.778\end{bmatrix}$	$\begin{bmatrix}4.400\\1.600\end{bmatrix}$	$\begin{bmatrix}4.364\\1.455\end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0.8\end{bmatrix}$	$\begin{bmatrix}1\\0.667\end{bmatrix}$	$\begin{bmatrix}1\\0.571\end{bmatrix}$	$\begin{bmatrix}1\\0.500\end{bmatrix}$	$\begin{bmatrix}1\\0.444\end{bmatrix}$	$\begin{bmatrix}1\\0.400\end{bmatrix}$	$\begin{bmatrix}1\\0.364\end{bmatrix}$	$\begin{bmatrix}1\\0.333\end{bmatrix}$
m_k	1	5	4.8	4.667	4.571	4.500	4.444	4.400	4.364

Since $\lambda_1 = \lambda_2 = 4, \mathbf{v}_1 = \begin{bmatrix}1\\0\end{bmatrix}$, m_k is converging slowly to the exact answer.

23.

k	0	1	2	3	4	5	6	7	8
$\mathbf{x}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}5\\4\\1\end{bmatrix}$	$\begin{bmatrix}4.2\\3.2\\0.2\end{bmatrix}$	$\begin{bmatrix}4.048\\3.048\\0.048\end{bmatrix}$	$\begin{bmatrix}4.012\\3.012\\0.012\end{bmatrix}$	$\begin{bmatrix}4.003\\3.003\\0.003\end{bmatrix}$	$\begin{bmatrix}4.001\\3.001\\0.001\end{bmatrix}$	$\begin{bmatrix}4.000\\3.000\\0.000\end{bmatrix}$	$\begin{bmatrix}4.000\\3.000\\0.000\end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0.8\\0.2\end{bmatrix}$	$\begin{bmatrix}1\\0.762\\0.048\end{bmatrix}$	$\begin{bmatrix}1\\0.753\\0.012\end{bmatrix}$	$\begin{bmatrix}1\\0.751\\0.003\end{bmatrix}$	$\begin{bmatrix}1\\0.750\\0.001\end{bmatrix}$	$\begin{bmatrix}1\\0.750\\0\end{bmatrix}$	$\begin{bmatrix}1\\0.750\\0\end{bmatrix}$	$\begin{bmatrix}1\\0.750\\0\end{bmatrix}$
m_k	1	5	4.2	4.048	4.012	4.003	4.001	4.000	4.000

In this case, $\lambda_1 = \lambda_2 = 4$ and $E_4 = \text{span}\left(\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}\right)$. Clearly, m_k is converging to 4 and $\mathbf{y}_k$ is converging to a vector in the eigenspace E_4—namely, $\begin{bmatrix}1\\0\\0\end{bmatrix} + 0.75\begin{bmatrix}0\\1\\0\end{bmatrix}$.

25.

k	0	1	2	3	4	5
$\mathbf{x}_k$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}-1\\-1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}-1\\-1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$
m_k	1	1	−1	1	−1	1

The exact eigenvalues are complex (i and $-i$), so the power method cannot possibly converge to either the dominant eigenvalue or the dominant eigenvector if we start with a *real* initial iterate. Instead, the power method oscillates between two sets of real vectors.

27.

k	0	1	2	3	4	5
$\mathbf{x}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}3\\4\\3\end{bmatrix}$	$\begin{bmatrix}2.500\\4.000\\2.500\end{bmatrix}$	$\begin{bmatrix}2.250\\4.000\\2.250\end{bmatrix}$	$\begin{bmatrix}2.125\\4.000\\2.125\end{bmatrix}$	$\begin{bmatrix}2.063\\4.000\\2.063\end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix}1\\1\\1\end{bmatrix}$	$\begin{bmatrix}0.750\\1\\0.750\end{bmatrix}$	$\begin{bmatrix}0.625\\1\\0.625\end{bmatrix}$	$\begin{bmatrix}0.562\\1\\0.562\end{bmatrix}$	$\begin{bmatrix}0.531\\1\\0.531\end{bmatrix}$	$\begin{bmatrix}0.516\\1\\0.516\end{bmatrix}$
m_k	1	4	4	4	4	4

The eigenvalues are $\lambda_1 = -12$, $\lambda_2 = 4$, $\lambda_3 = 2$, with corresponding eigenvectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Since $\mathbf{x}_0 = \frac{1}{2}\mathbf{v}_2 + \frac{1}{2}\mathbf{v}_3$, the initial vector $\mathbf{x}_0$ has a zero component in the direction of the dominant eigenvector, so the power method cannot converge to the dominant eigenvalue/eigenvector. Instead, it converges to a *second* eigenvalue/eigenvector pair, as the calculations show.

29. Apply the power method to $A - 18I = \begin{bmatrix} -4 & 12 \\ 5 & -15 \end{bmatrix}$.

k	0	1	2	3
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 8 \\ -10 \end{bmatrix}$	$\begin{bmatrix} 15.2 \\ -19 \end{bmatrix}$	$\begin{bmatrix} 15.2 \\ -19 \end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.8 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.8 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.8 \\ 1 \end{bmatrix}$
m_k	1	-10	-19	-19

Thus, -19 is the dominant eigenvalue of $A - 18I$, and $\lambda_2 = -19 + 18 = -1$ is the second eigenvalue of A.

31. Apply the power method to $A - 17I = \begin{bmatrix} -8 & 4 & 8 \\ 4 & -2 & -4 \\ 8 & -4 & -8 \end{bmatrix}$.

k	0	1	2	3
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 4 \\ -2 \\ -4 \end{bmatrix}$	$\begin{bmatrix} -18 \\ 9 \\ 18 \end{bmatrix}$	$\begin{bmatrix} -18 \\ 9 \\ 18 \end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -0.5 \\ -1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -0.5 \\ -1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -0.5 \\ -1 \end{bmatrix}$
m_k	1	4	-18	-18
$R(\mathbf{x}_k)$	-0.667	-18	-18	-18

In this case, there is no dominant eigenvalue. (We could choose either 18 or -18 for m_k, $k \geq 2$.) However, the Rayleigh quotient method (Exercises 17–20) converges to -18. Thus, -18 is the dominant eigenvalue of $A - 17I$, and $\lambda_2 = -18 + 17 = -1$ is the second eigenvalue of A.

33.

k	0	1	2	3	4	5
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$	$\begin{bmatrix} -0.833 \\ 1.056 \end{bmatrix}$	$\begin{bmatrix} 0.798 \\ -0.997 \end{bmatrix}$	$\begin{bmatrix} 0.800 \\ -1.000 \end{bmatrix}$	$\begin{bmatrix} 0.800 \\ -1.000 \end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$	$\begin{bmatrix} -0.789 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.801 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.800 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.800 \\ 1 \end{bmatrix}$
m_k	1	0.5	1.056	-0.997	-1.000	-1.000

Thus, the eigenvalue of A that is smallest in magnitude is $1/(-1) = -1$.

35.

k	0	1	2	3	4	5
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$	$\begin{bmatrix} -0.500 \\ 0.000 \\ 0.500 \end{bmatrix}$	$\begin{bmatrix} 0.500 \\ 0.333 \\ -0.500 \end{bmatrix}$	$\begin{bmatrix} 0.500 \\ 0.111 \\ -0.500 \end{bmatrix}$	$\begin{bmatrix} 0.500 \\ 0.259 \\ -0.500 \end{bmatrix}$	$\begin{bmatrix} 0.500 \\ 0.160 \\ -0.500 \end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$	$\begin{bmatrix} -1.000 \\ 0.000 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} -1.000 \\ -0.667 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} -1.000 \\ -0.222 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} -1.000 \\ -0.518 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} -1.000 \\ -0.321 \\ 1.000 \end{bmatrix}$
m_k	1	−0.500	−0.500	−0.500	−0.500	−0.500

Clearly, m_k converges to -0.5, so the smallest eigenvalue of A is $1/(-0.5) = -2$.

37. The calculations are the same as for Exercise 33.

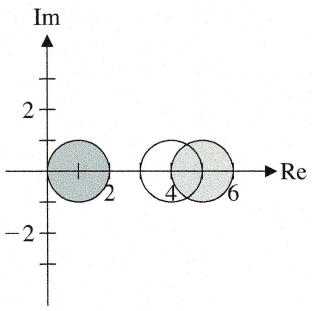
39. We apply the inverse power method to $A - 5I = \begin{bmatrix} -1 & 0 & 6 \\ -1 & -2 & 1 \\ 6 & 0 & -1 \end{bmatrix}$. Taking $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, we have

k	0	1	2	3
$\mathbf{x}_k$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.200 \\ -0.500 \\ 0.200 \end{bmatrix}$	$\begin{bmatrix} -0.080 \\ -0.500 \\ -0.080 \end{bmatrix}$	$\begin{bmatrix} 0.032 \\ -0.500 \\ 0.032 \end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.400 \\ 1 \\ -0.400 \end{bmatrix}$	$\begin{bmatrix} 0.160 \\ 1 \\ 0.160 \end{bmatrix}$	$\begin{bmatrix} -0.064 \\ 1 \\ -0.064 \end{bmatrix}$
m_k	1	−0.500	−0.500	−0.500

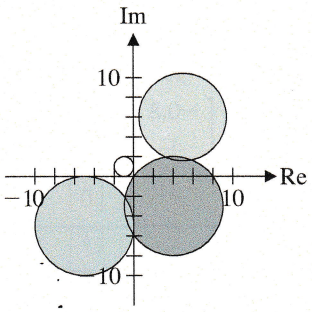
Clearly, m_k converges to -0.5, so the eigenvalue of A closest to 5 is $5 + 1/(-0.5) = 5 - 2 = 3$.

41. 0.732 **43.** −0.619

47.



49.



51. *Hint:* Show that 0 is not contained in any Gerschgorin disk and then apply Theorem 4.16.

53. Exercise 52 implies that $|\lambda|$ is less than or equal to all of the column sums of A for every eigenvalue λ. But for a stochastic matrix, all column sums are 1. Hence $|\lambda| \leq 1$.

Exercises 4.6

1. Not regular **3.** Regular

5. Not regular **7.** $L = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{4}{5} & \frac{4}{5} \end{bmatrix}$

9. $L = \begin{bmatrix} 0.304 & 0.304 & 0.304 \\ 0.354 & 0.354 & 0.354 \\ 0.342 & 0.342 & 0.342 \end{bmatrix}$

11. 1, $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ **13.** 2, $\begin{bmatrix} 16 \\ 4 \\ 1 \end{bmatrix}$

15. The population is increasing, decreasing, and constant, respectively.

17.

$$P^{-1}LP = \begin{bmatrix} b_1 & b_2s_1 & b_3s_1s_2 & \cdots & b_{n-1}s_1s_2\cdots s_{n-2} & b_ns_1s_2\cdots s_{n-1} \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

The characteristic polynomial of L is $(\lambda^n - b_1\lambda^{n-1} - b_2s_1\lambda^{n-2} - b_3s_1s_2\lambda^{n-3} - \cdots - b_ns_1s_2\cdots s_{n+1})(-1)^n$.

19. $\lambda \approx 1.746$, $\mathbf{p} \approx \begin{bmatrix} 0.660 \\ 0.264 \\ 0.076 \end{bmatrix}$

21. $\lambda \approx 1.092$, $\mathbf{p} \approx \begin{bmatrix} 0.535 \\ 0.147 \\ 0.094 \\ 0.078 \\ 0.064 \\ 0.053 \\ 0.029 \end{bmatrix}$

25. (a) $h \approx 0.082$ **29.** 3, $\begin{bmatrix} \frac{3}{5} \\ \frac{2}{5} \end{bmatrix}$

31. 3, $\begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}$ **33.** Reducible

35. Irreducible **43.** 1, 2, 4, 8, 16

45. 0, 1, 1, 0, −1 **47.** $x_n = 4^n - (-1)^n$

49. $y_n = (n - \frac{1}{2})2^n$

51. $b_n = \dfrac{1}{2\sqrt{3}}[(1 + \sqrt{3})^n - (1 - \sqrt{3})^n]$

57. (a) $d_1 = 1, d_2 = 2, d_3 = 3, d_4 = 5, d_5 = 8$

(b) $d_n = d_{n-1} + d_{n-2}$

(c) $d_n = \dfrac{1}{\sqrt{5}}\left[\left(\dfrac{1 + \sqrt{5}}{2}\right)^{n+1} - \left(\dfrac{1 - \sqrt{5}}{2}\right)^{n+1}\right]$

59. The general solution is $x(t) = -3C_1e^{-t} + C_2e^{4t}$, $y(t) = 2C_1e^{-t} + C_2e^{4t}$. The specific solution is $x(t) = -3e^{-t} + 3e^{4t}$, $y(t) = 2e^{-t} + 3e^{4t}$.

61. The general solution is $x_1(t) = (1 + \sqrt{2})C_1e^{\sqrt{2}t} + (1 - \sqrt{2})C_2e^{-\sqrt{2}t}$, $x_2(t) = C_1e^{\sqrt{2}t} + C_2e^{-\sqrt{2}t}$. The specific solution is $x_1(t) = (2 + \sqrt{2})e^{\sqrt{2}t}/4 + (2 - \sqrt{2})e^{-\sqrt{2}t}/4$, $x_2(t) = \sqrt{2}e^{\sqrt{2}t}/4 - \sqrt{2}e^{-\sqrt{2}t}/4$.

