Define

$$\langle x, y\rangle = \sum_{i=1}^{n} a_i \bar{b}_i.$$

(a) Prove that $\langle \cdot, \cdot \rangle$ is an inner product on V and that β is an orthonormal basis for V. Thus every real or complex vector space may be regarded as an inner product space.

(b) Prove that if $\mathsf{V} = \mathsf{R}^n$ or $\mathsf{V} = \mathsf{C}^n$ and β is the standard ordered basis, then the inner product defined above is the standard inner product.

23. Let $\mathsf{V} = \mathsf{F}^n$, and let $A \in \mathsf{M}_{n\times n}(F)$.

(a) Prove that $\langle x, Ay\rangle = \langle A^* x, y\rangle$ for all $x, y \in \mathsf{V}$.

(b) Suppose that for some $B \in \mathsf{M}_{n\times n}(F)$, we have $\langle x, Ay\rangle = \langle Bx, y\rangle$ for all $x, y \in \mathsf{V}$. Prove that $B = A^*$.

(c) Let α be the standard ordered basis for V. For any orthonormal basis β for V, let Q be the $n \times n$ matrix whose columns are the vectors in β. Prove that $Q^* = Q^{-1}$.

(d) Define linear operators T and U on V by $\mathsf{T}(x) = Ax$ and $\mathsf{U}(x) = A^* x$. Show that $[\mathsf{U}]_\beta = [\mathsf{T}]_\beta^*$ for any orthonormal basis β for V.

24. Let V be a complex inner product space with an inner product $\langle \cdot, \cdot \rangle$. Let $[\cdot, \cdot]$ be the real-valued function such that $[x, y]$ is the real part of the complex number $\langle x, y\rangle$ for all $x, y \in \mathsf{V}$. Prove that $[\cdot, \cdot]$ is an inner product for V, where V is regarded as a vector space over R. Prove, furthermore, that $[x, ix] = 0$ for all $x \in \mathsf{V}$.

25. Let V be a vector space over C, and suppose that $[\cdot, \cdot]$ is a real inner product on V, where V is regarded as a vector space over R, such that $[x, ix] = 0$ for all $x \in \mathsf{V}$. Let $\langle \cdot, \cdot \rangle$ be the complex-valued function defined by

$$\langle x, y\rangle = [x, y] + i[x, iy] \qquad \text{for } x, y \in \mathsf{V}.$$

Prove that $\langle \cdot, \cdot \rangle$ is a complex inner product on V.

The following definition is used in Exercises 26–30.

Definition. *Let V be a vector space over F, where F is either R or C. Regardless of whether V is or is not an inner product space, we may still define a norm $\|\cdot\|_\mathsf{V}$ as a real-valued function on V satisfying the following three conditions for all $x, y \in \mathsf{V}$ and $a \in F$:*

(1) $\|x\|_\mathsf{V} \geq 0$, *and* $\|x\|_\mathsf{V} = 0$ *if and only if* $x = 0$.
(2) $\|ax\|_\mathsf{V} = |a| \cdot \|x\|_\mathsf{V}$.
(3) $\|x + y\|_\mathsf{V} \leq \|x\|_\mathsf{V} + \|y\|_\mathsf{V}$.

26. Prove that the following are norms on the given vector spaces V.

(a) $V = R^2$; $\|(a,b)\|_V = |a| + |b|$ for all $(a,b) \in V$

(b) $V = C([0,1])$; $\|f\|_V = \max_{t\in[0,1]} |f(t)|$ for all $f \in V$

(c) $V = C([0,1])$; $\|f\|_V = \int_0^1 |f(t)|\,dt$ for all $f \in V$

(d) $V = M_{m\times n}(F)$; $\|A\|_V = \max_{i,j} |A_{ij}|$ for all $A \in V$

27. Use Exercise 11 to show that there is no inner product $\langle \cdot, \cdot \rangle$ on R^2 such that $\|x\|_V^2 = \langle x, x \rangle$ for all $x \in R^2$ if the norm is defined as in Exercise 26(a).

28. Let $\|\cdot\|_V$ be a norm on a vector space V, and define, for each ordered pair of vectors, the scalar $d(x,y) = \|x - y\|_V$, called the **distance** between x and y. Prove the following results for all $x, y, z \in V$.

(a) $d(x,y) \geq 0$.
(b) $d(x,y) = d(y,x)$.
(c) $d(x,y) \leq d(x,z) + d(z,y)$.
(d) $d(x,x) = 0$ if and only if $x = 0$.
(e) $d(x,y) \neq 0$ if $x \neq y$.

29. Let $\|\cdot\|_V$ be a norm on a real vector space V satisfying the parallelogram law given in Exercise 11. Define

$$\langle x, y \rangle = \frac{1}{4}\left[\|x+y\|_V^2 - \|x-y\|_V^2\right].$$

Prove that $\langle \cdot, \cdot \rangle$ defines an inner product on V such that $\|x\|_V^2 = \langle x, x \rangle$ for all $x \in V$. *Hints:*

(a) Prove $\langle x, 2y \rangle = 2\langle x, y \rangle$ for all $x, y \in V$.
(b) Prove $\langle x + u, y \rangle = \langle x, y \rangle + \langle u, y \rangle$ for all $x, u, y \in V$.
(c) Prove $\langle nx, y \rangle = n\langle x, y \rangle$ for every positive integer n and every $x, y \in V$.
(d) Prove $m\left\langle \frac{1}{m}x, y \right\rangle = \langle x, y \rangle$ for every positive integer m and every $x, y \in V$.
(e) Prove $\langle rx, y \rangle = r\langle x, y \rangle$ for every rational number r and every $x, y \in V$.
(f) Prove $|\langle x, y \rangle| \leq \|x\|_V \|y\|_V$ for every $x, y \in V$. *Hint:* Condition (3) in the definition of norm can be helpful.
(g) Prove that for every $c \in R$, every rational number r, and every $x, y \in V$,

$$\begin{aligned} |c\langle x, y \rangle - \langle cx, y \rangle| &= |(c-r)\langle x, y \rangle - \langle (c-r)x, y \rangle| \\ &\leq 2|c-r|\|x\|_V \|y\|_V. \end{aligned}$$

(h) Use the fact that for any $c \in R$, $|c - r|$ can be made arbitrarily small, where r varies over the set of rational numbers, to establish item (b) of the definition of *inner product*.

30. Let $\|\cdot\|_{\mathsf{V}}$ be a norm (as defined on page 337) on a complex vector space V satisfying the parallelogram law given in Exercise 11. Prove that there is an inner product $\langle \cdot, \cdot \rangle$ on V such that $\|x\|_{\mathsf{V}}^2 = \langle x, x \rangle$ for all $x \in \mathsf{V}$. *Hint:* Apply Exercise 29 to V regarded as a vector space over R. Then apply Exercise 25.

6.2 THE GRAM–SCHMIDT ORTHOGONALIZATION PROCESS AND ORTHOGONAL COMPLEMENTS

In previous chapters, we have seen the special role of the standard ordered bases for C^n and R^n. The special properties of these bases stem from the fact that the basis vectors form an orthonormal set. Just as bases are the building blocks of vector spaces, bases that are also orthonormal sets are the building blocks of inner product spaces. We now name such bases.

Definition. *Let V be an inner product space. A subset of V is an* ***orthonormal basis*** *for V if it is an ordered basis that is orthonormal.*

Example 1

The standard ordered basis for F^n is an orthonormal basis for F^n. ◆

Example 2

The set

$$\left\{ \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}} \right), \left(\frac{2}{\sqrt{5}}, \frac{-1}{\sqrt{5}} \right) \right\}$$

is an orthonormal basis for R^2. ◆

The next theorem and its corollaries illustrate why orthonormal sets and, in particular, orthonormal bases are so important.

Theorem 6.3. *Let V be an inner product space and $S = \{v_1, v_2, \ldots, v_k\}$ be an orthogonal subset of V consisting of nonzero vectors. If $y \in \operatorname{span}(S)$, then*

$$y = \sum_{i=1}^{k} \frac{\langle y, v_i \rangle}{\|v_i\|^2} v_i.$$

Proof. Write $y = \sum_{i=1}^{k} a_i v_i$, where $a_1, a_2, \ldots, a_k \in F$. Then, for $1 \le j \le k$, we have

$$\langle y, v_j \rangle = \left\langle \sum_{i=1}^{k} a_i v_i, v_j \right\rangle = \sum_{i=1}^{k} a_i \langle v_i, v_j \rangle = a_j \langle v_j, v_j \rangle = a_j \|v_j\|^2.$$

So $a_j = \dfrac{\langle y, v_j \rangle}{\|v_j\|^2}$, and the result follows. ∎

The next corollary follows immediately from Theorem 6.3.

Corollary 1. *If, in addition to the hypotheses of Theorem 6.3, S is orthonormal and $y \in \text{span}(S)$, then*

$$y = \sum_{i=1}^{k} \langle y, v_i \rangle v_i.$$

If V possesses a finite orthonormal basis, then Corollary 1 allows us to compute the coefficients in a linear combination very easily. (See Example 3.)

Corollary 2. *Let V be an inner product space, and let S be an orthogonal subset of V consisting of nonzero vectors. Then S is linearly independent.*

Proof. Suppose that $v_1, v_2, \ldots, v_k \in S$ and

$$\sum_{i=1}^{k} a_i v_i = 0.$$

As in the proof of Theorem 6.3 with $y = 0$, we have $a_j = \langle 0, v_j \rangle / \|v_j\|^2 = 0$ for all j. So S is linearly independent. ∎

Example 3

By Corollary 2, the orthonormal set

$$\left\{ \frac{1}{\sqrt{2}}(1,1,0), \frac{1}{\sqrt{3}}(1,-1,1), \frac{1}{\sqrt{6}}(-1,1,2) \right\}$$

obtained in Example 8 of Section 6.1 is an orthonormal basis for R^3. Let $x = (2,1,3)$. The coefficients given by Corollary 1 to Theorem 6.3 that express x as a linear combination of the basis vectors are

$$a_1 = \frac{1}{\sqrt{2}}(2+1) = \frac{3}{\sqrt{2}}, \quad a_2 = \frac{1}{\sqrt{3}}(2-1+3) = \frac{4}{\sqrt{3}},$$

and

$$a_3 = \frac{1}{\sqrt{6}}(-2+1+6) = \frac{5}{\sqrt{6}}.$$

As a check, we have

$$(2,1,3) = \frac{3}{2}(1,1,0) + \frac{4}{3}(1,-1,1) + \frac{5}{6}(-1,1,2). \quad \blacklozenge$$

Corollary 2 tells us that the vector space H in Section 6.1 contains an infinite linearly independent set, and hence H is not a finite-dimensional vector space.

Of course, we have not yet shown that every finite-dimensional inner product space possesses an orthonormal basis. The next theorem takes us most of the way in obtaining this result. It tells us how to construct an orthogonal set from a linearly independent set of vectors in such a way that both sets generate the same subspace.

Before stating this theorem, let us consider a simple case. Suppose that $\{w_1, w_2\}$ is a linearly independent subset of an inner product space (and hence a basis for some two-dimensional subspace). We want to construct an orthogonal set from $\{w_1, w_2\}$ that spans the same subspace. Figure 6.1 suggests that the set $\{v_1, v_2\}$, where $v_1 = w_1$ and $v_2 = w_2 - cw_1$, has this property if c is chosen so that v_2 is orthogonal to w_1.

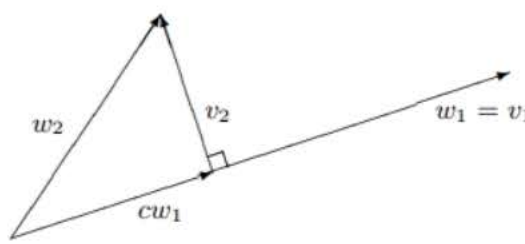


Figure 6.1

To find c, we need only solve the following equation:

$$0 = \langle v_2, w_1 \rangle = \langle w_2 - cw_1, w_1 \rangle = \langle w_2, w_1 \rangle - c \langle w_1, w_1 \rangle .$$

So

$$c = \frac{\langle w_2, w_1 \rangle}{\|w_1\|^2}.$$

Thus

$$v_2 = w_2 - \frac{\langle w_2, w_1 \rangle}{\|w_1\|^2} w_1.$$

The next theorem shows us that this process can be extended to any finite linearly independent subset.

Theorem 6.4. *Let* V *be an inner product space and* $S = \{w_1, w_2, \ldots, w_n\}$ *be a linearly independent subset of* V. *Define* $S' = \{v_1, v_2, \ldots, v_n\}$, *where* $v_1 = w_1$ *and*

$$v_k = w_k - \sum_{j=1}^{k-1} \frac{\langle w_k, v_j \rangle}{\|v_j\|^2} v_j \quad \text{for } 2 \le k \le n. \tag{1}$$

Then S' *is an orthogonal set of nonzero vectors such that* $\operatorname{span}(S') = \operatorname{span}(S)$.

Proof. The proof is by mathematical induction on n, the number of vectors in S. For $k = 1, 2, \ldots, n$, let $S_k = \{w_1, w_2, \ldots, w_k\}$. If $n = 1$, then the theorem is proved by taking $S_1' = S_1$; i.e., $v_1 = w_1 \neq 0$. Assume then that the set $S_{k-1}' = \{v_1, v_2, \ldots, v_{k-1}\}$ with the desired properties has been constructed by the repeated use of (1). We show that the set $S_k' = \{v_1, v_2, \ldots, v_{k-1}, v_k\}$ also has the desired properties, where v_k is obtained from S_{k-1}' by (1). If $v_k = 0$, then (1) implies that $w_k \in \text{span}(S_{k-1}') = \text{span}(S_{k-1})$, which contradicts the assumption that S_k is linearly independent. For $1 \leq i \leq k-1$, it follows from (1) that

$$\langle v_k, v_i \rangle = \langle w_k, v_i \rangle - \sum_{j=1}^{k-1} \frac{\langle w_k, v_j \rangle}{\|v_j\|^2} \langle v_j, v_i \rangle = \langle w_k, v_i \rangle - \frac{\langle w_k, v_i \rangle}{\|v_i\|^2} \|v_i\|^2 = 0,$$

since $\langle v_j, v_i \rangle = 0$ if $i \neq j$ by the induction assumption that S_{k-1}' is orthogonal. Hence S_k' is an orthogonal set of nonzero vectors. Now, by (1), we have that $\text{span}(S_k') \subseteq \text{span}(S_k)$. But by Corollary 2 to Theorem 6.3, S_k' is linearly independent; so $\dim(\text{span}(S_k')) = \dim(\text{span}(S_k)) = k$. Therefore $\text{span}(S_k') = \text{span}(S_k)$. ∎

The construction of $\{v_1, v_2, \ldots, v_n\}$ by the use of Theorem 6.4 is called the **Gram–Schmidt process.**

Example 4

In R^4, let $w_1 = (1, 0, 1, 0)$, $w_2 = (1, 1, 1, 1)$, and $w_3 = (0, 1, 2, 1)$. Then $\{w_1, w_2, w_3\}$ is linearly independent. We use the Gram–Schmidt process to compute the orthogonal vectors v_1, v_2, and v_3, and then we normalize these vectors to obtain an orthonormal set.

Take $v_1 = w_1 = (1, 0, 1, 0)$. Then

$$\begin{aligned} v_2 &= w_2 - \frac{\langle w_2, v_1 \rangle}{\|v_1\|^2} v_1 \\ &= (1, 1, 1, 1) - \frac{2}{2}(1, 0, 1, 0) \\ &= (0, 1, 0, 1). \end{aligned}$$

Finally,

$$\begin{aligned} v_3 &= w_3 - \frac{\langle w_3, v_1 \rangle}{\|v_1\|^2} v_1 - \frac{\langle w_3, v_2 \rangle}{\|v_2\|^2} v_2 \\ &= (0, 1, 2, 1) - \frac{2}{2}(1, 0, 1, 0) - \frac{2}{2}(0, 1, 0, 1) \\ &= (-1, 0, 1, 0). \end{aligned}$$

These vectors can be normalized to obtain the orthonormal basis $\{u_1, u_2, u_3\}$, where

$$u_1 = \frac{1}{\|v_1\|} v_1 = \frac{1}{\sqrt{2}}(1, 0, 1, 0),$$

$$u_2 = \frac{1}{\|v_2\|} v_2 = \frac{1}{\sqrt{2}}(0, 1, 0, 1),$$

and

$$u_3 = \frac{v_3}{\|v_3\|} = \frac{1}{\sqrt{2}}(-1, 0, 1, 0). \quad \blacklozenge$$

Example 5

Let $\mathsf{V} = \mathsf{P}(R)$ with the inner product $\langle f(x), g(x) \rangle = \int_{-1}^{1} f(t)g(t)\,dt$, and consider the subspace $\mathsf{P}_2(R)$ with the standard ordered basis β. We use the Gram–Schmidt process to replace β by an orthogonal basis $\{v_1, v_2, v_3\}$ for $\mathsf{P}_2(R)$, and then use this orthogonal basis to obtain an orthonormal basis for $\mathsf{P}_2(R)$.

Take $v_1 = 1$. Then $\|v_1\|^2 = \displaystyle\int_{-1}^{1} 1^2\,dt = 2$, and $\langle x, v_1 \rangle = \displaystyle\int_{-1}^{1} t \cdot 1\,dt = 0$.

Thus

$$v_2 = x - \frac{\langle v_1, x \rangle}{\|v_1\|^2} = x - \frac{0}{2} = x.$$

Furthermore,

$$\langle x^2, v_1 \rangle = \int_{-1}^{1} t^2 \cdot 1\,dt = \frac{2}{3} \quad \text{and} \quad \langle x^2, v_2 \rangle = \int_{-1}^{1} t^2 \cdot t\,dt = 0.$$

Therefore

$$v_3 = x^2 - \frac{\langle x^2, v_1 \rangle}{\|v_1\|^2} v_1 - \frac{\langle x^2, v_2 \rangle}{\|v_2\|^2} v_2$$

$$= x^2 - \frac{1}{3} \cdot 1 - 0 \cdot x$$

$$= x^2 - \frac{1}{3}.$$

We conclude that $\{1, x, x^2 - \frac{1}{3}\}$ is an orthogonal basis for $\mathsf{P}_2(R)$.

To obtain an orthonormal basis, we normalize v_1, v_2, and v_3 to obtain

$$u_1 = \frac{1}{\sqrt{\int_{-1}^{1} 1^2\,dt}} = \frac{1}{\sqrt{2}},$$

$$u_2 = \frac{x}{\sqrt{\int_{-1}^{1} t^2\,dt}} = \sqrt{\frac{3}{2}}\,x,$$

and similarly,

$$u_3 = \frac{v_3}{\|v_3\|} = \sqrt{\frac{5}{8}}\,(3x^2 - 1).$$

Thus $\{u_1, u_2, u_3\}$ is the desired orthonormal basis for $\mathsf{P}_2(R)$. ◆

Continuing to apply the Gram–Schmidt orthogonalization process to the basis $\{1, x, x^2, \ldots\}$ for $\mathsf{P}(R)$, we obtain an orthogonal basis $\{v_1, v_2, v_3 \ldots\}$. For each n, the polynomial $(1/v_k(1))v_k$ is called the kth *Legendre polynomial*. The first three Legendre polynomials are 1, x and $\frac{1}{2}(3x^2 - 1)$. The set of Legendre polynomials is also an orthogonal basis for $\mathsf{P}(R)$.

The following result gives us a simple method of representing a vector as a linear combination of the vectors in an orthonormal basis.

Theorem 6.5. *Let V be a nonzero finite-dimensional inner product space. Then V has an orthonormal basis β. Furthermore, if $\beta = \{v_1, v_2, \ldots, v_n\}$ and $x \in \mathsf{V}$, then*

$$x = \sum_{i=1}^{n} \langle x, v_i \rangle\, v_i.$$

Proof. Let β_0 be an ordered basis for V. Apply Theorem 6.4 to obtain an orthogonal set β' of nonzero vectors with $\operatorname{span}(\beta') = \operatorname{span}(\beta_0) = \mathsf{V}$. By normalizing each vector in β', we obtain an orthonormal set β that generates V. By Corollary 2 to Theorem 6.3, β is linearly independent; therefore β is an orthonormal basis for V. The remainder of the theorem follows from Corollary 1 to Theorem 6.3. ∎

Example 6

We use Theorem 6.5 to represent the polynomial $f(x) = 1 + 2x + 3x^2$ as a linear combination of the vectors in the orthonormal basis $\{u_1, u_2, u_3\}$ for $\mathsf{P}_2(R)$ obtained in Example 5. Observe that

$$\langle f(x), u_1 \rangle = \int_{-1}^{1} \frac{1}{\sqrt{2}}(1 + 2t + 3t^2)\,dt = 2\sqrt{2},$$

$$\langle f(x), u_2 \rangle = \int_{-1}^{1} \sqrt{\frac{3}{2}}\,t(1 + 2t + 3t^2)\,dt = \frac{2\sqrt{6}}{3},$$

and

$$\langle f(x), u_3 \rangle = \int_{-1}^{1} \sqrt{\frac{5}{8}}(3t^2 - 1)(1 + 2t + 3t^2)\,dt = \frac{2\sqrt{10}}{5}.$$

Therefore $f(x) = 2\sqrt{2}\,u_1 + \frac{2\sqrt{6}}{3}\,u_2 + \frac{2\sqrt{10}}{5}\,u_3$. ◆

Theorem 6.5 gives us a simple method for computing the entries of the matrix representation of a linear operator with respect to an orthonormal basis.

Corollary. *Let V be a finite-dimensional inner product space with an orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$. Let T be a linear operator on V, and let $A = [\mathsf{T}]_\beta$. Then for any i and j, $A_{ij} = \langle \mathsf{T}(v_j), v_i \rangle$.*

Proof. From Theorem 6.5, we have

$$\mathsf{T}(v_j) = \sum_{i=1}^{n} \langle \mathsf{T}(v_j), v_i \rangle\, v_i.$$

Hence $A_{ij} = \langle \mathsf{T}(v_j), v_i \rangle$. ∎

The scalars $\langle x, v_i \rangle$ given in Theorem 6.5 have been studied extensively for special inner product spaces. Although the vectors $v_1, v_2, \ldots, v_n$ were chosen from an orthonormal basis, we introduce a terminology associated with orthonormal sets β in more general inner product spaces.

Definition. *Let β be an orthonormal subset (possibly infinite) of an inner product space V, and let $x \in \mathsf{V}$. We define the* ***Fourier coefficients*** *of x relative to β to be the scalars $\langle x, y \rangle$, where $y \in \beta$.*

In the first half of the 19th century, the French mathematician Jean Baptiste Fourier was associated with the study of the scalars

$$\int_0^{2\pi} f(t) \sin nt\, dt \quad \text{and} \quad \int_0^{2\pi} f(t) \cos nt\, dt,$$

or in the complex case,

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int}\, dt,$$

for a function f. In the context of Example 9 of Section 6.1, we see that $c_n = \langle f, f_n \rangle$, where $f_n(t) = e^{int}$; that is, c_n is the nth Fourier coefficient for a continuous function $f \in \mathsf{V}$ relative to S. The coefficients c_n are the "classical" Fourier coefficients of a function, and the literature concerning their behavior is extensive. We learn more about Fourier coefficients in the remainder of this chapter.

Example 7

Let $S = \{e^{int} : n \text{ is an integer}\}$. In Example 9 of Section 6.1, S was shown to be an orthonormal set in H. We compute the Fourier coefficients of $f(t) = t$ relative to S. Using integration by parts, we have, for $n \neq 0$,

$$\langle f, f_n \rangle = \frac{1}{2\pi} \int_0^{2\pi} t\overline{e^{int}}\, dt = \frac{1}{2\pi} \int_0^{2\pi} te^{-int}\, dt = \frac{-1}{in},$$

and, for $n = 0$,

$$\langle f, 1 \rangle = \frac{1}{2\pi} \int_0^{2\pi} t(1)\, dt = \pi.$$

As a result of these computations, and using Exercise 16 of this section, we obtain an upper bound for the sum of a special infinite series as follows:

$$\begin{aligned} \|f\|^2 &\geq \sum_{n=-k}^{-1} |\langle f, f_n \rangle|^2 + |\langle f, 1 \rangle|^2 + \sum_{n=1}^{k} |\langle f, f_n \rangle|^2 \\ &= \sum_{n=-k}^{-1} \frac{1}{n^2} + \pi^2 + \sum_{n=1}^{k} \frac{1}{n^2} \\ &= 2\sum_{n=1}^{k} \frac{1}{n^2} + \pi^2 \end{aligned}$$

for every k. Now, using the fact that $\|f\|^2 = \frac{4}{3}\pi^2$, we obtain

$$\frac{4}{3}\pi^2 \geq 2\sum_{n=1}^{k} \frac{1}{n^2} + \pi^2,$$

or

$$\frac{\pi^2}{6} \geq \sum_{n=1}^{k} \frac{1}{n^2}.$$

Because this inequality holds for all k, we may let $k \to \infty$ to obtain

$$\frac{\pi^2}{6} \geq \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

Additional results may be produced by replacing f by other functions. ♦

We are now ready to proceed with the concept of an *orthogonal complement*.

Definition. *Let S be a nonempty subset of an inner product space V. We define $S^{\perp}$ (read "S perp") to be the set of all vectors in V that are orthogonal to every vector in S; that is, $S^{\perp} = \{x \in \mathsf{V}: \langle x, y\rangle = 0 \text{ for all } y \in S\}$. The set $S^{\perp}$ is called the* **orthogonal complement** *of S.*

It is easily seen that $S^{\perp}$ is a subspace of V for any subset S of V.

Example 8

The reader should verify that $\{0\}^{\perp} = \mathsf{V}$ and $\mathsf{V}^{\perp} = \{0\}$ for any inner product space V. ◆

Example 9

If $\mathsf{V} = \mathsf{R}^3$ and $S = \{e_3\}$, then $S^{\perp}$ equals the xy-plane (see Exercise 5). ◆

Exercise 18 provides an interesting example of an orthogonal complement in an infinite-dimensional inner product space.

Consider the problem in R^3 of finding the distance from a point P to a plane W. (See Figure 6.2.) Problems of this type arise in many settings. If we let y be the vector determined by 0 and P, we may restate the problem as follows: Determine the vector u in W that is "closest" to y. The desired distance is clearly given by $\|y - u\|$. Notice from the figure that the vector $z = y - u$ is orthogonal to every vector in W, and so $z \in \mathsf{W}^{\perp}$.

