

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

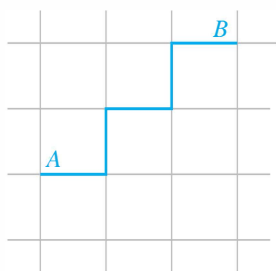


Figure 7.1

Taxicab distance

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 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, \dots, 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

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

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

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

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 $\langle \mathbf{u}, \mathbf{u} \rangle = \mathbf{u}^T A \mathbf{u} = 0$ 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

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

The matrix A is positive definite, by Theorem 5.24, 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 $\mathcal{C}[a, b]$, 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 $\mathcal{C}[a, b]$.

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 $\mathcal{C}[a, b]$, 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 $\mathcal{C}[a, b]$.



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

- $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$
- $\langle \mathbf{u}, c\mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$
- $\langle \mathbf{u}, \mathbf{0} \rangle = \langle \mathbf{0}, \mathbf{v} \rangle = 0$

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 .

- The **length** (or **norm**) of \mathbf{v} is $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$.
- The **distance** between \mathbf{u} and \mathbf{v} is $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$.
- \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 $\mathcal{C}[0, 1]$ given in Example 7.5. If $f(x) = x$ and $g(x) = 3x - 2$, find

- $\|f\|$
- $d(f, g)$
- $\langle f, g \rangle$

Solution (a) We find that

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

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

(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

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

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.

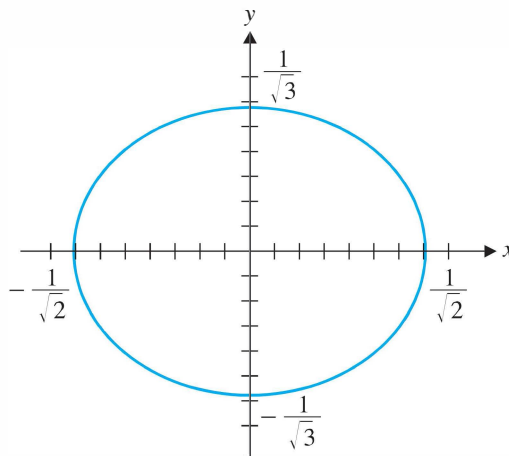


Figure 7.2

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 $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ 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.)



$\frac{dy}{dx}$

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

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



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

$$\begin{aligned} \langle \mathbf{v}_1, \mathbf{x}_3 \rangle &= \int_{-1}^1 x^2 dx = \left. \frac{x^3}{3} \right|_{-1}^1 = \frac{2}{3}, & \langle \mathbf{v}_2, \mathbf{x}_3 \rangle &= \int_{-1}^1 x^3 dx = \left. \frac{x^4}{4} \right|_{-1}^1 = 0, \\ \langle \mathbf{v}_2, \mathbf{v}_2 \rangle &= \int_{-1}^1 x^2 dx = \frac{2}{3} \end{aligned}$$

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

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

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 $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ 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

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

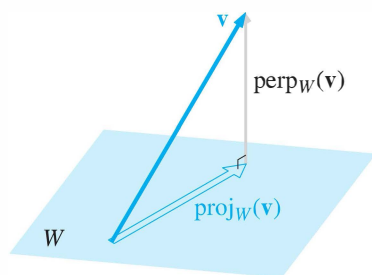


Figure 7.3

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

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

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

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

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,

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

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

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

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

$$A = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}. \text{ Compute}$$

(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. $\langle \mathbf{u}, \mathbf{v} \rangle = 4u_1 v_1 + u_1 v_2 + u_2 v_1 + 4u_2 v_2$

20. $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 + 2u_1 v_2 + 2u_2 v_1 + 5u_2 v_2$

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. $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 + \frac{1}{4} u_2 v_2$

22. $\langle \mathbf{u}, \mathbf{v} \rangle = 4u_1 v_1 + u_1 v_2 + u_2 v_1 + 4u_2 v_2$

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

$$\begin{aligned} \langle \mathbf{u}, \mathbf{v} \rangle &= 1, & \langle \mathbf{u}, \mathbf{w} \rangle &= 5, & \langle \mathbf{v}, \mathbf{w} \rangle &= 0 \\ \|\mathbf{u}\| &= 1, & \|\mathbf{v}\| &= \sqrt{3}, & \|\mathbf{w}\| &= 2 \end{aligned}$$

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

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

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

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

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

- (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) $\mathbf{u} \cdot \mathbf{u} \geq 0$ 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 $U^*AU = D$.

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

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

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

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

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



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.

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

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

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

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

$$\sqrt{xy} \leq \frac{x + y}{2} \quad (1)$$

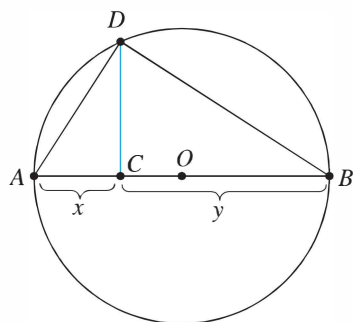


Figure 7.4

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 CD is perpendicular to 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, \dots, 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

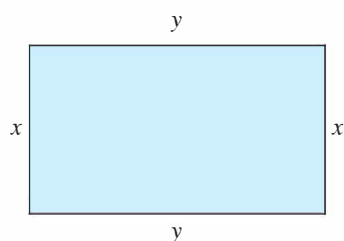


Figure 7.5

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

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:

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

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

$$V = xyz$$

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,

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

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

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

which is equivalent to $S \geq 24$. 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 $f(x) = x + \frac{1}{x}$ for $x > 0$?

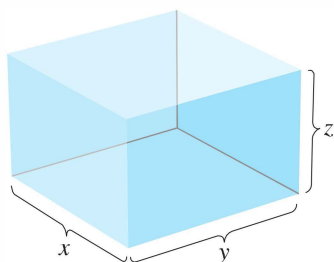


Figure 7.6



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 $x + \frac{8}{y(x - y)}$. [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 multi-variable 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 $f(x, y, z) = x^2 + y^2 + \frac{z^2}{2}$ 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, \dots, x_n is defined as

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

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

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

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.)

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.