63. The general solution is $x(t) = -C_1 + C_3e^{-t}$, $y(t) = C_1 + C_2e^{t} - C_3e^{-t}$, $z(t) = C_1 + C_2e^{t}$. The specific solution is $x(t) = 2 - e^{-t}$, $y(t) = -2 + e^{t} + e^{-t}$, $z(t) = -2 + e^{t}$.

65. (a) $x(t) = -120e^{8t/5} + 520e^{11t/10}$, $y(t) = 240e^{8t/5} + 260^{11t/10}$. Strain X dies out after approximately 2.93 days; strain Y continues to grow.

67. $a = 10, b = 20$; $x(t) = 10e^t(\cos t + \sin t) + 10$, $y(t) = 10e^t(\cos t - \sin t) + 20$. Species Y dies out when $t \approx 1.22$.

71. $x(t) = C_1e^{2t} + C_2e^{3t}$

77. (a) $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 9 \\ 9 \end{bmatrix}, \begin{bmatrix} 27 \\ 27 \end{bmatrix}$ **(c)** Repeller

79. (a) $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ **(c)** Neither

81. (a) $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 \\ -1 \end{bmatrix}, \begin{bmatrix} 1.75 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 3.125 \\ -1.75 \end{bmatrix}$ **(c)** Saddle point

83. (a) $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}, \begin{bmatrix} 0.36 \\ 0.36 \end{bmatrix}, \begin{bmatrix} 0.216 \\ 0.216 \end{bmatrix}$ **(c)** Attractor

85. $r = \sqrt{2}, \theta = 45°$, spiral repeller

87. $r = 2, \theta = -60°$, spiral repeller

89. $P = \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix}$, $C = \begin{bmatrix} 0.2 & -0.1 \\ 0.1 & 0.2 \end{bmatrix}$, spiral attractor

91. $P = \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ 1 & 0 \end{bmatrix}$, $C = \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{bmatrix}$, orbital center

Review Questions

1. (a) F **(c)** F **(e)** F **(g)** T **(i)** F

3. −18

5. Since $A^T = -A$, we have $\det A = \det(A^T) = \det(-A) = (-1)^n \det A = -\det A$ by Theorem 4.7 and the fact that n is odd. It follows that $\det A = 0$.

7. $A\mathbf{x} = \begin{bmatrix} 5 \\ 10 \end{bmatrix} = 5\mathbf{x}, \lambda = 5$

9. (a) $4 - 3\lambda^2 - \lambda^3$

(c) $E_1 = \text{span}\left(\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}\right)$, $E_{-2} = \text{span}\left(\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right)$

11. $\begin{bmatrix} 162 \\ 158 \end{bmatrix}$ **13.** Not similar **15.** Not similar

17. 0, 1, or −1

19. If $A\mathbf{x} = \lambda\mathbf{x}$, then $(A^2 - 5A + 2I)\mathbf{x} = A^2\mathbf{x} - 5A\mathbf{x} + 2\mathbf{x} = 3^2\mathbf{x} - 5(3\mathbf{x}) + 2\mathbf{x} = -4\mathbf{x}$.

Chapter 5

Exercises 5.1

1. Orthogonal **3.** Not orthogonal **5.** Orthogonal

7. $[\mathbf{w}]_{\mathcal{B}} = \begin{bmatrix} \frac{1}{2} \\ -1 \end{bmatrix}$ **9.** $[\mathbf{w}]_{\mathcal{B}} = \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix}$ **11.** Orthonormal

13. $\begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix}, \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \\ 0 \end{bmatrix}, \begin{bmatrix} 2/3\sqrt{5} \\ 4/3\sqrt{5} \\ -5/3\sqrt{5} \end{bmatrix}$

15. Orthonormal

17. Orthogonal, $\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

19. Orthogonal, $\begin{bmatrix} \cos\theta \sin\theta & \cos^2\theta & \sin\theta \\ -\cos\theta & \sin\theta & 0 \\ -\sin^2\theta & -\cos\theta \sin\theta & \cos\theta \end{bmatrix}$

21. Not orthogonal

27. $$\cos(\angle(Q\mathbf{x}, Q\mathbf{y})) = \frac{Q\mathbf{x} \cdot Q\mathbf{y}}{\|Q\mathbf{x}\|\|Q\mathbf{y}\|} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$
$$= \cos(\angle(\mathbf{x}, \mathbf{y})) \text{ by Theorem 5.6}$$

29. Rotation, $\theta = 45°$ **31.** Reflection, $y = \sqrt{3}x$

33. (a) $A(A^T + B^T)B = AA^TB + AB^TB = IB + AI = B + A = A + B$

(b) From part (a),

$$\begin{aligned} \det(A + B) &= \det(A(A^T + B^T)B) \\ &= \det A \det(A^T + B^T) \det B \\ &= \det A \det((A + B)^T) \det B \\ &= \det A \det(A + B) \det B \end{aligned}$$

Assume that $\det A + \det B = 0$ (so that $\det B = -\det A$) but that $A + B$ is invertible. Then $\det(A + B) \neq 0$, so $1 = \det A \det B = \det A(-\det A) = -(\det A)^2$. This is impossible, so we conclude that $A + B$ cannot be invertible.

Exercises 5.2

1. $W^{\perp} = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : x + 2y = 0 \right\}, \mathcal{B}^{\perp} = \left\{ \begin{bmatrix} -2 \\ 1 \end{bmatrix} \right\}$

3. $W^{\perp} = \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} : x = t, y = t, z = -t \right\}$,

$\mathcal{B}^{\perp} = \left\{ \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \right\}$

5. $W^{\perp} = \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} : x - y + 3z = 0 \right\}, \mathcal{B}^{\perp} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} \right\}$

7. row(A): {[1 0 1], [0 1 −2]}, null(A): $\left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \right\}$

9. col(A): $\left\{ \begin{bmatrix} 1 \\ 5 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ 1 \\ -1 \end{bmatrix} \right\}$, null($A^T$): $\left\{ \begin{bmatrix} 1 \\ 0 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} -5 \\ 1 \\ -7 \\ 0 \end{bmatrix} \right\}$

11. $\left\{ \begin{bmatrix} 1 \\ -10 \\ -4 \end{bmatrix} \right\}$ **13.** $\left\{ \begin{bmatrix} -4 \\ 1 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right\}$

15. $\begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix}$ **17.** $\begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ 3 \end{bmatrix}$

19. $\mathbf{v} = \begin{bmatrix} -\frac{2}{5} \\ -\frac{6}{5} \end{bmatrix} + \begin{bmatrix} \frac{12}{5} \\ -\frac{4}{5} \end{bmatrix}$ **21.** $\mathbf{v} = \begin{bmatrix} \frac{7}{2} \\ -2 \\ \frac{7}{2} \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ 0 \\ -\frac{1}{2} \end{bmatrix}$

25. No

Exercises 5.3

1. $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -1/2 \\ 1/2 \end{bmatrix}; \mathbf{q}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \mathbf{q}_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

3. $\mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}; \mathbf{q}_1 = \begin{bmatrix} 1/\sqrt{3} \\ -1/\sqrt{3} \\ -1/\sqrt{3} \end{bmatrix}$,

$\mathbf{q}_2 = \begin{bmatrix} 2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}, \mathbf{q}_3 = \begin{bmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

5. $\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ 2 \end{bmatrix} \right\}$ **7.** $\mathbf{v} = \begin{bmatrix} -\frac{2}{9} \\ \frac{2}{9} \\ \frac{8}{9} \end{bmatrix} + \begin{bmatrix} \frac{38}{9} \\ -\frac{38}{9} \\ \frac{19}{9} \end{bmatrix}$

9. $\left\{ \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \\ -\frac{2}{3} \end{bmatrix} \right\}$.

11. $\left\{\begin{bmatrix}3\\1\\5\end{bmatrix}, \begin{bmatrix}-\frac{3}{35}\\ \frac{34}{35}\\ -\frac{1}{7}\end{bmatrix}, \begin{bmatrix}-\frac{15}{34}\\ 0\\ \frac{9}{34}\end{bmatrix}\right\}$

13. $Q = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \\ 0 & 1/\sqrt{3} & -2/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \end{bmatrix}$

15. $\begin{bmatrix} 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \end{bmatrix}\begin{bmatrix} \sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 3/\sqrt{6} & 1/\sqrt{6} \\ 0 & 0 & 2/\sqrt{3} \end{bmatrix}$

17. $R = \begin{bmatrix} 3 & 9 & \frac{1}{3} \\ 0 & 6 & \frac{2}{3} \\ 0 & 0 & \frac{7}{3} \end{bmatrix}$

19. $A = AI$

21. $A^{-1} = (QR)^{-1} = R^{-1}Q^{-1} = R^{-1}Q^T =$
$\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{6} & -1/2\sqrt{3} \\ 0 & 2/\sqrt{6} & -1/2\sqrt{3} \\ 0 & 0 & 3/2\sqrt{3} \end{bmatrix} \cdot$
$\begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 2/\sqrt{6} & -1/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{3} & -1/\sqrt{3} \end{bmatrix}$

23. Let $R\mathbf{x} = \mathbf{0}$. Then $A\mathbf{x} = QR\mathbf{x} = Q\mathbf{0} = \mathbf{0}$. Since $A\mathbf{x}$ represents a linear combination of the columns of A (which are linearly independent), we must have $\mathbf{x} = \mathbf{0}$. Hence, R is invertible, by the Fundamental Theorem.

Exercises 5.4

1. $Q = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}, D = \begin{bmatrix} 5 & 0 \\ 0 & 3 \end{bmatrix}$

3. $Q = \begin{bmatrix} 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{3} & -2/\sqrt{6} \end{bmatrix}, D = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$

5. $Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}, D = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -2 \end{bmatrix}$

7. $Q = \begin{bmatrix} -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix}, D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

9. $Q = \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} & 0 \\ 0 & 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix},$

$D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

11. $Q^TAQ = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}\begin{bmatrix} a & b \\ b & a \end{bmatrix} \cdot$
$\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} a+b & 0 \\ 0 & a-b \end{bmatrix} = D$

13. (a) If A and B are orthogonally diagonalizable, then each is symmetric, by the Spectral Theorem. Therefore, $A + B$ is symmetric, by Exercise 35 in Section 3.2, and so is orthogonally diagonalizable, by the Spectral Theorem.

15. If A and B are orthogonally diagonalizable, then each is symmetric, by the Spectral Theorem. Since $AB = BA$, AB is also symmetric, by Exercise 36 in Section 3.2. Hence, AB is orthogonally diagonalizable, by the Spectral Theorem.

17. $A = \begin{bmatrix} \frac{5}{2} & \frac{5}{2} \\ \frac{5}{2} & \frac{5}{2} \end{bmatrix} + \begin{bmatrix} \frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{3}{2} \end{bmatrix}$

19. $A = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix}$

21. $\begin{bmatrix} \frac{1}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{1}{2} \end{bmatrix}$ **23.** $\begin{bmatrix} \frac{5}{3} & -\frac{2}{3} & -\frac{1}{3} \\ -\frac{2}{3} & \frac{5}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & \frac{8}{3} \end{bmatrix}$

Exercises 5.5

1. $2x^2 + 6xy + 4y^2$ **3.** 123 **5.** -5

7. $\begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$ **9.** $\begin{bmatrix} 3 & -\frac{3}{2} \\ -\frac{3}{2} & -1 \end{bmatrix}$ **11.** $\begin{bmatrix} 5 & 1 & -2 \\ 1 & -1 & 2 \\ -2 & 2 & 2 \end{bmatrix}$

13. $Q = \begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix}, y_1^2 + 6y_2^2$

15. $Q = \begin{bmatrix} 2/\sqrt{5} & 2/3\sqrt{5} & -1/3 \\ 0 & 5/3\sqrt{5} & 2/3 \\ 1/\sqrt{5} & -4/3\sqrt{5} & 2/3 \end{bmatrix}, 9y_1^2 + 9y_2^2 - 9y_3^2$

17. $Q = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ -1/\sqrt{3} & 0 & 2/\sqrt{6} \\ -1/\sqrt{3} & 1/\sqrt{2} & -1/\sqrt{6} \end{bmatrix}, 2(x')^2 +$
$(y')^2 - (z')^2$

19. Positive definite **21.** Negative definite

23. Positive definite **25.** Indefinite

27. For any vector $\mathbf{x}$, we have $\mathbf{x}^TA\mathbf{x} = \mathbf{x}^TB^TB\mathbf{x} = (B\mathbf{x})^T(B\mathbf{x}) = \|B\mathbf{x}\|^2 \geq 0$. If $\mathbf{x}^TA\mathbf{x} = 0$, then $\|B\mathbf{x}\|^2 = 0$, so $B\mathbf{x} = \mathbf{0}$. Since B is invertible, this implies that $\mathbf{x} = \mathbf{0}$. Therefore, $\mathbf{x}^TA\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, and hence $A = B^TB$ is positive definite.