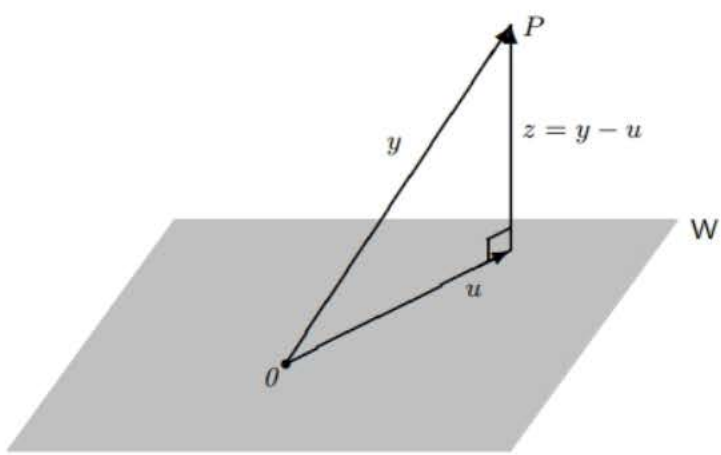


Figure 6.2

The next result presents a practical method of finding u in the case that W is a finite-dimensional subspace of an inner product space.

Theorem 6.6. *Let W be a finite-dimensional subspace of an inner product space V, and let $y \in \mathsf{V}$. Then there exist unique vectors $u \in \mathsf{W}$ and $z \in \mathsf{W}^{\perp}$*

such that $y = u + z$. *Furthermore, if* $\{v_1, v_2, \ldots, v_k\}$ *is an orthonormal basis for* W, *then*

$$u = \sum_{i=1}^{k} \langle y, v_i \rangle v_i.$$

Proof. Let $\{v_1, v_2, \ldots, v_k\}$ be an orthonormal basis for W, let u be as defined in the preceding equation, and let $z = y - u$. Clearly $u \in \mathsf{W}$ and $y = u + z$.

To show that $z \in \mathsf{W}^{\perp}$, it suffices to show, by Exercise 7, that z is orthogonal to each v_j. For any j, we have

$$\begin{aligned}\langle z, v_j \rangle &= \left\langle \left(y - \sum_{i=1}^{k} \langle y, v_i \rangle v_i \right), v_j \right\rangle = \langle y, v_j \rangle - \sum_{i=1}^{k} \langle y, v_i \rangle \langle v_i, v_j \rangle \\ &= \langle y, v_j \rangle - \langle y, v_j \rangle = 0.\end{aligned}$$

To show uniqueness of u and z, suppose that $y = u + z = u' + z'$, where $u' \in \mathsf{W}$ and $z' \in \mathsf{W}^{\perp}$. Then $u - u' = z' - z \in \mathsf{W} \cap \mathsf{W}^{\perp} = \{0\}$. Therefore $u = u'$ and $z = z'$. ∎

Corollary. *In the notation of Theorem 6.6, the vector* u *is the unique vector in* W *that is "closest" to* y; *that is, for any* $x \in \mathsf{W}$, $\|y - x\| \geq \|y - u\|$, *and this inequality is an equality if and only if* $x = u$.

Proof. As in Theorem 6.6, we have that $y = u + z$, where $z \in \mathsf{W}^{\perp}$. Let $x \in \mathsf{W}$. Then $u - x$ is orthogonal to z, so, by Exercise 10 of Section 6.1, we have

$$\begin{aligned}\|y - x\|^2 &= \|u + z - x\|^2 = \|(u - x) + z\|^2 = \|u - x\|^2 + \|z\|^2 \\ &\geq \|z\|^2 = \|y - u\|^2.\end{aligned}$$

Now suppose that $\|y - x\| = \|y - u\|$. Then the inequality above becomes an equality, and therefore $\|u - x\|^2 + \|z\|^2 = \|z\|^2$. It follows that $\|u - x\| = 0$, and hence $x = u$. The proof of the converse is obvious. ∎

The vector u in the corollary is called the **orthogonal projection** of y on W. We will see the importance of orthogonal projections of vectors in the application to least squares in Section 6.3.

Example 10

Let $\mathsf{V} = \mathsf{P}_3(R)$ with the inner product

$$\langle f(x), g(x) \rangle = \int_{-1}^{1} f(t)g(t)\,dt \quad \text{for all } f(x), g(x) \in \mathsf{V}.$$

We compute the orthogonal projection $f_1(x)$ of $f(x) = x^3$ on $\mathsf{P}_2(R)$.

By Example 5,

$$\{u_1, u_2, u_3\} = \left\{ \frac{1}{\sqrt{2}}, \sqrt{\frac{3}{2}}\, x, \sqrt{\frac{5}{8}}\,(3x^2 - 1) \right\}$$

is an orthonormal basis for $\mathsf{P}_2(R)$. For these vectors, we have

$$\langle f(x), u_1 \rangle = \int_{-1}^{1} t^3 \frac{1}{\sqrt{2}}\, dt = 0, \qquad \langle f(x), u_2 \rangle = \int_{-1}^{1} t^3 \sqrt{\frac{3}{2}}\, t\, dt = \frac{\sqrt{6}}{5},$$

and

$$\langle f(x), u_3 \rangle = \int_{-1}^{1} t^3 \sqrt{\frac{5}{8}}\,(3t^2 - 1)\, dt = 0.$$

Hence

$$f_1(x) = \langle f(x), u_1 \rangle\, u_1 + \langle f(x), u_2 \rangle\, u_2 + \langle f(x), u_3 \rangle\, u_3 = \frac{3}{5}x. \qquad \blacklozenge$$

It was shown (Corollary 2 to the replacement theorem, p. 48) that any linearly independent set in a finite-dimensional vector space can be extended to a basis. The next theorem provides an interesting analog for an orthonormal subset of a finite-dimensional inner product space.

Theorem 6.7. *Suppose that $S = \{v_1, v_2, \ldots, v_k\}$ is an orthonormal set in an n-dimensional inner product space V. Then*

(a) *S can be extended to an orthonormal basis $\{v_1, v_2, \ldots, v_k, v_{k+1}, \ldots, v_n\}$ for V.*

(b) *If $\mathsf{W} = \operatorname{span}(S)$, then $S_1 = \{v_{k+1}, v_{k+2}, \ldots, v_n\}$ is an orthonormal basis for $\mathsf{W}^{\perp}$ (using the preceding notation).*

(c) *If W is any subspace of V, then $\dim(\mathsf{V}) = \dim(\mathsf{W}) + \dim(\mathsf{W}^{\perp})$.*

Proof. (a) By Corollary 2 to the replacement theorem (p. 48), S can be extended to an ordered basis $S' = \{v_1, v_2, \ldots, v_k, w_{k+1}, \ldots, w_n\}$ for V. Now apply the Gram–Schmidt process to S'. The first k vectors resulting from this process are the vectors in S by Exercise 8, and this new set spans V. Normalizing the last $n - k$ vectors of this set produces an orthonormal set that spans V. The result now follows.

(b) Because S_1 is a subset of a basis, it is linearly independent. Since S_1 is clearly a subset of $\mathsf{W}^{\perp}$, we need only show that it spans $\mathsf{W}^{\perp}$. Note that, for any $x \in \mathsf{V}$, we have

$$x = \sum_{i=1}^{n} \langle x, v_i \rangle\, v_i.$$

If $x \in \mathsf{W}^\perp$, then $\langle x, v_i \rangle = 0$ for $1 \le i \le k$. Therefore

$$x = \sum_{i=k+1}^{n} \langle x, v_i \rangle \, v_i \in \text{span}(S_1).$$

(c) Let W be a subspace of V. It is a finite-dimensional inner product space because V is, and so it has an orthonormal basis $\{v_1, v_2, \ldots, v_k\}$. By (a) and (b), we have

$$\dim(\mathsf{V}) = n = k + (n-k) = \dim(\mathsf{W}) + \dim(\mathsf{W}^\perp). \qquad \blacksquare$$

Example 11

Let $\mathsf{W} = \text{span}(\{e_1, e_2\})$ in F^3. Then $x = (a, b, c) \in \mathsf{W}^\perp$ if and only if $0 = \langle x, e_1 \rangle = a$ and $0 = \langle x, e_2 \rangle = b$. So $x = (0, 0, c)$, and therefore $\mathsf{W}^\perp = \text{span}(\{e_3\})$. One can deduce the same result by noting that $e_3 \in \mathsf{W}^\perp$ and, from (c), that $\dim(\mathsf{W}^\perp) = 3 - 2 = 1$. ◆

EXERCISES

1. Label the following statements as true or false.

- **(a)** The Gram–Schmidt orthogonalization process produces an orthonormal set from an arbitrary linearly independent set.
- **(b)** Every nonzero finite-dimensional inner product space has an orthonormal basis.
- **(c)** The orthogonal complement of any set is a subspace.
- **(d)** If $\{v_1, v_2, \ldots, v_n\}$ is a basis for an inner product space V, then for any $x \in \mathsf{V}$ the scalars $\langle x, v_i \rangle$ are the Fourier coefficients of x.
- **(e)** An orthonormal basis must be an ordered basis.
- **(f)** Every orthogonal set is linearly independent.
- **(g)** Every orthonormal set is linearly independent.

2. In each part, apply the Gram–Schmidt process to the given subset S of the inner product space V to obtain an orthogonal basis for $\text{span}(S)$. Then normalize the vectors in this basis to obtain an orthonormal basis β for $\text{span}(S)$, and compute the Fourier coefficients of the given vector relative to β. Finally, use Theorem 6.5 to verify your result.

- **(a)** $\mathsf{V} = \mathsf{R}^3$, $S = \{(1,0,1), (0,1,1), (1,3,3)\}$, and $x = (1,1,2)$
- **(b)** $\mathsf{V} = \mathsf{R}^3$, $S = \{(1,1,1), (0,1,1), (0,0,1)\}$, and $x = (1,0,1)$
- **(c)** $\mathsf{V} = \mathsf{P}_2(R)$ with the inner product $\langle f(x), g(x) \rangle = \int_0^1 f(t)g(t)\,dt$, $S = \{1, x, x^2\}$, and $h(x) = 1 + x$
- **(d)** $\mathsf{V} = \text{span}(S)$, where $S = \{(1, i, 0), (1-i, 2, 4i)\}$, and $x = (3+i, 4i, -4)$

(e) $\mathsf{V} = \mathsf{R}^4$, $S = \{(2,-1,-2,4),(-2,1,-5,5),(-1,3,7,11)\}$, and $x = (-11,8,-4,18)$

(f) $\mathsf{V} = \mathsf{R}^4$, $S = \{(1,-2,-1,3),(3,6,3,-1),(1,4,2,8)\}$, and $x = (-1,2,1,1)$

(g) $\mathsf{V} = \mathsf{M}_{2\times 2}(R)$, $S = \left\{\begin{pmatrix} 3 & 5 \\ -1 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 9 \\ 5 & -1 \end{pmatrix}, \begin{pmatrix} 7 & -17 \\ 2 & -6 \end{pmatrix}\right\}$, and $A = \begin{pmatrix} -1 & 27 \\ -4 & 8 \end{pmatrix}$

(h) $\mathsf{V} = \mathsf{M}_{2\times 2}(R)$, $S = \left\{\begin{pmatrix} 2 & 2 \\ 2 & 1 \end{pmatrix}, \begin{pmatrix} 11 & 4 \\ 2 & 5 \end{pmatrix}, \begin{pmatrix} 4 & -12 \\ 3 & -16 \end{pmatrix}\right\}$, and $A = \begin{pmatrix} 8 & 6 \\ 25 & -13 \end{pmatrix}$

(i) $\mathsf{V} = \text{span}(S)$ with the inner product $\langle f,g\rangle = \int_0^{\pi} f(t)g(t)\,dt$, $S = \{\sin t, \cos t, 1, t\}$, and $h(t) = 2t + 1$

(j) $\mathsf{V} = \mathsf{C}^4$, $S = \{(1,i,2-i,-1),(2+3i,3i,1-i,2i),$ $(-1+7i,6+10i,11-4i,3+4i)\}$, and $x = (-2+7i,6+9i,9-3i,4+4i)$

(k) $\mathsf{V} = \mathsf{C}^4$, $S = \{(-4,3-2i,i,1-4i),$ $(-1-5i,5-4i,-3+5i,7-2i),(-27-i,-7-6i,-15+25i,-7-6i)\}$, and $x = (-13-7i,-12+3i,-39-11i,-26+5i)$

(l) $\mathsf{V} = \mathsf{M}_{2\times 2}(C)$, $S = \left\{\begin{pmatrix} 1-i & -2-3i \\ 2+2i & 4+i \end{pmatrix}, \begin{pmatrix} 8i & 4 \\ -3-3i & -4+4i \end{pmatrix}, \begin{pmatrix} -25-38i & -2-13i \\ 12-78i & -7+24i \end{pmatrix}\right\}$, and $A = \begin{pmatrix} -2+8i & -13+i \\ 10-10i & 9-9i \end{pmatrix}$

(m) $\mathsf{V} = \mathsf{M}_{2\times 2}(C)$, $S = \left\{\begin{pmatrix} -1+i & -i \\ 2-i & 1+3i \end{pmatrix}, \begin{pmatrix} -1-7i & -9-8i \\ 1+10i & -6-2i \end{pmatrix}, \begin{pmatrix} -11-132i & -34-31i \\ 7-126i & -71-5i \end{pmatrix}\right\}$, and $A = \begin{pmatrix} -7+5i & 3+18i \\ 9-6i & -3+7i \end{pmatrix}$

3. In R^2, let

$$\beta = \left\{\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right),\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}}\right)\right\}.$$

Find the Fourier coefficients of $(3,4)$ relative to β.

4. Let $S = \{(1,0,i),(1,2,1)\}$ in C^3. Compute $S^{\perp}$.

5. Let $S_0 = \{x_0\}$, where x_0 is a nonzero vector in R^3. Describe $S_0^{\perp}$ geometrically. Now suppose that $S = \{x_1, x_2\}$ is a linearly independent subset of R^3. Describe $S^{\perp}$ geometrically.

6. Let V be an inner product space, and let W be a finite-dimensional subspace of V. If $x \notin \mathsf{W}$, prove that there exists $y \in \mathsf{V}$ such that $y \in \mathsf{W}^\perp$, but $\langle x, y\rangle \neq 0$. *Hint:* Use Theorem 6.6.

7. Let β be a basis for a subspace W of an inner product space V, and let $z \in \mathsf{V}$. Prove that $z \in \mathsf{W}^\perp$ if and only if $\langle z, v\rangle = 0$ for every $v \in \beta$.

8. Prove that if $\{w_1, w_2, \ldots, w_n\}$ is an orthogonal set of nonzero vectors, then the vectors $v_1, v_2, \ldots, v_n$ derived from the Gram–Schmidt process satisfy $v_i = w_i$ for $i = 1, 2, \ldots, n$. *Hint:* Use mathematical induction.

9. Let $\mathsf{W} = \text{span}(\{(i, 0, 1)\})$ in C^3. Find orthonormal bases for W and $\mathsf{W}^\perp$.

10. Let W be a finite-dimensional subspace of an inner product space V. Prove that $\mathsf{V} = \mathsf{W} \oplus \mathsf{W}^\perp$. Using the definition on page 76, prove that there exists a projection T on W along $\mathsf{W}^\perp$ that satisfies $\mathsf{N}(\mathsf{T}) = \mathsf{W}^\perp$. In addition, prove that $\|\mathsf{T}(x)\| \leq \|x\|$ for all $x \in \mathsf{V}$. *Hint:* Use Theorem 6.6 and Exercise 10 of Section 6.1.

11. Let A be an $n \times n$ matrix with complex entries. Prove that $AA^* = I$ if and only if the rows of A form an orthonormal basis for C^n. Visit `goo.gl/iKcC4S` for a solution.

12. Prove that for any matrix $A \in \mathsf{M}_{m\times n}(F)$, $(\mathsf{R}(\mathsf{L}_{A^*}))^\perp = \mathsf{N}(\mathsf{L}_A)$.

13. Let V be an inner product space, S and S_0 be subsets of V, and W be a finite-dimensional subspace of V. Prove the following results.

(a) $S_0 \subseteq S$ implies that $S^\perp \subseteq S_0^\perp$.
(b) $S \subseteq (S^\perp)^\perp$; so $\text{span}(S) \subseteq (S^\perp)^\perp$.
(c) $\mathsf{W} = (\mathsf{W}^\perp)^\perp$. *Hint:* Use Exercise 6.
(d) $\mathsf{V} = \mathsf{W} \oplus \mathsf{W}^\perp$. (See the exercises of Section 1.3.)

14. Let W_1 and W_2 be subspaces of a finite-dimensional inner product space. Prove that $(\mathsf{W}_1 + \mathsf{W}_2)^\perp = \mathsf{W}_1^\perp \cap \mathsf{W}_2^\perp$ and $(\mathsf{W}_1 \cap \mathsf{W}_2)^\perp = \mathsf{W}_1^\perp + \mathsf{W}_2^\perp$. (See the definition of the sum of subsets of a vector space on page 22.) *Hint for the second equation:* Apply Exercise 13(c) to the first equation.

15. Let V be a finite-dimensional inner product space over F.

(a) *Parseval's Identity.* Let $\{v_1, v_2, \ldots, v_n\}$ be an orthonormal basis for V. For any $x, y \in \mathsf{V}$ prove that

$$\langle x, y\rangle = \sum_{i=1}^{n} \langle x, v_i\rangle \, \overline{\langle y, v_i\rangle}.$$

(b) Use (a) to prove that if β is an orthonormal basis for V with inner product $\langle \cdot, \cdot \rangle$, then for any $x, y \in \mathsf{V}$

$$\langle \phi_\beta(x), \phi_\beta(y)\rangle' = \langle [x]_\beta, [y]_\beta\rangle' = \langle x, y\rangle,$$

where $\langle \cdot, \cdot \rangle'$ is the standard inner product on F^n.

16. **(a)** *Bessel's Inequality.* Let V be an inner product space, and let $S = \{v_1, v_2, \ldots, v_n\}$ be an orthonormal subset of V. Prove that for any $x \in \mathsf{V}$ we have

$$\|x\|^2 \geq \sum_{i=1}^{n} |\langle x, v_i \rangle|^2.$$

Hint: Apply Theorem 6.6 to $x \in \mathsf{V}$ and $\mathsf{W} = \text{span}(S)$. Then use Exercise 10 of Section 6.1.

(b) In the context of (a), prove that Bessel's inequality is an equality if and only if $x \in \text{span}(S)$.

17. Let T be a linear operator on an inner product space V. If $\langle \mathsf{T}(x), y \rangle = 0$ for all $x, y \in \mathsf{V}$, prove that $\mathsf{T} = \mathsf{T}_0$. In fact, prove this result if the equality holds for all x and y in some basis for V.

18. Let $\mathsf{V} = \mathsf{C}([-1, 1])$. Suppose that W_e and W_o denote the subspaces of V consisting of the even and odd functions, respectively. (See Exercise 22 of Section 1.3.) Prove that $\mathsf{W}_e^{\perp} = \mathsf{W}_o$, where the inner product on V is defined by

$$\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)\, dt.$$

19. In each of the following parts, find the orthogonal projection of the given vector on the given subspace W of the inner product space V.

(a) $\mathsf{V} = \mathsf{R}^2$, $u = (2, 6)$, and $\mathsf{W} = \{(x, y) : y = 4x\}$.

(b) $\mathsf{V} = \mathsf{R}^3$, $u = (2, 1, 3)$, and $\mathsf{W} = \{(x, y, z) : x + 3y - 2z = 0\}$.

(c) $\mathsf{V} = \mathsf{P}(R)$ with the inner product $\langle f(x), g(x) \rangle = \int_0^1 f(t)g(t)\, dt$, $h(x) = 4 + 3x - 2x^2$, and $\mathsf{W} = \mathsf{P}_1(R)$.

20. In each part of Exercise 19, find the distance from the given vector to the subspace W.

21. Let $\mathsf{V} = \mathsf{C}([-1, 1])$ with the inner product $\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)\, dt$, and let W be the subspace $\mathsf{P}_2(R)$, viewed as a space of functions. Use the orthonormal basis obtained in Example 5 to compute the "best" (closest) second-degree polynomial approximation of the function $h(t) = e^t$ on the interval $[-1, 1]$.

22. Let $\mathsf{V} = \mathsf{C}([0, 1])$ with the inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)\, dt$. Let W be the subspace spanned by the linearly independent set $\{t, \sqrt{t}\}$.

(a) Find an orthonormal basis for W.

(b) Let $h(t) = t^2$. Use the orthonormal basis obtained in (a) to obtain the "best" (closest) approximation of h in W.

23. Let V be the vector space defined in Example 5 of Section 1.2, the space of all sequences σ in F (where $F = R$ or $F = C$) such that $\sigma(n) \neq 0$ for only finitely many positive integers n. For $\sigma, \mu \in \mathsf{V}$, we define $\langle \sigma, \mu \rangle = \sum_{n=1}^{\infty} \sigma(n)\overline{\mu(n)}$. Since all but a finite number of terms of the series are zero, the series converges.

(a) Prove that $\langle \cdot, \cdot \rangle$ is an inner product on V, and hence V is an inner product space.

(b) For each positive integer n, let e_n be the sequence defined by $e_n(k) = \delta_{nk}$, where δ_{nk} is the Kronecker delta. Prove that $\{e_1, e_2, \ldots\}$ is an orthonormal basis for V.

(c) Let $\sigma_n = e_1 + e_n$ and $\mathsf{W} = \text{span}(\{\sigma_n : n \geq 2\})$.

(i) Prove that $e_1 \notin \mathsf{W}$, so $\mathsf{W} \neq \mathsf{V}$.

(ii) Prove that $\mathsf{W}^{\perp} = \{0\}$, and conclude that $\mathsf{W} \neq (\mathsf{W}^{\perp})^{\perp}$.

Thus the assumption in Exercise 13(c) that W is finite-dimensional is essential.

6.3 THE ADJOINT OF A LINEAR OPERATOR

In Section 6.1, we defined the conjugate transpose A^* of a matrix A. For a linear operator T on an inner product space V, we now define a related linear operator on V called the *adjoint* of T, whose matrix representation with respect to any orthonormal basis β for V is $[\mathsf{T}]_{\beta}^*$. The analogy between conjugation of complex numbers and adjoints of linear operators will become apparent. We first need a preliminary result.

Let V be an inner product space, and let $y \in \mathsf{V}$. The function $\mathsf{g}: \mathsf{V} \to F$ defined by $\mathsf{g}(x) = \langle x, y \rangle$ is clearly linear. More interesting is the fact that if V is finite-dimensional, every linear transformation from V into F is of this form.

Theorem 6.8. *Let V be a finite-dimensional inner product space over F, and let $\mathsf{g}: \mathsf{V} \to F$ be a linear transformation. Then there exists a unique vector $y \in \mathsf{V}$ such that $\mathsf{g}(x) = \langle x, y \rangle$ for all $x \in \mathsf{V}$.*

Proof. Let $\beta = \{v_1, v_2, \ldots, v_n\}$ be an orthonormal basis for V, and let

$$y = \sum_{i=1}^{n} \overline{\mathsf{g}(v_i)} v_i.$$

Define $\mathsf{h}: \mathsf{V} \to F$ by $\mathsf{h}(x) = \langle x, y \rangle$, which is clearly linear. Furthermore, for $1 \leq j \leq n$ we have

$$\mathsf{h}(v_j) = \langle v_j, y \rangle = \left\langle v_j, \sum_{i=1}^{n} \overline{\mathsf{g}(v_i)} v_i \right\rangle$$

$$= \sum_{i=1}^{n} \mathsf{g}(v_i)\,\langle v_j, v_i\rangle = \sum_{i=1}^{n} \mathsf{g}(v_i)\delta_{ji} = \mathsf{g}(v_j).$$

Since g and h agree on β, we have that $\mathsf{g} = \mathsf{h}$ by the corollary to Theorem 2.6 (p. 73).

To show that y is unique, suppose that $\mathsf{g}(x) = \langle x, y'\rangle$ for all x. Then $\langle x, y\rangle = \langle x, y'\rangle$ for all x; so by Theorem 6.1(e) (p. 331), we have $y = y'$. ∎

Example 1

Define $\mathsf{g}\colon \mathsf{R}^2 \to R$ by $\mathsf{g}(a_1, a_2) = 2a_1 + a_2$; clearly g is a linear transformation. Let $\beta = \{e_1, e_2\}$, and let $y = \mathsf{g}(e_1)e_1 + \mathsf{g}(e_2)e_2 = 2e_1 + e_2 = (2, 1)$, as in the proof of Theorem 6.8. Then $\mathsf{g}(a_1, a_2) = \langle (a_1, a_2), (2, 1)\rangle = 2a_1 + a_2$. ◆

Theorem 6.9. *Let V be a finite-dimensional inner product space, and let T be a linear operator on V. Then there exists a unique function $\mathsf{T}^*\colon \mathsf{V} \to \mathsf{V}$ such that $\langle \mathsf{T}(x), y\rangle = \langle x, \mathsf{T}^*(y)\rangle$ for all $x, y \in \mathsf{V}$. Furthermore, T^* is linear.*

Proof. Let $y \in \mathsf{V}$. Define $\mathsf{g}\colon \mathsf{V} \to F$ by $\mathsf{g}(x) = \langle \mathsf{T}(x), y\rangle$ for all $x \in \mathsf{V}$. We first show that g is linear. Let $x_1, x_2 \in \mathsf{V}$ and $c \in F$. Then

$$\begin{aligned}\mathsf{g}(cx_1 + x_2) &= \langle \mathsf{T}(cx_1 + x_2), y\rangle = \langle c\mathsf{T}(x_1) + \mathsf{T}(x_2), y\rangle \\ &= c\,\langle \mathsf{T}(x_1), y\rangle + \langle \mathsf{T}(x_2), y\rangle = c\mathsf{g}(x_1) + \mathsf{g}(x_2).\end{aligned}$$

Hence g is linear.