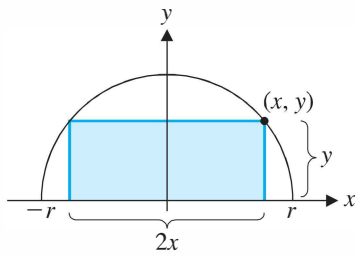


Figure 7.7

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 $c\mathbf{v} = [cv_1 \cdots cv_n]^T$, 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 $\mathbf{u} = [u_1 \ \cdots \ u_n]^T$, 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 $\mathbf{v} = [v_1 \ \cdots \ v_n]^T$, then

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

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 $|v_1|, \dots, |v_n|$ 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 ,

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

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

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

(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

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

then $\mathbf{u} + \mathbf{v} = [1 \ 0 \ 1 \ 0 \ 1]^T$, 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

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

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,

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

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

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

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

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

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 :

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

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_n 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_n 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

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



where the inequality arises from the Cauchy-Schwarz Inequality applied to the dot products of the row vectors 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 i th column of B . Using the matrix-column representation of the product AB , we have

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

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

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

The expression $\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$ measures the “stretching capability” of A . If we normalize each

nonzero vector \mathbf{x} by dividing it by its norm, we get unit vectors $\hat{\mathbf{x}} = \frac{1}{\|\mathbf{x}\|} \mathbf{x}$ and thus

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

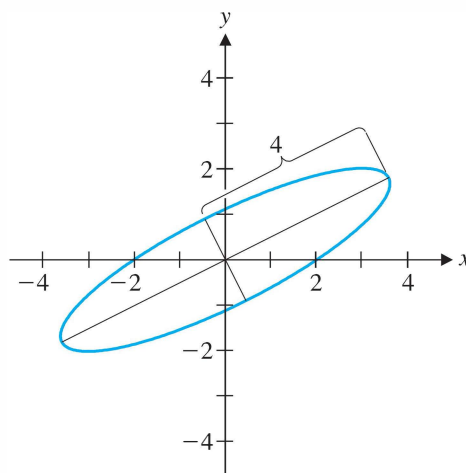


Figure 7.8

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_m 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 i th 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_m 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,

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

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}_j and row vectors \mathbf{A}_i for $i = 1, \dots, n$.

$$\begin{aligned} \text{a. } \|A\|_1 &= \max_{j=1, \dots, n} \{\|\mathbf{a}_j\|_s\} = \max_{j=1, \dots, n} \left\{ \sum_{i=1}^n |a_{ij}| \right\} \\ \text{b. } \|A\|_\infty &= \max_{i=1, \dots, n} \{\|\mathbf{A}_i\|_s\} = \max_{i=1, \dots, n} \left\{ \sum_{j=1}^n |a_{ij}| \right\} \end{aligned}$$

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

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

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

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

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, $\|A\|_1 = \max_{\|\mathbf{x}\|_s=1} \|A\mathbf{x}\|_s = M = \max_{j=1, \dots, n} \{\|\mathbf{a}_j\|_s\}$, 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

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

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

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

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),

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

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

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

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

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

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

$$\|\mathbf{x}_{k+1} - \mathbf{x}\| = \|M(\mathbf{x}_k - \mathbf{x})\| \leq \|M\| \|\mathbf{x}_k - \mathbf{x}\| \quad (9)$$

If we can show that $\|M\| < 1$, then we will have $\|\mathbf{x}_{k+1} - \mathbf{x}\| < \|\mathbf{x}_k - \mathbf{x}\|$ for all $k \geq 0$, 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 k th 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 $\|\mathbf{x}_k - \mathbf{x}\| \rightarrow 0$, 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 $\mathbf{u} = [1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1]^T$ and $\mathbf{v} = [0 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1]^T$.


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

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 $A^n \rightarrow \mathbf{0}$ as $n \rightarrow \infty$.
- (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.

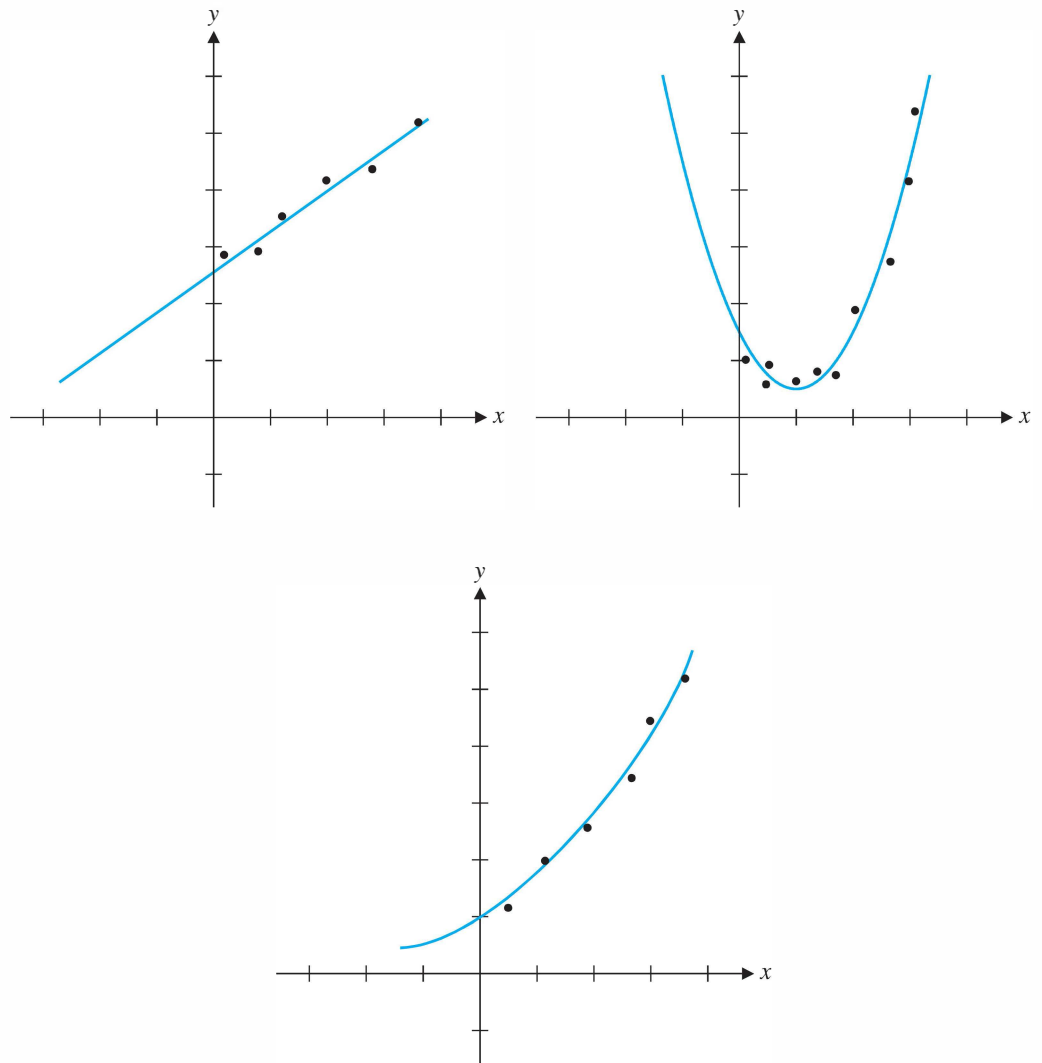


Figure 7.9
Curves of “best fit”

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

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

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.

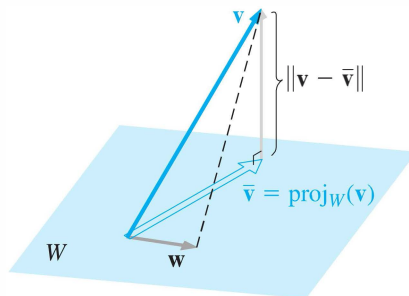


Figure 7.10

If $\bar{\mathbf{v}} = \text{proj}_W(\mathbf{v})$, then
 $\|\mathbf{v} - \bar{\mathbf{v}}\| \leq \|\mathbf{v} - \mathbf{w}\|$ 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,

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

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

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

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

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

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

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

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

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.