29. **(a)** Every eigenvalue of cA is of the form $c\lambda$ for some eigenvalue λ of A. By Theorem 5.24, $\lambda > 0$, so $c\lambda > 0$, since c is positive. Hence, cA is positive definite, by Theorem 5.24.

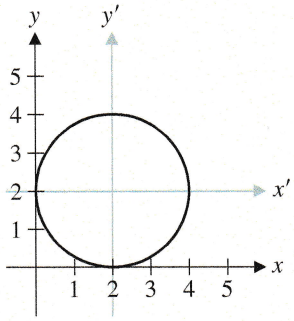
(c) Let $\mathbf{x} \neq \mathbf{0}$. Then $\mathbf{x}^TA\mathbf{x} > 0$ and $\mathbf{x}^TB\mathbf{x} > 0$, since A and B are positive definite. But then $\mathbf{x}^T(A + B)\mathbf{x} = \mathbf{x}^TA\mathbf{x} + \mathbf{x}^TB\mathbf{x} > 0$, so $A + B$ is positive definite.

31. The maximum value of $f(\mathbf{x})$ is 2 when $\mathbf{x} = \pm\begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$; the minimum value of $f(\mathbf{x})$ is 0 when $\mathbf{x} = \pm\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$.

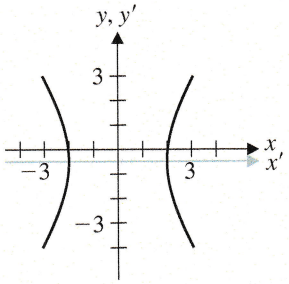
33. The maximum value of $f(\mathbf{x})$ is 4 when $\mathbf{x} = \pm\begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}$; the minimum value of $f(\mathbf{x})$ is 1 when $\mathbf{x} = \pm\begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix}$ or $\pm\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}$.

35. Ellipse **37.** Parabola **39.** Hyperbola

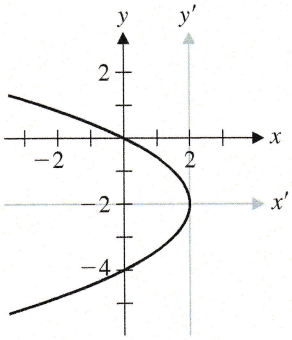
41. Circle, $x' = x - 2, y' = y - 2, (x')^2 + (y')^2 = 4$



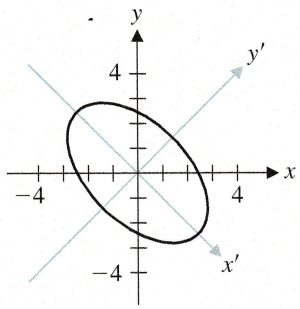
43. Hyperbola, $x' = x, y' = y + \frac{1}{2}, (x')^2/4 - (y')^2/9 = 1$



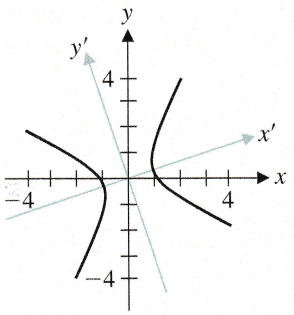
45. Parabola, $x' = x - 2, y' = y + 2, x' = -\frac{1}{2}(y')^2$



47. Ellipse, $(x')^2/4 + (y')^2/12 = 1$



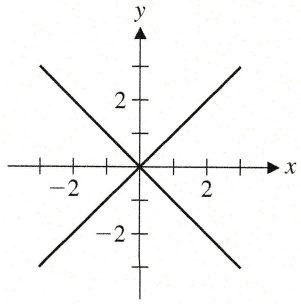
49. Hyperbola, $(x')^2 - (y')^2 = 1$



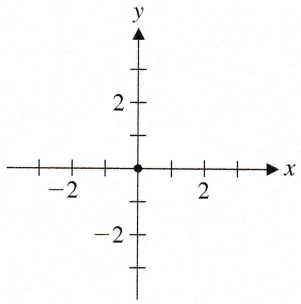
51. Ellipse, $(x'')^2/50 + (y'')^2/10 = 1$

53. Hyperbola, $(x'')^2 - (y'')^2 = 1$

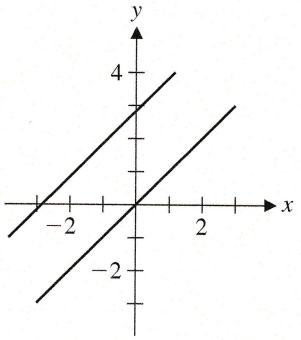
55. Degenerate (two lines)



57. Degenerate (a point)



59. Degenerate (two lines)



61. Hyperboloid of one sheet, $(x')^2 - (y')^2 + 3(z')^2 = 1$

63. Hyperbolic paraboloid, $z = -(x')^2 + (y')^2$

65. Hyperbolic paraboloid, $x' = -\sqrt{3}(y')^2 + \sqrt{3}(z')^2$

67. Ellipsoid, $3(x'')^2 + (y'')^2 + 2(z'')^2 = 4$

Review Questions

1. **(a)** T **(c)** T **(e)** F **(g)** F **(i)** F

3. $\begin{bmatrix} 9/2 \\ 2/3 \\ -11/6 \end{bmatrix}$

5. Verify that $Q^TQ = I$.

7. Theorem 5.6(c) shows that if $\mathbf{v}_i \cdot \mathbf{v}_j = 0$, then $Q\mathbf{v}_i \cdot Q\mathbf{v}_j = 0$. Theorem 5.6(b) shows that $\{Q\mathbf{v}_1, \ldots, Q\mathbf{v}_k\}$ consists of unit vectors, because $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ does. Hence, $\{Q\mathbf{v}_1, \ldots, Q\mathbf{v}_k\}$ is an orthonormal set.

9. $\left\{\begin{bmatrix} 2 \\ -5 \end{bmatrix}\right\}$ **11.** $\left\{\begin{bmatrix} -1 \\ 3 \\ 1 \end{bmatrix}\right\}$

13. row(A): $\{[1 \quad 0 \quad 2 \quad 3 \quad 4], [0 \quad 1 \quad 0 \quad 2 \quad 1]\}$

col(A): $\left\{\begin{bmatrix} 1 \\ -1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ 1 \\ -5 \end{bmatrix}\right\}$

null(A): $\left\{\begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ -2 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix}\right\}$

null(A^T): $\left\{\begin{bmatrix} -5 \\ -3 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ 0 \\ 1 \end{bmatrix}\right\}$

15. **(a)** $\left\{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ -\frac{3}{4} \end{bmatrix}, \begin{bmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \end{bmatrix}\right\}$

17. $\left\{\begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ -1 \end{bmatrix}\right\}$

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

Chapter 6

Exercises 6.1

1. Vector space

3. Not a vector space; axiom 1 fails.

5. Not a vector space; axiom 8 fails.

7. Vector space **9.** Vector space

11. Vector space **15.** Complex vector space

17. Not a complex vector space; axiom 6 fails.

19. Not a vector space; axioms 1, 4, and 6 fail.

21. Not a vector space; the operations of addition and multiplication are not even the same.

25. Subspace **27.** Not a subspace

29. Not a subspace **31.** Subspace

33. Subspace **35.** Subspace

37. Not a subspace **39.** Subspace

41. Subspace **43.** Not a subspace

45. Not a subspace

47. Take U to be the x-axis and W the y-axis, for example. Then $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ are in $U \cup W$, but $\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix}$ is not.

51. No

53. Yes; $s(x) = (3 + 2t)p(x) + (1 + t)q(x) + tr(x)$ for any scalar t.

55. Yes; $h(x) = f(x) + g(x)$

57. No

59. No

61. Yes

Exercises 6.2

1. Linearly independent

3. Linearly dependent; $\begin{bmatrix}-1 & 0\\-1 & 7\end{bmatrix} = 4\begin{bmatrix}-1 & 1\\-2 & 2\end{bmatrix} + \begin{bmatrix}3 & 0\\1 & 1\end{bmatrix} - 2\begin{bmatrix}0 & 2\\-3 & 1\end{bmatrix}$

5. Linearly independent

7. Linearly dependent; $3x + 2x^2 = 7x - 2(2x - x^2)$

9. Linearly independent

11. Linearly dependent; $1 = \sin^2 x + \cos^2 x$

13. Linearly dependent; $\ln(x^2) = -2\ln 2 \cdot 1 + 2 \cdot \ln(2x)$

17. **(a)** Linearly independent

(b) Linearly dependent

19. Basis **21.** Not a basis

23. Not a basis **25.** Not a basis

27. $[A]_{\mathcal{B}} = \begin{bmatrix}-1\\-1\\-1\\4\end{bmatrix}$ **29.** $[p(x)]_{\mathcal{B}} = \begin{bmatrix}6\\-1\\3\end{bmatrix}$

35. $\dim V = 2, \mathcal{B} = \{1 - x, 1 - x^2\}$

37. $\dim V = 3, \mathcal{B} = \left\{\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}, \begin{bmatrix}0 & 1\\0 & 0\end{bmatrix}, \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}\right\}$

39. $\dim V = 2, \mathcal{B} = \left\{\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \begin{bmatrix}0 & 1\\0 & 0\end{bmatrix}\right\}$

41. $(n^2 - n)/2$

43. **(a)** $\dim(U \times V) = \dim U + \dim V$

(b) Show that if $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ is a basis for W, then $\{(\mathbf{w}_1, \mathbf{w}_1), \ldots, (\mathbf{w}_n, \mathbf{w}_n)\}$ is a basis for Δ.

45. $\{1 + x, 1 + x + x^2, 1\}$

47. $\left\{\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \begin{bmatrix}0 & 1\\1 & 0\end{bmatrix}, \begin{bmatrix}0 & -1\\1 & 0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}\right\}$

49. $\{1, 1 + x\}$

51. $\{1 - x, x - x^2\}$

53. $\{\sin^2 x, \cos^2 x\}$

59. **(a)** $p_0(x) = \frac{1}{2}x^2 - \frac{5}{2}x + 3, p_1(x) = -x^2 + 4x - 3, p_2(x) = \frac{1}{2}x^2 - \frac{3}{2}x + 1$

61. **(c)** (i) $3x^2 - 16x + 19$ (ii) $x^2 - 4x + 5$

63. $(p^n - 1)(p^n - p)(p^n - p^2)\cdots(p^n - p^{n-1})$

Exercises 6.3

1. $[\mathbf{x}]_{\mathcal{B}} = \begin{bmatrix}2\\3\end{bmatrix}, [\mathbf{x}]_{\mathcal{C}} = \begin{bmatrix}\frac{5}{2}\\-\frac{1}{2}\end{bmatrix}, P_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\\frac{1}{2} & -\frac{1}{2}\end{bmatrix}, P_{\mathcal{B}\leftarrow\mathcal{C}} = \begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}$

3. $[\mathbf{x}]_{\mathcal{B}} = \begin{bmatrix}1\\0\\-1\end{bmatrix}, [\mathbf{x}]_{\mathcal{C}} = \begin{bmatrix}1\\-1\\-1\end{bmatrix}, P_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}1 & 0 & 0\\-1 & 1 & 0\\0 & -1 & 1\end{bmatrix}, P_{\mathcal{B}\leftarrow\mathcal{C}} = \begin{bmatrix}1 & 0 & 0\\1 & 1 & 0\\1 & 1 & 1\end{bmatrix}$

5. $[p(x)]_{\mathcal{B}} = \begin{bmatrix}2\\-1\end{bmatrix}, [p(x)]_{\mathcal{C}} = \begin{bmatrix}-3\\2\end{bmatrix}, P_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}-1 & 1\\1 & 0\end{bmatrix}, P_{\mathcal{B}\leftarrow\mathcal{C}} = \begin{bmatrix}0 & 1\\1 & 1\end{bmatrix}$

7. $[p(x)]_{\mathcal{B}} = \begin{bmatrix}1\\-1\\1\end{bmatrix}, [p(x)]_{\mathcal{C}} = \begin{bmatrix}1\\0\\1\end{bmatrix}, P_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}1 & 0 & 0\\1 & 1 & 0\\1 & 1 & 1\end{bmatrix}, P_{\mathcal{B}\leftarrow\mathcal{C}} = \begin{bmatrix}1 & 0 & 0\\-1 & 1 & 0\\0 & -1 & 1\end{bmatrix}$.