We now apply Theorem 6.8 to obtain a unique vector $y' \in \mathsf{V}$ such that $\mathsf{g}(x) = \langle x, y'\rangle$; that is, $\langle \mathsf{T}(x), y\rangle = \langle x, y'\rangle$ for all $x \in \mathsf{V}$. Defining $\mathsf{T}^*\colon \mathsf{V} \to \mathsf{V}$ by $\mathsf{T}^*(y) = y'$, we have $\langle \mathsf{T}(x), y\rangle = \langle x, \mathsf{T}^*(y)\rangle$.

To show that T^* is linear, let $y_1, y_2 \in \mathsf{V}$ and $c \in F$. Then for any $x \in \mathsf{V}$, we have

$$\begin{aligned}\langle x, \mathsf{T}^*(cy_1 + y_2)\rangle &= \langle \mathsf{T}(x), cy_1 + y_2\rangle \\ &= \bar{c}\,\langle \mathsf{T}(x), y_1\rangle + \langle \mathsf{T}(x), y_2\rangle \\ &= \bar{c}\,\langle x, \mathsf{T}^*(y_1)\rangle + \langle x, \mathsf{T}^*(y_2)\rangle \\ &= \langle x, c\mathsf{T}^*(y_1) + \mathsf{T}^*(y_2)\rangle\,.\end{aligned}$$

Since x is arbitrary, $\mathsf{T}^*(cy_1 + y_2) = c\mathsf{T}^*(y_1) + \mathsf{T}^*(y_2)$ by Theorem 6.1(e) (p. 331).

Finally, we need to show that T^* is unique. Suppose that $\mathsf{U}\colon \mathsf{V} \to \mathsf{V}$ is linear and that it satisfies $\langle \mathsf{T}(x), y\rangle = \langle x, \mathsf{U}(y)\rangle$ for all $x, y \in \mathsf{V}$. Then $\langle x, \mathsf{T}^*(y)\rangle = \langle x, \mathsf{U}(y)\rangle$ for all $x, y \in \mathsf{V}$, so $\mathsf{T}^* = \mathsf{U}$. ∎

The linear operator T^* described in Theorem 6.9 is called the **adjoint** of the operator T. The symbol T^* is read "T star."

Thus T^* is the unique operator on V satisfying $\langle \mathsf{T}(x), y\rangle = \langle x, \mathsf{T}^*(y)\rangle$ for all $x, y \in \mathsf{V}$. Note that we also have

$$\langle x, \mathsf{T}(y)\rangle = \overline{\langle \mathsf{T}(y), x\rangle} = \overline{\langle y, \mathsf{T}^*(x)\rangle} = \langle \mathsf{T}^*(x), y\rangle\,;$$

so $\langle x, \mathsf{T}(y)\rangle = \langle \mathsf{T}^*(x), y\rangle$ for all $x, y \in \mathsf{V}$. We may view these equations symbolically as adding a * to T when shifting its position inside the inner product symbol.

For an infinite-dimensional inner product space, the adjoint of a linear operator T may be defined to be the function T^* such that $\langle \mathsf{T}(x), y\rangle = \langle x, \mathsf{T}^*(y)\rangle$ for all $x, y \in \mathsf{V}$, provided it exists. Although the uniqueness and linearity of T^* follow as before, the existence of the adjoint is not guaranteed (see Exercise 24). The reader should observe the necessity of the hypothesis of finite-dimensionality in the proof of Theorem 6.8. Many of the theorems we prove about adjoints, nevertheless, do not depend on V being finite-dimensional.

Theorem 6.10 is a useful result for computing adjoints.

Theorem 6.10. *Let V be a finite-dimensional inner product space, and let β be an orthonormal basis for V. If T is a linear operator on V, then*

$$[\mathsf{T}^*]_\beta = [\mathsf{T}]_\beta^*.$$

Proof. Let $A = [\mathsf{T}]_\beta$, $B = [\mathsf{T}^*]_\beta$, and $\beta = \{v_1, v_2, \ldots, v_n\}$. Then from the corollary to Theorem 6.5 (p. 344), we have

$$B_{ij} = \langle \mathsf{T}^*(v_j), v_i\rangle = \overline{\langle v_i, \mathsf{T}^*(v_j)\rangle} = \overline{\langle \mathsf{T}(v_i), v_j\rangle} = \overline{A}_{ji} = (A^*)_{ij}.$$

Hence $B = A^*$. ∎

Corollary. *Let A be an $n \times n$ matrix. Then $\mathsf{L}_{A^*} = (\mathsf{L}_A)^*$.*

Proof. If β is the standard ordered basis for F^n, then, by Theorem 2.16 (p. 94), we have $[\mathsf{L}_A]_\beta = A$. Hence $[(\mathsf{L}_A)^*]_\beta = [\mathsf{L}_A]_\beta^* = A^* = [\mathsf{L}_{A^*}]_\beta$, and so $(\mathsf{L}_A)^* = \mathsf{L}_{A^*}$. ∎

As an illustration of Theorem 6.10, we compute the adjoint of a specific linear operator.

Example 2

Let T be the linear operator on C^2 defined by $\mathsf{T}(a_1, a_2) = (2ia_1 + 3a_2, a_1 - a_2)$. If β is the standard ordered basis for C^2, then

$$[\mathsf{T}]_\beta = \begin{pmatrix} 2i & 3 \\ 1 & -1 \end{pmatrix}.$$

So

$$[\mathsf{T}^*]_\beta = [\mathsf{T}]_\beta^* = \begin{pmatrix} -2i & 1 \\ 3 & -1 \end{pmatrix}.$$

Hence

$$\mathsf{T}^*(a_1, a_2) = (-2ia_1 + a_2, 3a_1 - a_2). \quad \blacklozenge$$

The following theorem suggests an analogy between the conjugates of complex numbers and the adjoints of linear operators.

Theorem 6.11. *Let V be an inner product space, and let T and U be linear operators on V whose adjoints exist. Then*

(a) *$\mathsf{T} + \mathsf{U}$ has an adjoint, and $(\mathsf{T} + \mathsf{U})^* = \mathsf{T}^* + \mathsf{U}^*$.*
(b) *$c\mathsf{T}$ has an adjoint, and $(c\mathsf{T})^* = \bar{c}\,\mathsf{T}^*$ for any $c \in F$.*
(c) *TU has an adjoint, and $(\mathsf{TU})^* = \mathsf{U}^*\mathsf{T}^*$.*
(d) *T^* has an adjoint, and $\mathsf{T}^{**} = \mathsf{T}$.*
(e) *I has an adjoint, and $\mathsf{I}^* = \mathsf{I}$.*

Proof. We prove (a) and (d); the rest are proved similarly. Let $x, y \in \mathsf{V}$.

(a) Because

$$\begin{aligned} \langle (\mathsf{T} + \mathsf{U})(x), y \rangle &= \langle \mathsf{T}(x) + \mathsf{U}(x), y \rangle \\ &= \langle x, \mathsf{T}^*(y) \rangle + \langle x, \mathsf{U}^*(y) \rangle \\ &= \langle x, \mathsf{T}^*(y) + \mathsf{U}^*(y) \rangle = \langle x, (\mathsf{T}^* + \mathsf{U}^*)(y) \rangle, \end{aligned}$$

it follows that $(\mathsf{T} + \mathsf{U})^*$ exists and is equal to $\mathsf{T}^* + \mathsf{U}^*$.

(d) Similarly, since

$$\langle \mathsf{T}^*(x), y \rangle, = \langle x, \mathsf{T}(y) \rangle,$$

(d) follows. ∎

Unless stated otherwise, for the remainder of this chapter we adopt the convention that a reference to the adjoint of a linear operator on an infinite-dimensional inner product space assumes its existence.

Corollary. *Let A and B be $n \times n$ matrices. Then*

(a) *$(A + B)^* = A^* + B^*$.*
(b) *$(cA)^* = \bar{c}A^*$ for all $c \in F$.*
(c) *$(AB)^* = B^*A^*$.*
(d) *$A^{**} = A$.*
(e) *$I^* = I$.*

Proof. We prove only (c); the remaining parts can be proved similarly.

Since $\mathsf{L}_{(AB)^*} = (\mathsf{L}_{AB})^* = (\mathsf{L}_A\mathsf{L}_B)^* = (\mathsf{L}_B)^*(\mathsf{L}_A)^* = \mathsf{L}_{B^*}\mathsf{L}_{A^*} = \mathsf{L}_{B^*A^*}$, we have $(AB)^* = B^*A^*$. ∎

In the preceding proof, we relied on the corollary to Theorem 6.10. An alternative proof, which holds even for nonsquare matrices, can be given by appealing directly to the definition of the conjugate transpose of a matrix (see Exercise 5).

Least Squares Approximation

Consider the following problem: An experimenter collects data by taking measurements $y_1, y_2, \ldots, y_m$ at times $t_1, t_2, \ldots, t_m$, respectively. For example, he or she may be measuring unemployment at various times during some period. Suppose that the data $(t_1, y_1), (t_2, y_2), \ldots, (t_m, y_m)$ are plotted as points in the plane. (See Figure 6.3.) From this plot, the experimenter feels that there exists an essentially linear relationship between y and t, say $y = ct + d$, and would like to find the constants c and d so that the line $y = ct + d$ represents the best possible *fit* to the data collected. One such estimate of fit is to calculate the error E that represents the sum of the squares of the vertical distances from the points to the line; that is,

$$E = \sum_{i=1}^{m} (y_i - ct_i - d)^2.$$

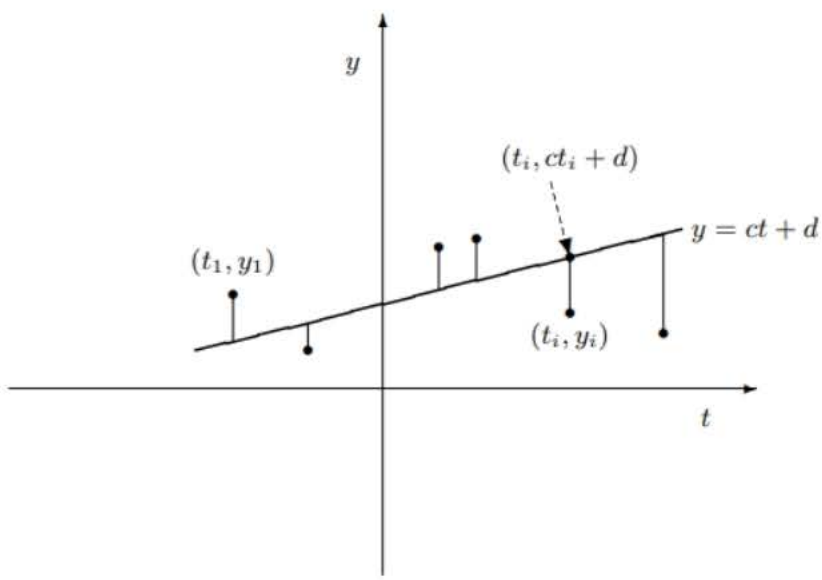


Figure 6.3

Thus the problem is reduced to finding the constants c and d that minimize E. (For this reason the line $y = ct + d$ is called the **least squares line**.) If

we let

$$A = \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{pmatrix}, \quad x = \begin{pmatrix} c \\ d \end{pmatrix}, \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix},$$

then it follows that $E = \|y - Ax\|^2$.

We develop a general method for finding an explicit vector $x_0 \in \mathsf{F}^n$ that minimizes E; that is, given an $m \times n$ matrix A, we find $x_0 \in \mathsf{F}^n$ such that $\|y - Ax_0\| \le \|y - Ax\|$ for all vectors $x \in \mathsf{F}^n$. This method not only allows us to find the linear function that best fits the data, but also, for any positive integer k, the best fit using a polynomial of degree at most k.

First, we need some notation and two simple lemmas. For $x, y \in \mathsf{F}^n$, let $\langle x, y \rangle_n$ denote the standard inner product of x and y in F^n. Recall that if x and y are regarded as column vectors, then $\langle x, y \rangle_n = y^* x$.

Lemma 1. *Let $A \in \mathsf{M}_{m \times n}(F)$, $x \in \mathsf{F}^n$, and $y \in \mathsf{F}^m$. Then*

$$\langle Ax, y \rangle_m = \langle x, A^* y \rangle_n .$$

Proof. By a generalization of the corollary to Theorem 6.11 (see Exercise 5(b)), we have

$$\langle Ax, y \rangle_m = y^*(Ax) = (y^* A)x = (A^* y)^* x = \langle x, A^* y \rangle_n . \qquad \blacksquare$$

Lemma 2. *Let $A \in \mathsf{M}_{m \times n}(F)$. Then* $\operatorname{rank}(A^* A) = \operatorname{rank}(A)$.

Proof. By the dimension theorem, we need only show that, for $x \in \mathsf{F}^n$, we have $A^* Ax = 0$ if and only if $Ax = 0$. Clearly, $Ax = 0$ implies that $A^* Ax = 0$. So assume that $A^* Ax = 0$. Then

$$0 = \langle A^* Ax, x \rangle_n = \langle Ax, A^{**} x \rangle_m = \langle Ax, Ax \rangle_m ,$$

so that $Ax = 0$. $\blacksquare$

Corollary. *If A is an $m \times n$ matrix such that* $\operatorname{rank}(A) = n$, *then $A^* A$ is invertible.*

Now let A be an $m \times n$ matrix and $y \in \mathsf{F}^m$. Define $\mathsf{W} = \{Ax : x \in \mathsf{F}^n\}$; that is, $\mathsf{W} = \mathsf{R}(\mathsf{L}_A)$. By the corollary to Theorem 6.6 (p. 347), there exists a unique vector in W that is closest to y. Call this vector Ax_0, where $x_0 \in \mathsf{F}^n$. Then $\|Ax_0 - y\| \le \|Ax - y\|$ for all $x \in \mathsf{F}^n$; so x_0 has the property that $E = \|Ax_0 - y\|$ is minimal, as desired.

To develop a practical method for finding such an x_0, we note from Theorem 6.6 and its corollary that $Ax_0 - y \in \mathsf{W}^\perp$; so $\langle Ax, Ax_0 - y \rangle_m = 0$ for

all $x \in \mathsf{F}^n$. Thus, by Lemma 1, we have that $\langle x, A^*(Ax_0 - y)\rangle_n = 0$ for all $x \in \mathsf{F}^n$; that is, $A^*(Ax_0 - y) = 0$. So we need only find a solution x_0 to $A^*Ax = A^*y$. If, in addition, we assume that $\text{rank}(A) = n$, then by Lemma 2 we have $x_0 = (A^*A)^{-1}A^*y$. We summarize this discussion in the following theorem.

Theorem 6.12. *Let $A \in \mathsf{M}_{m\times n}(F)$ and $y \in \mathsf{F}^m$. Then there exists $x_0 \in \mathsf{F}^n$ such that $(A^*A)x_0 = A^*y$ and $\|Ax_0 - y\| \le \|Ax - y\|$ for all $x \in \mathsf{F}^n$. Furthermore, if $\text{rank}(A) = n$, then $x_0 = (A^*A)^{-1}A^*y$.*

To return to our experimenter, let us suppose that the data collected are $(1,2), (2,3), (3,5)$, and $(4,7)$. Then

$$A = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} 2 \\ 3 \\ 5 \\ 7 \end{pmatrix};$$

hence

$$A^*A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 30 & 10 \\ 10 & 4 \end{pmatrix}.$$

Thus

$$(A^*A)^{-1} = \frac{1}{20}\begin{pmatrix} 4 & -10 \\ -10 & 30 \end{pmatrix}.$$

Therefore

$$\begin{pmatrix} c \\ d \end{pmatrix} = x_0 = \frac{1}{20}\begin{pmatrix} 4 & -10 \\ -10 & 30 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \\ 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 1.7 \\ 0 \end{pmatrix}.$$

It follows that the line $y = 1.7t$ is the least squares line. The error E may be computed directly as $\|Ax_0 - y\|^2 = 0.3$.

Suppose that the experimenter chose the times t_i $(1 \le i \le m)$ to satisfy

$$\sum_{i=1}^{m} t_i = 0.$$

Then the two columns of A would be orthogonal, so A^*A would be a diagonal matrix (see Exercise 19). In this case, the computations are greatly simplified.

In practice, the $m \times 2$ matrix A in our least squares application has rank equal to two, and hence A^*A is invertible by the corollary to Lemma 2. For, otherwise, the first column of A is a multiple of the second column, which

consists only of ones. But this would occur only if the experimenter collects all the data at exactly one time.

Finally, the method above may also be applied if, for some k, the experimenter wants to fit a polynomial of degree at most k to the data. For instance, if a polynomial $y = ct^2 + dt + e$ of degree at most 2 is desired, the appropriate model is

$$x = \begin{pmatrix} c \\ d \\ e \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \text{and} \quad A = \begin{pmatrix} t_1^2 & t_1 & 1 \\ \vdots & \vdots & \vdots \\ t_m^2 & t_m & 1 \end{pmatrix}.$$

Minimal Solutions to Systems of Linear Equations

Even when a system of linear equations $Ax = b$ is consistent, there may be no unique solution. In such cases, it may be desirable to find a solution of minimal norm. A solution s to $Ax = b$ is called a **minimal solution** if $\|s\| \leq \|u\|$ for all other solutions u. The next theorem assures that every consistent system of linear equations has a unique minimal solution and provides a method for computing it.

Theorem 6.13. *Let $A \in \mathsf{M}_{m\times n}(F)$ and $b \in \mathsf{F}^m$. Suppose that $Ax = b$ is consistent. Then the following statements are true.*

(a) *There exists exactly one minimal solution s of $Ax = b$, and $s \in \mathsf{R}(\mathsf{L}_{A^*})$.*

(b) *The vector s is the only solution to $Ax = b$ that lies in $\mathsf{R}(\mathsf{L}_{A^*})$; in fact, if u satisfies $(AA^*)u = b$, then $s = A^*u$.*

Proof. (a) For simplicity of notation, we let $\mathsf{W} = \mathsf{R}(\mathsf{L}_{A^*})$ and $\mathsf{W}' = \mathsf{N}(\mathsf{L}_A)$. Let x be any solution to $Ax = b$. By Theorem 6.6 (p. 347), $x = s + y$ for some $s \in \mathsf{W}$ and $y \in \mathsf{W}^\perp$. But $\mathsf{W}^\perp = \mathsf{W}'$ by Exercise 12, and therefore $b = Ax = As + Ay = As$. So s is a solution to $Ax = b$ that lies in W. To prove (a), we need only show that s is the unique minimal solution. Let v be any solution to $Ax = b$. By Theorem 3.9 (p. 172), we have that $v = s + u$, where $u \in \mathsf{W}'$. Since $s \in \mathsf{W}$, which equals $\mathsf{W}'^\perp$ by Exercise 12, we have

$$\|v\|^2 = \|s + u\|^2 = \|s\|^2 + \|u\|^2 \geq \|s\|^2$$

by Exercise 10 of Section 6.1. Thus s is a minimal solution. We can also see from the preceding calculation that if $\|v\| = \|s\|$, then $u = 0$; hence $v = s$. Therefore s is the unique minimal solution to $Ax = b$, proving (a).

(b) Assume that v is also a solution to $Ax = b$ that lies in W. Then

$$v - s \in \mathsf{W} \cap \mathsf{W}' = \mathsf{W} \cap \mathsf{W}^\perp = \{0\};$$

so $v = s$.

Finally, suppose that $(AA^*)u = b$, and let $v = A^*u$. Then $v \in \mathsf{W}$ and $Av = b$. Therefore $s = v = A^*u$ by the discussion above. ∎

Example 3

Consider the system

$$\begin{aligned} x + 2y + z &= 4 \\ x - y + 2z &= -11 \\ x + 5y &= 19. \end{aligned}$$

Let

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 2 \\ 1 & 5 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 4 \\ -11 \\ 19 \end{pmatrix}.$$

To find the minimal solution to this system, we must first find some solution u to $AA^*x = b$. Now

$$AA^* = \begin{pmatrix} 6 & 1 & 11 \\ 1 & 6 & -4 \\ 11 & -4 & 26 \end{pmatrix};$$

so we consider the system

$$\begin{aligned} 6x + y + 11z &= 4 \\ x + 6y - 4z &= -11 \\ 11x - 4y + 26z &= 19, \end{aligned}$$

for which one solution is

$$u = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}.$$

(Any solution will suffice.) Hence

$$s = A^*u = \begin{pmatrix} -1 \\ 4 \\ -3 \end{pmatrix}$$

is the minimal solution to the given system. ◆

EXERCISES

1. Label the following statements as true or false. Assume that the underlying inner product spaces are finite-dimensional.

- **(a)** Every linear operator has an adjoint.
- **(b)** Every linear operator on V has the form $x \to \langle x, y \rangle$ for some $y \in \mathsf{V}$.
- **(c)** For every linear operator T on V and every ordered basis β for V, we have $[\mathsf{T}^*]_\beta = ([\mathsf{T}]_\beta)^*$.
- **(d)** The adjoint of a linear operator is unique.

(e) For any linear operators T and U and scalars a and b,

$$(a\mathsf{T} + b\mathsf{U})^* = a\mathsf{T}^* + b\mathsf{U}^*.$$

(f) For any $n \times n$ matrix A, we have $(\mathsf{L}_A)^* = \mathsf{L}_{A^*}$.

(g) For any linear operator T, we have $(\mathsf{T}^*)^* = \mathsf{T}$.

2. For each of the following inner product spaces V (over F) and linear transformations $\mathsf{g}: \mathsf{V} \to F$, find a vector y such that $\mathsf{g}(x) = \langle x, y\rangle$ for all $x \in \mathsf{V}$.

(a) $\mathsf{V} = \mathsf{R}^3$, $\mathsf{g}(a_1, a_2, a_3) = a_1 - 2a_2 + 4a_3$

(b) $\mathsf{V} = \mathsf{C}^2$, $\mathsf{g}(z_1, z_2) = z_1 - 2z_2$

(c) $\mathsf{V} = \mathsf{P}_2(R)$ with $\langle f(x), h(x)\rangle = \int_0^1 f(t)h(t)\,dt$, $\mathsf{g}(f) = f(0) + f'(1)$

3. For each of the following inner product spaces V and linear operators T on V, evaluate T^* at the given vector in V.

(a) $\mathsf{V} = \mathsf{R}^2$, $\mathsf{T}(a, b) = (2a + b, a - 3b)$, $x = (3, 5)$.

(b) $\mathsf{V} = \mathsf{C}^2$, $\mathsf{T}(z_1, z_2) = (2z_1 + iz_2, (1 - i)z_1)$, $x = (3 - i, 1 + 2i)$.

(c) $\mathsf{V} = \mathsf{P}_1(R)$ with $\langle f(x), g(x)\rangle = \int_{-1}^1 f(t)g(t)\,dt$, $\mathsf{T}(f) = f' + 3f$, $f(t) = 4 - 2t$

4. Complete the proof of Theorem 6.11.

5. **(a)** Complete the proof of the corollary to Theorem 6.11 by using Theorem 6.11, as in the proof of (c).

(b) State a result for nonsquare matrices that is analogous to the corollary to Theorem 6.11, and prove it using a matrix argument.

6. Let T be a linear operator on an inner product space V. Let $\mathsf{U}_1 = \mathsf{T} + \mathsf{T}^*$ and $\mathsf{U}_2 = \mathsf{T}\mathsf{T}^*$. Prove that $\mathsf{U}_1 = \mathsf{U}_1^*$ and $\mathsf{U}_2 = \mathsf{U}_2^*$.

7. Give an example of a linear operator T on an inner product space V such that $\mathsf{N}(\mathsf{T}) \neq \mathsf{N}(\mathsf{T}^*)$.

8. Let V be a finite-dimensional inner product space, and let T be a linear operator on V. Prove that if T is invertible, then T^* is invertible and $(\mathsf{T}^*)^{-1} = (\mathsf{T}^{-1})^*$.

9. Prove that if $\mathsf{V} = \mathsf{W} \oplus \mathsf{W}^\perp$ and T is the projection on W along $\mathsf{W}^\perp$, then $\mathsf{T} = \mathsf{T}^*$. *Hint:* Recall that $\mathsf{N}(\mathsf{T}) = \mathsf{W}^\perp$. (For definitions, see the exercises of Sections 1.3 and 2.1.)

10. Let T be a linear operator on an inner product space V. Prove that $\|\mathsf{T}(x)\| = \|x\|$ for all $x \in \mathsf{V}$ if and only if $\langle \mathsf{T}(x), \mathsf{T}(y)\rangle = \langle x, y\rangle$ for all $x, y \in \mathsf{V}$. *Hint:* Use Exercise 20 of Section 6.1.

11. For a linear operator T on an inner product space V, prove that $T^*T = T_0$ implies $T = T_0$. Is the same result true if we assume that $TT^* = T_0$?

12. Let V be an inner product space, and let T be a linear operator on V. Prove the following results.

(a) $R(T^*)^\perp = N(T)$.
(b) If V is finite-dimensional, then $R(T^*) = N(T)^\perp$. *Hint:* Use Exercise 13(c) of Section 6.2.

13. Let T be a linear operator on a finite-dimensional inner product space V. Prove the following results.

(a) $N(T^*T) = N(T)$. Deduce that $\text{rank}(T^*T) = \text{rank}(T)$.
(b) $\text{rank}(T) = \text{rank}(T^*)$. Deduce from (a) that $\text{rank}(TT^*) = \text{rank}(T)$.
(c) For any $n \times n$ matrix A, $\text{rank}(A^*A) = \text{rank}(AA^*) = \text{rank}(A)$.

14. Let V be an inner product space, and let $y, z \in V$. Define $T: V \to V$ by $T(x) = \langle x, y\rangle z$ for all $x \in V$. First prove that T is linear. Then show that T^* exists, and find an explicit expression for it.

The following definition is used in Exercises 15–17 and is an extension of the definition of the *adjoint* of a linear operator.

Definition. *Let $T: V \to W$ be a linear transformation, where V and W are finite-dimensional inner product spaces with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, respectively. A function $T^*: W \to V$ is called an* **adjoint** *of T if $\langle T(x), y\rangle_2 = \langle x, T^*(y)\rangle_1$ for all $x \in V$ and $y \in W$.*

15. Let $T: V \to W$ be a linear transformation, where V and W are finite-dimensional inner product spaces with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, respectively. Prove the following results.