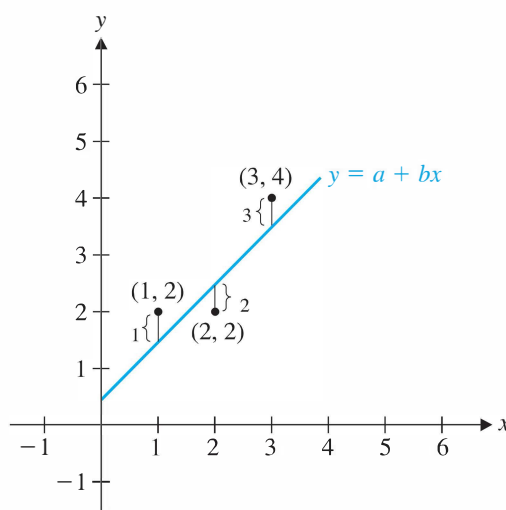


Figure 7.11

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	$4 - (1 + 3) = 0$	$4 - (-2 + 6) = 0$	$4 - (\frac{2}{3} + 3) = \frac{1}{3}$
$\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2$	$0^2 + (-1)^2 + 0^2 = 1$	$2^2 + 0^2 + 0^2 = 4$	$(\frac{1}{3})^2 + (-\frac{2}{3})^2 + (\frac{1}{3})^2 = \frac{2}{3}$
$\ e\ $	1	2	$\sqrt{\frac{2}{3}} \approx 0.816$

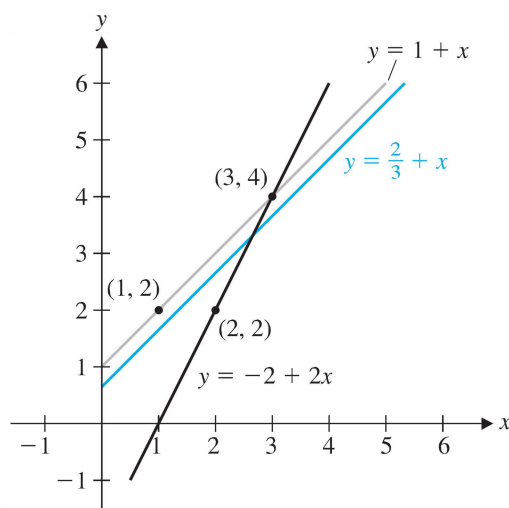


Figure 7.12

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 $(x_1, y_1), \dots, (x_n, y_n)$ 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 $\varepsilon_1^2 + \dots + \varepsilon_n^2$ is called the **least squares approximating line** (or the **line of best fit**) for the points $(x_1, y_1), \dots, (x_n, y_n)$. 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

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

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:

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

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

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

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

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

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

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

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

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 $(A^T A)^{-1}$ exists, and we could calculate $\bar{\mathbf{x}}$ as $\bar{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$. 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,

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

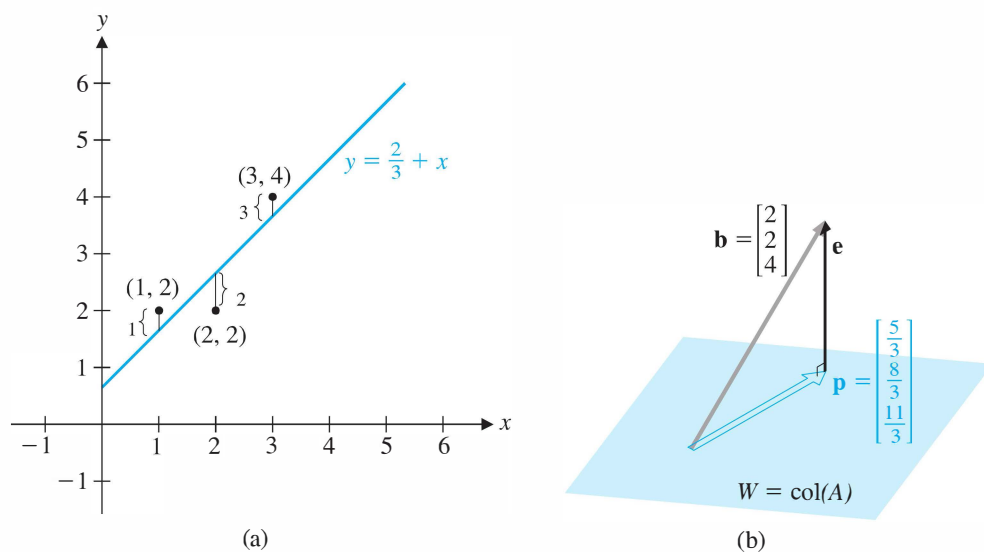


Figure 7.13

➡ 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

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

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.

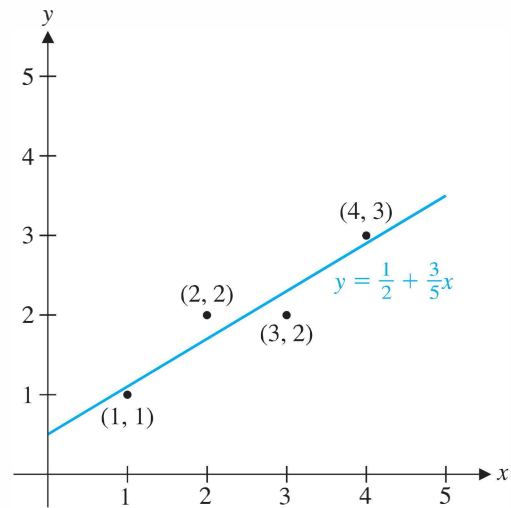


Figure 7.14

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

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

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

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

as shown in Figure 7.15.

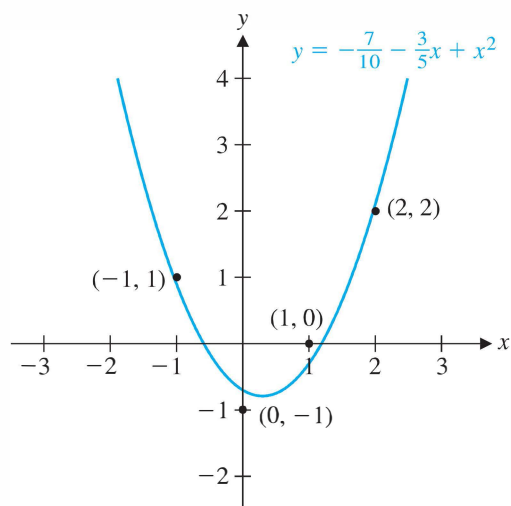


Figure 7.15

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**

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

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.

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

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

Therefore, $k = \bar{x} = 9.80/55 \approx 0.178$. 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?)