9. $[A]_{\mathcal{B}} = \begin{bmatrix} 4 \\ 2 \\ 0 \\ -1 \end{bmatrix}$, $[A]_{\mathcal{C}} = \begin{bmatrix} \frac{5}{2} \\ 0 \\ -3 \\ \frac{9}{2} \end{bmatrix}$, $P_{\mathcal{C}\leftarrow\mathcal{B}} =$

$$\begin{bmatrix} \frac{1}{2} & 0 & -1 & -\frac{1}{2} \\ 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 \\ \frac{3}{2} & -1 & -2 & -\frac{1}{2} \end{bmatrix}, P_{\mathcal{B}\leftarrow\mathcal{C}} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 1 \end{bmatrix}$$

11. $[f(x)]_{\mathcal{B}} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}$, $[f(x)]_{\mathcal{C}} = \begin{bmatrix} -1/2 \\ 5/2 \end{bmatrix}$,

$P_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix} 1 & 1/2 \\ 0 & -1/2 \end{bmatrix}$, $P_{\mathcal{B}\leftarrow\mathcal{C}} = \begin{bmatrix} 1 & 1 \\ 0 & -2 \end{bmatrix}$

13. (a) $\begin{bmatrix} (3 - 2\sqrt{3})/2 \\ (-3\sqrt{3} + 2)/2 \end{bmatrix} \approx \begin{bmatrix} 3.232 \\ -1.598 \end{bmatrix}$

(b) $\begin{bmatrix} 2 + 2\sqrt{3} \\ 2\sqrt{3} - 2 \end{bmatrix} \approx \begin{bmatrix} 5.464 \\ 1.464 \end{bmatrix}$

15. $\mathcal{B} = \left\{ \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix} \right\}$

17. $-2 - 8(x - 1) - 5(x - 1)^2$

19. $-1 + 3(x + 1) - 3(x + 1)^2 + (x + 1)^3$

Exercises 6.4

1. Linear transformation **3.** Linear transformation

5. Linear transformation

7. Not a linear transformation

9. Linear transformation

11. Not a linear transformation

13. We have

$$\begin{aligned} S(p(x) + q(x)) &= S((p + q)(x)) = x((p + q)(x)) \\ &= x(p(x) + q(x)) = xp(x) + xq(x) \\ &= S(p(x)) + S(q(x)) \end{aligned}$$

and
$$\begin{aligned} S(cp(x)) &= S((cp)(x)) = x((cp)(x)) \\ &= x(cp(x)) = cxp(x) = cS(p(x)) \end{aligned}$$

Therefore, S is linear. Similarly,

$$\begin{aligned} T\left(\begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix}\right) &= T\begin{bmatrix} a + c \\ b + d \end{bmatrix} \\ &= (a + c) + ((a + c) + (b + d))x \\ &= (a + (a + b)x) + (c + (c + d)x) \\ &= T\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) + T\left(\begin{bmatrix} c \\ d \end{bmatrix}\right) \end{aligned}$$

and
$$\begin{aligned} T\left(k\begin{bmatrix} a \\ b \end{bmatrix}\right) &= T\begin{bmatrix} ka \\ kb \end{bmatrix} = (ka) + (ka + kb)x \\ &= k(a + (a + b)x) = kT\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) \end{aligned}$$

Therefore, T is linear.

15. $T\begin{bmatrix} -7 \\ 9 \end{bmatrix} = 5 - 14x - 8x^2$, $T\begin{bmatrix} a \\ b \end{bmatrix} = \left(\frac{a + 3b}{4}\right) - \left(\frac{a + 7b}{4}\right)x + \left(\frac{a - b}{2}\right)x^2$

17. $T(4 - x + 3x^2) = 4 + 3x + 5x^2$, $T(a + bx + cx^2) = a + cx + \left(\frac{3a - b - c}{2}\right)x^2$

19. *Hint:* Let $a = T(E_{11})$, $b = T(E_{12})$, $c = T(E_{21})$, $d = T(E_{22})$.

23. *Hint:* Consider the effect of T and D on the standard basis for $\mathcal{P}_n$.

25. $(S \circ T)\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 & -1 \\ 0 & 6 \end{bmatrix}$, $(S \circ T)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2x & -y \\ 0 & 2x + 2y \end{bmatrix}$.

$(T \circ S)\begin{bmatrix} x \\ y \end{bmatrix}$ does not make sense.

27. $(S \circ T)(p(x)) = p'(x + 1)$, $(T \circ S)(p(x)) = (p(x + 1))' = p'(x + 1)$

29. $(S \circ T)\begin{bmatrix} x \\ y \end{bmatrix} = S\left(T\begin{bmatrix} x \\ y \end{bmatrix}\right) = S\left(\begin{bmatrix} x - y \\ -3x + 4y \end{bmatrix}\right) = \begin{bmatrix} 4(x - y) + (-3x + 4y) \\ 3(x - y) + (-3x + 4y) \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$

$(T \circ S)\begin{bmatrix} x \\ y \end{bmatrix} = T\left(S\begin{bmatrix} x \\ y \end{bmatrix}\right) = T\left(\begin{bmatrix} 4x + y \\ 3x + y \end{bmatrix}\right) = \begin{bmatrix} (4x + y) - (3x + y) \\ -3(4x + y) + 4(3x + y) \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$

Therefore, $S \circ T = I$ and $T \circ S = I$, so S and T are inverses.

Exercises 6.5

1. (a) Only (ii) is in $\ker(T)$.

(b) Only (iii) is in $\text{range}(T)$.

(c) $\ker(T) = \left\{ \begin{bmatrix} 0 & b \\ c & 0 \end{bmatrix} \right\}$, $\text{range}(T) = \left\{ \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \right\}$

3. (a) Only (iii) is in $\ker(T)$.

(b) All of them are in $\text{range}(T)$.

(c) $\ker(T) = \{a + bx + cx^2 : a = -c, b = -c\} = \{t + tx - tx^2\}$, $\text{range}(T) = \mathbb{R}^2$

5. A basis for $\ker(T)$ is $\left\{\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}, \begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}\right\}$, and a basis for $\text{range}(T)$ is $\left\{\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}, \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\right\}$; $\text{rank}(T) = \text{nullity}(T) = 2$, and $\text{rank}(T) + \text{nullity}(T) = 4 = \dim M_{22}$.

7. A basis for $\ker(T)$ is $\{1 + x - x^2\}$, and a basis for $\text{range}(T)$ is $\left\{\begin{bmatrix}1\\ 0\end{bmatrix}, \begin{bmatrix}0\\ 1\end{bmatrix}\right\}$; $\text{rank}(T) = 2$, $\text{nullity}(T) = 1$, and $\text{rank}(T) + \text{nullity}(T) = 3 = \dim \mathcal{P}_2$.

9. $\text{rank}(T) = \text{nullity}(T) = 2$

11. $\text{rank}(T) = \text{nullity}(T) = 2$

13. $\text{rank}(T) = 1$, $\text{nullity}(T) = 2$

15. One-to-one and onto

17. Neither one-to-one nor onto

19. One-to-one but not onto

21. Isomorphic, $T\begin{bmatrix}a & 0 & 0\\ 0 & b & 0\\ 0 & 0 & c\end{bmatrix} = \begin{bmatrix}a\\ b\\ c\end{bmatrix}$

23. Not isomorphic

25. Isomorphic, $T(a + bi) = \begin{bmatrix}a\\ b\end{bmatrix}$

31. *Hint:* Define $T : \mathcal{C}[0, 1] \to \mathcal{C}[0, 2]$ by letting $T(f)$ be the function whose value at x is $(T(f))(x) = f(x/2)$ for x in $[0, 2]$.

33. (a) Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be in V and let $(S \circ T)(\mathbf{v}_1) = (S \circ T)(\mathbf{v}_2)$. Then $S(T(\mathbf{v}_1)) = S(T(\mathbf{v}_2))$, so $T(\mathbf{v}_1) = T(\mathbf{v}_2)$, since S is one-to-one. But now $\mathbf{v}_1 = \mathbf{v}_2$, since T is one-to-one. Hence, $S \circ T$ is one-to-one.

35. (a) By the Rank Theorem, $\text{rank}(T) + \text{nullity}(T) = \dim V$. If T is onto, then $\text{range}(T) = W$, so $\text{rank}(T) = \dim(\text{range}(T)) = \dim W$. Therefore,

$$\dim V + \text{nullity}(T) < \dim W + \text{nullity}(T) = \text{rank}(T) + \text{nullity}(T) = \dim V$$

so $\text{nullity}(T) < 0$, which is impossible. Therefore, T cannot be onto.

Exercises 6.6

1. $[T]_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}$, $[T]_{\mathcal{C}\leftarrow\mathcal{B}}[4 + 2x]_{\mathcal{B}} =$

$$\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{bmatrix}4\\ 2\end{bmatrix} = \begin{bmatrix}2\\ -4\end{bmatrix} = [2 - 4x]_{\mathcal{C}} = [T(4 + 2x)]_{\mathcal{C}}$$

3. $[T]_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$, $[T]_{\mathcal{C}\leftarrow\mathcal{B}}[a + bx + cx^2]_{\mathcal{B}} =$

$$\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}a\\ b\\ c\end{bmatrix} = [a + b(x + 2) + c(x + 2)^2)]_{\mathcal{C}} = [T(a + bx + cx^2)]_{\mathcal{C}}$$

5. $[T]_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}1 & 0 & 0\\ 1 & 1 & 1\end{bmatrix}$, $[T]_{\mathcal{C}\leftarrow\mathcal{B}}[a + bx + cx^2]_{\mathcal{B}} =$

$$\begin{bmatrix}1 & 0 & 0\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}a\\ a + b + c\end{bmatrix} = \begin{bmatrix}a + b\cdot 0 + c\cdot 0^2\\ a + b\cdot 1 + c\cdot 1^2\end{bmatrix}_{\mathcal{C}} = [T(a + bx + cx^2)]_{\mathcal{C}}$$

7. $[T]_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}6 & 4\\ -3 & -2\\ 2 & -1\end{bmatrix}$, $[T]_{\mathcal{C}\leftarrow\mathcal{B}}\begin{bmatrix}-7\\ 7\end{bmatrix}_{\mathcal{B}} =$

$$\begin{bmatrix}6 & 4\\ -3 & -2\\ 2 & -1\end{bmatrix}\begin{bmatrix}2\\ -3\end{bmatrix} = \begin{bmatrix}0\\ 0\\ 7\end{bmatrix} = \begin{bmatrix}7\\ 7\\ 7\end{bmatrix}_{\mathcal{C}} = \begin{bmatrix}7\\ 7\\ 7\end{bmatrix}_{\mathcal{C}} = \left[T\begin{bmatrix}-7\\ 7\end{bmatrix}\right]_{\mathcal{C}}$$

9. $[T]_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}$, $[T]_{\mathcal{C}\leftarrow\mathcal{B}}[A]_{\mathcal{B}} =$

$$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}a\\ b\\ c\\ d\end{bmatrix} = \begin{bmatrix}a\\ c\\ b\\ d\end{bmatrix} = \left[\begin{bmatrix}a & c\\ b & d\end{bmatrix}\right]_{\mathcal{C}} = [T(A)]_{\mathcal{C}}$$

11. $[T]_{\mathcal{C}\leftarrow\mathcal{B}} = \begin{bmatrix}0 & -1 & 1 & 0\\ -1 & 0 & 0 & 1\\ 1 & 0 & 0 & -1\\ 0 & 1 & -1 & 0\end{bmatrix}$, $[T]_{\mathcal{C}\leftarrow\mathcal{B}}[A]_{\mathcal{B}} =$

$$\begin{bmatrix}0 & -1 & 1 & 0\\ -1 & 0 & 0 & 1\\ 1 & 0 & 0 & -1\\ 0 & 1 & -1 & 0\end{bmatrix}\begin{bmatrix}a\\ b\\ c\\ d\end{bmatrix} = \begin{bmatrix}c - b\\ d - a\\ a - d\\ b - c\end{bmatrix} = \left[\begin{bmatrix}c - b & d - a\\ a - d & b - c\end{bmatrix}\right]_{\mathcal{C}} = [AB - BA]_{\mathcal{C}} = [T(A)]_{\mathcal{C}}$$

13. (b) $[D]_{\mathcal{B}} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

(c) $[D]_{\mathcal{B}}[3\sin x - 5\cos x]_{\mathcal{B}} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 3 \\ -5 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix} = [3\cos x + 5\sin x]_{\mathcal{B}} = [D(3\sin x - 5\cos x)]_{\mathcal{B}}$

15. (a) $[D]_{\mathcal{B}} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}$

17. $[S \circ T]_{\mathcal{D}\leftarrow\mathcal{B}} = \begin{bmatrix} -1 & -2 \\ 1 & -1 \end{bmatrix}$

19. Invertible, $T^{-1}(a + bx) = -b + ax$

21. Invertible, $T^{-1}(p(x)) = p(x - 2)$

23. Invertible, $T^{-1}(a + bx + cx^2) = (a - b + 2c) + (b - 2c)x + cx^2$ or $T^{-1}(p(x)) = p(x) - p'(x) + p''(x)$

25. Not invertible **27.** $-3\sin x - \cos x + C$

29. $\frac{4}{5}e^{2x}\cos x - \frac{3}{5}e^{2x}\sin x + C$

31. $\mathcal{C} = \left\{\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -4 \\ 1 \end{bmatrix}\right\}$ **33.** $\mathcal{C} = \{1 - x, 2 + x\}$

35. $\mathcal{C} = \{1, x\}$

37. $[T]_{\mathcal{E}} = \begin{bmatrix} (d_1^2 - d_2^2)/(d_1^2 + d_2^2) & 2d_1d_2/(d_1^2 + d_2^2) \\ 2d_1d_2/(d_1^2 + d_2^2) & (d_2^2 - d_1^2)/(d_1^2 + d_2^2) \end{bmatrix}$

Exercises 6.7

1. $y(t) = 2e^{3t}/e^3$

3. $y(t) = ((1 - e^4)e^{3t} + (e^3 - 1)e^{4t})/(e^3 - e^4)$

5. $f(t) = \left(\frac{e^{(\sqrt{5}-1)/2}}{e^{\sqrt{5}} - 1}\right)[e^{(1+\sqrt{5})t/2} - e^{(1-\sqrt{5})t/2}]$

7. $y(t) = e^t - (1 - e^{-1})te^t$

9. $y(t) = ((k + 1)e^{kt} + (k - 1)e^{-kt})/2k$

11. $y(t) = e^t\cos(2t)$

13. (a) $p(t) = 100e^{\ln(16)t/3} \approx 100e^{0.924t}$

(b) 45 minutes **(c)** In 9.968 hours

15. (a) $m(t) = 50e^{-ct}$, where $c = \ln 2/1590 \approx 4.36 \times 10^{-4}$; 32.33 mg remain after 1000 years.