(a) There is a unique adjoint T^* of T, and T^* is linear.
(b) If β and γ are orthonormal bases for V and W, respectively, then $[T^*]_\gamma^\beta = ([T]_\beta^\gamma)^*$.
(c) $\text{rank}(T^*) = \text{rank}(T)$.
(d) $\langle T^*(x), y\rangle_1 = \langle x, T(y)\rangle_2$ for all $x \in W$ and $y \in V$.
(e) For all $x \in V$, $T^*T(x) = 0$ if and only if $T(x) = 0$.

16. State and prove a result that extends the first four parts of Theorem 6.11 using the preceding definition.

17. Let $T: V \to W$ be a linear transformation, where V and W are finite-dimensional inner product spaces. Prove that $(R(T^*))^\perp = N(T)$, using the preceding definition.

18.† Let A be an $n \times n$ matrix. Prove that $\det(A^*) = \overline{\det(A)}$. Visit `goo.gl/csqoFY` for a solution.

19. Suppose that A is an $m \times n$ matrix in which no two columns are identical. Prove that A^*A is a diagonal matrix if and only if every pair of columns of A is orthogonal.

20. For each of the sets of data that follows, use the least squares approximation to find the best fits with both (i) a linear function and (ii) a quadratic function. Compute the error E in both cases.

(a) $\{(-3,9),(-2,6),(0,2),(1,1)\}$
(b) $\{(1,2),(3,4),(5,7),(7,9),(9,12)\}$
(c) $\{(-2,4),(-1,3),(0,1),(1,-1),(2,-3)\}$

21. In physics, *Hooke's law* states that (within certain limits) there is a linear relationship between the length x of a spring and the force y applied to (or exerted by) the spring. That is, $y = cx + d$, where c is called the **spring constant**. Use the following data to estimate the spring constant (the length is given in inches and the force is given in pounds).

Length x	Force y
3.5	1.0
4.0	2.2
4.5	2.8
5.0	4.3

22. Find the minimal solution to each of the following systems of linear equations.

(a) $x + 2y - z = 12$

(b)
$$\begin{aligned} x + 2y - z &= 1 \\ 2x + 3y + z &= 2 \\ 4x + 7y - z &= 4 \end{aligned}$$

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

(d)
$$\begin{aligned} x + y + z - w &= 1 \\ 2x - y \quad\; + w &= 1 \end{aligned}$$

23. Consider the problem of finding the least squares line $y = ct + d$ corresponding to the m observations $(t_1, y_1), (t_2, y_2), \ldots, (t_m, y_m)$.

(a) Show that the equation $(A^*A)x_0 = A^*y$ of Theorem 6.12 takes the form of the *normal equations:*

$$\left(\sum_{i=1}^{m} t_i^2\right) c + \left(\sum_{i=1}^{m} t_i\right) d = \sum_{i=1}^{m} t_i y_i$$

and

$$\left(\sum_{i=1}^{m} t_i\right) c + md = \sum_{i=1}^{m} y_i.$$

These equations may also be obtained from the error E by setting the partial derivatives of E with respect to both c and d equal to zero.

(b) Use the second normal equation of (a) to show that the least squares line must pass through the *center of mass*, $(\bar{t}, \bar{y})$, where

$$\bar{t} = \frac{1}{m}\sum_{i=1}^{m} t_i \qquad \text{and} \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i.$$

24. Let V and $\{e_1, e_2, \ldots\}$ be defined as in Exercise 23 of Section 6.2. Define $\mathsf{T}: \mathsf{V} \to \mathsf{V}$ by

$$\mathsf{T}(\sigma)(k) = \sum_{i=k}^{\infty} \sigma(i) \quad \text{for every positive integer } k.$$

Notice that the infinite series in the definition of T converges because $\sigma(i) \neq 0$ for only finitely many i.

(a) Prove that T is a linear operator on V.

(b) Prove that for any positive integer n, $\mathsf{T}(e_n) = \sum_{i=1}^{n} e_i$.

(c) Prove that T has no adjoint. *Hint:* By way of contradiction, suppose that T^* exists. Prove that for any positive integer n, $\mathsf{T}^*(e_n)(k) \neq 0$ for infinitely many k.

6.4 NORMAL AND SELF-ADJOINT OPERATORS

We have seen the importance of diagonalizable operators in Chapter 5. For an operator on a vector space V to be diagonalizble, it is necessary and sufficient for V to contain a basis of eigenvectors for this operator. As V is an inner product space in this chapter, it is reasonable to seek conditions that guarantee that V has an orthonormal basis of eigenvectors. A very important result that helps achieve our goal is Schur's theorem (Theorem 6.14). The formulation that follows is in terms of linear operators. The next section contains the more familiar matrix form. We begin with a lemma.

Lemma. *Let T be a linear operator on a finite-dimensional inner product space V. If T has an eigenvector, then so does T^*.*

Proof. Suppose that v is an eigenvector of T with corresponding eigenvalue λ. Then for any $x \in \mathsf{V}$,

$$0 = \langle 0, x\rangle = \langle (\mathsf{T} - \lambda \mathsf{I})(v), x\rangle = \langle v, (\mathsf{T} - \lambda \mathsf{I})^*(x)\rangle = \left\langle v, (\mathsf{T}^* - \overline{\lambda}\mathsf{I})(x)\right\rangle,$$

and hence v is orthogonal to the range of $\mathsf{T}^* - \overline{\lambda}\mathsf{I}$. So $\mathsf{T}^* - \overline{\lambda}\mathsf{I}$ is not onto and hence is not one-to-one. Thus $\mathsf{T}^* - \overline{\lambda}\mathsf{I}$ has a nonzero null space, and any nonzero vector in this null space is an eigenvector of T^* with corresponding eigenvalue $\overline{\lambda}$. ∎

Recall (see the exercises of Section 2.1 and see Section 5.4) that a subspace W of V is said to be **T-invariant** if $\mathsf{T}(\mathsf{W})$ is contained in W. If W is T-invariant, we may define the restriction $\mathsf{T}_\mathsf{W}\colon \mathsf{W} \to \mathsf{W}$ by $\mathsf{T}_\mathsf{W}(x) = \mathsf{T}(x)$ for all $x \in \mathsf{W}$. It is clear that T_W is a linear operator on W. Recall from Section 5.2 that a polynomial is said to **split** if it factors into linear polynomials.

Theorem 6.14 (Schur). *Let T be a linear operator on a finite-dimensional inner product space V. Suppose that the characteristic polynomial of T splits. Then there exists an orthonormal basis γ for V such that the matrix $[\mathsf{T}]_\gamma$ is upper triangular.*

Proof. By Exercise 12(a) of Section 5.2, there exists an ordered basis $\beta = \{w_1, w_2, \ldots, w_n\}$ for V such that $[\mathsf{T}]_\beta$ is upper triangular. Now apply the Gram-Schmidt process to β to obtain an orthogonal basis $\beta' = \{v_1, v_2, \ldots, v_n\}$ for V. For each k, $1 \le k \le n$, let

$$S_k = \{w_1, w_2, \ldots, w_k\} \quad \text{and} \quad S_k' = \{v_1, v_2, \ldots, v_k\}.$$

As in the proof of Theorem 6.4, $\operatorname{span}(S_k) = \operatorname{span}(S_k')$ for all k. By Exercise 12 of Section 2.2, $\mathsf{T}(w_k) \in \operatorname{span}(S_k)$ for all k. Hence $\mathsf{T}(v_k) \in \operatorname{span}(S_k')$ for all k, and so $[\mathsf{T}]_{\beta'}$ is upper triangular by the same exercise. Finally, let $z_i = \frac{1}{\|v_i\|} v_i$ for all $1 \le i \le n$ and $\gamma = \{z_1, z_2, \ldots, z_n\}$. Then γ is an orthonormal basis for V, and $[\mathsf{T}]_\gamma$ is upper triangular. ∎

We now return to our original goal of finding an orthonormal basis of eigenvectors of a linear operator T on a finite-dimensional inner product space V. Note that if such an orthonormal basis β exists, then $[\mathsf{T}]_\beta$ is a diagonal matrix, and hence $[\mathsf{T}^*]_\beta = [\mathsf{T}]_\beta^*$ is also a diagonal matrix. Because diagonal matrices commute, we conclude that T and T^* commute. Thus *if V possesses an orthonormal basis of eigenvectors of T, then* $\mathsf{T}\mathsf{T}^* = \mathsf{T}^*\mathsf{T}$.

Definitions. *Let V be an inner product space, and let T be a linear operator on V. We say that T is **normal** if $\mathsf{T}\mathsf{T}^* = \mathsf{T}^*\mathsf{T}$. An $n \times n$ real or complex matrix A is **normal** if $AA^* = A^*A$.*

It follows immediately from Theorem 6.10 (p. 356) that T is normal if and only if $[\mathsf{T}]_\beta$ is normal, where β is an orthonormal basis.

Example 1

Let $\mathsf{T}: \mathsf{R}^2 \to \mathsf{R}^2$ be rotation by θ, where $0 < \theta < \pi$. The matrix representation of T in the standard ordered basis is given by

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Note that $AA^* = I = A^*A$; so A, and hence T, is normal. ♦

Example 2

Suppose that A is a real *skew-symmetric* matrix; that is, $A^t = -A$. Then A is normal because both AA^t and A^tA are equal to $-A^2$. ♦

Clearly, the operator T in Example 1 does not even possess one eigenvector. So in the case of a real inner product space, we see that normality is not sufficient to guarantee an orthonormal basis of eigenvectors. All is not lost, however. We show that normality suffices if V is a *complex* inner product space.

Before we prove the promised result for normal operators, we need some general properties of normal operators.

Theorem 6.15. *Let V be an inner product space, and let T be a normal operator on V. Then the following statements are true.*

(a) $\|\mathsf{T}(x)\| = \|\mathsf{T}^*(x)\|$ *for all* $x \in \mathsf{V}$.
(b) $\mathsf{T} - c\mathsf{I}$ *is normal for every* $c \in F$.
(c) *If x is an eigenvector of T corresponding to eigenvalue λ, then x is also an eigenvector of T^* corresponding to eigenvalue $\overline{\lambda}$. That is, if $\mathsf{T}(x) = \lambda x$, then $\mathsf{T}^*(x) = \overline{\lambda}x$.*
(d) *If λ_1 and λ_2 are distinct eigenvalues of T with corresponding eigenvectors x_1 and x_2, then x_1 and x_2 are orthogonal.*

Proof. (a) For any $x \in \mathsf{V}$, we have

$$\begin{aligned} \|\mathsf{T}(x)\|^2 &= \langle \mathsf{T}(x), \mathsf{T}(x) \rangle = \langle \mathsf{T}^*\mathsf{T}(x), x \rangle = \langle \mathsf{T}\mathsf{T}^*(x), x \rangle \\ &= \langle \mathsf{T}^*(x), \mathsf{T}^*(x) \rangle = \|\mathsf{T}^*(x)\|^2. \end{aligned}$$

The proof of (b) is left as an exercise.

(c) Suppose that $\mathsf{T}(x) = \lambda x$ for some $x \in \mathsf{V}$. Let $\mathsf{U} = \mathsf{T} - \lambda\mathsf{I}$. Then $\mathsf{U}(x) = 0$, and U is normal by (b). Thus (a) implies that

$$0 = \|\mathsf{U}(x)\| = \|\mathsf{U}^*(x)\| = \|(\mathsf{T}^* - \overline{\lambda}\mathsf{I})(x)\| = \|\mathsf{T}^*(x) - \overline{\lambda}x\|.$$

Hence $\mathsf{T}^*(x) = \overline{\lambda}x$. So x is an eigenvector of T^*.

(d) Let λ_1 and λ_2 be distinct eigenvalues of T with corresponding eigenvectors x_1 and x_2. Then, using (c), we have

$$\lambda_1 \langle x_1, x_2 \rangle = \langle \lambda_1 x_1, x_2 \rangle = \langle \mathsf{T}(x_1), x_2 \rangle = \langle x_1, \mathsf{T}^*(x_2) \rangle$$

$$= \langle x_1, \overline{\lambda}_2 x_2 \rangle = \lambda_2 \langle x_1, x_2 \rangle .$$

Since $\lambda_1 \neq \lambda_2$, we conclude that $\langle x_1, x_2 \rangle = 0$. ∎

Theorem 6.16. *Let T be a linear operator on a finite-dimensional complex inner product space V. Then T is normal if and only if there exists an orthonormal basis for V consisting of eigenvectors of T.*

Proof. Suppose that T is normal. By the fundamental theorem of algebra (Theorem D.4), the characteristic polynomial of T splits. So we may apply Schur's theorem to obtain an orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$ for V such that $[\mathsf{T}]_\beta = A$ is upper triangular. We know that v_1 is an eigenvector of T because A is upper triangular. Assume that $v_1, v_2, \ldots, v_{k-1}$ are eigenvectors of T. We claim that v_k is also an eigenvector of T. It then follows by mathematical induction on k that all of the v_i's are eigenvectors of T. Consider any $j < k$, and let λ_j denote the eigenvalue of T corresponding to v_j. By Theorem 6.15, $\mathsf{T}^*(v_j) = \overline{\lambda}_j v_j$. Since A is upper triangular,

$$\mathsf{T}(v_k) = A_{1k}v_1 + A_{2k}v_2 + \cdots + A_{jk}v_j + \cdots + A_{kk}v_k.$$

Furthermore, by the corollary to Theorem 6.5 (p. 345),

$$A_{jk} = \langle \mathsf{T}(v_k), v_j \rangle = \langle v_k, \mathsf{T}^*(v_j) \rangle = \langle v_k, \overline{\lambda}_j v_j \rangle = \lambda_j \langle v_k, v_j \rangle = 0.$$

It follows that $\mathsf{T}(v_k) = A_{kk}v_k$, and hence v_k is an eigenvector of T. So by induction, all the vectors in β are eigenvectors of T.

The converse was already proved on page 367. ∎

Interestingly, as the next example shows, Theorem 6.16 does not extend to infinite-dimensional complex inner product spaces.

Example 3

Consider the inner product space H with the orthonormal set S from Example 9 in Section 6.1. Let $\mathsf{V} = \text{span}(S)$, and let T and U be the linear operators on V defined by $\mathsf{T}(f) = f_1 f$ and $\mathsf{U}(f) = f_{-1} f$. Then

$$\mathsf{T}(f_n) = f_{n+1} \quad \text{and} \quad \mathsf{U}(f_n) = f_{n-1}$$

for all integers n. Thus

$$\langle \mathsf{T}(f_m), f_n \rangle = \langle f_{m+1}, f_n \rangle = \delta_{(m+1),n} = \delta_{m,(n-1)} = \langle f_m, f_{n-1} \rangle = \langle f_m, \mathsf{U}(f_n) \rangle .$$

It follows that $\mathsf{U} = \mathsf{T}^*$. Furthermore, $\mathsf{T}\mathsf{T}^* = \mathsf{I} = \mathsf{T}^*\mathsf{T}$; so T is normal.

We show that T has no eigenvectors. Suppose that f is an eigenvector of T, say, $\mathsf{T}(f) = \lambda f$ for some λ. Since V equals the span of S, we may write

$$f = \sum_{i=n}^{m} a_i f_i, \quad \text{where } a_m \neq 0.$$

Hence

$$\sum_{i=n}^{m} a_i f_{i+1} = \mathsf{T}(f) = \lambda f = \sum_{i=n}^{m} \lambda a_i f_i.$$

Since $a_m \neq 0$, we can write f_{m+1} as a linear combination of $f_n, f_{n+1}, \ldots, f_m$. But this is a contradiction because S is linearly independent. ♦

Example 1 illustrates that normality is not sufficient to guarantee the existence of an orthonormal basis of eigenvectors for real inner product spaces. For real inner product spaces, we must replace normality by the stronger condition that $\mathsf{T} = \mathsf{T}^*$ in order to guarantee such a basis.

Definitions. *Let T be a linear operator on an inner product space V. We say that T is* ***self-adjoint (or Hermitian)*** *if $\mathsf{T} = \mathsf{T}^*$. An $n \times n$ real or complex matrix A is* ***self-adjoint (or Hermitian)*** *if $A = A^*$.*

It follows immediately that if β is an orthonormal basis, then T is self-adjoint if and only if $[\mathsf{T}]_\beta$ is self-adjoint. For real matrices, this condition reduces to the requirement that A be symmetric.

Before we state our main result for self-adjoint operators, we need some preliminary work.

By definition, a linear operator on a real inner product space has only real eigenvalues. The lemma that follows shows that the same can be said for self-adjoint operators on a complex inner product space. Similarly, the characteristic polynomial of every linear operator on a complex inner product space splits, and the same is true for self-adjoint operators on a real inner product space.

Lemma. *Let T be a self-adjoint operator on a finite-dimensional inner product space V. Then*

(a) *Every eigenvalue of T is real.*

(b) *Suppose that V is a real inner product space. Then the characteristic polynomial of T splits.*

Proof. (a) Suppose that $\mathsf{T}(x) = \lambda x$ for $x \neq 0$. Because a self-adjoint operator is also normal, we can apply Theorem 6.15(c) to obtain

$$\lambda x = \mathsf{T}(x) = \mathsf{T}^*(x) = \overline{\lambda} x.$$

So $\lambda = \overline{\lambda}$; that is, λ is real.

(b) Let $n = \dim(\mathsf{V})$, β be an orthonormal basis for V, and $A = [\mathsf{T}]_\beta$. Then A is self-adjoint. Let T_A be the linear operator on C^n defined by $\mathsf{T}_A(x) = Ax$ for all $x \in \mathsf{C}^n$. Note that T_A is self-adjoint because $[\mathsf{T}_A]_\gamma = A$, where γ is the standard ordered (orthonormal) basis for C^n. So, by (a), the eigenvalues of T_A are real. By the fundamental theorem of algebra, the characteristic polynomial of T_A splits into factors of the form $t - \lambda$. Since each

λ is real, the characteristic polynomial splits over R. But T_A has the same characteristic polynomial as A, which has the same characteristic polynomial as T. Therefore the characteristic polynomial of T splits. ∎

We are now able to establish one of the major results of this chapter.

Theorem 6.17. *Let T be a linear operator on a finite-dimensional real inner product space V. Then T is self-adjoint if and only if there exists an orthonormal basis β for V consisting of eigenvectors of T.*

Proof. Suppose that T is self-adjoint. By the lemma, we may apply Schur's theorem to obtain an orthonormal basis β for V such that the matrix $A = [\mathsf{T}]_\beta$ is upper triangular. But

$$A^* = [\mathsf{T}]_\beta^* = [\mathsf{T}^*]_\beta = [\mathsf{T}]_\beta = A.$$

So A and A^* are both upper triangular, and therefore A is a diagonal matrix. Thus β must consist of eigenvectors of T.

The converse is left as an exercise. ∎

We restate this theorem in matrix form in the next section (as Theorem 6.20 on p. 381).

Example 4

As we noted earlier, real symmetric matrices are self-adjoint, and self-adjoint matrices are normal. The following matrix A is complex and symmetric:

$$A = \begin{pmatrix} i & i \\ i & 1 \end{pmatrix} \quad \text{and} \quad A^* = \begin{pmatrix} -i & -i \\ -i & 1 \end{pmatrix}.$$

But A is not normal, because $(AA^*)_{12} = 1+i$ and $(A^*A)_{12} = 1-i$. Therefore complex symmetric matrices need not be normal. ◆

EXERCISES

1. Label the following statements as true or false. Assume that the underlying inner product spaces are finite-dimensional.
   - **(a)** Every self-adjoint operator is normal.
   - **(b)** Operators and their adjoints have the same eigenvectors.
   - **(c)** If T is an operator on an inner product space V, then T is normal if and only if $[\mathsf{T}]_\beta$ is normal, where β is any ordered basis for V.
   - **(d)** A real or complex matrix A is normal if and only if L_A is normal.
   - **(e)** The eigenvalues of a self-adjoint operator must all be real.
   - **(f)** The identity and zero operators are self-adjoint.

(g) Every normal operator is diagonalizable.
(h) Every self-adjoint operator is diagonalizable.

2. For each linear operator T on an inner product space V, determine whether T is normal, self-adjoint, or neither. If possible, produce an orthonormal basis of eigenvectors of T for V and list the corresponding eigenvalues.

(a) $\mathsf{V} = \mathsf{R}^2$ and T is defined by $\mathsf{T}(a, b) = (2a - 2b, -2a + 5b)$.
(b) $\mathsf{V} = \mathsf{R}^3$ and T is defined by $\mathsf{T}(a, b, c) = (-a + b, 5b, 4a - 2b + 5c)$.
(c) $\mathsf{V} = \mathsf{C}^2$ and T is defined by $\mathsf{T}(a, b) = (2a + ib, a + 2b)$.
(d) $\mathsf{V} = \mathsf{P}_2(R)$ and T is defined by $\mathsf{T}(f) = f'$, where

$$\langle f(x), g(x) \rangle = \int_0^1 f(t)g(t)\, dt.$$

(e) $\mathsf{V} = \mathsf{M}_{2\times 2}(R)$ and T is defined by $\mathsf{T}(A) = A^t$.
(f) $\mathsf{V} = \mathsf{M}_{2\times 2}(R)$ and T is defined by $\mathsf{T}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ a & b \end{pmatrix}$.

3. Give an example of a linear operator T on R^2 and an ordered basis for R^2 that provides a counterexample to the statement in Exercise 1(c).

4. Let T and U be self-adjoint operators on an inner product space V. Prove that TU is self-adjoint if and only if $\mathsf{TU} = \mathsf{UT}$.

5. Prove (b) of Theorem 6.15.

6. Let V be a complex inner product space, and let T be a linear operator on V. Define

$$\mathsf{T}_1 = \frac{1}{2}(\mathsf{T} + \mathsf{T}^*) \quad \text{and} \quad \mathsf{T}_2 = \frac{1}{2i}(\mathsf{T} - \mathsf{T}^*).$$

(a) Prove that T_1 and T_2 are self-adjoint and that $\mathsf{T} = \mathsf{T}_1 + i\,\mathsf{T}_2$.
(b) Suppose also that $\mathsf{T} = \mathsf{U}_1 + i\mathsf{U}_2$, where U_1 and U_2 are self-adjoint. Prove that $\mathsf{U}_1 = \mathsf{T}_1$ and $\mathsf{U}_2 = \mathsf{T}_2$.
(c) Prove that T is normal if and only if $\mathsf{T}_1\mathsf{T}_2 = \mathsf{T}_2\mathsf{T}_1$.

7. Let T be a linear operator on an inner product space V, and let W be a T-invariant subspace of V. Prove the following results.

(a) If T is self-adjoint, then T_W is self-adjoint.
(b) $\mathsf{W}^\perp$ is T^*-invariant.
(c) If W is both T- and T^*-invariant, then $(\mathsf{T}_\mathsf{W})^* = (\mathsf{T}^*)_\mathsf{W}$.
(d) If W is both T- and T^*-invariant and T is normal, then T_W is normal.

8. Let T be a normal operator on a finite-dimensional complex inner product space V, and let W be a subspace of V. Prove that if W is T-invariant, then W is also T^*-invariant. *Hint:* Use Exercise 24 of Section 5.4.

9. Let T be a normal operator on a finite-dimensional inner product space V. Prove that $\mathsf{N}(\mathsf{T}) = \mathsf{N}(\mathsf{T}^*)$ and $\mathsf{R}(\mathsf{T}) = \mathsf{R}(\mathsf{T}^*)$. *Hint:* Use Theorem 6.15 and Exercise 12 of Section 6.3.

10. Let T be a self-adjoint operator on a finite-dimensional inner product space V. Prove that for all $x \in \mathsf{V}$

$$\|\mathsf{T}(x) \pm ix\|^2 = \|\mathsf{T}(x)\|^2 + \|x\|^2.$$

Deduce that $\mathsf{T} - i\mathsf{I}$ is invertible and that the adjoint of $(\mathsf{T} - i\mathsf{I})^{-1}$ is $(\mathsf{T} + i\mathsf{I})^{-1}$.

11. Assume that T is a linear operator on a complex (not necessarily finite-dimensional) inner product space V with an adjoint T^*. Prove the following results.

(a) If T is self-adjoint, then $\langle \mathsf{T}(x), x\rangle$ is real for all $x \in \mathsf{V}$.

(b) If T satisfies $\langle \mathsf{T}(x), x\rangle = 0$ for all $x \in \mathsf{V}$, then $\mathsf{T} = \mathsf{T}_0$. *Hint:* Replace x by $x + y$ and then by $x + iy$, and expand the resulting inner products.

(c) If $\langle \mathsf{T}(x), x\rangle$ is real for all $x \in \mathsf{V}$, then T is self-adjoint.

12. Let T be a normal operator on a finite-dimensional real inner product space V whose characteristic polynomial splits. Prove that V has an orthonormal basis of eigenvectors of T. Hence prove that T is self-adjoint.

13. An $n \times n$ real matrix A is said to be a **Gramian** matrix if there exists a real (square) matrix B such that $A = B^t B$. Prove that A is a Gramian matrix if and only if A is symmetric and all of its eigenvalues are nonnegative. *Hint:* Apply Theorem 6.17 to $\mathsf{T} = \mathsf{L}_A$ to obtain an orthonormal basis $\{v_1, v_2, \ldots, v_n\}$ of eigenvectors with the associated eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Define the linear operator U by $\mathsf{U}(v_i) = \sqrt{\lambda_i} v_i$.

14. *Simultaneous Diagonalization.* Let V be a finite-dimensional real inner product space, and let U and T be self-adjoint linear operators on V such that $\mathsf{UT} = \mathsf{TU}$. Prove that there exists an orthonormal basis for V consisting of vectors that are eigenvectors of both U and T. (The complex version of this result appears as Exercise 10 of Section 6.6.) *Hint:* For any eigenspace $\mathsf{W} = \mathsf{E}_\lambda$ of T, we have that W is both T- and U-invariant. By Exercise 7, we have that $\mathsf{W}^\perp$ is both T- and U-invariant. Apply Theorem 6.17 and Theorem 6.6 (p. 347).

15. Let A and B be symmetric $n \times n$ matrices such that $AB = BA$. Use Exercise 14 to prove that there exists an orthogonal matrix P such that P^tAP and P^tBP are both diagonal matrices.