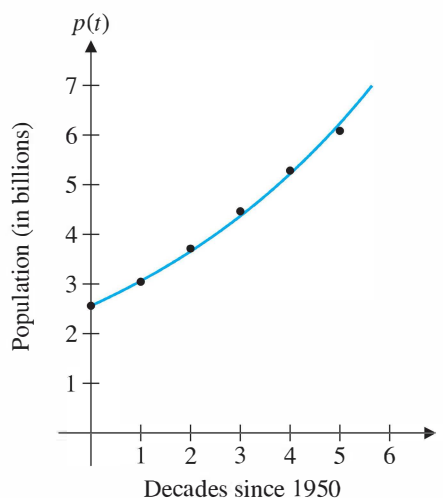


Figure 7.16

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

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



since $Q^T Q = I$. (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

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

The linear transformation $P: \mathbb{R}^m \rightarrow \mathbb{R}^m$ that projects \mathbb{R}^m onto W has $A(A^T A)^{-1} A^T$ 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

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

By Equation (1),

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

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

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

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

so $(A^T A)^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$

By Theorem 7.11, the standard matrix of the orthogonal projection transformation onto W is

$$A(A^T A)^{-1} A^T = A \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & \frac{1}{6} & -\frac{1}{3} \\ \frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

so the orthogonal projection of \mathbf{v} onto W is

$$\text{proj}_W(\mathbf{v}) = A(A^T A)^{-1} A^T \mathbf{v} = \begin{bmatrix} \frac{5}{6} & \frac{1}{6} & -\frac{1}{3} \\ \frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ \frac{1}{3} \\ -\frac{2}{3} \end{bmatrix}$$

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

$$A^+ = (A^T A)^{-1} A^T$$

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

$$A^+ = (A^T A)^{-1} A^T = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} \frac{4}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

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 :

- $AA^+A = A$
- $A^+AA^+ = A^+$
- 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.

- $y = -2 + 2x$
- $y = x$
- $y = -3 + \frac{5}{2}x$

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.

- $y = 3 - \frac{1}{3}x$
- $y = \frac{5}{2}$
- $y = 2 - \frac{1}{5}x$

In Exercises 7–14, find the least squares approximating line for the given points and compute the corresponding least squares error.

- $(1, 0)$, $(2, 1)$, $(3, 5)$
- $(1, 6)$, $(2, 3)$, $(3, 1)$
- $(0, 4)$, $(1, 1)$, $(2, 0)$
- $(0, 3)$, $(1, 3)$, $(2, 5)$
- $(-5, -1)$, $(0, 1)$, $(5, 2)$, $(10, 4)$

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.

$$\begin{array}{ll} 25. & x + y - z = 2 \\ & -y + 2z = 6 \\ & 3x + 2y - z = 11 \\ & -x + \quad z = 0 \end{array} \quad \begin{array}{ll} 26. & 2x + 3y + z = 21 \\ & x + y + z = 7 \\ & -x + y - z = 14 \\ & 2y + z = 0 \end{array}$$

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.

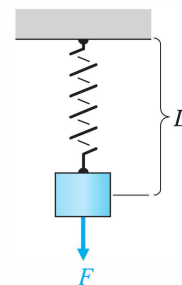


Figure 7.17

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

$$s(t) = s_0 + v_0 t + \frac{1}{2}gt^2$$

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 $P = A(A^T A)^{-1} A^T$ be the matrix of orthogonal projection onto $\text{col}(A)$.

- (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 $(x_1, y_1), \dots, (x_n, y_n)$ 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 $(x_1, y_1), \dots, (x_n, y_n)$ be given. Generalize Exercise 57 to show that if at least $k + 1$ of x_1, \dots, 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 $A^T A$ is symmetric and hence can be orthogonally diagonalized, by the Spectral Theorem. Not only are the eigenvalues of $A^T A$ all real (Theorem 5.18), they are all *nonnegative*. To show this, let λ be an eigenvalue of $A^T A$ 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 $A^T A$ and are denoted by $\sigma_1, \dots, \sigma_n$. It is conventional to arrange the singular values so that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$.

Example 7.33

Find the singular values of

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

Solution The matrix

$$A^T A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

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 $A^T A$. Since $A^T A$ is symmetric, we know that there is an *orthonormal* basis for \mathbb{R}^n that consists of eigenvectors of $A^T A$. Let $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be such a basis corresponding to the eigenvalues of $A^T A$, ordered so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. 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 $A\mathbf{v}_1, \dots, A\mathbf{v}_n$.

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

$$\begin{aligned} \|A\mathbf{x}\|^2 &= (A\mathbf{x}) \cdot (A\mathbf{x}) = (A\mathbf{x})^T(A\mathbf{x}) = \mathbf{x}^T A^T A \mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 + 2x_1x_2 + 2x_2^2 \end{aligned}$$

which we recognize is a quadratic form. By Theorem 5.25, 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 $A^T A$ —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

$$\|A\mathbf{v}_i\|^2 = \mathbf{v}_i^T A^T A \mathbf{v}_i = \lambda_i$$

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.

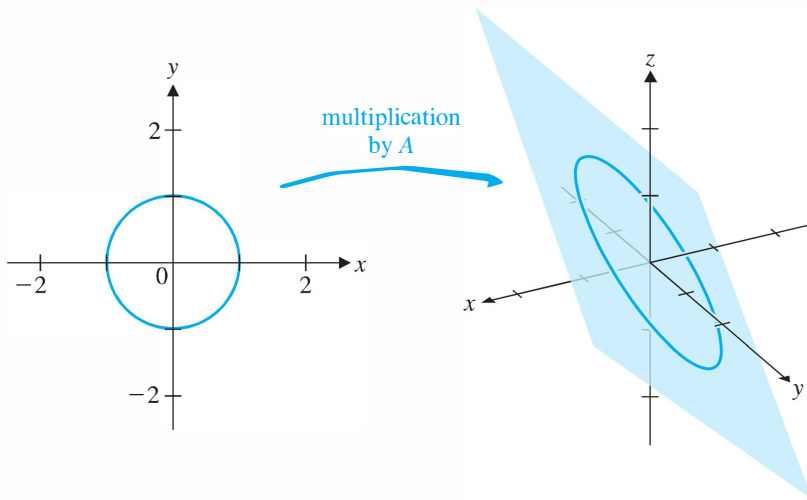


Figure 7.18

The matrix A transforms the unit circle in \mathbb{R}^2 into an ellipse in \mathbb{R}^3

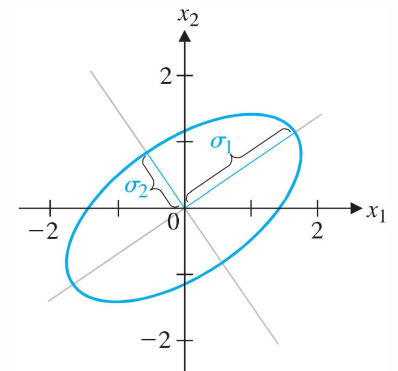


Figure 7.19

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 $Av_i = \sigma_i u_i$ for $i = 1, \dots, r$

and $\|Av_i\| = \sigma_i = 0$ for $i = r + 1, \dots, n$. Hence,

$$Av_i = \mathbf{0} \quad \text{for } i = r + 1, \dots, n$$

Therefore,

$$\begin{aligned} AV &= A[\mathbf{v}_1 \ \cdots \ \mathbf{v}_n] \\ &= [A\mathbf{v}_1 \ \cdots \ A\mathbf{v}_n] \\ &= [A\mathbf{v}_1 \ \cdots \ A\mathbf{v}_r \ \mathbf{0} \ \cdots \ \mathbf{0}] \\ &= [\sigma_1 \mathbf{u}_1 \ \cdots \ \sigma_r \mathbf{u}_r \ \mathbf{0} \ \cdots \ \mathbf{0}] \\ &= [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m] \begin{bmatrix} \sigma_1 & \cdots & 0 & \vdots & \vdots & \vdots \\ 0 & \cdots & \sigma_r & \vdots & \vdots & \vdots \\ \hline & & & 0 & \cdots & 0 \end{bmatrix} \\ &= U\Sigma \end{aligned}$$