(b) After 3691.9 years

17. $x(t) = \dfrac{5 - 10\cos(10\sqrt{K})}{\sin(10\sqrt{K})}\sin(\sqrt{K}t) + 10\cos(\sqrt{K}t)$

19. (b) No

Review Questions

1. (a) F **(c)** T **(e)** F **(g)** F **(i)** T

3. Subspace **5.** Subspace

7. Let $c_1A + c_2B = O$. Then $c_1A - c_2B = c_1A^T + c_2B^T = (c_1A + c_2B)^T = O$. Adding, we have $2c_1A = O$, so $c_1 = 0$ because A is nonzero. Hence $c_2B = O$, and so $c_2 = 0$. Thus, $\{A, B\}$ is linearly independent.

9. $\{1, x^2, x^4\}$, dim $W = 3$

11. Linear transformation

13. Linear transformation **15.** $n^2 - 1$

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

19. $S \circ T$ is the zero transformation.

Chapter 7

Exercises 7.1

1. (a) -10 **(b)** $\sqrt{14}$ **(c)** $\sqrt{93}$

3. Any nonzero scalar multiple of $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$

5. (a) 1 **(b)** $\sqrt{13}$ **(c)** $\sqrt{14}$

7. x^2 is one possibility

9. (a) π **(b)** $\sqrt{\pi}$ **(c)** $\sqrt{\pi}$

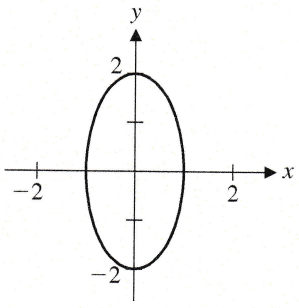
13. Axiom (4) fails: $\mathbf{u} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \neq \mathbf{0}$, but $\langle \mathbf{u}, \mathbf{u} \rangle = 0$.

15. Axiom (4) fails: $\mathbf{u} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \neq \mathbf{0}$, but $\langle \mathbf{u}, \mathbf{u} \rangle = 0$.

17. Axiom (4) fails: $p(x) = 1 - x$ is not the zero polynomial, but $\langle p(x), p(x) \rangle = 0$.

19. $A = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$

21.



25. -8 **27.** $\sqrt{6}$

29.
$$\begin{aligned} \|\mathbf{u} + \mathbf{v} - \mathbf{w}\|^2 &= \langle \mathbf{u} + \mathbf{v} - \mathbf{w}, \mathbf{u} + \mathbf{v} - \mathbf{w} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle \\ &\quad + 2\langle \mathbf{u}, \mathbf{v} \rangle - 2\langle \mathbf{u}, \mathbf{w} \rangle - 2\langle \mathbf{v}, \mathbf{w} \rangle \\ &= 1 + 3 + 4 + 2 - 10 - 0 = 0 \end{aligned}$$

Therefore, $\|\mathbf{u}+\mathbf{v}-\mathbf{w}\| = 0$, so, by axiom (4), $\mathbf{u}+\mathbf{v}-\mathbf{w} = \mathbf{0}$ or $\mathbf{u}+\mathbf{v} = \mathbf{w}$.

31. $\langle \mathbf{u}+\mathbf{v}, \mathbf{u}-\mathbf{v}\rangle = \langle \mathbf{u},\mathbf{u}\rangle - \langle \mathbf{u},\mathbf{v}\rangle + \langle \mathbf{v},\mathbf{u}\rangle - \langle \mathbf{v},\mathbf{v}\rangle = \|\mathbf{u}\|^2 - \langle \mathbf{u},\mathbf{v}\rangle + \langle \mathbf{u},\mathbf{v}\rangle - \|\mathbf{v}\|^2 = \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2$

33. Using Exercise 32 and a similar identity for $\|\mathbf{u}-\mathbf{v}\|^2$, we have

$$\begin{aligned}\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 &= \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}\rangle + \langle \mathbf{u}-\mathbf{v}, \mathbf{u}-\mathbf{v}\rangle \\ &= \|\mathbf{u}\|^2 + 2\langle \mathbf{u},\mathbf{v}\rangle + \|\mathbf{v}\|^2 \\ &\quad + \|\mathbf{u}\|^2 - 2\langle \mathbf{u},\mathbf{v}\rangle + \|\mathbf{v}\|^2 \\ &= 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2\end{aligned}$$

Dividing by 2 yields the identity we want.

35. $\|\mathbf{u}+\mathbf{v}\| = \|\mathbf{u}-\mathbf{v}\| \Leftrightarrow \|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}-\mathbf{v}\|^2 \Leftrightarrow \|\mathbf{u}\|^2 + 2\langle \mathbf{u},\mathbf{v}\rangle + \|\mathbf{v}\|^2 = \|\mathbf{u}\|^2 - 2\langle \mathbf{u},\mathbf{v}\rangle + \|\mathbf{v}\|^2 \Leftrightarrow 2\langle \mathbf{u},\mathbf{v}\rangle = -2\langle \mathbf{u},\mathbf{v}\rangle \Leftrightarrow \langle \mathbf{u},\mathbf{v}\rangle = 0$

37. $\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}\right\}$ **39.** $\{1, x, x^2\}$

41. **(a)** $1/\sqrt{2}, \sqrt{3}x/\sqrt{2}, \sqrt{5}(3x^2-1)/2\sqrt{2}$

(b) $\sqrt{7}(5x^3-3x)/2\sqrt{2}$

Exercises 7.2

1. $\|\mathbf{u}\|_E = \sqrt{42}$, $\|\mathbf{u}\|_s = 10$, $\|\mathbf{u}\|_m = 5$

3. $d_E(\mathbf{u},\mathbf{v}) = \sqrt{70}$, $d_s(\mathbf{u},\mathbf{v}) = 14$, $d_m(\mathbf{u},\mathbf{v}) = 6$

5. $\|\mathbf{u}\|_H = 4$, $\|\mathbf{v}\|_H = 5$

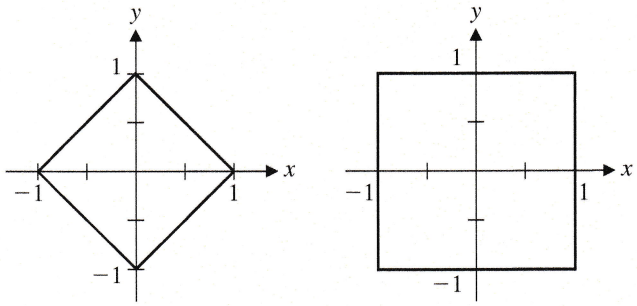
7. **(a)** At most one component of $\mathbf{v}$ is nonzero.

9. Suppose $\|\mathbf{v}\|_m = |v_k|$. Then $\|\mathbf{v}\|_E = \sqrt{v_1^2 + \cdots + v_k^2 + \cdots + v_n^2} \geq \sqrt{v_k^2} = |v_k| = \|\mathbf{v}\|_m$.

11. Suppose $\|\mathbf{v}\|_m = |v_k|$. Then $|v_i| \leq |v_k|$ for $i = 1, \ldots, n$, so

$$\begin{aligned}\|\mathbf{v}\|_s &= |v_1| + \cdots + |v_n| \leq |v_k| + \cdots + |v_k| \\ &= n|v_k| = n\|\mathbf{v}\|_m\end{aligned}$$

13.



21. $\|A\|_F = \sqrt{19}$, $\|A\|_1 = 4$, $\|A\|_\infty = 6$

23. $\|A\|_F = \sqrt{31}$, $\|A\|_1 = 6$, $\|A\|_\infty = 6$

25. $\|A\|_F = 2\sqrt{11}$, $\|A\|_1 = 7$, $\|A\|_\infty = 7$

27. $\mathbf{x} = \begin{bmatrix}0\\1\end{bmatrix}$, $\mathbf{y} = \begin{bmatrix}-1\\1\end{bmatrix}$ **29.** $\mathbf{x} = \begin{bmatrix}0\\0\\1\end{bmatrix}$, $\mathbf{y} = \begin{bmatrix}1\\1\\1\end{bmatrix}$

31. $\mathbf{x} = \begin{bmatrix}1\\0\\0\end{bmatrix}$, $\mathbf{y} = \begin{bmatrix}1\\-1\\-1\end{bmatrix}$

33. **(a)** By the definition of an operator norm, $\|I\| = \max_{\|\mathbf{x}\|=1}\|I\mathbf{x}\| = \max_{\|\mathbf{x}\|=1}\|\mathbf{x}\| = 1$.

35. $\text{cond}_1(A) = \text{cond}_\infty(A) = 21$; well-conditioned

37. $\text{cond}_1(A) = \text{cond}_\infty(A) = 400$; ill-conditioned

39. $\text{cond}_1(A) = 77$, $\text{cond}_\infty(A) = 128$; moderately ill-conditioned

41. **(a)** $\text{cond}_\infty(A) = (\max\{|k|+1, 2\}) \cdot \left(\max\left\{\left|\frac{k}{k-1}\right| + \left|\frac{1}{k-1}\right|, \left|\frac{2}{k-1}\right|\right\}\right)$

43. **(a)** $\text{cond}_\infty(A) = 40$

(b) At most 400% relative change

45. Using Exercise 33(a), we have $\text{cond}(A) = \|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|I\| = 1$.

49. $k \geq 6$ **51.** $k \geq 10$

Exercises 7.3

1. $\|\mathbf{e}\| = \sqrt{2} \approx 1.414$ **3.** $\|\mathbf{e}\| = \sqrt{6}/2 \approx 1.225$

5. $\|\mathbf{e}\| = \sqrt{7} \approx 2.646$

7. $y = -3 + \frac{5}{2}x$, $\|\mathbf{e}\| \approx 1.225$

9. $y = \frac{11}{3} - 2x$, $\|\mathbf{e}\| \approx 0.816$

11. $y = \frac{7}{10} + \frac{8}{25}x$, $\|\mathbf{e}\| \approx 0.447$

13. $y = -\frac{1}{5} + \frac{7}{5}x$, $\|\mathbf{e}\| \approx 0.632$

15. $y = 3 - \frac{18}{5}x + x^2$ **17.** $y = \frac{18}{5} - \frac{17}{10}x - \frac{1}{2}x^2$

19. $\bar{\mathbf{x}} = \begin{bmatrix}\frac{1}{5}\\ \frac{7}{15}\end{bmatrix}$ **21.** $\bar{\mathbf{x}} = \begin{bmatrix}\frac{4}{3}\\ -\frac{5}{6}\end{bmatrix}$

23. $\bar{\mathbf{x}} = \begin{bmatrix}4+t\\ -5-t\\ -5-2t\\ t\end{bmatrix}$ **25.** $\begin{bmatrix}\frac{42}{11}\\ \frac{19}{11}\\ \frac{42}{11}\end{bmatrix}$

27. $\bar{\mathbf{x}} = \begin{bmatrix} \frac{5}{3} \\ -2 \end{bmatrix}$ **29.** $y = 0.92 + 0.73x$

31. (a) If we let the year 1920 correspond to $t = 0$, then $y = 56.6 + 2.9t$; 79.9 years

33. (a) $p(t) = 150e^{0.131t}$

35. 139 days

37. $\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{7}{2} \\ \frac{7}{2} \end{bmatrix}$ **39.** $\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$

41. $\begin{bmatrix} \frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{6} & \frac{1}{3} & \frac{5}{6} \end{bmatrix}, \begin{bmatrix} \frac{5}{6} \\ \frac{1}{3} \\ -\frac{1}{6} \end{bmatrix}$ **45.** $A^+ = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} \end{bmatrix}$

47. $A^+ = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}$ **49.** $A^+ = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$

51. $A^+ = \begin{bmatrix} \frac{2}{3} & 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & 1 & \frac{1}{3} & -\frac{1}{3} \end{bmatrix}$

53. (a) If A is invertible, so is A^T, and we have $A^+ = (A^TA)^{-1}A^T = A^{-1}(A^T)^{-1}A^T = A^{-1}$.