16. Prove the *Cayley–Hamilton theorem* for a complex $n \times n$ matrix A. That is, if $f(t)$ is the characteristic polynomial of A, prove that $f(A) = O$. *Hint:* Use Schur's theorem to show that A may be assumed to be upper triangular, in which case

$$f(t) = \prod_{i=1}^{n} (A_{ii} - t).$$

Now if $\mathsf{T} = \mathsf{L}_A$, we have $(A_{jj}\mathsf{I} - \mathsf{T})(e_j) \in \text{span}(\{e_1, e_2, \ldots, e_{j-1}\})$ for $j \geq 2$, where $\{e_1, e_2, \ldots, e_n\}$ is the standard ordered basis for C^n. (The general case is proved in Section 5.4.)

The following definitions are used in Exercises 17 through 23.

Definitions. *A linear operator* T *on a finite-dimensional inner product space is called* ***positive definite*** [***positive semidefinite***] *if* T *is self-adjoint and* $\langle \mathsf{T}(x), x \rangle > 0$ [$\langle \mathsf{T}(x), x \rangle \geq 0$] *for all* $x \neq 0$.

An $n \times n$ *matrix* A *with entries from* R *or* C *is called* ***positive definite*** [***positive semidefinite***] *if* L_A *is positive definite* [*positive semidefinite*].

17. Let T and U be self-adjoint linear operators on an n-dimensional inner product space V, and let $A = [\mathsf{T}]_\beta$, where β is an orthonormal basis for V. Prove the following results.

(a) T is positive definite [semidefinite] if and only if all of its eigenvalues are positive [nonnegative].

(b) T is positive definite if and only if

$$\sum_{i,j} A_{ij} a_j \overline{a}_i > 0 \text{ for all nonzero } n\text{-tuples } (a_1, a_2, \ldots, a_n).$$

(c) T is positive semidefinite if and only if $A = B^*B$ for some square matrix B.

(d) If T and U are positive semidefinite operators such that $\mathsf{T}^2 = \mathsf{U}^2$, then $\mathsf{T} = \mathsf{U}$.

(e) If T and U are positive definite operators such that $\mathsf{TU} = \mathsf{UT}$, then TU is positive definite.

(f) T is positive definite [semidefinite] if and only if A is positive definite [semidefinite].

Because of (f), results analogous to items (a) through (d) hold for matrices as well as operators.

18. Let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be a linear transformation, where V and W are finite-dimensional inner product spaces. Prove the following results.

(a) $\mathsf{T}^*\mathsf{T}$ and $\mathsf{T}\mathsf{T}^*$ are positive semidefinite. (See Exercise 15 of Section 6.3.)

(b) $\operatorname{rank}(\mathsf{T}^*\mathsf{T}) = \operatorname{rank}(\mathsf{T}\mathsf{T}^*) = \operatorname{rank}(\mathsf{T})$.

19. Let T and U be positive definite operators on an inner product space V. Prove the following results.

(a) $\mathsf{T} + \mathsf{U}$ is positive definite.

(b) If $c > 0$, then $c\mathsf{T}$ is positive definite.

(c) T^{-1} is positive definite.

Visit `goo.gl/cQch7i` for a solution.

20. Let V be an inner product space with inner product $\langle \cdot, \cdot \rangle$, and let T be a positive definite linear operator on V. Prove that $\langle x, y \rangle' = \langle \mathsf{T}(x), y \rangle$ defines another inner product on V.

21. Let V be a finite-dimensional inner product space, and let T and U be self-adjoint operators on V such that T is positive definite. Prove that both $\mathsf{T}\mathsf{U}$ and $\mathsf{U}\mathsf{T}$ are diagonalizable linear operators that have only real eigenvalues. *Hint:* Show that $\mathsf{U}\mathsf{T}$ is self-adjoint with respect to the inner product $\langle x, y \rangle' = \langle \mathsf{T}(x), y \rangle$. To show that $\mathsf{T}\mathsf{U}$ is self-adjoint, repeat the argument with T^{-1} in place of T.

22. This exercise provides a converse to Exercise 20. Let V be a finite-dimensional inner product space with inner product $\langle \cdot, \cdot \rangle$, and let $\langle \cdot, \cdot \rangle'$ be any other inner product on V.

(a) Prove that there exists a unique linear operator T on V such that $\langle x, y \rangle' = \langle \mathsf{T}(x), y \rangle$ for all x and y in V. *Hint:* Let $\beta = \{v_1, v_2, \ldots, v_n\}$ be an orthonormal basis for V with respect to $\langle \cdot, \cdot \rangle$, and define a matrix A by $A_{ij} = \langle v_j, v_i \rangle'$ for all i and j. Let T be the unique linear operator on V such that $[\mathsf{T}]_\beta = A$.

(b) Prove that the operator T of (a) is positive definite with respect to both inner products.

23. Let U be a diagonalizable linear operator on a finite-dimensional inner product space V such that all of the eigenvalues of U are real. Prove that there exist positive definite linear operators T_1 and T_1' and self-adjoint linear operators T_2 and T_2' such that $\mathsf{U} = \mathsf{T}_2\mathsf{T}_1 = \mathsf{T}_1'\mathsf{T}_2'$. *Hint:* Let $\langle \cdot, \cdot \rangle$ be the inner product associated with V, β a basis of eigenvectors for U, $\langle \cdot, \cdot \rangle'$ the inner product on V with respect to which β is orthonormal (see Exercise 22(a) of Section 6.1), and T_1 the positive definite operator according to Exercise 22. Show that U is self-adjoint with respect to $\langle \cdot, \cdot \rangle'$ and $\mathsf{U} = \mathsf{T}_1^{-1}\mathsf{U}^*\mathsf{T}_1$ (the adjoint is with respect to $\langle \cdot, \cdot \rangle$). Let $\mathsf{T}_2 = \mathsf{T}_1{}^{-1}\mathsf{U}^*$.

6.5 UNITARY AND ORTHOGONAL OPERATORS AND THEIR MATRICES

In this section, we continue our analogy between complex numbers and linear operators. Recall that the adjoint of a linear operator acts similarly to the conjugate of a complex number (see, for example, Theorem 6.11 p. 357). A complex number z has length 1 if $z\overline{z} = 1$. In this section, we study those linear operators T on an inner product space V such that $\mathsf{TT}^* = \mathsf{T}^*\mathsf{T} = \mathsf{I}$. We will see that these are precisely the linear operators that "preserve length" in the sense that $\|\mathsf{T}(x)\| = \|x\|$ for all $x \in \mathsf{V}$. As another characterization, we prove that, on a finite-dimensional complex inner product space, these are the normal operators whose eigenvalues all have absolute value 1.

In past chapters, we were interested in studying those functions that preserve the structure of the underlying space. In particular, linear operators preserve the operations of vector addition and scalar multiplication, and isomorphisms preserve all the vector space structure. It is now natural to consider those linear operators T on an inner product space that preserve length. We will see that this condition guarantees, in fact, that T preserves the inner product.

Definitions. *Let T be a linear operator on a finite-dimensional inner product space V (over F). If $\|\mathsf{T}(x)\| = \|x\|$ for all $x \in \mathsf{V}$, we call T a* ***unitary operator*** *if $F = C$ and an* ***orthogonal operator*** *if $F = R$.*

It should be noted that in the infinite-dimensional case, an operator that preserves the norm is one-to-one, but not necessarily onto. If it is also onto, then we call it a **unitary** or **orthogonal operator**.

Clearly, any rotation or reflection in R^2 preserves length and hence is an orthogonal operator. We study these operators in much more detail in Section 6.11.

Example 1

Recall the inner product space H defined on page 330. Let $h \in \mathsf{H}$ satisfy $|h(x)| = 1$ for all x. Define the linear operator T on H by $\mathsf{T}(f) = hf$. Then

$$\|\mathsf{T}(f)\|^2 = \|hf\|^2 = \frac{1}{2\pi}\int_0^{2\pi} h(t)f(t)\overline{h(t)f(t)}\,dt = \|f\|^2$$

since $|h(t)|^2 = 1$ for all t. So T is a unitary operator. ♦

Theorem 6.18. *Let T be a linear operator on a finite-dimensional inner product space V. Then the following statements are equivalent.*

(a) $\mathsf{T}^*\mathsf{T} = \mathsf{I}$.
(b) $\mathsf{TT}^* = \mathsf{I}$.
(c) $\langle \mathsf{T}(x), \mathsf{T}(y)\rangle = \langle x, y\rangle$ *for all* $x, y \in \mathsf{V}$.

(d) *If β is an orthonormal basis for V, then $\mathsf{T}(\beta)$ is an orthonormal basis for V.*
(e) *There exists an orthonormal basis β for V such that $\mathsf{T}(\beta)$ is an orthonormal basis for V.*
(f) $\|\mathsf{T}(x)\| = \|x\|$ *for all* $x \in \mathsf{V}$.

Thus all the conditions above are equivalent to the definition of a unitary or orthogonal operator. From (a) and (b), it follows that unitary or orthogonal operators are normal.

Before proving the theorem, we first prove a lemma. Compare this lemma to Exercise 11(b) of Section 6.4.

Lemma. *Let U be a self-adjoint operator on an inner product space V, and suppose that $\langle x, \mathsf{U}(x)\rangle = 0$ for all $x \in \mathsf{V}$. Then $\mathsf{U} = \mathsf{T}_0$.*

Proof. For any $x \in \mathsf{V}$,

$$\begin{aligned}
0 &= \langle x + \mathsf{U}(x), \mathsf{U}(x + \mathsf{U}(x))\rangle \\
&= \langle x + \mathsf{U}(x), \mathsf{U}(x) + \mathsf{U}^2(x)\rangle \\
&= \langle x, \mathsf{U}(x)\rangle + \langle x, \mathsf{U}^2(x)\rangle + \langle \mathsf{U}(x), \mathsf{U}(x)\rangle + \langle \mathsf{U}(x), \mathsf{U}^2(x)\rangle \\
&= 0 + \langle x, \mathsf{U}^2(x)\rangle + \langle \mathsf{U}(x), \mathsf{U}(x)\rangle + 0 \\
&= \langle x, \mathsf{U}^*\mathsf{U}(x)\rangle + \|\mathsf{U}(x)\|^2 \\
&= 2\|\mathsf{U}(x)\|^2.
\end{aligned}$$

So for any $x \in \mathsf{V}$, $\|\mathsf{U}(x)\| = 0$. It follows that $\mathsf{U} = \mathsf{T}_0$. ∎

Proof of Theorem 6.18. Part (a) implies (b) by Theorem 6.10 and Exercise 10(c) of Section 2.4.

To prove that (b) implies (c), let $x, y \in \mathsf{V}$. Then $\langle x, y\rangle = \langle \mathsf{T}^*\mathsf{T}(x), y\rangle = \langle \mathsf{T}(x), \mathsf{T}(y)\rangle$.

Next, we prove that (c) implies (d). Let $\beta = \{v_1, v_2, \ldots, v_n\}$ be an orthonormal basis for V; so $\mathsf{T}(\beta) = \{\mathsf{T}(v_1), \mathsf{T}(v_2), \ldots, \mathsf{T}(v_n)\}$. It follows that $\langle \mathsf{T}(v_i), \mathsf{T}(v_j)\rangle = \langle v_i, v_j\rangle = \delta_{ij}$. Therefore $\mathsf{T}(\beta)$ is an orthonormal basis for V.

That (d) implies (e) is obvious.

Next we prove that (e) implies (f). Let $x \in \mathsf{V}$, and let $\beta = \{v_1, v_2, \ldots, v_n\}$. Now

$$x = \sum_{i=1}^{n} a_i v_i$$

for some scalars a_i, and so

$$\|x\|^2 = \left\langle \sum_{i=1}^{n} a_i v_i, \sum_{j=1}^{n} a_j v_j \right\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \overline{a}_j \langle v_i, v_j\rangle$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{a}_j \delta_{ij} = \sum_{i=1}^{n} |a_i|^2$$

since β is orthonormal.

Applying the same manipulations to

$$\mathsf{T}(x) = \sum_{i=1}^{n} a_i \mathsf{T}(v_i)$$

and using the fact that $\mathsf{T}(\beta)$ is also orthonormal, we obtain

$$\|\mathsf{T}(x)\|^2 = \sum_{i=1}^{n} |a_i|^2.$$

Hence $\|\mathsf{T}(x)\| = \|x\|$.

Finally, we prove that (f) implies (a). For any $x \in \mathsf{V}$, we have

$$\langle x, x \rangle = \|x\|^2 = \|\mathsf{T}(x)\|^2 = \langle \mathsf{T}(x), \mathsf{T}(x) \rangle = \langle x, \mathsf{T}^*\mathsf{T}(x) \rangle .$$

So $\langle x, (\mathsf{I} - \mathsf{T}^*\mathsf{T})(x) \rangle = 0$ for all $x \in \mathsf{V}$. Let $\mathsf{U} = \mathsf{I} - \mathsf{T}^*\mathsf{T}$; then U is self-adjoint, and $\langle x, \mathsf{U}(x) \rangle = 0$ for all $x \in \mathsf{V}$. Hence, by the lemma, we have $\mathsf{T}_0 = \mathsf{U} = \mathsf{I} - \mathsf{T}^*\mathsf{T}$, and therefore $\mathsf{T}^*\mathsf{T} = \mathsf{I}$. ∎

In the case that T satisfies (c), we say that T *preserves inner products.* In the case that T satisfies (f), we say that T *preserves norms.*

It follows immediately from the definition that every eigenvalue of a unitary or orthogonal operator has absolute value 1. In fact, even more is true.

Corollary 1. *Let T be a linear operator on a finite-dimensional real inner product space V. Then V has an orthonormal basis of eigenvectors of T with corresponding eigenvalues of absolute value 1 if and only if T is both self-adjoint and orthogonal.*

Proof. Suppose that V has an orthonormal basis $\{v_1, v_2, \ldots, v_n\}$ such that $\mathsf{T}(v_i) = \lambda_i v_i$ and $|\lambda_i| = 1$ for all i. By Theorem 6.17 (p. 371), T is self-adjoint. Thus $(\mathsf{T}\mathsf{T}^*)(v_i) = \mathsf{T}(\lambda_i v_i) = \lambda_i \lambda_i v_i = \lambda_i^2 v_i = v_i$ for each i. So $\mathsf{T}\mathsf{T}^* = \mathsf{I}$, and again by Exercise 10 of Section 2.4, T is orthogonal by Theorem 6.18(a).

If T is self-adjoint, then, by Theorem 6.17, we have that V possesses an orthonormal basis $\{v_1, v_2, \ldots, v_n\}$ such that $\mathsf{T}(v_i) = \lambda_i v_i$ for all i. If T is also orthogonal, we have

$$|\lambda_i| \cdot \|v_i\| = \|\lambda_i v_i\| = \|\mathsf{T}(v_i)\| = \|v_i\|;$$

so $|\lambda_i| = 1$ for every i. ∎

Corollary 2. *Let* T *be a linear operator on a finite-dimensional complex inner product space* V. *Then* V *has an orthonormal basis of eigenvectors of* T *with corresponding eigenvalues of absolute value 1 if and only if* T *is unitary.*

Proof. The proof is similar to the proof of Corollary 1. ∎

Example 2

Let $T: R^2 \to R^2$ be a rotation by θ, where $0 < \theta < \pi$. It is clear geometrically that T "preserves length," that is, that $\|T(x)\| = \|x\|$ for all $x \in R^2$. The fact that rotations by a fixed angle preserve perpendicularity not only can be seen geometrically but now follows from (b) of Theorem 6.18. Perhaps the fact that such a transformation preserves the inner product is not so obvious; however, we obtain this fact from (b) also. Finally, an inspection of the matrix representation of T with respect to the standard ordered basis, which is

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

reveals that T is not self-adjoint for the given restriction on θ. As we mentioned earlier, this fact also follows from the geometric observation that T has no eigenvectors and from Theorem 6.15 (p. 368). It is seen easily from the preceding matrix that T^* is the rotation by $-\theta$. ♦

Definition. *Let* L *be a one-dimensional subspace of* R^2. *We may view* L *as a line in the plane through the origin. A linear operator* T *on* R^2 *is called a* **reflection of** R^2 **about** L *if* $T(x) = x$ *for all* $x \in L$ *and* $T(x) = -x$ *for all* $x \in L^\perp$.

As an example of a reflection, consider the operator defined in Example 3 of Section 2.5.

Example 3

Let T be a reflection of R^2 about a line L through the origin. We show that T is an orthogonal operator. Select vectors $v_1 \in L$ and $v_2 \in L^\perp$ such that $\|v_1\| = \|v_2\| = 1$. Then $T(v_1) = v_1$ and $T(v_2) = -v_2$. Thus v_1 and v_2 are eigenvectors of T with corresponding eigenvalues 1 and -1, respectively. Furthermore, $\{v_1, v_2\}$ is an orthonormal basis for R^2. It follows that T is an orthogonal operator by Corollary 1 to Theorem 6.18. ♦

We now examine the matrices that represent unitary and orthogonal transformations.

Definitions. *A square matrix* A *is called an* ***orthogonal matrix*** *if* $A^t A = AA^t = I$ *and* ***unitary*** *if* $A^* A = AA^* = I$.

Since for a real matrix A we have $A^* = A^t$, a real unitary matrix is also orthogonal. In this case, we call A **orthogonal** rather than unitary.

Note that the condition $AA^* = I$ is equivalent to the statement that the rows of A form an orthonormal basis for F^n because

$$\delta_{ij} = I_{ij} = (AA^*)_{ij} = \sum_{k=1}^{n} A_{ik}(A^*)_{kj} = \sum_{k=1}^{n} A_{ik}\overline{A}_{jk},$$

and the last term represents the inner product of the ith and jth rows of A.

A similar remark can be made about the columns of A and the condition $A^*A = I$.

It also follows from the definition above and from Theorem 6.10 (p. 356) that a linear operator T on an inner product space V is unitary [orthogonal] if and only if $[\mathsf{T}]_\beta$ is unitary [orthogonal] for some orthonormal basis β for V.

Example 4

From Example 2, the matrix

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is clearly orthogonal. One can easily see that the rows of the matrix form an orthonormal basis for R^2. Similarly, the columns of the matrix form an orthonormal basis for R^2. ◆

Example 5

Let T be a reflection of R^2 about a line L through the origin, let β be the standard ordered basis for R^2, and let $A = [\mathsf{T}]_\beta$. Then $\mathsf{T} = \mathsf{L}_A$. Since T is an orthogonal operator and β is an orthonormal basis, A is an orthogonal matrix. We describe A.

Suppose that α is the angle from the positive x-axis to L. Let $v_1 = (\cos\alpha, \sin\alpha)$ and $v_2 = (-\sin\alpha, \cos\alpha)$. Then $\|v_1\| = \|v_2\| = 1$, $v_1 \in \mathsf{L}$, and $v_2 \in \mathsf{L}^\perp$. Hence $\gamma = \{v_1, v_2\}$ is an orthonormal basis for R^2. Because $\mathsf{T}(v_1) = v_1$ and $\mathsf{T}(v_2) = -v_2$, we have

$$[\mathsf{T}]_\gamma = [\mathsf{L}_A]_\gamma = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Let

$$Q = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

By the corollary to Theorem 2.23 (p. 115),

$$A = Q[\mathsf{L}_A]_\gamma Q^{-1}$$

$$= \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

$$= \begin{pmatrix} \cos^2\alpha - \sin^2\alpha & 2\sin\alpha\cos\alpha \\ 2\sin\alpha\cos\alpha & -(\cos^2\alpha - \sin^2\alpha) \end{pmatrix}$$

$$= \begin{pmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{pmatrix}. \quad \blacklozenge$$

We know that, for a complex normal [real symmetric] matrix A, there exists an orthonormal basis β for F^n consisting of eigenvectors of A. Hence A is similar to a diagonal matrix D. By the corollary to Theorem 2.23 (p. 115), the matrix Q whose columns are the vectors in β is such that $D = Q^{-1}AQ$. But since the columns of Q are an orthonormal basis for F^n, it follows that Q is unitary [orthogonal]. In this case, we say that A is **unitarily equivalent** [**orthogonally equivalent**] to D. It is easily seen (see Exercise 18) that this relation is an equivalence relation on $\mathsf{M}_{n\times n}(C)$ [$\mathsf{M}_{n\times n}(R)$]. More generally, *$A$ and B are unitarily equivalent [orthogonally equivalent] if and only if there exists a unitary [orthogonal] matrix P such that $A = P^*BP$.*

The preceding paragraph has proved half of each of the next two theorems.

Theorem 6.19. *Let A be a complex $n \times n$ matrix. Then A is normal if and only if A is unitarily equivalent to a diagonal matrix.*

Proof. By the preceding remarks, we need only prove that if A is unitarily equivalent to a diagonal matrix, then A is normal.

Suppose that $A = P^*DP$, where P is a unitary matrix and D is a diagonal matrix. Then

$$AA^* = (P^*DP)(P^*DP)^* = (P^*DP)(P^*D^*P) = P^*DID^*P = P^*DD^*P.$$

Similarly, $A^*A = P^*D^*DP$. Since D is a diagonal matrix, however, we have $DD^* = D^*D$. Thus $AA^* = A^*A$. ∎

Theorem 6.20. *Let A be a real $n \times n$ matrix. Then A is symmetric if and only if A is orthogonally equivalent to a real diagonal matrix.*

Proof. The proof is similar to the proof of Theorem 6.19 and is left as an exercise. ∎

Theorem 6.20 is used extensively in many areas of mathematics and statistics. See, for example, `goo.gl/cbqApK`.

Example 6

Let

$$A = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

Since A is symmetric, Theorem 6.20 tells us that A is orthogonally equivalent to a diagonal matrix. We find an orthogonal matrix P and a diagonal matrix D such that $P^tAP = D$.

To find P, we obtain an orthonormal basis of eigenvectors. It is easy to show that the eigenvalues of A are 2 and 8. The set $\{(-1,1,0),(-1,0,1)\}$ is a basis for the eigenspace corresponding to 2. Because this set is not orthogonal, we apply the Gram–Schmidt process to obtain the orthogonal set $\{(-1,1,0),-\frac{1}{2}(1,1,-2)\}$. The set $\{(1,1,1)\}$ is a basis for the eigenspace corresponding to 8. Notice that $(1,1,1)$ is orthogonal to the preceding two vectors, as predicted by Theorem 6.15(d) (p. 368). Taking the union of these two bases and normalizing the vectors, we obtain the following orthonormal basis for R^3 consisting of eigenvectors of A:

$$\left\{\frac{1}{\sqrt{2}}(-1,1,0),\frac{1}{\sqrt{6}}(1,1,-2),\frac{1}{\sqrt{3}}(1,1,1)\right\}.$$

Thus one possible choice for P is

$$P = \begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}, \quad \text{and} \quad D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 8 \end{pmatrix}. \quad \blacklozenge$$

Because of Schur's theorem (Theorem 6.14 p. 367), the next result is immediate. As it is the matrix form of Schur's theorem, we also refer to it as Schur's theorem.

Theorem 6.21 (Schur). *Let $A \in \mathsf{M}_{n\times n}(F)$ be a matrix whose characteristic polynomial splits over F.*

(a) *If $F = C$, then A is unitarily equivalent to a complex upper triangular matrix.*

(b) *If $F = R$, then A is orthogonally equivalent to a real upper triangular matrix.*

Rigid Motions*

The purpose of this application is to characterize the so-called *rigid motions* of a finite-dimensional real inner product space. One may think intuitively of such a motion as a transformation that does not affect the shape of

a figure under its action, hence the term *rigid*. The key requirement for such a transformation is that it preserves distances.

Definition. *Let* V *be a real inner product space. A function* $f: \mathsf{V} \to \mathsf{V}$ *is called a* ***rigid motion*** *if*

$$\|f(x) - f(y)\| = \|x - y\|$$

for all $x, y \in \mathsf{V}$.

For example, any orthogonal operator on a finite-dimensional real inner product space is a rigid motion.

Another class of rigid motions are the *translations*. A function $g: \mathsf{V} \to \mathsf{V}$, where V is a real inner product space, is called a **translation** if there exists a vector $v_0 \in \mathsf{V}$ such that $g(x) = x + v_0$ for all $x \in \mathsf{V}$. We say that g is the *translation by* v_0. It is a simple exercise to show that translations, as well as composites of rigid motions on a real inner product space, are also rigid motions. (See Exercise 22.) Thus an orthogonal operator on a finite-dimensional real inner product space V followed by a translation on V is a rigid motion on V. Remarkably, every rigid motion on V may be characterized in this way.

Theorem 6.22. *Let* $f: \mathsf{V} \to \mathsf{V}$ *be a rigid motion on a finite-dimensional real inner product space* V. *Then there exists a unique orthogonal operator* T *on* V *and a unique translation* g *on* V *such that* $f = g \circ \mathsf{T}$.

Any orthogonal operator is a special case of this composite, in which the translation is by 0. Any translation is also a special case, in which the orthogonal operator is the identity operator.

Proof. Let $\mathsf{T}: \mathsf{V} \to \mathsf{V}$ be defined by

$$\mathsf{T}(x) = f(x) - f(0)$$

for all $x \in \mathsf{V}$. Note that $f = g \circ \mathsf{T}$, where g is the translation by $f(0)$. Moreover, T is the composite of f and the translation by $-f(0)$; hence T is a rigid motion. We begin by showing that T is a linear operator. For any $x \in \mathsf{V}$, we have

$$\|\mathsf{T}(x)\|^2 = \|f(x) - f(0)\|^2 = \|x - 0\|^2 = \|x\|^2,$$

and consequently $\|\mathsf{T}(x)\| = \|x\|$ for any $x \in \mathsf{V}$. Thus for any $x, y \in \mathsf{V}$,

$$\begin{aligned} \|\mathsf{T}(x) - \mathsf{T}(y)\|^2 &= \|\mathsf{T}(x)\|^2 - 2\langle \mathsf{T}(x), \mathsf{T}(y)\rangle + \|\mathsf{T}(y)\|^2 \\ &= \|x\|^2 - 2\langle \mathsf{T}(x), \mathsf{T}(y)\rangle + \|y\|^2 \end{aligned}$$

and

$$\|x - y\|^2 = \|x\|^2 - 2\langle x, y\rangle + \|y\|^2.$$

But $\|\mathsf{T}(x) - \mathsf{T}(y)\|^2 = \|x - y\|^2$; so $\langle \mathsf{T}(x), \mathsf{T}(y)\rangle = \langle x, y\rangle$ for all $x, y \in \mathsf{V}$.