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:

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

Solution (a) We compute

$$A^T A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

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

$$\mathbf{u}_1 = \frac{1}{\sigma_1} A \mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

and

$$\mathbf{u}_2 = \frac{1}{\sigma_2} A \mathbf{v}_2 = \frac{1}{1} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

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 $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{e}_3\}$ (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

$$\begin{aligned}
 A = U\Sigma V^T &= [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_m] \begin{bmatrix} \sigma_1 & \cdots & 0 & | & 0 \\ \vdots & \ddots & \vdots & | & \vdots \\ 0 & \cdots & \sigma_r & | & 0 \\ \hline & & O & & O \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\
 &= [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_r \quad \mathbf{u}_{r+1} \quad \cdots \quad \mathbf{u}_m] \begin{bmatrix} \sigma_1 & \cdots & 0 & | & 0 \\ \vdots & \ddots & \vdots & | & \vdots \\ 0 & \cdots & \sigma_r & | & 0 \\ \hline & & O & & O \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \\ \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\
 &= [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_r] \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} + [\mathbf{u}_{r+1} \quad \cdots \quad \mathbf{u}_m] [O] \begin{bmatrix} \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\
 &= [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_r] \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} \\
 &= [\sigma_1 \mathbf{u}_1 \quad \cdots \quad \sigma_r \mathbf{u}_r] \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} \\
 &= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T
 \end{aligned}$$

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 $\mathbf{u}_1, \dots, \mathbf{u}_r$ be left singular vectors and let $\mathbf{v}_1, \dots, \mathbf{v}_r$ 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 $\sigma_1, \dots, \sigma_r$ be all the nonzero singular values of A . Then:

- The rank of A is r .
- $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is an orthonormal basis for $\text{col}(A)$.
- $\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_m\}$ is an orthonormal basis for $\text{null}(A^T)$.
- $\{\mathbf{v}_1, \dots, \mathbf{v}_r\}$ is an orthonormal basis for $\text{row}(A)$.
- $\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_n\}$ is an orthonormal basis for $\text{null}(A)$.

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 $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ 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 $i = 1, \dots, r$, each \mathbf{u}_i is in the column space of A . (Why?) Furthermore,

$$r = \text{rank}(A) = \dim(\text{col}(A))$$

Therefore, $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is an orthonormal basis for $\text{col}(A)$, by Theorem 6.10(c).

(c) Since $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ is an orthonormal basis for \mathbb{R}^m and $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is a basis for $\text{col}(A)$, by property (b), it follows that $\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_m\}$ is an orthonormal basis for the orthogonal complement of $\text{col}(A)$. But $(\text{col}(A))^\perp = \text{null}(A^T)$, by Theorem 5.10.

(e) Since

$$A\mathbf{v}_{r+1} = \cdots = A\mathbf{v}_n = \mathbf{0}$$

the set $\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_n\}$ is an orthonormal set contained in the null space of A . Therefore, $\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_n\}$ is a linearly independent set of $n - r$ vectors in $\text{null}(A)$. But

$$\dim(\text{null}(A)) = n - r$$

by the Rank Theorem, so $\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_n\}$ is an orthonormal basis for $\text{null}(A)$, 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

- the surface of an ellipsoid in \mathbb{R}^m if $r = n$.
- 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 $\mathbf{u}_1, \dots, \mathbf{u}_m$ and $\mathbf{v}_1, \dots, \mathbf{v}_n$, respectively. Since $\text{rank}(A) = r$, the singular values of A satisfy

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0 \quad \text{and} \quad \sigma_{r+1} = \sigma_{r+2} = \dots = \sigma_n = 0$$

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

$$V^T\mathbf{x} = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{v}_1^T \mathbf{x} \\ \vdots \\ \mathbf{v}_n^T \mathbf{x} \end{bmatrix}$$

$$\text{so } (\mathbf{v}_1^T \mathbf{x})^2 + \dots + (\mathbf{v}_n^T \mathbf{x})^2 = 1.$$

By the outer product form of the SVD, we have $A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \dots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T$. Therefore,

$$\begin{aligned} A\mathbf{x} &= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T \mathbf{x} + \dots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T \mathbf{x} \\ &= (\sigma_1 \mathbf{v}_1^T \mathbf{x}) \mathbf{u}_1 + \dots + (\sigma_r \mathbf{v}_r^T \mathbf{x}) \mathbf{u}_r \\ &= y_1 \mathbf{u}_1 + \dots + y_r \mathbf{u}_r \end{aligned}$$

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

$$\begin{aligned} A\mathbf{x} &= y_1 \mathbf{u}_1 + \dots + y_n \mathbf{u}_n \\ &= U\mathbf{y} \end{aligned}$$

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

$$\left(\frac{y_1}{\sigma_1}\right)^2 + \dots + \left(\frac{y_n}{\sigma_n}\right)^2 = (\mathbf{v}_1^T \mathbf{x})^2 + \dots + (\mathbf{v}_n^T \mathbf{x})^2 = 1$$



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

$$\left(\frac{y_1}{\sigma_1}\right)^2 + \dots + \left(\frac{y_r}{\sigma_r}\right)^2 \leq 1$$

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 $y_1 y_2$ 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.

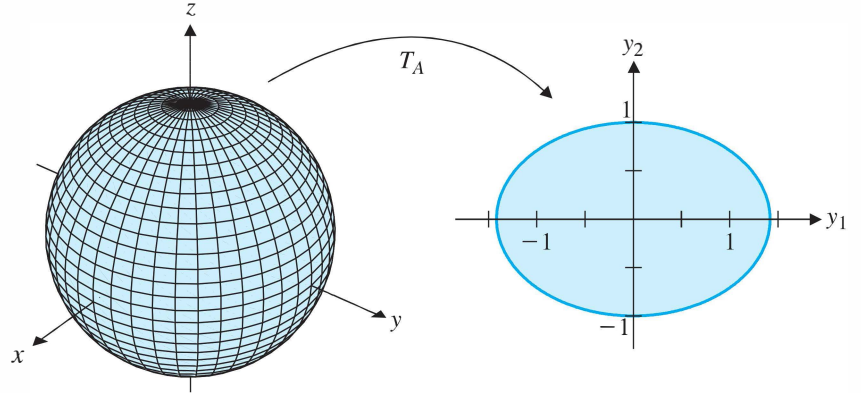


Figure 7.20

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 $\sigma_1, \dots, \sigma_r$ then distort into an ellipsoid. The orthogonal matrix U then aligns the axes of this ellipsoid with the orthonormal basis vectors $\mathbf{u}_1, \dots, \mathbf{u}_r$ 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.