Exercises 7.4

1. 2, 3 **3.** $\sqrt{2}$, 0 **5.** 5 **7.** 2, 3

9. $\sqrt{5}$, 2, 0

11. $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$

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

15. $A = \begin{bmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix}\begin{bmatrix} 5 \\ 0 \end{bmatrix}[1]$

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

19. $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}\begin{bmatrix} 2/\sqrt{5} & 0 & 1/\sqrt{5} \\ 0 & 1 & 0 \\ 1/\sqrt{5} & 0 & -2/\sqrt{5} \end{bmatrix}$

21. $A = \sqrt{2}\begin{bmatrix} 1 \\ 0 \end{bmatrix}[1/\sqrt{2} \quad 1/\sqrt{2}] + 0\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
$[-1/\sqrt{2} \quad 1/\sqrt{2}]$ (Exercise 3)

23. (Exercise 7) $A = 3\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}[0 \quad 1] + 2\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}[1 \quad 0]$

33. The line segment $[-1, 1]$

35. The solid ellipse $\frac{y_1^2}{5} + \frac{y_2^2}{4} \le 1$

37. (a) $\|A\|_2 = \sqrt{2}$ **(b)** $\text{cond}_2(A) = \infty$

39. (a) $\|A\|_2 = 1.95$ **(b)** $\text{cond}_2(A) = 38.11$

41. $A^+ = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 0 \end{bmatrix}$ **43.** $A^+ = \begin{bmatrix} \frac{2}{5} & 0 \\ 0 & \frac{1}{2} \\ \frac{1}{5} & 0 \end{bmatrix}$

45. $A^+ = \begin{bmatrix} \frac{1}{25} & \frac{2}{25} \\ \frac{2}{25} & \frac{4}{25} \end{bmatrix}, \bar{\mathbf{x}} = \begin{bmatrix} 0.52 \\ 1.04 \end{bmatrix}$

47. $A^+ = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}, \bar{\mathbf{x}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

61. $\begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

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

Exercises 7.5

1. $g(x) = \frac{1}{3}$ **3.** $g(x) = \frac{3}{5}x$

5. $g(x) = \frac{3}{16} + \frac{15}{16}x^2$ **7.** $\{1, x - \frac{1}{2}\}$

9. $g(x) = x - \frac{1}{6}$

11. $g(x) = (4e - 10) + (18 - 6e)x \approx 0.87 + 1.69x$

13. $g(x) = \frac{1}{20} - \frac{3}{5}x + \frac{3}{2}x^2$

15. $g(x) = 39e - 105 + (588 - 216e)x + (210e - 570)x^2 \approx 1.01 + 0.85x + 0.84x^2$

21. $\frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{\cos 3x}{9}\right)$

23. $a_0 = \frac{1}{2}, a_k = 0, b_k = \dfrac{1 - (-1)^k}{k\pi}$

25. $a_0 = \pi, a_k = 0, b_k = \dfrac{2(-1)^k}{k}$

Review Questions

1. (a) T **(c)** F **(e)** T **(g)** T **(i)** T

3. Inner product **5.** $\sqrt{3}$ **7.** $\left\{\begin{bmatrix} \frac{1}{1} \end{bmatrix}\right\}, \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$

9. Not a norm

11. $\text{cond}_\infty(A) \approx 2432$

13. $y = 1.7x$

15. $\begin{bmatrix} \frac{7}{3} \\ \frac{2}{3} \\ \frac{5}{3} \end{bmatrix}$

17. (a) $\sqrt{2}, \sqrt{2}$

(b) $A = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

(c) $A^+ = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix}$

19. The singular values of PAQ are the square roots of the eigenvalues of $(PAQ)^T(PAQ) = Q^TA^TP^TPAQ = Q^T(A^TA)Q$. But $Q^T(A^TA)Q$ is similar to A^TA because $Q^T = Q^{-1}$, and hence it has the same eigenvalues as A^TA. Thus, PAQ and A have the same singular values.

Index

A

Abel, Niels Henrik, 311, D**8**
Absolute value, C3
Addition
  closure under, 192, 429
  of complex numbers, C**1**
  of matrices, 140
  of polynomials, D**2**
  of vectors, 5–6, 9, 429
Adjacency matrix, 242, 244
Adjoint (adjugate) of a matrix, 276
Al-Khwarizmi, Abu Ja'far Muhammad ibn Musa, 85
Algebraic multiplicity, 294
Algebraic properties of vectors, 10
Algorithm, 85
Allocation of resources, 99–101
Altitude of a triangle, 33
Angle between vectors, 24–26
Argand, Jean-Robert, C**1**
Argand plane, C**1**
Argument of a complex number, C**4**
Arithmetic mean, 548
Arithmetic Mean—Geometric Mean Inequality, 548
Associativity, 10, 154, 158, 223, 429
Attractor, 350
Augmented matrix, 61, 64
Axioms
  inner product space, 531
  vector space, 429

B

Back substitution, 61
Balanced chemical equation, 101–102
Basis, 198, 446–448
  change of, 463–470, 507–509
  coordinates with respect to, 208, 448–452
  orthogonal, 370, 537
  orthonormal, 372, 537
  standard, 198, 447
Basis step, B**1**
Basis Theorem, 202, 453
Best Approximation Theorem, 570
Best approximation, to a vector, 570–571
Binary vector, 14
Binet, Jacques, 338
Binet's formula, 339, 428
Bipartite graph, 251, 254
Block, 145
Block multiplication, 148
Block triangular form, 283
Bunyakovsky, Viktor Yakovlevitch, 539

C

$\mathscr{C}$, 435
$\mathbb{C}^2$, 432
$\mathbb{C}^n$, 543
Carroll, Lewis, 141, 284
Cassini, Giovanni Domenico, 362
Cassini's identity, 362
Cauchy, Augustin-Louis, 273, 280
Cauchy-Schwarz Inequality, 22, 539–540
Cayley, Arthur, 300
Cayley-Hamilton Theorem, 300
Centroid of a triangle, 32
Change of basis, 463–470
Characteristic equation, 292
Characteristic polynomial, 292
Circuit, 242
Circumcenter of a triangle, 33
Closure
  under addition, 192, 429
  under linear combinations, 192
  under scalar multiplication, 192, 429
Codomain, 212
Coefficient(s)
  Fourier, 615
  of a linear combination, 12, 154
  of a linear equation, 58
  matrix, 64
  method of undetermined, D7
  of a polynomial, D**1**
Cofactor, 266
Cofactor expansion, 266–269
Column matrix, 138
Column-row representation of a matrix product, 147
Column space, 195
Column vector, 3, 138
Commutativity, 10, 19, 154, 429
Companion matrix, 299
Complete bipartite graph, 254
Complex dot product, 543
Complex numbers, C**1**–C**11**
  absolute value of, C**3**
  addition of, C**1**
  argument of, C**4**
  conjugate of, C**2**
  division of, C**2**, C**5**
  equality of, C**1**
  imaginary part of, C**1**
  modulus of, C**3**
  multiplication of, C**1**–C**2**, C**5**
  negative of, C**2**
  polar form of, C**3**–C**6**
  powers of, C**6**–C**7**
  principal argument of, C**4**
  real part of, C**1**
  roots of, C**7**–C**8**
Complex plane, C**1**
Complex vector space, 429, 543
Component of a vector, 3
  orthogonal to a subspace, 382, 538
Composition of linear transformations, 219, 476–478
Condensation method, 284–285
Condition number, 562, 602
Conic sections, 415–416
Conjugate of complex numbers, C**2**–C**3**
Conjugate transpose of a matrix, 544–545
Connected graph, 361
Conservation of flow, 102
Consistent linear system, 60
Constant polynomial, D**1**
Constrained optimization, 413–415, 547–551
Consumption matrix, 236
Contradiction, proof by, A**8**
Contrapositive, proof by, A**8**
Convergence of iterative methods, 125, 316, 563–566
Coordinate grid, 13
Coordinate vector, 208, 448–452
Coordinates, 207–209
Cotes, Roger, 569
Cramer, Gabriel, 274
Cramer's Rule, 274–275
Cross product, 48–49, 286–287
Crystallographic restriction, 517
Curve fitting, 290–291

D

$\mathscr{D}$, 435
De Moivre, Abraham, C**6**
De Moivre's Theorem, C**6**–C**9**
Degenerate conic, 415, 424
Degree of a polynomial, D**1**
Demand vector, 236
Descartes, René, 3, D**9**

Descartes' Rule of Signs, D9–D10
Determinant(s), 165, 263–265
  cofactor expansion of, 266–269
  of elementary matrices, 271–272
  geometric applications of, 286–291
  history of, 280–281
  and matrix operations, 272–274
  of $n \times n$ matrices, 265–269
  properties of, 269–274
  Vandermonde, 291
Diagonal entries of a matrix, 139
Diagonal matrix, 139
Diagonalizable linear transformation, 509
Diagonalizable matrix, 303
  orthogonally, 400
  unitarily, 546–547
Diagonalization, 303–309
  orthogonal, 400–407
Diagonalization Theorem, 307
Diagonalizing a quadratic form, 411
Diagonally dominant matrix, 128, 324
Difference
  of complex numbers, C2
  of matrices, 140
  of polynomials, D2
  of vectors, 8, 433
Differential equation(s), 363, 436, 518
  boundary conditions for, 523
  homogeneous, 436, 518–525
  initial conditions for, 340, 343, 344, 363
  solution of, 518
  system of linear, 340–348
Differential operator, 473
Digital image compression, 607–608
Digraph, 243
Dimension, 203, 452–456
Direct proof, A7
Direction vector, 35, 39
Disjoint sets, A4
Distance
  Hamming, 554
  from a point to a line, 41–43
  from a point to a plane, 43–44
  taxicab, 529–531
  between vectors, 23–24, 535
Distance functions, 554–555
Distributivity, 10, 19, 154, 158, 429
Divergence, 127
Division algorithm, D4
Dodgson, Charles Lutwidge, 281, 284
Domain, 212
Dominant eigenvalue, 311
Dominant eigenvector, 311
Dot product, 18–20, 49
  complex, 543
  weighted, 532
Dual space, 514
Dynamical system, 253, 348–355
  trajectory of, 349

E

Echelon form of a matrix
  reduced row, 73
  row, 65
Edge of a graph, 242
Eigenspace, 256
Eigenvalue(s), 254
  algebraic multiplicity of, 294
  dominant, 311
  geometric multiplicity of, 294
  inverse power method for computing, 317–318
  power method for computing, 311–316
  shifted inverse power method for computing, 318–319
  shifted power method for computing, 316–317
Eigenvector(s), 254
  dominant, 311
  orthogonal, 402
Electrical network, 104–107
Elementary matrix, 170
Elementary reflector, 397
Elementary row operations, 66
Elements of a matrix, 138
Elimination
  Gauss-Jordan, 72–76
  Gaussian, 68–72
Empty set, A2
Equality
  of complex numbers, C1
  of matrices, 139
  of polynomials, D2
  of sets, A1–A2
  of vectors, 4
Equation(s)
  linear, 58
  normal, 575
  system of linear, 59
Equilibrium, 50, 107
Equivalence relation, 302
Error vector, 565, 572
Euclidean norm, 553
Euler, Leonhard, C9
Euler's formula, C9–C11
Even function, 617
Exchange matrix, 235
Expansion by cofactors, 266–269
Exponential of a matrix, 346

F

$\mathcal{F}$, 431
Factor Theorem, D4
Factorization
  LU, 180–186
  modified *QR*, 396–398
  QR, 392–394
Feasible solution, 236
Fibonacci, 336
Fibonacci numbers, 335, 338–339, 427
Field, 429
Finite-dimensional vector space, 453
Finite linear games, 109–113
Floating-point form, 83
Force vectors, 50–53
Fourier approximation, 615
Fourier coefficients, 615
Fourier, Jean-Baptiste Joseph, 616
Fourier series, 617
Free variable, 71
Frobenius, Georg, 204
Frobenius norm, 556
Fundamental subspaces of a matrix, 380
Fundamental Theorem of Algebra, D8
Fundamental Theorem of Invertible Matrices, 172, 206, 296, 512, 605–606

G

Galilei, Galileo, 526
Galois, Evariste, 311, D8
Gauss, Carl Friedrich, 69, 125, 538, 569, D8
Gauss-Jordan elimination, 72–76
Gauss-Jordan inverse method, 175–178
Gauss-Seidel method, 124–131
Gaussian elimination, 68–72
General form of the equation of a line, 34, 36, 41
General form of the equation of a plane, 38, 41
Geometric mean, 548
Geometric multiplicity, 294
Gerschgorin disk, 319
Gerschgorin Disk Theorem, 321
Gerschgorin, Semyon Aranovich, 319
Gerschgorin's theorem, 319–322
Gibbs, Josiah Willard, 49
Global Positioning System (GPS), 121–123
Google, 358
Gram, Jörgen Pedersen, 390
Gram-Schmidt Process, 388–392
Graph, 242, 253–254
  adjacency matrix of, 242, 244
  bipartite, 251
  complete, 253
  complete bipartite, 254
  connected, 361
  cycle, 254
  directed (digraph), 243
  edges of, 242
  k-regular, 361
  path in a, 242
  Petersen, 254
  vertices of, 242
Grassmann, Hermann, 429
Grassmann's Identity, 458, 496