We are now in a position to show that T is a linear transformation. Let $x, y \in \mathsf{V}$, and let $a \in R$. Then

$$\begin{aligned}
\|\mathsf{T}(x+ay) - \mathsf{T}(x) - a\mathsf{T}(y)\|^2 &= \|[\mathsf{T}(x+ay) - \mathsf{T}(x)] - a\mathsf{T}(y)\|^2 \\
&= \|\mathsf{T}(x+ay) - \mathsf{T}(x)\|^2 + a^2\|\mathsf{T}(y)\|^2 - 2a\langle \mathsf{T}(x+ay) - \mathsf{T}(x), \mathsf{T}(y)\rangle \\
&= \|(x+ay) - x\|^2 + a^2\|y\|^2 - 2a[\langle \mathsf{T}(x+ay), \mathsf{T}(y)\rangle - \langle \mathsf{T}(x), \mathsf{T}(y)\rangle] \\
&= a^2\|y\|^2 + a^2\|y\|^2 - 2a[\langle x+ay, y\rangle - \langle x, y\rangle] \\
&= 2a^2\|y\|^2 - 2a[\langle x, y\rangle + a\|y\|^2 - \langle x, y\rangle] \\
&= 0.
\end{aligned}$$

Thus $\mathsf{T}(x+ay) = \mathsf{T}(x) + a\mathsf{T}(y)$, and hence T is linear. Since we have already shown that T preserves inner products, T is an orthogonal operator.

To prove uniqueness, suppose that u_0 and v_0 are in V and T and U are orthogonal operators on V such that

$$f(x) = \mathsf{T}(x) + u_0 = \mathsf{U}(x) + v_0$$

for all $x \in \mathsf{V}$. Substituting $x = 0$ in the preceding equation yields $u_0 = v_0$, and hence the translation is unique. This equation, therefore, reduces to $\mathsf{T}(x) = \mathsf{U}(x)$ for all $x \in \mathsf{V}$, and hence $\mathsf{T} = \mathsf{U}$. ∎

Orthogonal Operators on R^2

Because of Theorem 6.22, an understanding of rigid motions requires a characterization of orthogonal operators. The next result characterizes orthogonal operators on R^2. We postpone the case of orthogonal operators on more general spaces to Section 6.11.

Theorem 6.23. *Let T be an orthogonal operator on R^2, and let $A = [\mathsf{T}]_\beta$, where β is the standard ordered basis for R^2. Then exactly one of the following conditions is satisfied:*

(a) *T is a rotation, and $\det(A) = 1$.*
(b) *T is a reflection about a line through the origin, and $\det(A) = -1$.*

Proof. Because T is an orthogonal operator, $\mathsf{T}(\beta) = \{\mathsf{T}(e_1), \mathsf{T}(e_2)\}$ is an orthonormal basis for R^2 by Theorem 6.18(c). Since $\mathsf{T}(e_1)$ is a unit vector, there is a unique angle θ, $0 \le \theta < 2\pi$, such that $\mathsf{T}(e_1) = (\cos\theta, \sin\theta)$. Since $\mathsf{T}(e_2)$ is a unit vector and is orthogonal to $\mathsf{T}(e_1)$, there are only two possible choices for $\mathsf{T}(e_2)$. Either

$$\mathsf{T}(e_2) = (-\sin\theta, \cos\theta) \quad \text{or} \quad \mathsf{T}(e_2) = (\sin\theta, -\cos\theta).$$

First, suppose that $\mathsf{T}(e_2) = (-\sin\theta, \cos\theta)$. Then $A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

It follows from Example 1 of Section 6.4 that T is a rotation by the angle θ. Also

$$\det(A) = \cos^2\theta + \sin^2\theta = 1.$$

Now suppose that $\mathsf{T}(e_2) = (\sin\theta, -\cos\theta)$. Then $A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$. Comparing this matrix to the matrix A of Example 5, we see that T is the reflection of R^2 about a line L such that $\alpha = \theta/2$ is the angle from the positive x-axis to L. Furthermore,

$$\det(A) = -\cos^2\theta - \sin^2\theta = -1. \quad \blacksquare$$

Combining Theorems 6.22 and 6.23, we obtain the following characterization of rigid motions on R^2.

Corollary. *Any rigid motion on R^2 is either a rotation followed by a translation or a reflection about a line through the origin followed by a translation.*

Example 7

Let

$$A = \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \end{pmatrix}.$$

We show that L_A is the reflection of R^2 about a line L through the origin, and then describe L.

Clearly $AA^* = A^*A = I$, and therefore A is an orthogonal matrix. Hence L_A is an orthogonal operator. Furthermore,

$$\det(A) = -\frac{1}{5} - \frac{4}{5} = -1,$$

and thus L_A is a reflection of R^2 about a line L through the origin by Theorem 6.23. Since L is the one-dimensional eigenspace corresponding to the eigenvalue 1 of L_A, it suffices to find an eigenvector of L_A corresponding to 1. One such vector is $v = (2, \sqrt{5} - 1)$. Thus L is the span of $\{v\}$. Alternatively, L is the line through the origin with slope $(\sqrt{5} - 1)/2$, and hence is the line with the equation

$$y = \frac{\sqrt{5} - 1}{2} x. \quad \blacklozenge$$

Conic Sections

As an application of Theorem 6.20, we consider the quadratic equation

$$ax^2 + 2bxy + cy^2 + dx + ey + f = 0. \tag{2}$$

For special choices of the coefficients in (2), we obtain the various conic sections. For example, if $a = c = 1$, $b = d = e = 0$, and $f = -1$, we obtain the circle $x^2 + y^2 = 1$ with center at the origin. The remaining conic sections, namely, the ellipse, parabola, and hyperbola, are obtained by other choices of the coefficients. If $b = 0$, then it is easy to graph the equation by the method of completing the square because the xy-term is absent. For example, the equation $x^2+2x+y^2+4y+2 = 0$ may be rewritten as $(x+1)^2+(y+2)^2 = 3$, which describes a circle with radius $\sqrt{3}$ and center at $(-1,-2)$ in the xy-coordinate system. If we consider the transformation of coordinates $(x,y) \to (x',y')$, where $x' = x+1$ and $y' = y+2$, then our equation simplifies to $(x')^2 + (y')^2 = 3$. This change of variable allows us to eliminate the x- and y-terms.

We now concentrate solely on the elimination of the xy-term. To accomplish this, we consider the expression

$$ax^2 + 2bxy + cy^2, \tag{3}$$

which is called the **associated quadratic form** of (2). Quadratic forms are studied in more generality in Section 6.8.

If we let

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} x \\ y \end{pmatrix},$$

then (3) may be written as $X^tAX = \langle AX, X\rangle$. For example, the quadratic form $3x^2 + 4xy + 6y^2$ may be written as

$$X^t \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} X.$$

The fact that A is symmetric is crucial in our discussion. For, by Theorem 6.20, we may choose an orthogonal matrix P and a diagonal matrix D with real diagonal entries λ_1 and λ_2 such that $P^tAP = D$. Now define

$$X' = \begin{pmatrix} x' \\ y' \end{pmatrix}$$

by $X' = P^tX$ or, equivalently, by $PX' = PP^tX = X$. Then

$$X^tAX = (PX')^tA(PX') = X'^t(P^tAP)X' = X'^tDX' = \lambda_1(x')^2 + \lambda_2(y')^2.$$

Thus the transformation $(x,y) \to (x',y')$ allows us to eliminate the xy-term in (3), and hence in (2).

Furthermore, since P is orthogonal, we have by Theorem 6.23 (with $\mathsf{T} = \mathsf{L}_P$) that $\det(P) = \pm 1$. If $\det(P) = -1$, we may interchange the columns of P to obtain a matrix Q. Because the columns of P form an orthonormal basis of eigenvectors of A, the same is true of the columns of Q. Therefore

$$Q^tAQ = \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{pmatrix}.$$

Notice that $\det(Q) = -\det(P) = 1$. So, if $\det(P) = -1$, we can take Q for our new P; consequently, we may always choose P so that $\det(P) = 1$. By Lemma 4 to Theorem 6.22 (with $\mathsf{T} = \mathsf{L}_P$), it follows that matrix P represents a rotation.

In summary, the xy-term in (2) may be eliminated by a rotation of the x-axis and y-axis to new axes x' and y' given by $X = PX'$, where P is an orthogonal matrix and $\det(P) = 1$. Furthermore, the coefficients of $(x')^2$ and $(y')^2$ are the eigenvalues of

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

This result is a restatement of a result known as the *principal axis theorem* for R^2. The arguments above, of course, are easily extended to quadratic equations in n variables. For example, in the case $n = 3$, by special choices of the coefficients, we obtain the quadratic surfaces—the elliptic cone, the ellipsoid, the hyperbolic paraboloid, etc.

As an illustration of the preceding transformation, consider the quadratic equation

$$2x^2 - 4xy + 5y^2 - 36 = 0,$$

for which the associated quadratic form is $2x^2 - 4xy + 5y^2$. In the notation we have been using,

$$A = \begin{pmatrix} 2 & -2 \\ -2 & 5 \end{pmatrix},$$

so that the eigenvalues of A are 1 and 6 with associated eigenvectors

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

As expected (from Theorem 6.15(d) p. 368), these vectors are orthogonal. The corresponding orthonormal basis of eigenvectors

$$\beta = \left\{ \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}, \begin{pmatrix} \frac{-1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix} \right\}$$

determines new axes x' and y' as in Figure 6.4. Hence if

$$P = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix},$$

then

$$P^tAP = \begin{pmatrix} 1 & 0 \\ 0 & 6 \end{pmatrix}.$$

Under the transformation $X = PX'$ or

$$x = \frac{2}{\sqrt{5}}x' - \frac{1}{\sqrt{5}}y'$$
$$y = \frac{1}{\sqrt{5}}x' + \frac{2}{\sqrt{5}}y',$$

we have the new quadratic form $(x')^2 + 6(y')^2$. Thus the original equation $2x^2 - 4xy + 5y^2 = 36$ may be written in the form $(x')^2 + 6(y')^2 = 36$ relative to a new coordinate system with the x'- and y'-axes in the directions of the first and second vectors of β, respectively. It is clear that this equation represents

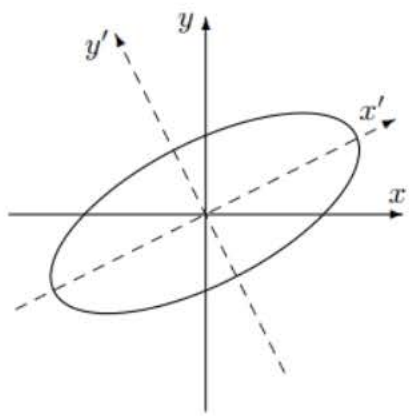


Figure 6.4

an ellipse. (See Figure 6.4.) Note that the preceding matrix P has the form

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

where $\theta = \cos^{-1}\dfrac{2}{\sqrt{5}} \approx 26.6°$. So P is the matrix representation of a rotation of R^2 through the angle θ. Thus the change of variable $X = PX'$ can be accomplished by this rotation of the x- and y-axes. There is another possibility for P, however. If the eigenvector of A corresponding to the eigenvalue 6 is taken to be $(1, -2)$ instead of $(-1, 2)$, and the eigenvalues are interchanged, then we obtain the matrix

$$\begin{pmatrix} \dfrac{1}{\sqrt{5}} & \dfrac{2}{\sqrt{5}} \\ \dfrac{-2}{\sqrt{5}} & \dfrac{1}{\sqrt{5}} \end{pmatrix},$$

which is the matrix representation of a rotation through the angle $\theta = \sin^{-1}\left(-\frac{2}{\sqrt{5}}\right) \approx -63.4°$. This possibility produces the same ellipse as the one in Figure 6.4, but interchanges the names of the x'- and y'-axes.

EXERCISES

1. Label the following statements as true or false. Assume that the underlying inner product spaces are finite-dimensional.

- **(a)** Every unitary operator is normal.
- **(b)** Every orthogonal operator is diagonalizable.
- **(c)** A matrix is unitary if and only if it is invertible.
- **(d)** If two matrices are unitarily equivalent, then they are also similar.
- **(e)** The sum of unitary matrices is unitary.
- **(f)** The adjoint of a unitary operator is unitary.
- **(g)** If T is an orthogonal operator on V, then $[\mathsf{T}]_\beta$ is an orthogonal matrix for any ordered basis β for V.
- **(h)** If all the eigenvalues of a linear operator are 1, then the operator must be unitary or orthogonal.
- **(i)** A linear operator may preserve norms without preserving inner products.

2. For each of the following matrices A, find an orthogonal or unitary matrix P and a diagonal matrix D such that $P^*AP = D$.

(a) $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ **(b)** $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ **(c)** $\begin{pmatrix} 2 & 3-3i \\ 3+3i & 5 \end{pmatrix}$

(d) $\begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}$ **(e)** $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$

3. Prove that the composite of unitary [orthogonal] operators is unitary [orthogonal].

4. For $z \in C$, define $\mathsf{T}_z : C \to C$ by $\mathsf{T}_z(u) = zu$. Characterize those z for which T_z is normal, self-adjoint, or unitary.

5. Which of the following pairs of matrices are unitarily equivalent?

(a) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ **(b)** $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$

(c) $\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

(d) $\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 & 0 \\ 0 & i & 0 \\ 0 & 0 & -i \end{pmatrix}$

(e) $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$

6. Let V be the inner product space of complex-valued continuous functions on $[0, 1]$ with the inner product

$$\langle f, g \rangle = \int_0^1 f(t)\overline{g(t)}\, dt.$$

Let $h \in \mathsf{V}$, and define $\mathsf{T}: \mathsf{V} \to \mathsf{V}$ by $\mathsf{T}(f) = hf$. Prove that T is a unitary operator if and only if $|h(t)| = 1$ for $0 \le t \le 1$.

red *Hint for the "only if" part:* Suppose that T is unitary. Set $f(t) = 1 - |h(t)|^2$ and $g(t) = 1$. Show that

$$\int_0^1 (1 - |h(t)|^2)^2\, dt = 0,$$

and use the fact that if the integral of a nonnegative continuous function is zero, then the function is identically zero.

7. Prove that if T is a unitary operator on a finite-dimensional inner product space V, then T has a unitary *square root*; that is, there exists a unitary operator U such that $\mathsf{T} = \mathsf{U}^2$. Visit `goo.gl/jADTaS` for a solution.

8. Let T be a self-adjoint linear operator on a finite-dimensional inner product space. Prove that $(\mathsf{T}+i\mathsf{I})(\mathsf{T}-i\mathsf{I})^{-1}$ is unitary using Exercise 10 of Section 6.4.

9. Let U be a linear operator on a finite-dimensional inner product space V. If $\|\mathsf{U}(x)\| = \|x\|$ for all x in some orthonormal basis for V, must U be unitary? Justify your answer with a proof or a counterexample.

10. Let A be an $n \times n$ real symmetric or complex normal matrix. Prove that

$$\operatorname{tr}(A) = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad \operatorname{tr}(A^* A) = \sum_{i=1}^{n} |\lambda_i|^2,$$

where the λ_i's are the (not necessarily distinct) eigenvalues of A.

11. Find an orthogonal matrix whose first row is $(\frac{1}{3}, \frac{2}{3}, \frac{2}{3})$.

12. Let A be an $n \times n$ real symmetric or complex normal matrix. Prove that

$$\det(A) = \prod_{i=1}^{n} \lambda_i,$$

where the λ_i's are the (not necessarily distinct) eigenvalues of A.

13. Suppose that A and B are diagonalizable matrices. Prove or disprove that A is similar to B if and only if A and B are unitarily equivalent.

14. Prove that if A and B are unitarily equivalent matrices, then A is positive definite [semidefinite] if and only if B is positive definite [semidefinite]. (See the definitions in the exercises in Section 6.4.)

15. Let U be a unitary operator on an inner product space V, and let W be a finite-dimensional U-invariant subspace of V. Prove that

(a) $\mathsf{U}(\mathsf{W}) = \mathsf{W}$;
(b) $\mathsf{W}^\perp$ is U-invariant.

Contrast (b) with Exercise 16.

16. Find an example of a unitary operator U on an inner product space and a U-invariant subspace W such that $\mathsf{W}^\perp$ is not U-invariant.

17. Prove that a matrix that is both unitary and upper triangular must be a diagonal matrix.

18. Show that "is unitarily equivalent to" is an equivalence relation on $\mathsf{M}_{n\times n}(C)$.

19. Let W be a finite-dimensional subspace of an inner product space V. By Theorem 6.7 (p. 349) and the exercises of Section 1.3, $\mathsf{V} = \mathsf{W} \oplus \mathsf{W}^\perp$. Define $\mathsf{U}\colon \mathsf{V} \to \mathsf{V}$ by $\mathsf{U}(v_1 + v_2) = v_1 - v_2$, where $v_1 \in \mathsf{W}$ and $v_2 \in \mathsf{W}^\perp$. Prove that U is a self-adjoint unitary operator.

20. Let V be a finite-dimensional inner product space. A linear operator U on V is called a **partial isometry** if there exists a subspace W of V such that $\|\mathsf{U}(x)\| = \|x\|$ for all $x \in \mathsf{W}$ and $\mathsf{U}(x) = 0$ for all $x \in \mathsf{W}^\perp$. Observe that W need *not* be U-invariant. Suppose that U is such an operator and $\{v_1, v_2, \ldots, v_k\}$ is an orthonormal basis for W. Prove the following results.

(a) $\langle \mathsf{U}(x), \mathsf{U}(y) \rangle = \langle x, y \rangle$ for all $x, y \in \mathsf{W}$. *Hint:* Use Exercise 20 of Section 6.1.
(b) $\{\mathsf{U}(v_1), \mathsf{U}(v_2), \ldots, \mathsf{U}(v_k)\}$ is an orthonormal basis for $\mathsf{R}(\mathsf{U})$.

(c) There exists an orthonormal basis γ for V such that the first k columns of $[\mathsf{U}]_\gamma$ form an orthonormal set and the remaining columns are zero.

(d) Let $\{w_1, w_2, \ldots, w_j\}$ be an orthonormal basis for $\mathsf{R}(\mathsf{U})^\perp$ and $\beta = \{\mathsf{U}(v_1), \mathsf{U}(v_2), \ldots, \mathsf{U}(v_k), w_1, \ldots, w_j\}$. Then β is an orthonormal basis for V.

(e) Let T be the linear operator on V that satisfies $\mathsf{T}(\mathsf{U}(v_i)) = v_i$ $(1 \le i \le k)$ and $\mathsf{T}(w_i) = 0$ $(1 \le i \le j)$. Then T is well defined, and $\mathsf{T} = \mathsf{U}^*$. *Hint:* Show that $\langle \mathsf{U}(x), y\rangle = \langle x, \mathsf{T}(y)\rangle$ for all $x, y \in \beta$. There are four cases.

(f) U^* is a partial isometry.

This exercise is continued in Exercise 9 of Section 6.6.

21. Let A and B be $n \times n$ matrices that are unitarily equivalent.

(a) Prove that $\operatorname{tr}(A^*A) = \operatorname{tr}(B^*B)$.

(b) Use (a) to prove that

$$\sum_{i,j=1}^{n} |A_{ij}|^2 = \sum_{i,j=1}^{n} |B_{ij}|^2.$$

(c) Use (b) to show that the matrices

$$\begin{pmatrix} 1 & 2 \\ 2 & i \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} i & 4 \\ 1 & 1 \end{pmatrix}$$

are *not* unitarily equivalent.

22. Let V be a real inner product space.

(a) Prove that any translation on V is a rigid motion.

(b) Prove that the composite of any two rigid motions on V is a rigid motion on V.

23. Prove the following variant of Theorem 6.22: If $f: \mathsf{V} \to \mathsf{V}$ is a rigid motion on a finite-dimensional real inner product space V, then there exists a unique orthogonal operator T on V and a unique translation g on V such that $f = \mathsf{T} \circ g$. (Note that the conclusion of Theorem 6.22 has $f = g \circ \mathsf{T}$).

24. Let T and U be orthogonal operators on R^2. Use Theorem 6.23 to prove the following results.

(a) If T and U are both reflections about lines through the origin, then UT is a rotation.

(b) If T is a rotation and U is a reflection about a line through the origin, then both UT and TU are reflections about lines through the origin.

25. Suppose that T and U are reflections of R^2 about the respective lines L and L′ through the origin and that ϕ and ψ are the angles from the positive x-axis to L and L′, respectively. By Exercise 24, UT is a rotation. Find its angle of rotation.

26. Suppose that T and U are orthogonal operators on R^2 such that T is the rotation by the angle ϕ and U is the reflection about the line L through the origin. Let ψ be the angle from the positive x-axis to L. By Exercise 24, both UT and TU are reflections about lines L_1 and L_2, respectively, through the origin.

(a) Find the angle θ from the positive x-axis to L_1.
(b) Find the angle θ from the positive x-axis to L_2.

27. Find new coordinates x', y' so that the following quadratic forms can be written as $\lambda_1(x')^2 + \lambda_2(y')^2$.

(a) $x^2 + 4xy + y^2$
(b) $2x^2 + 2xy + 2y^2$
(c) $x^2 - 12xy - 4y^2$
(d) $3x^2 + 2xy + 3y^2$
(e) $x^2 - 2xy + y^2$

28. Consider the expression X^tAX, where $X^t = (x, y, z)$ and A is as defined in Exercise 2(e). Find a change of coordinates x', y', z' so that the preceding expression is of the form $\lambda_1(x')^2 + \lambda_2(y')^2 + \lambda_3(z')^2$.

29. *QR-Factorization.* Let $w_1, w_2, \ldots, w_n$ be linearly independent vectors in F^n, and let $v_1, v_2, \ldots, v_n$ be the orthogonal vectors obtained from $w_1, w_2, \ldots, w_n$ by the Gram–Schmidt process. Let $u_1, u_2, \ldots, u_n$ be the orthonormal basis obtained by normalizing the v_i's.

(a) Solving (1) in Section 6.2 for w_k in terms of u_k, show that

$$w_k = \|v_k\| u_k + \sum_{j=1}^{k-1} \langle w_k, u_j \rangle u_j \quad (1 \le k \le n).$$

(b) Let A and Q denote the $n \times n$ matrices in which the kth columns are w_k and u_k, respectively. Define $R \in \mathsf{M}_{n\times n}(F)$ by

$$R_{jk} = \begin{cases} \|v_j\| & \text{if } j = k \\ \langle w_k, u_j \rangle & \text{if } j < k \\ 0 & \text{if } j > k. \end{cases}$$

Prove $A = QR$.

(c) Compute Q and R as in (b) for the 3×3 matrix whose columns are the vectors $(1, 1, 0)$, $(2, 0, 1)$, and $(2, 2, 1)$.

(d) Since Q is unitary [orthogonal] and R is upper triangular in (b), we have shown that every invertible matrix is the product of a unitary [orthogonal] matrix and an upper triangular matrix. Suppose that $A \in \mathsf{M}_{n\times n}(F)$ is invertible and $A = Q_1R_1 = Q_2R_2$, where $Q_1, Q_2 \in \mathsf{M}_{n\times n}(F)$ are unitary and $R_1, R_2 \in \mathsf{M}_{n\times n}(F)$ are upper triangular. Prove that $D = R_2R_1^{-1}$ is a unitary diagonal matrix. *Hint:* Use Exercise 17.

(e) The QR factorization described in (b) provides an orthogonalization method for solving a linear system $Ax = b$ when A is invertible. Decompose A to QR, by the Gram–Schmidt process or other means, where Q is unitary and R is upper triangular. Then $QRx = b$, and hence $Rx = Q^*b$. This last system can be easily solved since R is upper triangular. [1]

Use the orthogonalization method and (c) to solve the system

$$\begin{aligned} x_1 + 2x_2 + 2x_3 &= 1 \\ x_1 \qquad + 2x_3 &= 11 \\ x_2 + x_3 &= -1. \end{aligned}$$

30. Suppose that β and γ are ordered bases for an n-dimensional real [complex] inner product space V. Prove that if Q is an orthogonal [unitary] $n \times n$ matrix that changes γ-coordinates into β-coordinates, then β is orthonormal if and only if γ is orthonormal.

The following definition is used in Exercises 31 and 32.

Definition. *Let V be a finite-dimensional complex [real] inner product space, and let u be a unit vector in V. Define the* **Householder** *operator $\mathsf{H}_u\colon \mathsf{V} \to \mathsf{V}$ by $\mathsf{H}_u(x) = x - 2\langle x, u\rangle u$ for all $x \in \mathsf{V}$.*

31. Let H_u be a Householder operator on a finite-dimensional inner product space V. Prove the following results.

(a) H_u is linear.

(b) $\mathsf{H}_u(x) = x$ if and only if x is orthogonal to u.

(c) $\mathsf{H}_u(u) = -u$.

(d) $\mathsf{H}_u^* = \mathsf{H}_u$ and $\mathsf{H}_u^2 = \mathsf{I}$, and hence H_u is a unitary [orthogonal] operator on V.

(*Note:* If V is a real inner product space, then in the language of Section 6.11, H_u is a reflection.)

[1] At one time, because of its great stability, this method for solving large systems of linear equations with a computer was being advocated as a better method than Gaussian elimination even though it requires about three times as much work. (Later, however, J. H. Wilkinson showed that if Gaussian elimination is done "properly," then it is nearly as stable as the orthogonalization method.)

32. Let V be a finite-dimensional inner product space over F. Let x and y be linearly independent vectors in V such that $\|x\| = \|y\|$.

(a) If $F = C$, prove that there exists a unit vector u in V and a complex number θ with $|\theta| = 1$ such that $\mathsf{H}_u(x) = \theta y$. *Hint:* Choose θ so that $\langle x, \theta y \rangle$ is real, and set $u = \dfrac{1}{\|x - \theta y\|}(x - \theta y)$.

(b) If $F = R$, prove that there exists a unit vector u in V such that $\mathsf{H}_u(x) = y$.

6.6 ORTHOGONAL PROJECTIONS AND THE SPECTRAL THEOREM

In this section, we rely heavily on Theorems 6.16 (p. 369) and 6.17 (p. 371) to develop an elegant representation of a normal (if $F = C$) or a self-adjoint (if $F = R$) operator T on a finite-dimensional inner product space. We prove that T can be written in the form $\lambda_1\mathsf{T}_1 + \lambda_2\mathsf{T}_2 + \cdots + \lambda_k\mathsf{T}_k$, where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the distinct eigenvalues of T and $\mathsf{T}_1, \mathsf{T}_2, \ldots, \mathsf{T}_k$ are *orthogonal projections*. We must first develop some results about these special projections.