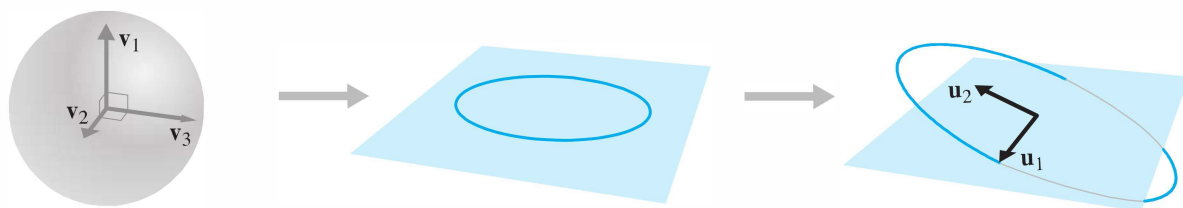


Figure 7.21

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 $\sigma_1, \dots, \sigma_r$ be all the nonzero singular values of A . Then

$$\|A\|_F = \sqrt{\sigma_1^2 + \dots + \sigma_r^2}$$

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

$$\|QA\|_F = \|A\|_F \quad (2)$$

To show that this is true, we compute

$$\begin{aligned} \|QA\|_F^2 &= \|[Q\mathbf{a}_1 \ \cdots \ Q\mathbf{a}_n]\|_F^2 \\ &= \|Q\mathbf{a}_1\|_E^2 + \cdots + \|Q\mathbf{a}_n\|_E^2 \\ &= \|\mathbf{a}_1\|_E^2 + \cdots + \|\mathbf{a}_n\|_E^2 \\ &= \|A\|_F^2 \end{aligned}$$

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 $1/\sigma_1, \dots, 1/\sigma_n$ (why?), and

$$1/\sigma_n \geq \dots \geq 1/\sigma_1$$

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 $A^+ = (A^T A)^{-1} A^T$ for the pseudoinverse of a matrix A . Clearly, this formula is valid only if $A^T A$ 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 $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ 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

$$A^+ = V\Sigma^+U^T = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 0 & 1 \end{bmatrix}$$

(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

$$\begin{aligned} A^+ &= V\Sigma^+U^T = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix} \\ &= \begin{bmatrix} 1/3 & 2/3 & -1/3 \\ 1/3 & -1/3 & 2/3 \end{bmatrix} \end{aligned}$$

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.

Jerry Bauer

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 $(A^T A)^{-1} A^T = V\Sigma^+ U^T$. (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 $A^T A\mathbf{x} = A^T \mathbf{b}$ have the unique solution

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

When the columns of A are linearly dependent, then $A^T A$ 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

$$\begin{aligned} \|\mathbf{b} - A\mathbf{x}\|^2 &= \|U^T(\mathbf{b} - A\mathbf{x})\|^2 = \|U^T(\mathbf{b} - U\Sigma V^T \mathbf{x})\|^2 = \|U^T \mathbf{b} - U^T U \Sigma V^T \mathbf{x}\|^2 \\ &= \|\mathbf{c} - \Sigma \mathbf{y}\|^2 = \left\| \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} - \begin{bmatrix} D & O \\ O & O \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} \mathbf{c}_1 - D\mathbf{y}_1 \\ \mathbf{c}_2 \end{bmatrix} \right\|^2 \end{aligned}$$

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 $T: V \rightarrow W$ be a linear transformation whose matrix $[T]_{C \leftarrow B}$ with respect to bases B and C of V and W , respectively, is A . The following statements are equivalent:

- A is invertible.
- $A\mathbf{x} = \mathbf{b}$ has a unique solution for every \mathbf{b} in \mathbb{R}^n .
- $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
- The reduced row echelon form of A is I_n .
- A is a product of elementary matrices.
- $\text{rank}(A) = n$
- $\text{nullity}(A) = 0$
- The column vectors of A are linearly independent.
- The column vectors of A span \mathbb{R}^n .
- The column vectors of A form a basis for \mathbb{R}^n .
- The row vectors of A are linearly independent.
- 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 $A^T A$.

(a) \Rightarrow (u) If A is invertible, so is A^T , and hence $A^T A$ is as well. Therefore, property (o) implies that 0 is not an eigenvalue of $A^T A$, 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 $A^T A$. Therefore, $A^T A$ 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 $\mathbf{u}_1, \dots, \mathbf{u}_{20}$ in \mathbb{R}^{340} and the 20 vectors $\mathbf{v}_1, \dots, \mathbf{v}_{20}$ in \mathbb{R}^{280} , for a total of

$$20 + 20 \cdot 340 + 20 \cdot 280 = 12,420$$

numbers. This represents a substantial saving!

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 7.22

Original, $k = r = 280$



$k = 2$



$k = 4$



$k = 8$



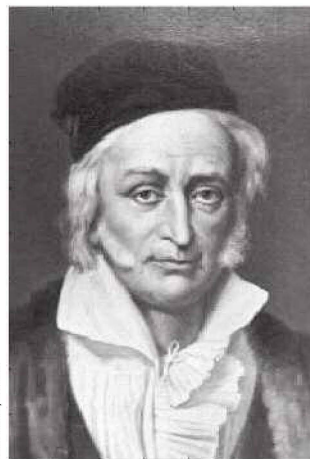
$k = 16$



$k = 32$



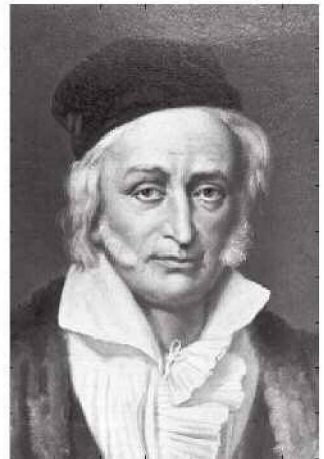
$k = 64$



$k = 128$



$k = 256$



Beitmann/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 516. A in Exercise 617. A in Exercise 718. A in Exercise 819. 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 338. 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 $AA' = (A')^T A^T$ 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 $A^T A$. Now use Exercise 34 in Section 7.2.]

 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

$$A = U\Sigma V^T = U\Sigma(U^T U)V^T = (U\Sigma U^T)(UV^T)$$

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



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 $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is an orthogonal basis for W , then

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

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

$$\begin{aligned} g(x) = \text{proj}_W(e^x) &= \frac{\langle 1, e^x \rangle}{\langle 1, 1 \rangle} 1 + \frac{\langle x, e^x \rangle}{\langle x, x \rangle} x \\ &= \frac{\int_{-1}^1 (1 \cdot e^x) dx}{\int_{-1}^1 (1 \cdot 1) dx} + \frac{\int_{-1}^1 x e^x dx}{\int_{-1}^1 x^2 dx} x \\ &= \frac{e - e^{-1}}{2} + \frac{2e^{-1}}{\frac{2}{3}} x \\ &= \frac{1}{2}(e - e^{-1}) + 3e^{-1}x \approx 1.18 + 1.10x \end{aligned}$$

➡ where we have used integration by parts to evaluate $\int_{-1}^1 xe^x dx$. (Check these calculations.) See Figure 7.24.