H

Half-life, 520
Hamilton, William Rowan, 2, 300
Hamming distance, 554
Hamming norm, 554
Harmonic mean, 551
Head of a vector, 3
Head-to-tail rule, 6
Hermitian matrix, 545
Hilbert, David, 403
Hoëné-Wronski, Jósef Maria, 457
Homogeneous linear differential equations, 518–525
Homogeneous linear system, 76
Hooke's Law, 524
Householder, Alston Scott, 397
Householder matrix, 397
Hyperplane, 40

I
i, C1
Idempotent matrix, 179
Identity matrix, 139
Identity transformation, 221, 474
Ill-conditioned linear system, 84
Ill-conditioned matrix, 561
Image, 212
Imaginary axis, C1
Imaginary conic, 424
Imaginary part of a complex number, C1
Inconsistent linear system, 60
Indefinite matrix, 413
  quadratic form of, 413
Index of summation, A5
Indirect proof, A7
Induction hypothesis, B1
Induction step, B1
Infinite-dimensional vector space, 453
Initial point of a vector, 3
Inner product, 531
Inner product space, 531–534
  and Cauchy-Schwarz and Triangle Inequalities, 539–540
  distance between vectors in, 535
  length of vectors in, 535
  orthogonal vectors in, 535
  properties of, 535
Integers modulo m, 14–16
Interior of a matrix, 284
Intersection of sets, A4
Inverse
  Gauss-Jordan method of computing, 175–178
  of a linear transformation, 221–222, 478–479
  of a matrix, 163
Inverse power method, 317–318
  shifted, 318–319
Invertible linear transformation, 221–222, 478–479
Invertible matrix, 163–170
Irreducible matrix, 335
Irreducible polynomial, D7
Isometry, 375
Isomorphism, 493–495
Iterative method(s)
  convergence of, 125, 316, 563–566
  Gauss-Seidel method, 124–131
  inverse power method, 317–318
  Jacobi's method, 124–131
  power method, 311–316
  shifted inverse power method, 318–319
  shifted power method, 316–317

J
Jacobi, Carl Gustav, 124
Jacobi's method, 124–131
Jordan, Wilhelm, 72

K
Kernel, 482
Kirchhoff's Laws, 104

L
Lagrange interpolation formula, 459
Lagrange, Joseph-Louis, 458
Lagrange polynomials, 458
Laplace Expansion Theorem, 266, 277–280
Laplace, Pierre Simon, 267
Lattice, 516
Leading entry, 65
Leading 1, 73
Leading variable, 71
Least squares approximating line, 574
Least squares approximation, 568–569, 571–582
  Best Approximation Theorem and, 570–571
  and orthogonal projection, 583–585
  and the pseudoinverse of a matrix, 585–586
  via the QR factorization, 582–583
  via the singular value decomposition, 603–605
Least squares error, 572
Least squares solution, 574
  of minimal length, 603–604
Least Squares Theorem, 575
Left singular vectors, 593
Legendre, Adrien Marie, 538
Legendre polynomials, 538
Leibniz, Gottfried Wilhelm von, 281
Lemma, 271
Length
  of a binary vector, 14
  of an m-ary vector, 16
  of a path, 242
  of a vector, 20, 535
Leonardo of Pisa, 336
Leontief closed model, 108, 235
Leontief open model, 108, 236
Leontief, Wassily, 107
Leslie matrix, 240
Leslie model, 239–241, 330–332
Line, 34–38
  of best fit, 574
  equation(s) of, 34, 36, 41
  least squares approximating, 574
Linear combination, 12, 154, 433
Linear dependence, 92–93, 157, 443, 446
Linear economic models, 107–109, 235–236
Linear equation(s), 58, 59. *See also* Linear system(s)
Linear independence, 92–97, 157, 443–446
Linear recurrence relations, 335–336
Linear system(s), 58–62
  augmented matrix of, 61, 64
  coefficient matrix of, 64
  consistent, 60
  direct method for solving, 64–79
  equivalent, 60
  homogeneous, 76
  ill-conditioned, 84
  inconsistent, 60
  iterative methods for solving, 124–131
  solution (set) of, 59
  over $\mathbb{R}_p$, 77–79
Linear transformation(s), 213–214, 472–474
  onto, 488
  composition of, 219, 476–478
  diagonalizable, 509
  identity, 221, 474
  inverse of, 221–222, 478–479
  invertible, 221–222, 478–479
  kernel of, 482
  matrix of, 216, 497–503
  nullity of, 484
  one-to-one, 488
  properties of, 475–476
  zero, 474
Linearly dependent matrices, 157
Linearly dependent vectors, 93, 443, 446
Linearly independent matrices, 157
Linearly independent vectors, 93, 443, 446
Long range transition matrix, 329
LU factorization, 180–186
Lucas, Edouard, 336, 428

M
m-ary vector, 16
M_{mn}, 430
Maclaurin, Colin, 274, 280
Magic square, 460–462
  classical, 460
  weight of a, 460
Mantissa, 83
Markov, Andrei Andreyevich, 230
Markov chain, 230–235, 325–330
Mathematical induction, B1–B7
  first principle of, B1
  second principle of, B5
Matrix (matrices), 61, 138
  addition of, 140
  adjacency, 242, 244
  adjoint (adjugate), 276
  associated with a quadratic form, 409
  augmented, 61, 64
  change-of-basis, 465
  characteristic equation of, 292
  characteristic polynomial of, 292
  coefficient, 64
  column space of, 195
  companion, 299
  condition number of, 562
  conjugate transpose of, 544–545
  consumption, 236
  determinant of, 165, 264, 265–269
  diagonal, 139
  diagonalizable, 303–309
  difference of, 140
  eigenspace of, 256
  eigenvalue of, 254
  eigenvector of, 254
  elementary, 170
  elements of, 138
  entries of, 138
  equality of, 139
  exchange, 235
  exponential of, 346
  factorization of, 180
  fundamental subspaces of, 380
  Hermitian, 545
  idempotent, 179
  identity, 139
  ill-conditioned, 561
  indefinite, 413
  interior, 284
  inverse of, 163
  invertible, 163–170

Matrix (*Continued*)
irreducible, 335
Leslie, 240
of a linear transformation, 216, 497–503
multiplication of, 141–143
negative definite, 413
negative of, 140
negative semidefinite, 413
nilpotent, 282
norm of, 555–561
normal, 547
null space of, 197
nullity of, 204
orthogonal, 373–376
orthogonally diagonalizable, 400
partitioned, 145–149
permutation, 187
positive, 325
positive definite, 413
positive semidefinite, 413
powers of, 149–150
primitive, 335
productive, 237–238
projection, 218–219, 366, 586
pseudoinverse of, 585–586, 602–603
rank of, 72, 204
reduced row echelon form of, 73
reducible, 334
regular, 325
row echelon form of, 65
row equivalent, 68
row space of, 195
scalar, 139
scalar multiple of, 140
similar, 301–303
singular values of, 590–591
singular vectors of, 593
size of, 138
skew-symmetric, 162
square, 139
standard, 216
stochastic, 232
strictly diagonally dominant, 128
sum of, 140
symmetric, 151–152, 160–161
trace of, 162
transition, 231
transpose of, 151, 159–160
unit lower triangular, 181
unitarily diagonalizable, 546–547
unitary, 545–546
upper triangular, 162
zero, 141
Matrix-column representation of a matrix product, 146
Matrix factorization, 180. *See also* Singular value decomposition (SVD)
and diagonalization, 303–309
LU, 180–186
modified *QR*, 396–398
$P^T LU$, 186–187
QR, 392–394
and Schur's Triangularization Theorem, 408
Matrix transformation, 211–216, 472
projection, 218–219, 509–510
reflection, 215
rotation, 216–218
Max norm, 553
Mean
arithmetic, 548
geometric, 548
harmonic, 551
quadratic, 550
Median of a triangle, 32
Metric, 555
Metric space, 555
Minimum length least squares solution, 603–604
Minor, 264
Modified *QR* factorization, 396–398
Modular arithmetic, 13–16
Modulus of a complex number, C3
Moore, Eliakim Hastings, 602
Moore-Penrose inverse, 602
Muir, Thomas, 281
Multiplication
of complex numbers, C1–C2, C5
of matrices, 141–143
of polynomials, D2–D3
scalar, 7–8, 140, 429
Multiplicity of an eigenvalue
algebraic, 294
geometric, 294

N

Negative
of a complex number, C2
of a matrix, 140
of a vector, 8, 429
Negative definite matrix, 413
quadratic form of, 413
Negative semidefinite matrix, 413
quadratic form of, 413
Net reproduction rate, 360
Network, 102
Network analysis, 102–103
Newton's Second Law of Motion, 524
Nilpotent matrix, 282
Node, 102
Nondegenerate conic, 415–416
Norm of a matrix, 555–561
1-, 559
2-, 559
7-, 559
compatible, 556
Frobenius, 556
operator, 559
Norm of a vector, 20, 535, 552
1-, 553
2-, 553
7-, 553
Euclidean, 553
Hamming, 554
max, 553
sum, 552
taxicab, 530
uniform, 553
Normal equations, 575
Normal form of the equation of a line, 34, 36, 41
Normal form of the equation of a plane, 38, 41
Normal matrix, 547
Normal vector, 34, 38
Normalizing a vector, 21
Normed linear space, 552
Null space, 197
Nullity
of a linear transformation, 484
of a matrix, 204

O

Odd function, 617
Ohm's Law, 104
One-to-one, 488
Onto, 488
Operator norm, 559
Optimization
constrained, 413–415
geometric inequalities and, 547–551
Orbital center, 355
Ordered *n*-tuple, 9
Ordered pair, 3
Ordered triple, 8
Orthocenter of a triangle, 33
Orthogonal basis, 370, 537
Orthogonal complement, 378–382
Orthogonal Decomposition Theorem, 384–385
Orthogonal diagonalization, 400–407
Orthogonal matrix, 373–376
Orthogonal projection, 382–387, 538
least squares approximation, 583–585
Orthogonal set of vectors, 369–373, 537
Orthogonal vectors, 26, 535
Orthonormal basis, 372, 537
Orthonormal set of vectors, 372, 537
Outer product, 147
Outer product expansion, 147
Outer product form of the SVD, 596

P

$\mathcal{P}$, 431
$\mathcal{P}_n$, 431
Parallel vectors, 8
Parallelogram rule, 6
Parameter, 36
Parametric equation
of a line, 36, 41
of a plane, 39, 41
Partial fractions, 119
Partial pivoting, 84–85
Partitioned matrix, 145–149
Path(s)
k-, 243
length of, 242
number of, 242–245
simple, 242
Peano, Giuseppe, 429
Penrose conditions, 586
Penrose, Roger, 603
Permutation matrix, 187
Perpendicular bisector, 33

Perron eigenvector, 335
Perron-Frobenius Theorem, 332–335
Perron, Oskar, 332
Perron root, 335
Perron's Theorem, 333
Petersen graph, 254
Pivot, 66
Pivoting, 66
  partial, 84–85
Plane, 38–41
  Argand, C1
  complex, C1
  equation of, 38, 39, 41
Polar decomposition, 610
Polar form of a complex number, C3–C6
Pólya, George, A7
Polynomial, D1–D10
  characteristic, 292
  constant, D1
  degree of, D1
  irreducible, D7
  Lagrange, 458
  Legendre, 538
  Taylor, 472
  trigonometric, 614
  zero of, D4
Population distribution vector, 239
Population growth, 239–241, 330–332
Positive definite matrix, 413
  quadratic form of, 413
Positive matrix, 325
Positive semidefinite matrix, 413
  quadratic form of, 413
Power method, 311–316
  inverse, 317–318
  shifted, 316–317
  shifted inverse, 318–319
Predator-prey model, 343
Price vector(s), 235
Primitive matrix, 335
Principal argument of a complex number, C4
Principal Axes Theorem, 411
Probability vector, 231
Product
  of complex numbers, C1–C2
  of matrices, 141–143
  of polynomials, D2–D3
Production vector, 236
Projection
  orthogonal, 382–387, 538
  into a subspace, 382
  onto a vector, 27–28
Projection form of the Spectral Theorem, 405
Projection matrix, 218–219, 366, 586
Proof
  by contradiction, A8
  by contrapositive, A8
  direct, A7
  indirect, A7
  by mathematical induction, B1–B7
Pseudoinverse of a matrix, 585–586, 602–603
Pythagoras' Theorem, 26, 537

Q
QR algorithm, 398–399
QR factorization, 392–394
  least squares and, 582–583
  modified, 396–398
Quadratic equation(s), D6
  graphing, 415–423
Quadratic form, 408–416
  indefinite, 413
  matrix associated with, 409
  negative definite, 413
  negative semidefinite, 413
  positive definite, 413
  positive semidefinite, 413
Quadratic mean, 550
Quadric surface, 420
Quotient of complex numbers, C2, C5