We assume that the reader is familiar with the results about direct sums developed at the end of Section 5.2. The special case where V is a direct sum of two subspaces is considered in the exercises of Section 1.3.

Recall from the exercises of Section 2.1 that if $\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2$, then a linear operator T on V is the **projection on W_1 along W_2** if, whenever $x = x_1 + x_2$, with $x_1 \in \mathsf{W}_1$ and $x_2 \in \mathsf{W}_2$, we have $\mathsf{T}(x) = x_1$. By Exercise 27 of Section 2.1, we have

$$\mathsf{R}(\mathsf{T}) = \mathsf{W}_1 = \{x \in \mathsf{V} : \mathsf{T}(x) = x\} \quad \text{and} \quad \mathsf{N}(\mathsf{T}) = \mathsf{W}_2.$$

So $\mathsf{V} = \mathsf{R}(\mathsf{T}) \oplus \mathsf{N}(\mathsf{T})$. Thus there is no ambiguity if we refer to T as a "projection on W_1" or simply as a "projection." In fact, it can be shown (see Exercise 17 of Section 2.3) that T is a projection if and only if $\mathsf{T} = \mathsf{T}^2$. Because $\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 = \mathsf{W}_1 \oplus \mathsf{W}_3$ does *not* imply that $\mathsf{W}_2 = \mathsf{W}_3$, we see that W_1 does not uniquely determine T. For an *orthogonal* projection T, however, T is uniquely determined by its range.

Definition. *Let V be an inner product space, and let $\mathsf{T} : \mathsf{V} \to \mathsf{V}$ be a projection. We say that T is an **orthogonal projection** if $\mathsf{R}(\mathsf{T})^\perp = \mathsf{N}(\mathsf{T})$ and $\mathsf{N}(\mathsf{T})^\perp = \mathsf{R}(\mathsf{T})$.*

Note that by Exercise 13(c) of Section 6.2, if V is finite-dimensional, we need only assume that one of the equalities in this definition holds. For example, if $\mathsf{R}(\mathsf{T})^\perp = \mathsf{N}(\mathsf{T})$, then $\mathsf{R}(\mathsf{T}) = \mathsf{R}(\mathsf{T})^{\perp\perp} = \mathsf{N}(\mathsf{T})^\perp$.

An orthogonal projection is *not* the same as an orthogonal operator. In Figure 6.5, T is an orthogonal projection, but T is clearly not an orthogonal operator because $\|\mathsf{T}(v)\| \neq \|v\|$.

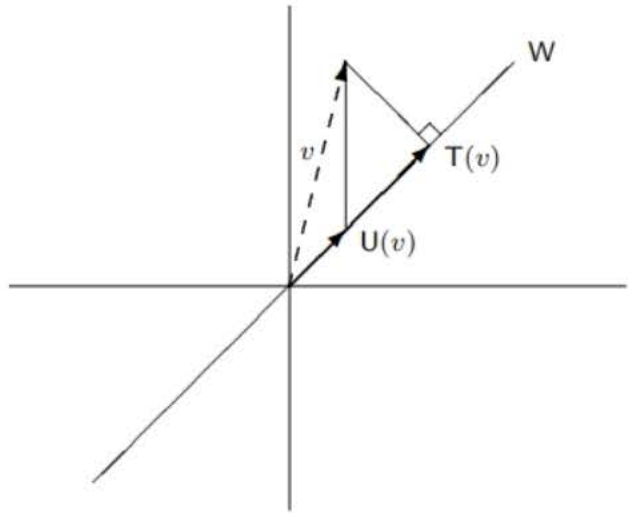


Figure 6.5

Now assume that W is a finite-dimensional subspace of an inner product space V. In the notation of Theorem 6.6 (p. 347), we can define a function $T: V \to V$ by $T(y) = u$. It is easy to show that T is an orthogonal projection on W. We can say even more—there exists exactly one orthogonal projection on W. For if T and U are orthogonal projections on W, then $R(T) = W = R(U)$. Hence $N(T) = R(T)^{\perp} = R(U)^{\perp} = N(U)$, and since every projection is uniquely determined by its range and null space, we have $T = U$. We call T the **orthogonal projection** of V on W.

To understand the geometric difference between an arbitrary projection on W and the orthogonal projection on W, let $V = R^2$ and $W = \text{span}\{(1,1)\}$. Define U and T as in Figure 6.5, where $T(v)$ is the foot of a perpendicular from v on the line $y = x$ and $U(a_1, a_2) = (a_1, a_1)$. Then T is the orthogonal projection of V on W, and U is a different projection on W. Note that $v - T(v) \in W^{\perp}$, whereas $v - U(v) \notin W^{\perp}$. From Figure 6.5, we see that $T(v)$ is the "best approximation in W to v"; that is, if $w \in W$, then $\|w - v\| \geq \|T(v) - v\|$. In fact, this approximation property characterizes T. These results follow immediately from the corollary to Theorem 6.6 (p. 348).

As an application to Fourier analysis, recall the inner product space H and the orthonormal set S in Example 9 of Section 6.1. Define a **trigonometric polynomial of degree** n to be a function $g \in H$ of the form

$$g(t) = \sum_{j=-n}^{n} a_j f_j(t) = \sum_{j=-n}^{n} a_j e^{ijt},$$

where a_n or a_{-n} is nonzero.

Let $f \in H$. We show that the best approximation to f by a trigonometric

polynomial of degree less than or equal to n is the trigonometric polynomial whose coefficients are the Fourier coefficients of f relative to the orthonormal set S. For this result, let $\mathsf{W} = \text{span}(\{f_j : |j| \leq n\})$, and let T be the orthogonal projection of H on W. The corollary to Theorem 6.6 (p. 348) tells us that the best approximation to f by a function in W is

$$\mathsf{T}(f) = \sum_{j=-n}^{n} \langle f, f_j \rangle f_j.$$

For an application of this material to electronic music, visit `goo.gl/EN5Fai`.

An algebraic characterization of orthogonal projections follows in the next theorem.

Theorem 6.24. *Let V be an inner product space, and let T be a linear operator on V. Then T is an orthogonal projection if and only if T has an adjoint T^* and $\mathsf{T}^2 = \mathsf{T} = \mathsf{T}^*$.*

Proof. Suppose that T is an orthogonal projection. Since $\mathsf{T}^2 = \mathsf{T}$ because T is a projection, we need only show that T^* exists and $\mathsf{T} = \mathsf{T}^*$. Now $\mathsf{V} = \mathsf{R}(\mathsf{T}) \oplus \mathsf{N}(\mathsf{T})$ and $\mathsf{R}(\mathsf{T})^\perp = \mathsf{N}(\mathsf{T})$. Let $x, y \in \mathsf{V}$. Then we can write $x = x_1 + x_2$ and $y = y_1 + y_2$, where $x_1, y_1 \in \mathsf{R}(\mathsf{T})$ and $x_2, y_2 \in \mathsf{N}(\mathsf{T})$. Hence

$$\langle x, \mathsf{T}(y) \rangle = \langle x_1 + x_2, y_1 \rangle = \langle x_1, y_1 \rangle + \langle x_2, y_1 \rangle = \langle x_1, y_1 \rangle$$

and

$$\langle \mathsf{T}(x), y \rangle = \langle x_1, y_1 + y_2 \rangle = \langle x_1, y_1 \rangle + \langle x_1, y_2 \rangle = \langle x_1, y_1 \rangle .$$

So $\langle x, \mathsf{T}(y) \rangle = \langle \mathsf{T}(x), y \rangle$ for all $x, y \in \mathsf{V}$; thus T^* exists and $\mathsf{T} = \mathsf{T}^*$.

Now suppose that $\mathsf{T}^2 = \mathsf{T} = \mathsf{T}^*$. Then T is a projection by Exercise 17 of Section 2.3, and hence we must show that $\mathsf{R}(\mathsf{T}) = \mathsf{N}(\mathsf{T})^\perp$ and $\mathsf{R}(\mathsf{T})^\perp = \mathsf{N}(\mathsf{T})$. Let $x \in \mathsf{R}(\mathsf{T})$ and $y \in \mathsf{N}(\mathsf{T})$. Then $x = \mathsf{T}(x) = \mathsf{T}^*(x)$, and so

$$\langle x, y \rangle = \langle \mathsf{T}^*(x), y \rangle = \langle x, \mathsf{T}(y) \rangle = \langle x, 0 \rangle = 0.$$

Therefore $x \in \mathsf{N}(\mathsf{T})^\perp$, from which it follows that $\mathsf{R}(\mathsf{T}) \subseteq \mathsf{N}(\mathsf{T})^\perp$.

Let $y \in \mathsf{N}(\mathsf{T})^\perp$. We must show that $y \in \mathsf{R}(\mathsf{T})$, that is, $\mathsf{T}(y) = y$. Now

$$\begin{aligned} \|y - \mathsf{T}(y)\|^2 &= \langle y - \mathsf{T}(y), y - \mathsf{T}(y) \rangle \\ &= \langle y, y - \mathsf{T}(y) \rangle - \langle \mathsf{T}(y), y - \mathsf{T}(y) \rangle . \end{aligned}$$

Since $y - \mathsf{T}(y) \in \mathsf{N}(\mathsf{T})$, the first term must equal zero. But also

$$\langle \mathsf{T}(y), y - \mathsf{T}(y) \rangle = \langle y, \mathsf{T}^*(y - \mathsf{T}(y)) \rangle = \langle y, \mathsf{T}(y - \mathsf{T}(y)) \rangle = \langle y, 0 \rangle = 0.$$

Thus $y - \mathsf{T}(y) = 0$; that is, $y = \mathsf{T}(y) \in \mathsf{R}(\mathsf{T})$. Hence $\mathsf{R}(\mathsf{T}) = \mathsf{N}(\mathsf{T})^\perp$.

Using the preceding results, we have $\mathsf{R}(\mathsf{T})^\perp = \mathsf{N}(\mathsf{T})^{\perp\perp} \supseteq \mathsf{N}(\mathsf{T})$ by Exercise 13(b) of Section 6.2. Now suppose that $x \in \mathsf{R}(\mathsf{T})^\perp$. For any $y \in \mathsf{V}$, we have $\langle \mathsf{T}(x), y \rangle = \langle x, \mathsf{T}^*(y) \rangle = \langle x, \mathsf{T}(y) \rangle = 0$. So $\mathsf{T}(x) = 0$, and thus $x \in \mathsf{N}(\mathsf{T})$. Hence $\mathsf{R}(\mathsf{T})^\perp = \mathsf{N}(\mathsf{T})$. ∎

Let V be a finite-dimensional inner product space, W be a subspace of V, and T be the orthogonal projection of V on W. We may choose an orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$ for V such that $\{v_1, v_2, \ldots, v_k\}$ is a basis for W. Then $[\mathsf{T}]_\beta$ is a diagonal matrix with ones as the first k diagonal entries and zeros elsewhere. In fact, $[\mathsf{T}]_\beta$ has the form

$$\begin{pmatrix} I_k & O_1 \\ O_2 & O_3 \end{pmatrix}.$$

If U is any projection on W, we may choose a basis γ for V such that $[\mathsf{U}]_\gamma$ has the form above; however γ is not necessarily orthonormal.

We are now ready for the principal theorem of this section.

Theorem 6.25 (The Spectral Theorem). *Suppose that T is a linear operator on a finite-dimensional inner product space V over F with the distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Assume that T is normal if $F = C$ and that T is self-adjoint if $F = R$. For each i $(1 \le i \le k)$, let W_i be the eigenspace of T corresponding to the eigenvalue λ_i, and let T_i be the orthogonal projection of V on W_i. Then the following statements are true.*

(a) $\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 \oplus \cdots \oplus \mathsf{W}_k$.

(b) *If W_i' denotes the direct sum of the subspaces W_j for $j \ne i$, then* $\mathsf{W}_i^\perp = \mathsf{W}_i'$.

(c) $\mathsf{T}_i \mathsf{T}_j = \delta_{ij} \mathsf{T}_i$ *for* $1 \le i, j \le k$.

(d) $\mathsf{I} = \mathsf{T}_1 + \mathsf{T}_2 + \cdots + \mathsf{T}_k$.

(e) $\mathsf{T} = \lambda_1 \mathsf{T}_1 + \lambda_2 \mathsf{T}_2 + \cdots + \lambda_k \mathsf{T}_k$.

Proof. (a) By Theorems 6.16 (p. 369) and 6.17 (p. 371), T is diagonalizable; so

$$\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 \oplus \cdots \oplus \mathsf{W}_k$$

by Theorem 5.10 (p. 277).

(b) If $x \in \mathsf{W}_i$ and $y \in \mathsf{W}_j$ for some $i \ne j$, then $\langle x, y \rangle = 0$ by Theorem 6.15(d) (p. 368). It follows easily from this result that $\mathsf{W}_i' \subseteq \mathsf{W}_i^\perp$. From (a), we have

$$\dim(\mathsf{W}_i') = \sum_{j \ne i} \dim(\mathsf{W}_j) = \dim(\mathsf{V}) - \dim(\mathsf{W}_i).$$

On the other hand, we have $\dim(\mathsf{W}_i^\perp) = \dim(\mathsf{V}) - \dim(\mathsf{W}_i)$ by Theorem 6.7(c) (p. 349). Hence $\mathsf{W}_i' = \mathsf{W}_i^\perp$, proving (b).

(c) The proof of (c) is left as an exercise.

(d) Since T_i is the orthogonal projection of V on W_i, it follows from (b) that $\mathsf{N}(\mathsf{T}_i) = \mathsf{R}(\mathsf{T}_i)^\perp = \mathsf{W}_i^\perp = \mathsf{W}_i'$. Hence, for $x \in \mathsf{V}$, we have $x = x_1 + x_2 + \cdots + x_k$, where $\mathsf{T}_i(x) = x_i \in \mathsf{W}_i$, proving (d).

(e) For $x \in \mathsf{V}$, write $x = x_1 + x_2 + \cdots + x_k$, where $x_i \in \mathsf{W}_i$. Then

$$\mathsf{T}(x) = \mathsf{T}(x_1) + \mathsf{T}(x_2) + \cdots + \mathsf{T}(x_k)$$

$$\begin{aligned} &= \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k \\ &= \lambda_1 \mathsf{T}_1(x) + \lambda_2 \mathsf{T}_2(x) + \cdots + \lambda_k \mathsf{T}_k(x) \\ &= (\lambda_1 \mathsf{T}_1 + \lambda_2 \mathsf{T}_2 + \cdots + \lambda_k \mathsf{T}_k)(x). \end{aligned}$$ ∎

The set $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ of eigenvalues of T is called the **spectrum** of T, the sum $\mathsf{I} = \mathsf{T}_1 + \mathsf{T}_2 + \cdots + \mathsf{T}_k$ in (d) is called the **resolution of the identity operator** induced by T, and the sum $\mathsf{T} = \lambda_1 \mathsf{T}_1 + \lambda_2 \mathsf{T}_2 + \cdots + \lambda_k \mathsf{T}_k$ in (e) is called the **spectral decomposition** of T. The spectral decomposition of T is unique up to the order of its eigenvalues.

With the preceding notation, let β be the union of orthonormal bases of the W_i's and let $m_i = \dim(\mathsf{W}_i)$. (Thus m_i is the multiplicity of λ_i.) Then $[\mathsf{T}]_\beta$ has the form

$$\begin{pmatrix} \lambda_1 I_{m_1} & O & \cdots & O \\ O & \lambda_2 I_{m_2} & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & \lambda_k I_{m_k} \end{pmatrix};$$

that is, $[\mathsf{T}]_\beta$ is a diagonal matrix in which the diagonal entries are the eigenvalues λ_i of T, and each λ_i is repeated m_i times. If $\lambda_1 \mathsf{T}_1 + \lambda_2 \mathsf{T}_2 + \cdots + \lambda_k \mathsf{T}_k$ is the spectral decomposition of T, then it follows (from Exercise 7) that $g(\mathsf{T}) = g(\lambda_1)\mathsf{T}_1 + g(\lambda_2)\mathsf{T}_2 + \cdots + g(\lambda_k)\mathsf{T}_k$ for any polynomial g. This fact is used below.

We now list several interesting corollaries of the spectral theorem; many more results are found in the exercises. For what follows, we assume that T *is a linear operator on a finite-dimensional inner product space* V *over* F.

Corollary 1. *If* $F = C$, *then* T *is normal if and only if* $\mathsf{T}^* = g(\mathsf{T})$ *for some polynomial* g.

Proof. Suppose first that T is normal. Let $\mathsf{T} = \lambda_1 \mathsf{T}_1 + \lambda_2 \mathsf{T}_2 + \cdots + \lambda_k \mathsf{T}_k$ be the spectral decomposition of T. Taking the adjoint of both sides of the preceding equation, we have $\mathsf{T}^* = \overline{\lambda}_1 \mathsf{T}_1 + \overline{\lambda}_2 \mathsf{T}_2 + \cdots + \overline{\lambda}_k \mathsf{T}_k$ since each T_i is self-adjoint. Using the Lagrange interpolation formula (see page 53), we may choose a polynomial g such that $g(\lambda_i) = \overline{\lambda}_i$ for $1 \le i \le k$. Then

$$g(\mathsf{T}) = g(\lambda_1)\mathsf{T}_1 + g(\lambda_2)\mathsf{T}_2 + \cdots + g(\lambda_k)\mathsf{T}_k = \overline{\lambda}_1 \mathsf{T}_1 + \overline{\lambda}_2 \mathsf{T}_2 + \cdots + \overline{\lambda}_k \mathsf{T}_k = \mathsf{T}^*.$$

Conversely, if $\mathsf{T}^* = g(\mathsf{T})$ for some polynomial g, then T commutes with T^* since T commutes with every polynomial in T. So T is normal. ∎

Corollary 2. *If* $F = C$, *then* T *is unitary if and only if* T *is normal and* $|\lambda| = 1$ *for every eigenvalue* λ *of* T.

Proof. If T is unitary, then T is normal and every eigenvalue of T has absolute value 1 by Corollary 2 to Theorem 6.18 (p. 379).

Let $\mathsf{T} = \lambda_1\mathsf{T}_1 + \lambda_2\mathsf{T}_2 + \cdots + \lambda_k\mathsf{T}_k$ be the spectral decomposition of T. If $|\lambda| = 1$ for every eigenvalue λ of T, then by (c) of the spectral theorem,

$$\begin{aligned}\mathsf{T}\mathsf{T}^* &= (\lambda_1\mathsf{T}_1 + \lambda_2\mathsf{T}_2 + \cdots + \lambda_k\mathsf{T}_k)(\overline{\lambda}_1\mathsf{T}_1 + \overline{\lambda}_2\mathsf{T}_2 + \cdots + \overline{\lambda}_k\mathsf{T}_k)\\ &= |\lambda_1|^2\mathsf{T}_1 + |\lambda_2|^2\mathsf{T}_2 + \cdots + |\lambda_k|^2\mathsf{T}_k\\ &= \mathsf{T}_1 + \mathsf{T}_2 + \cdots + \mathsf{T}_k\\ &= \mathsf{I}.\end{aligned}$$

Hence T is unitary. ∎

Corollary 3. *If $F = C$, then T is self-adjoint if and only if T is normal and every eigenvalue of T is real.*

Proof. Suppose that T is normal and that its eigenvalues are real. Let $\mathsf{T} = \lambda_1\mathsf{T}_1 + \lambda_2\mathsf{T}_2 + \cdots + \lambda_k\mathsf{T}_k$ be the spectral decomposition of T. Then

$$\mathsf{T}^* = \overline{\lambda}_1\mathsf{T}_1 + \overline{\lambda}_2\mathsf{T}_2 + \cdots + \overline{\lambda}_k\mathsf{T}_k = \lambda_1\mathsf{T}_1 + \lambda_2\mathsf{T}_2 + \cdots + \lambda_k\mathsf{T}_k = \mathsf{T}.$$

Hence T is self-adjoint.

Now suppose that T is self-adjoint and hence normal. That its eigenvalues are real has been proved in the lemma to Theorem 6.17 (p. 371). ∎

Corollary 4. *Let T be as in the spectral theorem with spectral decomposition $\mathsf{T} = \lambda_1\mathsf{T}_1 + \lambda_2\mathsf{T}_2 + \cdots + \lambda_k\mathsf{T}_k$. Then each T_j is a polynomial in T.*

Proof. Choose a polynomial g_j $(1 \le j \le k)$ such that $g_j(\lambda_i) = \delta_{ij}$. Then

$$\begin{aligned}g_j(\mathsf{T}) &= g_j(\lambda_1)\mathsf{T}_1 + g_j(\lambda_2)\mathsf{T}_2 + \cdots + g_j(\lambda_k)\mathsf{T}_k\\ &= \delta_{1j}\mathsf{T}_1 + \delta_{2j}\mathsf{T}_2 + \cdots + \delta_{kj}\mathsf{T}_k = \mathsf{T}_j.\end{aligned}$$

∎

EXERCISES

1. Label the following statements as true or false. Assume that the underlying inner product spaces are finite-dimensional.
   - **(a)** All projections are self-adjoint.
   - **(b)** An orthogonal projection is uniquely determined by its range.
   - **(c)** Every self-adjoint operator is a linear combination of orthogonal projections.
   - **(d)** If T is a projection on W, then $\mathsf{T}(x)$ is the vector in W that is closest to x.

(e) Every orthogonal projection is a unitary operator.

2. Let $V = R^2$, $W = \text{span}(\{(1,2)\})$, and β be the standard ordered basis for V. Compute $[T]_\beta$, where T is the orthogonal projection of V on W. Do the same for $V = R^3$ and $W = \text{span}(\{(1,0,1)\})$.

3. For each of the matrices A in Exercise 2 of Section 6.5:

(1) Verify that L_A possesses a spectral decomposition.
(2) For each eigenvalue of L_A, explicitly define the orthogonal projection on the corresponding eigenspace.
(3) Verify your results using the spectral theorem.

4. Let W be a finite-dimensional subspace of an inner product space V. Show that if T is the orthogonal projection of V on W, then $I - T$ is the orthogonal projection of V on $W^\perp$.

5. Let T be a linear operator on a finite-dimensional inner product space V.

(a) If T is an orthogonal projection, prove that $\|T(x)\| \leq \|x\|$ for all $x \in V$. Give an example of a projection for which this inequality does not hold. What can be concluded about a projection for which the inequality is actually an equality for all $x \in V$?
(b) Suppose that T is a projection such that $\|T(x)\| \leq \|x\|$ for all $x \in V$. Prove that T is an orthogonal projection.

6. Let T be a normal operator on a finite-dimensional inner product space. Prove that if T is a projection, then T is also an orthogonal projection.

7. Let T be a normal operator on a finite-dimensional complex inner product space V. Use the spectral decomposition $\lambda_1 T_1 + \lambda_2 T_2 + \cdots + \lambda_k T_k$ of T to prove the following results.

(a) If g is a polynomial, then

$$g(T) = \sum_{i=1}^{k} g(\lambda_i) T_i.$$

(b) If $T^n = T_0$ for some n, then $T = T_0$.
(c) Let U be a linear operator on V. Then U commutes with T if and only if U commutes with each T_i.
(d) There exists a normal operator U on V such that $U^2 = T$.
(e) T is invertible if and only if $\lambda_i \neq 0$ for $1 \leq i \leq k$.
(f) T is a projection if and only if every eigenvalue of T is 1 or 0.
(g) $T = -T^*$ if and only if every λ_i is an imaginary number.

8. Use Corollary 1 of the spectral theorem to show that if T is a normal operator on a complex finite-dimensional inner product space and U is a linear operator that commutes with T, then U commutes with T^*.

9. Referring to Exercise 20 of Section 6.5, prove the following facts about a partial isometry U.

(a) $\mathsf{U}^*\mathsf{U}$ is an orthogonal projection on W.
(b) $\mathsf{U}\mathsf{U}^*\mathsf{U} = \mathsf{U}$.

10. *Simultaneous diagonalization.* Let U and T be normal operators on a finite-dimensional complex inner product space V such that $\mathsf{TU} = \mathsf{UT}$. Prove that there exists an orthonormal basis for V consisting of vectors that are eigenvectors of both T and U. *Hint:* Use the hint of Exercise 14 of Section 6.4 along with Exercise 8.

11. Prove (c) of the spectral theorem. Visit `goo.gl/utQ9Pb` for a solution.

6.7* THE SINGULAR VALUE DECOMPOSITION AND THE PSEUDOINVERSE

In Section 6.4, we characterized normal operators on complex spaces and self-adjoint operators on real spaces in terms of orthonormal bases of eigenvectors and their corresponding eigenvalues (Theorems 6.16, p. 369, and 6.17, p. 371). In this section, we establish a comparable theorem whose scope is the entire class of linear transformations on both complex and real finite-dimensional inner product spaces—*the singular value theorem for linear transformations* (Theorem 6.26). There are similarities and differences among these theorems. All rely on the use of orthonormal bases and numerical invariants. However, because of its general scope, the singular value theorem is concerned with two (usually distinct) inner product spaces and with two (usually distinct) orthonormal bases. If the two spaces and the two bases are identical, then the transformation would, in fact, be a normal or self-adjoint operator. Another difference is that the numerical invariants in the singular value theorem, the *singular values*, are nonnegative, in contrast to their counterparts, the eigenvalues, for which there is no such restriction. This property is necessary to guarantee the uniqueness of singular values.

The singular value theorem encompasses both real and complex spaces. For brevity, in this section we use the terms *unitary operator* and *unitary matrix* to include orthogonal operators and orthogonal matrices in the context of real spaces. Thus any operator T for which $\langle \mathsf{T}(x), \mathsf{T}(y)\rangle = \langle x, y\rangle$, or any matrix A for which $\langle Ax, Ay\rangle = \langle x, y\rangle$, for all x and y is called *unitary* for the purposes of this section.

In Exercise 15 of Section 6.3, the definition of the adjoint of an operator is extended to any linear transformation $\mathsf{T}: \mathsf{V} \to \mathsf{W}$, where V and W are

finite-dimensional inner product spaces. By this exercise, the adjoint T^* of T is a linear transformation from W to V and $[T^*]_\gamma^\beta = ([T]_\beta^\gamma)^*$, where β and γ are orthonormal bases for V and W, respectively. Furthermore, the linear operator T^*T on V is positive semidefinite and $\text{rank}(T^*T) = \text{rank}(T)$ by Exercise 18 of Section 6.4.