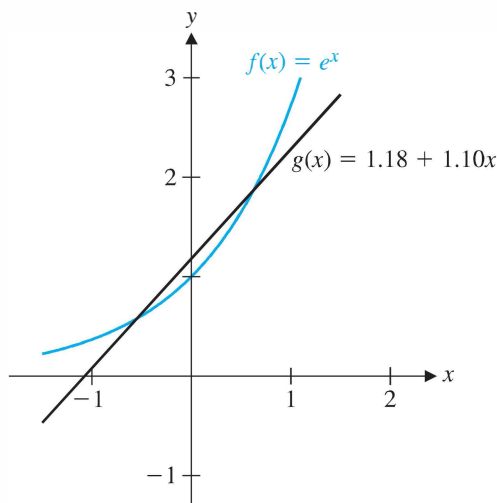


Figure 7.24

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 $\mathcal{C}[-1, 1]$. 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.)

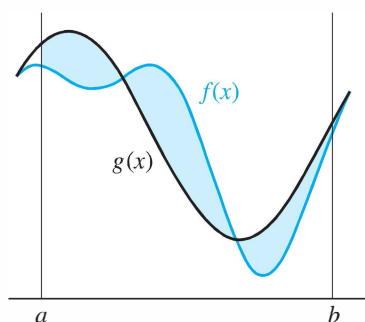


Figure 7.25

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 4.30

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

$$\left\{1, x, x^2 - \frac{1}{3}\right\}$$

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

$$\left\langle x^2 - \frac{1}{3}, e^x \right\rangle = \int_{-1}^1 \left(x^2 - \frac{1}{3}\right) e^x dx = \int_{-1}^1 x^2 e^x dx - \frac{1}{3} \int_{-1}^1 e^x dx = \frac{2}{3}(e - 7e^{-1})$$

$$\text{and } \left\langle x^2 - \frac{1}{3}, x^2 - \frac{1}{3} \right\rangle = \int_{-1}^1 \left(x^2 - \frac{1}{3}\right)^2 dx = \int_{-1}^1 \left(x^4 - \frac{2}{3}x^2 + \frac{1}{9}\right) dx = \frac{8}{45}$$

Then the best quadratic approximation to $f(x) = e^x$ on the interval $[-1, 1]$ is

$$\begin{aligned} g(x) = \text{proj}_W(e^x) &= \frac{\langle 1, e^x \rangle}{\langle 1, 1 \rangle} 1 + \frac{\langle x, e^x \rangle}{\langle x, x \rangle} x + \frac{\langle x^2 - \frac{1}{3}, e^x \rangle}{\langle x^2 - \frac{1}{3}, x^2 - \frac{1}{3} \rangle} \left(x^2 - \frac{1}{3}\right) \\ &= \frac{1}{2}(e - e^{-1}) + 3e^{-1}x + \frac{\frac{2}{3}(e - 7e^{-1})}{\frac{8}{45}} \left(x^2 - \frac{1}{3}\right) \\ &= \frac{3(11e^{-1} - e)}{4} + 3e^{-1}x + \frac{15(e - 7e^{-1})}{4} x^2 \approx 1.00 + 1.10x + 0.54x^2 \end{aligned}$$

(See Figure 7.26.)

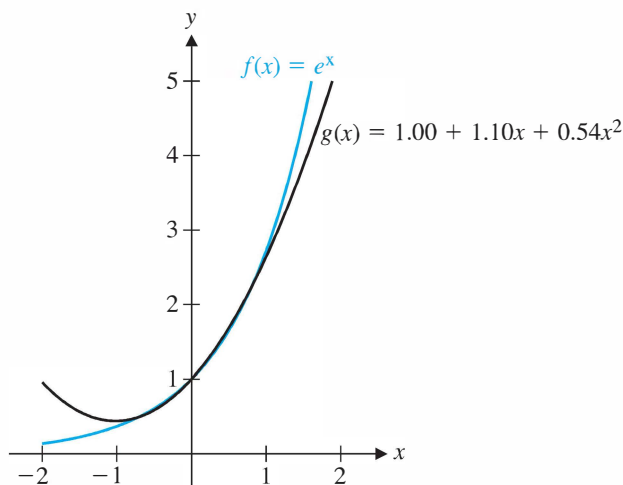


Figure 7.26

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

$$p(x) = a_0 + a_1 \cos x + a_2 \cos 2x + \cdots + a_n \cos nx + b_1 \sin x + b_2 \sin 2x + \cdots + b_n \sin nx \quad (1)$$

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

$$\mathcal{B} = \{1, \cos x, \dots, \cos nx, \sin x, \dots, \sin nx\}$$

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

$$\begin{aligned} \int_{-\pi}^{\pi} \sin jx \cos kx dx &= \frac{1}{2} \int_{-\pi}^{\pi} [\sin(j+k)x + \sin(j-k)x] dx \\ &= -\frac{1}{2} \left[\frac{\cos(j+k)x}{j+k} + \frac{\cos(j-k)x}{j-k} \right]_{-\pi}^{\pi} \\ &= 0 \end{aligned}$$

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

$$\begin{aligned} \langle \sin kx, \sin kx \rangle &= \int_{-\pi}^{\pi} \sin^2 kx \, dx \\ &= \frac{1}{2} \int_{-\pi}^{\pi} (1 - \cos 2kx) \, dx \\ &= \frac{1}{2} \left[x - \frac{\sin 2kx}{2k} \right]_{-\pi}^{\pi} \\ &= \pi \end{aligned}$$

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

$$\text{proj}_W(f) = a_0 + a_1 \cos x + \cdots + a_n \cos nx + b_1 \sin x + \cdots + b_n \sin nx \quad (2)$$

where

$$\begin{aligned} a_0 &= \frac{\langle 1, f \rangle}{\langle 1, 1 \rangle} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \, dx \\ a_k &= \frac{\langle \cos kx, f \rangle}{\langle \cos kx, \cos kx \rangle} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx \, dx \\ b_k &= \frac{\langle \sin kx, f \rangle}{\langle \sin kx, \sin kx \rangle} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx \, dx \end{aligned} \quad (3)$$

for $k \geq 1$. The approximation to f given by Equations (2) and (3) is called the ***n*th-order Fourier approximation** to f on $[-\pi, \pi]$. The coefficients $a_0, a_1, \dots, a_n, b_1, \dots, b_n$ 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.

$$\begin{aligned}
 \text{and} \quad b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} x \sin kx \, dx = \frac{1}{\pi} \left[-\frac{x}{k} \cos kx + \frac{1}{k^2} \sin kx \right]_{-\pi}^{\pi} \\
 &= \frac{1}{\pi} \left[\frac{-\pi \cos k\pi - \pi \cos(-k\pi)}{k} \right] \\
 &= \begin{cases} -\frac{2}{k} & \text{if } k \text{ is even} \\ \frac{2}{k} & \text{if } k \text{ is odd} \end{cases} \\
 &= \frac{2(-1)^{k+1}}{k}
 \end{aligned}$$

It follows that the fourth-order Fourier approximation to $f(x) = x$ on $[-\pi, \pi]$ is

$$2(\sin x - \frac{1}{2} \sin 2x + \frac{1}{3} \sin 3x - \frac{1}{4} \sin 4x)$$

Figure 7.27 shows the first four Fourier approximations to $f(x) = x$ on $[-\pi, \pi]$.