R
$\mathbb{R}$, 4
$\mathbb{R}^3$, 8
$\mathbb{R}^n$, 9–11
Racetrack game, 1–3
Range, 212, 482
Rank
  of a linear transformation, 484
  of a matrix, 72, 204
  singular value decomposition, 600
Rank Theorem, 72, 205, 386, 486
Ranking vector, 356–358
Rational Roots Theorem, D5
Rayleigh, Baron, 316
Rayleigh quotient, 316
Real axis, C1
Real part of a complex number, C1
Recurrence relation, 336
  solution of, 337
Reduced row echelon form, 73
Reducible matrix, 334
Reflection, 215
Regular graph, 361
Regular matrix, 325
Repeller, 352
Resolving a vector, 51
Resultants, 50
Right singular vectors, 593
Robotics, 226–229
Root, for a polynomial equation, D4
Root mean square error, 612
Rotation, 216–218
  center of, 516
Rotational symmetry, 516
Roundoff error, 62, 83–84
Row echelon form, 65
Row equivalent matrices, 68
Row matrix, 138
Row-matrix representation of a matrix product, 146
Row reduction, 66
Row space, 195
Row vector, 3, 138

S
Saddle point, 352
Scalar, 8
Scalar matrix, 139
Scalar multiple, 481
Scalar multiplication, 7–8, 9, 140, 429
  closure under, 192, 429
Scaling, 314
Schmidt, Erhard, 390
Schur complement, 283
Schur, Issai, 283
Schur's Triangularization Theorem, 408
Schwarz, Karl Herman Amandus, 539
Seidel, Philipp Ludwig, 125
Seki Kōwa, Takakazu, 280
Set(s), A1–A4
  disjoint, A4
  elements of, A1
  empty, A2
  intersection of, A4
  subset of, A2
  union of, A4
Shifted inverse power method, 318–319
Similar matrices, 301–303, 508
Simple path, 242
Singular value decomposition (SVD), 590–599
  applications of, 599–606
  and condition number, 602
  and least squares approximation, 603–605
  and matrix norms, 600–602
  outer product form of, 596
  and polar decomposition, 610
  and pseudoinverse, 602–603
  and rank, 600
Singular values, 590–591
Singular vectors, 593
Size of a matrix, 138
Skew lines, 76
Skew-symmetric matrix, 162
Solution
  of a differential equation, 518
  least squares, 574–582
  of a linear system, 59
  minimum length least squares, 603
  of a recurrence relation, 337
  of a system of differential equations, 340–342
Span, 90, 156, 193, 438
Spanning set of vectors, 88–92
Spanning sets, 438–441
Spectral decomposition, 405
Spectral Theorem, 403
  projection form of, 405
Spectrum, 403
Spiral attractor, 355
Spiral repeller, 355
Square matrix, 139, 374
Square root of a matrix, 424
Standard basis, 198, 447
Standard matrix, 216
Standard position, 4
Standard unit vectors, 22
State vector, 231
Steady state vector, 233

Stochastic matrix, 232
Strictly diagonally dominant matrix, 324
Strutt, John William, 316
Submatrices, 145
Subset, A2
Subspace(s), 192, 433–438
  fundamental, 380
  spanned by a set of vectors, 192–193, 441
  sum of, 442
  trivial, 437
  zero, 437
Subtraction
  of complex numbers, C2
  of matrices, 140
  of polynomials, D2
  of vectors, 8, 433
Sum
  of complex numbers, C1
  of linear transformations, 481
  of matrices, 140
  of polynomials, D2
  of subspaces, 442
  of vectors, 5–6, 9, 439
Sum norm, 552
Summation notation, A4–A7
Sustainable harvesting policy, 360
Sylvester, James Joseph, 206, 280
Symmetric matrix, 151–152, 160–161
System of linear differential equations, 340–348
System(s) of linear equations. *See* Linear system(s)

T
Tail of a vector, 3
Taussky-Todd, Olga, 320
Taxicab circle, 530
Taxicab distance, 529–531
Taxicab norm, 530
Taxicab perpendicular bisector, 530
Taxicab pi, 530
Taylor polynomial, 472
Terminal point of a vector, 3
Ternary vector, 16
Theorem, 10
Tiling, 515
Tournament, 244
Trace of a matrix, 162
Transformation, 212
  linear, 213–214, 472–474
  matrix, 211–216, 472
Transitional matrix, 231
Transitional probabilities, 230
Translational symmetry, 516
Transpose of a matrix, 151, 159–160
Triangle inequality, 22, 540, 552
Trigonometric polynomial, 614
Triple scalar product identity, 287
Trivial subspace, 437
Turing, Alan Mathison, 181

U
Uniform norm, 553
Union of sets, A4
Unit circle, 21
Unit lower triangular matrix, 181
Unit sphere, 535
Unit vector, 21, 535
Unitarily diagonalizable matrix, 546–547
Unitary matrix, 545–546
Upper triangular matrix, 162
  block, 283

V
Vandermonde, Alexandre-Théophile, 291
Vandermonde determinant, 291
Vector(s), 3, 9, 429, 439
  addition of, 5–6, 9, 439
  algebraic properties of, 10
  angle between, 24–26
  binary, 14
  column, 3, 138
  complex, 429, 432, 543–544
  complex dot product of, 543
  components of, 3
  coordinate, 208, 448–452
  cross product of, 48–49, 286–287
  demand, 236
  direction, 35, 39
  distance between, 23–24, 535
  dot product of, 18–20
  equality of, 3
  force, 50–53
  inner product of, 531
  length of, 20, 535
  linear combination of, 12, 433
  linearly dependent, 92–93, 443, 446
  linearly independent, 92–97, 443, 446
  norm of, 20, 535, 552
  normal, 34, 38
  orthogonal, 26, 369–373, 535, 537
  orthonormal, 372, 537
  parallel, 8
  population distribution, 240
  price, 235
  probability, 231
  production, 236
  ranking, 356–358
  resultant, 50
  row, 3, 138
  scalar multiplication of, 7–8, 9, 429
  span of, 438
  spanning sets of, 88–92
  state, 231
  steady-state, 233
  ternary, 16
  unit, 21, 535
  zero, 4, 429
Vector form of the equation of a line, 36, 41
Vector form of the equation of a plane, 39, 41
Vector space(s), 429
  basis for, 446
  complex, 429, 432, 543–544
  dimension of, 453
  finite-dimensional, 453
  infinite-dimensional, 453
  isomorphic, 493–495
  subspace of, 433–438
  over $\mathbb{Z}_p$, 429, 432
Venn diagram, A2–A3
Venn, John, A2
Vertex of a graph, 242

W
Weight of a magic square, 460
Weighted dot product, 532
Well-conditioned matrix, 561
Wessel, Caspar, C1
Weyl, Hermann, 429
Wheatstone bridge circuit, 105–106
Wilson, Edwin B., 49
Wronskian, 457

X
x-axis, 3
xy-plane, 8
xz-plane, 8

Y
y-axis, 3
yz-plane, 8

Z
$\mathbb{Z}$, 14
$\mathbb{Z}_2$, 14
$\mathbb{Z}_2^n$, 14
$\mathbb{Z}_m$, 16
$\mathbb{Z}_m^n$, 16
Z-axis, 8
Zero matrix, 141
Zero of a polynomial, D4
Zero subspace, 437
Zero transformation, 474
Zero vector, 4, 429

PEDAGOGICKÁ FAKULTA
ÚSTŘEDNÍ KNIHOVNA
BRNO

Index of Notation

$\mathbf{v}$ vector, 3, 439

$\overrightarrow{AB}$ vector as directed line segment, 3

$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ vector as ordered n-tuple, 9

$\mathbb{R}^n$ vector space of ordered n-tuples of real numbers, 9

$\mathbb{Z}_2$ the integers modulo 2, 14

$\mathbb{Z}_2^n$ binary vectors of length n, 14

$\mathbb{Z}_m$ the integers modulo m, 14

$\mathbb{Z}_m^n$ m-ary vectors of length n, 16

$\mathbf{0}$ zero vector, 4

$\mathbf{u} \cdot \mathbf{v}$ dot product of vectors, 18

$\|\mathbf{v}\|$ length (norm) of a vector, 20, 535, 552

$\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ standard unit (basis) vectors in $\mathbb{R}^n$, 22

$d(\mathbf{u}, \mathbf{v})$ distance between vectors, 23

$\text{proj}_{\mathbf{u}}(\mathbf{v})$ (orthogonal) projection of $\mathbf{v}$ onto $\mathbf{u}$, 27

$\mathbf{u} \times \mathbf{v}$ cross product of vectors, 48

$[A|\mathbf{b}]$ augmented matrix, 61

$\text{rank}(A)$ rank of a matrix, 72, 204

$\text{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ span of a set of vectors, 99, 438

$$A = [a_{aj}] = [a_{ij}]_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$m \times n$ matrix, 138–139

$I = I_n$ identity matrix, 139

$O = O_{m \times n}$ zero matrix, 141

A^k kth power of a (square) matrix, 149

A^T transpose of a matrix, 151

A^{-1} inverse of a matrix, 163

$\det A = |A|$ determinant of a matrix, 165, 264–265

$\text{row}(A)$ row space of a matrix, 195

$\text{col}(A)$ column space of a matrix, 195

$\text{null}(A)$ null space of a matrix, 197

$\mathcal{B}$ basis, 198, 446

$\dim V$ dimension of a vector space, 203, 454

$\text{nullity}(A)$ nullity of a matrix, 204

$[\mathbf{v}]_{\mathcal{B}}$ coordinate vector of $\mathbf{v}$ with respect to the basis $\mathcal{B}$, 208, 449

T linear transformation, 213, 472

T_A matrix transformation, 214

$S \circ T$ composition of linear transformations, 219, 477

$[T]$ standard matrix of a linear transformation, 216

λ eigenvalue, 254

E_λ eigenspace, 256

C_{ij} (i,j)-cofactor of a matrix, 266

$A_i(\mathbf{b})$ matrix A with column i replaced by $\mathbf{b}$, 274

$\text{adj}\, A$ adjoint of a matrix, 276

$C(p)$ companion matrix of a polynomial, 299

$c_A(\lambda)$ characteristic polynomial of matrix A, 292

$A \sim B$ is similar to B, 301

$A = PDP^{-1}$ eigenvalue decomposition (diagonalizable matrix A), 309

D_i ith Gerschgorin disk, 319

f_n nth Fibonacci number, 335, 427

e^A matrix exponential, 346

$W^\perp$ orthogonal complement of subspace W, 378

$\text{proj}_W(\mathbf{v})$ orthogonal projection of $\mathbf{v}$ onto subspace W, 382

$\text{perp}_W(\mathbf{v})$ component of $\mathbf{v}$ orthogonal to W, 382

$A = QDQ^T$ spectral decomposition (symmetric matrix A), 400, 405

$\mathbf{x}^T A\mathbf{x}$ quadratic form, 409

M_{mn} vector space of $m \times n$ matrices, 430

$\mathcal{P}_n$ vector space of polynomials of degree $\leq n$, 431

$\mathcal{P}$ vector space of all polynomials, 431

$\mathcal{F}$ vector space of functions $f: \mathbb{R} \to \mathbb{R}$, 431

$\mathcal{F}[a, b]$ vector space of functions $f: [a, b] \to \mathbb{R}$, 432

$\mathbb{C}^n$ vector space of ordered n-tuples of complex numbers, 432, 540

$\mathcal{C}$ vector space of continuous functions $f: \mathbb{R} \to \mathbb{R}$, 435

$\mathcal{C}[a,b]$ vector space of continuous functions $f: [a, b] \to \mathbb{R}$, 435

$\mathcal{D}$ vector space of differentiable functions $f: \mathbb{R} \to \mathbb{R}$, 435

$\mathcal{D}[a, b]$ vector space of differentiable functions $f: [a, b] \to \mathbb{R}$, 435

$P_{\mathcal{C} \leftarrow \mathcal{B}}$ change-of-basis matrix, 465

$\ker(T)$ kernel of a linear transformation, 482

$\text{range}(T)$ range of a linear transformation, 482

$V \cong W$ V is isomorphic to W, 493

$[T]_{\mathcal{C} \leftarrow \mathcal{B}}$ matrix of T with respect to bases $\mathcal{B}$ and $\mathcal{C}$, 498

$\langle \mathbf{u}, \mathbf{v} \rangle$ inner product of vectors, 531

$\overline{A}$ complex conjugate of a matrix, 544

A^* conjugate transpose of a matrix, 544

$\|\mathbf{v}\|_s$ $(= \|\mathbf{v}\|_1)$ sum norm (= 1-norm), 552

$\|\mathbf{v}\|_m$ $(= \|\mathbf{v}\|_\infty)$ max norm (= ∞-norm), 553

$\|\mathbf{v}\|_E$ $(= \|\mathbf{v}\|_2)$ Euclidean norm (= 2-norm), 553

$\|\mathbf{v}\|_H$ Hamming norm, 554

$\|A\|$ matrix norm, 556

$\|A\|_F$ Frobenius norm, 556

$\text{cond}(A)$ condition number of a matrix, 562

$\bar{\mathbf{x}}$ least squares solution of $A\mathbf{x} = \mathbf{b}$, 574

A^+ pseudoinverse of a matrix, 585, 602

σ singular value, 590

$A = U\Sigma V^T$ singular value decomposition (SVD), 593