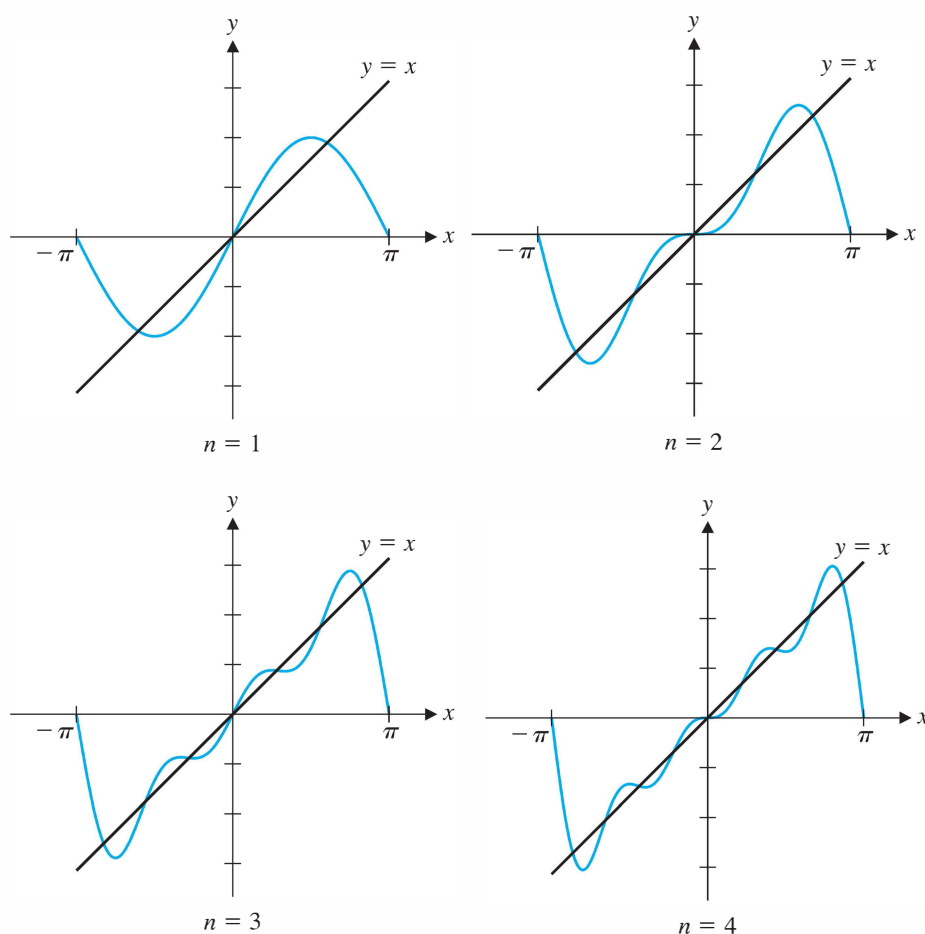


Figure 7.27



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 $[-\pi, \pi]$.

$$23. f(x) = \begin{cases} 0 & \text{if } -\pi \leq x < 0 \\ 1 & \text{if } 0 \leq x \leq \pi \end{cases}$$

$$24. f(x) = \begin{cases} -1 & \text{if } -\pi \leq x < 0 \\ 1 & \text{if } 0 \leq x \leq \pi \end{cases}$$

$$25. f(x) = \pi - x \qquad 26. f(x) = |x|$$

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 $\int_{-\pi}^{\pi} f(x) dx = 0$ 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 $\int_{-\pi}^{\pi} f(x) dx = 2 \int_0^{\pi} f(x) dx$ 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 | least squares error, 572 | orthonormal set of vectors, 537 |
| Cauchy-Schwarz Inequality, 539 | least squares solution, 574, 604 | pseudoinverse of a matrix, 585, 602 |
| condition number of a matrix, 562 | Least Squares Theorem, 575 | singular value decomposition (SVD), 593 |
| distance, 535 | matrix norm, 556 | singular values, 590 |
| Euclidean norm (2-norm), 553 | max norm (∞ -norm, uniform norm), 553 | singular vectors, 593 |
| Frobenius norm, 556 | norm, 535, 552 | sum norm (1-norm), 552 |
| Fundamental Theorem of Invertible Matrices, 605 | normed linear space, 552 | Triangle Inequality, 540 |
| Hamming distance, 554 | operator norm, 559 | unit sphere, 535 |
| Hamming norm, 554 | orthogonal basis, 537 | unit vector, 535 |
| ill-conditioned matrix, 561 | orthogonal projection, 538, 583 | well-conditioned matrix, 561 |
| inner product, 531 | orthogonal (set of) vectors, 537 | |
| inner product space, 531 | orthonormal basis, 537 | |

Review Questions

1. Mark each of the following statements true or false:

- If $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, then $\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 + \pi u_2 v_2$ defines an inner product on \mathbb{R}^2 .
- If $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, then $\langle \mathbf{u}, \mathbf{v} \rangle = 4u_1 v_1 - 2u_1 v_2 - 2u_2 v_1 + 4u_2 v_2$ defines an inner product on \mathbb{R}^2 .
- $\langle A, B \rangle = \text{tr}(A) + \text{tr}(B)$ defines an inner product on M_{22} .
- 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 $\|\mathbf{u} + \mathbf{v}\| = 5$.
- The sum norm, max norm, and Euclidean norm on \mathbb{R}^n are all equal to the absolute value function when $n = 1$.
- If a matrix A is well-conditioned, then $\text{cond}(A)$ is small.
- If $\text{cond}(A)$ is small, then the matrix A is well-conditioned.
- Every linear system has a unique least squares solution.
- If A is a matrix with orthonormal columns, then the standard matrix of an orthogonal projection onto the column space of A is $P = AA^T$.
- 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.

- $\langle p(x), q(x) \rangle = p(0)q(1) + p(1)q(0)$ for $p(x), q(x)$ in \mathcal{P}_1
- $\langle A, B \rangle = \text{tr}(A^T B)$ for A, B in M_{22}
- $\langle f, g \rangle = (\max_{0 \leq x \leq 1} f(x))(\max_{0 \leq x \leq 1} g(x))$ for f, g in $\mathcal{C}[0, 1]$

In Questions 5 and 6, compute the indicated quantity using the specified inner product.

- $\|1 + x + x^2\|$ if $\langle a_0 + a_1 x + a_2 x^2, b_0 + b_1 x + b_2 x^2 \rangle = a_0 b_0 + a_1 b_1 + a_2 b_2$
- $d(x, x^2)$ 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.

- $\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$ if $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T A \mathbf{v}$, where $A = \begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix}$
- $\{1, x, x^2\}$ 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.

- $\|\mathbf{v}\| = \mathbf{v}^T \mathbf{v}$ for \mathbf{v} in \mathbb{R}^n
- $\|p(x)\| = |p(0)| + |p(1) - p(0)|$ 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 $\text{span}(\mathbf{u}, \mathbf{v})$. [Hint: Show that $P = A(A^T A)^{-1}A^T$ 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 $(A^+)^2 = O$.