With these facts in mind, we begin with the principal result.

Theorem 6.26 (Singular Value Theorem for Linear Transformations). *Let V and W be finite-dimensional inner product spaces, and let $T: V \to W$ be a linear transformation of rank r. Then there exist orthonormal bases $\{v_1, v_2, \ldots, v_n\}$ for V and $\{u_1, u_2, \ldots, u_m\}$ for W and positive scalars $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ such that*

$$T(v_i) = \begin{cases} \sigma_i u_i & \text{if } 1 \leq i \leq r \\ 0 & \text{if } i > r. \end{cases} \tag{4}$$

Conversely, suppose that the preceding conditions are satisfied. Then for $1 \leq i \leq n$, v_i is an eigenvector of T^*T with corresponding eigenvalue σ_i^2 if $1 \leq i \leq r$ and 0 if $i > r$. Therefore the scalars $\sigma_1, \sigma_2, \ldots, \sigma_r$ are uniquely determined by T.

Proof. We first establish the existence of the bases and scalars. By Exercises 18 of Section 6.4 and 15(d) of Section 6.3, T^*T is a positive semidefinite linear operator of rank r on V; hence there is an orthonormal basis $\{v_1, v_2, \ldots, v_n\}$ for V consisting of eigenvectors of T^*T with corresponding eigenvalues λ_i, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$, and $\lambda_i = 0$ for $i > r$. For $1 \leq i \leq r$, define $\sigma_i = \sqrt{\lambda_i}$ and $u_i = \frac{1}{\sigma_i} T(v_i)$. We show that $\{u_1, u_2, \ldots, u_r\}$ is an orthonormal subset of W. Suppose $1 \leq i, j \leq r$. Then

$$\begin{aligned} \langle u_i, u_j \rangle &= \left\langle \frac{1}{\sigma_i} T(v_i), \frac{1}{\sigma_j} T(v_j) \right\rangle \\ &= \frac{1}{\sigma_i \sigma_j} \langle T^*T(v_i), v_j \rangle \\ &= \frac{1}{\sigma_i \sigma_j} \langle \lambda_i v_i, v_j \rangle \\ &= \frac{\sigma_i^2}{\sigma_i \sigma_j} \langle v_i, v_j \rangle \\ &= \delta_{ij}, \end{aligned}$$

and hence $\{u_1, u_2, \ldots, u_r\}$ is orthonormal. By Theorem 6.7(a) (p. 349), this set extends to an orthonormal basis $\{u_1, u_2, \ldots, u_r, \ldots, u_m\}$ for W. Clearly

$\mathsf{T}(v_i) = \sigma_i u_i$ if $1 \le i \le r$. If $i > r$, then $\mathsf{T}^*\mathsf{T}(v_i) = 0$, and so $\mathsf{T}(v_i) = 0$ by Exercise 15(d) of Section 6.3.

To establish uniqueness, suppose that $\{v_1, v_2, \ldots, v_n\}$, $\{u_1, u_2, \ldots, u_m\}$, and $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$ satisfy the properties stated in the first part of the theorem. Then for $1 \le i \le m$ and $1 \le j \le n$,

$$\langle \mathsf{T}^*(u_i), v_j \rangle = \langle u_i, \mathsf{T}(v_j) \rangle$$

$$= \begin{cases} \sigma_i & \text{if } i = j \le r \\ 0 & \text{otherwise,} \end{cases}$$

and hence for any $1 \le i \le m$,

$$\mathsf{T}^*(u_i) = \sum_{j=1}^{n} \langle \mathsf{T}^*(u_i), v_j \rangle v_j = \begin{cases} \sigma_i v_i & \text{if } i = j \le r \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

So for $i \le r$,

$$\mathsf{T}^*\mathsf{T}(v_i) = \mathsf{T}^*(\sigma_i u_i) = \sigma_i \mathsf{T}^*(u_i) = \sigma_i^2 v_i$$

and $\mathsf{T}^*\mathsf{T}(v_i) = \mathsf{T}^*(0) = 0$ for $i > r$. Therefore each v_i is an eigenvector of $\mathsf{T}^*\mathsf{T}$ with corresponding eigenvalue σ_i^2 if $i \le r$ and 0 if $i > r$. ∎

Definition. *The unique scalars $\sigma_1, \sigma_2, \ldots, \sigma_r$ in Theorem 6.26 are called the* ***singular values*** *of T. If r is less than both m and n, then the term* singular value *is extended to include $\sigma_{r+1} = \cdots = \sigma_k = 0$, where k is the minimum of m and n.*

Although the singular values of a linear transformation T are uniquely determined by T, the orthonormal bases given in the statement of Theorem 6.26 are not uniquely determined because there is more than one orthonormal basis of eigenvectors of $\mathsf{T}^*\mathsf{T}$.

In view of (5), the singular values of a linear transformation $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ and its adjoint T^* are identical. Furthermore, the orthonormal bases for V and W given in Theorem 6.26 are simply reversed for T^*.

Example 1

Let $\mathsf{P}_2(R)$ and $\mathsf{P}_1(R)$ be the polynomial spaces with inner products defined by

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

Let $\mathsf{T}\colon \mathsf{P}_2(R) \to \mathsf{P}_1(R)$ be the linear transformation defined by $\mathsf{T}(f(x)) = f'(x)$. Find orthonormal bases $\beta = \{v_1, v_2, v_3\}$ for $\mathsf{P}_2(R)$ and $\gamma = \{u_1, u_2\}$ for $\mathsf{P}_1(R)$ such that $\mathsf{T}(v_i) = \sigma_i u_i$ for $i = 1, 2$ and $\mathsf{T}(v_3) = 0$, where $\sigma_1 \geq \sigma_2 > 0$ are the nonzero singular values of T.

To facilitate the computations, we translate this problem into the corresponding problem for a matrix representation of T. Caution is advised here because not any matrix representation will do. Since the adjoint is defined in terms of inner products, we must use a matrix representation constructed from orthonormal bases for $\mathsf{P}_2(R)$ and $\mathsf{P}_1(R)$ to guarantee that the adjoint of the matrix representation of T is the same as the matrix representation of the adjoint of T. (See Exercise 15 of Section 6.3.) For this purpose, we use the results of Exercise 21(a) of Section 6.2 to obtain orthonormal bases

$$\alpha = \left\{\frac{1}{\sqrt{2}},\ \sqrt{\frac{3}{2}}\,x,\ \sqrt{\frac{5}{8}}\,(3x^2 - 1)\right\} \quad \text{and} \quad \alpha' = \left\{\frac{1}{\sqrt{2}},\ \sqrt{\frac{3}{2}}\,x\right\}$$

for $\mathsf{P}_2(R)$ and $\mathsf{P}_1(R)$, respectively.

Let

$$A = [\mathsf{T}]_{\alpha}^{\alpha'} = \begin{pmatrix} 0 & \sqrt{3} & 0 \\ 0 & 0 & \sqrt{15} \end{pmatrix}.$$

Then

$$A^*A = \begin{pmatrix} 0 & 0 \\ \sqrt{3} & 0 \\ 0 & \sqrt{15} \end{pmatrix} \begin{pmatrix} 0 & \sqrt{3} & 0 \\ 0 & 0 & \sqrt{15} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 15 \end{pmatrix},$$

which has eigenvalues (listed in descending order of size) $\lambda_1 = 15$, $\lambda_2 = 3$, and $\lambda_3 = 0$. These eigenvalues correspond, respectively, to the orthonormal eigenvectors $e_3 = (0, 0, 1)$, $e_2 = (0, 1, 0)$, and $e_1 = (1, 0, 0)$ in R^3. Translating everything into the context of T, $\mathsf{P}_2(R)$, and $\mathsf{P}_1(R)$, let

$$v_1 = \sqrt{\frac{5}{8}}\,(3x^2 - 1), \quad v_2 = \sqrt{\frac{3}{2}}\,x, \quad \text{and} \quad v_3 = \frac{1}{\sqrt{2}}.$$

Then $\beta = \{v_1, v_2, v_3\}$ is an orthonormal basis for $\mathsf{P}_2(R)$ consisting of eigenvectors of $\mathsf{T}^*\mathsf{T}$ with corresponding eigenvalues λ_1, λ_2, and λ_3. Now set $\sigma_1 = \sqrt{\lambda_1} = \sqrt{15}$ and $\sigma_2 = \sqrt{\lambda_2} = \sqrt{3}$, the nonzero singular values of T, and take

$$u_1 = \frac{1}{\sigma_1}\mathsf{T}(v_1) = \sqrt{\frac{3}{2}}\,x \quad \text{and} \quad u_2 = \frac{1}{\sigma_2}\mathsf{T}(v_2) = \frac{1}{\sqrt{2}},$$

to obtain the required basis $\gamma = \{u_1, u_2\}$ for $\mathsf{P}_1(R)$. ◆

We can use singular values to describe how a figure is distorted by a linear transformation. This is illustrated in the next example.

Example 2

Let T be an invertible linear operator on R^2 and $S = \{x \in \mathsf{R}^2 : \|x\| = 1\}$, the unit circle in R^2. We apply Theorem 6.26 to describe $S' = \mathsf{T}(S)$.

Since T is invertible, it has rank equal to 2 and hence has singular values $\sigma_1 \geq \sigma_2 > 0$. Let $\{v_1, v_2\}$ and $\beta = \{u_1, u_2\}$ be orthonormal bases for R^2 so that $\mathsf{T}(v_1) = \sigma_1 u_1$ and $\mathsf{T}(v_2) = \sigma_2 u_2$, as in Theorem 6.26. Then β determines a coordinate system, which we will call the $x'y'$-coordinate system for R^2, where the x'-axis contains u_1 and the y'-axis contains u_2. For any vector $u \in \mathsf{R}^2$, if $u = x_1' u_1 + x_2' u_2$, then $[u]_\beta = \begin{pmatrix} x_1' \\ x_2' \end{pmatrix}$ is the coordinate vector of u relative to β. We characterize S' in terms of an equation relating x_1' and x_2'.

For any vector $v = x_1 v_1 + x_2 v_2 \in \mathsf{R}^2$, the equation $u = \mathsf{T}(v)$ means that

$$u = \mathsf{T}(x_1 v_1 + x_2 v_2) = x_1 \mathsf{T}(v_1) + x_2 \mathsf{T}(v_2) = x_1 \sigma_1 u_1 + x_2 \sigma_2 u_2.$$

Thus for $u = x_1' u_1 + x_2' u_2$, we have $x_1' = x_1 \sigma_1$ and $x_2' = x_2 \sigma_2$. Furthermore, $u \in S'$ if and only if $v \in S$ if and only if

$$\frac{(x_1')^2}{\sigma_1^2} + \frac{(x_2')^2}{\sigma_2^2} = x_1^2 + x_2^2 = 1.$$

If $\sigma_1 = \sigma_2$, this is the equation of a circle of radius σ_1, and if $\sigma_1 > \sigma_2$, this is the equation of an ellipse with major axis and minor axis oriented along the x'-axis and the y'-axis, respectively. (See Figure 6.6.) ◆

The singular value theorem for linear transformations is useful in its matrix form because we can perform numerical computations on matrices. We begin with the definition of the singular values of a matrix.

Definition. *Let A be an $m \times n$ matrix. We define the* ***singular values*** *of A to be the singular values of the linear transformation L_A.*

Theorem 6.27 (Singular Value Decomposition Theorem for Matrices). *Let A be an $m \times n$ matrix of rank r with the positive singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, and let Σ be the $m \times n$ matrix defined by*

$$\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \leq r \\ 0 & \text{otherwise.} \end{cases}$$

Then there exists an $m \times m$ unitary matrix U and an $n \times n$ unitary matrix V such that

$$A = U \Sigma V^*.$$

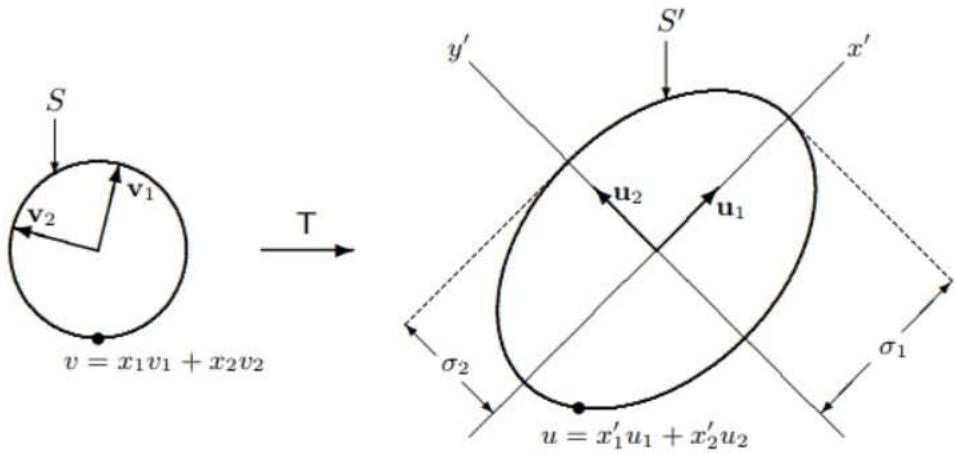


Figure 6.6

Proof. Let $T = \mathsf{L}_A\colon \mathsf{F}^n \to \mathsf{F}^m$. By Theorem 6.26, there exist orthonormal bases $\beta = \{v_1, v_2, \ldots, v_n\}$ for F^n and $\gamma = \{u_1, u_2, \ldots, u_m\}$ for F^m such that $\mathsf{T}(v_i) = \sigma_i u_i$ for $1 \le i \le r$ and $\mathsf{T}(v_i) = 0$ for $i > r$. Let U be the $m \times m$ matrix whose jth column is u_j for all j, and let V be the $n \times n$ matrix whose jth column is v_j for all j. Note that both U and V are unitary matrices.

By Theorem 2.13(a) (p. 91), the jth column of AV is $Av_j = \sigma_j u_j$. Observe that the jth column of Σ is $\sigma_j e_j$, where e_j is the jth standard vector of F^m. So by Theorem 2.13(a) and (b), the jth column of $U\Sigma$ is given by

$$U(\sigma_j e_j) = \sigma_j U e_j = \sigma_j u_j.$$

It follows that AV and $U\Sigma$ are $m \times n$ matrices whose corresponding columns are equal, and hence $AV = U\Sigma$. Therefore $A = AVV^* = U\Sigma V^*$. ∎

Definition. *Let A be an $m \times n$ matrix of rank r with positive singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$. A factorization $A = U\Sigma V^*$ where U and V are unitary matrices and Σ is the $m \times n$ matrix defined as in Theorem 6.27 is called a* ***singular value decomposition*** *of A.*

In the proof of Theorem 6.27, the columns of V are the vectors in β, and the columns of U are the vectors in γ. Furthermore, the nonzero singular values of A are the same as those of L_A; hence they are the square roots of the nonzero eigenvalues of A^*A or of AA^*. (See Exercise 9.)

Example 3

We find a singular value decomposition for $A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \end{pmatrix}$.

First observe that for

$$v_1 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \quad v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \text{and} \quad v_3 = \frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix},$$

the set $\beta = \{v_1, v_2, v_3\}$ is an orthonormal basis for R^3 consisting of eigenvectors of A^*A with corresponding eigenvalues $\lambda_1 = 6$, and $\lambda_2 = \lambda_3 = 0$. Consequently, $\sigma_1 = \sqrt{6}$ is the only nonzero singular value of A. Hence, as in the proof of Theorem 6.27, we let V be the matrix whose columns are the vectors in β. Then

$$\Sigma = \begin{pmatrix} \sqrt{6} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad V = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{pmatrix}.$$

Also, as in Theorem 6.27, we take

$$u_1 = \frac{1}{\sigma_i}\mathsf{L}_A(v_1) = \frac{1}{\sigma_i}Av_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Next choose $u_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, a unit vector orthogonal to u_1, to obtain the orthonormal basis $\gamma = \{u_1, u_2\}$ for R^2, and set

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}.$$

Then $A = U\Sigma V^*$ is the desired singular value decomposition. ♦

The Polar Decomposition of a Square Matrix

A singular value decomposition of a matrix can be used to factor a square matrix in a manner analogous to the factoring of a complex number as the product of a complex number of length 1 and a nonnegative number. In the case of matrices, the complex number of length 1 is replaced by a unitary matrix, and the nonnegative number is replaced by a positive semidefinite matrix.

Theorem 6.28 (Polar Decomposition). *For any square matrix A, there exists a unitary matrix W and a positive semidefinite matrix P such that*

$$A = WP.$$

Furthermore, if A is invertible, then the representation is unique.

Proof. By Theorem 6.27, there exist unitary matrices U and V and a diagonal matrix Σ with nonnegative diagonal entries such that $A = U\Sigma V^*$. So

$$A = U\Sigma V^* = UV^*V\Sigma V^* = WP,$$

where $W = UV^*$ and $P = V\Sigma V^*$. Since W is the product of unitary matrices, W is unitary, and since Σ is positive semidefinite and P is unitarily equivalent to Σ, P is positive semidefinite by Exercise 14 of Section 6.5.

Now suppose that A is invertible and factors as the products

$$A = WP = ZQ,$$

where W and Z are unitary and P and Q are positive semidefinite. Since A is invertible, it follows that P and Q are positive definite and invertible, and therefore $Z^*W = QP^{-1}$. Thus QP^{-1} is unitary, and so

$$I = (QP^{-1})^*(QP^{-1}) = P^{-1}Q^2P^{-1}.$$

Hence $P^2 = Q^2$. Since both P and Q are positive definite, it follows that $P = Q$ by Exercise 17 of Section 6.4. Therefore $W = Z$, and consequently the factorization is unique. ∎

The factorization of a square matrix A as WP where W is unitary and P is positive semidefinite, is called a **polar decomposition** of A.

Example 4

To find the polar decomposition of $A = \begin{pmatrix} 11 & -5 \\ -2 & 10 \end{pmatrix}$, we begin by finding a singular value decomposition $U\Sigma V^*$ of A. The object is to find an orthonormal basis β for R^2 consisting of eigenvectors of A^*A. It can be shown that

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

are orthonormal eigenvectors of A^*A with corresponding eigenvalues $\lambda_1 = 200$ and $\lambda_2 = 50$. So $\beta = \{v_1, v_2\}$ is an appropriate basis. Thus $\sigma_1 = \sqrt{200} = 10\sqrt{2}$ and $\sigma_2 = \sqrt{50} = 5\sqrt{2}$ are the singular values of A. So we have

$$V = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 10\sqrt{2} & 0 \\ 0 & 5\sqrt{2} \end{pmatrix}.$$

Next, we find the columns u_1 and u_2 of U:

$$u_1 = \frac{1}{\sigma_1}Av_1 = \frac{1}{5}\begin{pmatrix} 4 \\ -3 \end{pmatrix} \quad \text{and} \quad u_2 = \frac{1}{\sigma_2}Av_2 = \frac{1}{5}\begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

Thus

$$U = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ -\frac{3}{5} & \frac{4}{5} \end{pmatrix}.$$

Therefore, in the notation of Theorem 6.28, we have

$$W = UV^* = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ -\frac{3}{5} & \frac{4}{5} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{5\sqrt{2}} \begin{pmatrix} 7 & -1 \\ 1 & 7 \end{pmatrix},$$

and

$$P = V\Sigma V^* = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 10\sqrt{2} & 0 \\ 0 & 5\sqrt{2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{5}{\sqrt{2}} \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}.$$

◆

The Pseudoinverse

Let V and W be finite-dimensional inner product spaces over the same field, and let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be a linear transformation. It is desirable to have a linear transformation from W to V that captures some of the essence of an inverse of T even if T is not invertible. A simple approach to this problem is to focus on the "part" of T that is invertible, namely, the restriction of T to $\mathsf{N}(\mathsf{T})^\perp$. Let $\mathsf{L}: \mathsf{N}(\mathsf{T})^\perp \to \mathsf{R}(\mathsf{T})$ be the linear transformation defined by $\mathsf{L}(x) = \mathsf{T}(x)$ for all $x \in \mathsf{N}(\mathsf{T})^\perp$. Then L is invertible, and we can use the inverse of L to construct a linear transformation from W to V that salvages some of the benefits of an inverse of T.

Definition. *Let V and W be finite-dimensional inner product spaces over the same field, and let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be a linear transformation. Let $\mathsf{L}: \mathsf{N}(\mathsf{T})^\perp \to \mathsf{R}(\mathsf{T})$ be the linear transformation defined by $\mathsf{L}(x) = \mathsf{T}(x)$ for all $x \in \mathsf{N}(\mathsf{T})^\perp$. The* **pseudoinverse** *(or Moore-Penrose generalized inverse) of T, denoted by $T^\dagger$, is defined as the unique linear transformation from W to V such that*

$$\mathsf{T}^\dagger(y) = \begin{cases} \mathsf{L}^{-1}(y) & \text{for } y \in \mathsf{R}(\mathsf{T}) \\ 0 & \text{for } y \in \mathsf{R}(\mathsf{T})^\perp. \end{cases}$$

Note that $\mathsf{L}^{-1}\mathsf{T}(x) = x$ for all $x \in \mathsf{N}(\mathsf{T})^\perp$.

The pseudoinverse of a linear transformation T on a finite-dimensional inner product space exists even if T is not invertible. Furthermore, if T is invertible, then $\mathsf{T}^\dagger = \mathsf{T}^{-1}$ because $\mathsf{N}(\mathsf{T})^\perp = \mathsf{V}$, and L (as defined above) coincides with T.

As an extreme example, consider the zero transformation $\mathsf{T}_0: \mathsf{V} \to \mathsf{W}$ between two finite-dimensional inner product spaces V and W. Then $\mathsf{R}(\mathsf{T}_0) = \{0\}$, and therefore $\mathsf{T}^\dagger$ is the zero transformation from W to V.

We can use the singular value theorem to describe the pseudoinverse of a linear transformation. Suppose that V and W are finite-dimensional vector spaces and $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ is a linear transformation of rank r. Let $\{v_1, v_2, \ldots, v_n\}$ and $\{u_1, u_2, \ldots, u_m\}$ be orthonormal bases for V and W, respectively, and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ be the nonzero singular values of T satisfying (4) in Theorem 6.26. Then $\{v_1, v_2, \ldots, v_r\}$ is a basis for $\mathsf{N}(\mathsf{T})^\perp$, $\{v_{r+1}, v_{r+2}, \ldots, v_n\}$ is a basis for $\mathsf{N}(\mathsf{T})$, $\{u_1, u_2, \ldots, u_r\}$ is a basis for $\mathsf{R}(\mathsf{T})$, and $\{u_{r+1}, u_{r+2}, \ldots, u_m\}$ is a basis for $\mathsf{R}(\mathsf{T})^\perp$. Let L be the restriction of T to $\mathsf{N}(\mathsf{T})^\perp$, as in the definition of pseudoinverse. Then $\mathsf{L}^{-1}(u_i) = \dfrac{1}{\sigma_i} v_i$ for $1 \leq i \leq r$. Therefore

$$\mathsf{T}^\dagger(u_i) = \begin{cases} \dfrac{1}{\sigma_i} v_i & \text{if } 1 \leq i \leq r \\ 0 & \text{if } r < i \leq m. \end{cases} \tag{6}$$

Example 5

Let $\mathsf{T}: \mathsf{P}_2(R) \to \mathsf{P}_1(R)$ be the linear transformation defined by $\mathsf{T}(f(x)) = f'(x)$, as in Example 1. Let $\beta = \{v_1, v_2, v_3\}$ and $\gamma = \{u_1, u_2\}$ be the orthonormal bases for $\mathsf{P}_2(R)$ and $\mathsf{P}_1(R)$ in Example 1. Then $\sigma_1 = \sqrt{15}$ and $\sigma_2 = \sqrt{3}$ are the nonzero singular values of T. It follows that

$$\mathsf{T}^\dagger\left(\sqrt{\frac{3}{2}}x\right) = \mathsf{T}^\dagger(u_1) = \frac{1}{\sigma_1} v_1 = \frac{1}{\sqrt{15}}\sqrt{\frac{5}{8}}(3x^2 - 1),$$

and hence

$$\mathsf{T}^\dagger(x) = \frac{1}{6}(3x^2 - 1).$$

Similarly, $\mathsf{T}^\dagger(1) = x$. Thus, for any polynomial $a + bx \in \mathsf{P}_1(R)$,

$$\mathsf{T}^\dagger(a + bx) = a\mathsf{T}^\dagger(1) + b\mathsf{T}^\dagger(x) = ax + \frac{b}{6}(3x^2 - 1). \quad \blacklozenge$$

The Pseudoinverse of a Matrix

Let A be an $m \times n$ matrix. Then there exists a unique $n \times m$ matrix B such that $(\mathsf{L}_A)^\dagger: \mathsf{F}^m \to \mathsf{F}^n$ is equal to the left-multiplication transformation L_B. We call B the **pseudoinverse** of A and denote it by $B = A^\dagger$. Thus

$$(\mathsf{L}_A)^\dagger = \mathsf{L}_{A^\dagger}.$$

Let A be an $m \times n$ matrix of rank r. The pseudoinverse of A can be computed with the aid of a singular value decomposition $A = U\Sigma V^*$. Let β and γ be the ordered bases whose vectors are the columns of V and U, respectively, and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ be the nonzero singular values of

A. Then β and γ are orthonormal bases for F^n and F^m, respectively, and (4) and (6) are satisfied for $\mathsf{T} = \mathsf{L}_A$. Reversing the roles of β and γ in the proof of Theorem 6.27, we obtain the following result.

Theorem 6.29. *Let A be an $m \times n$ matrix of rank r with a singular value decomposition $A = U\Sigma V^*$ and nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. Let $\Sigma^\dagger$ be the $n \times m$ matrix defined by*

$$\Sigma^\dagger_{ij} = \begin{cases} \dfrac{1}{\sigma_i} & \text{if } i = j \leq r \\ 0 & \text{otherwise.} \end{cases}$$

Then $A^\dagger = V\Sigma^\dagger U^*$, and this is a singular value decomposition of $A^\dagger$.

Notice that $\Sigma^\dagger$ as defined in Theorem 6.29 is actually the pseudoinverse of Σ.

Example 6

We find $A^\dagger$ for the matrix $A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \end{pmatrix}$.

Since A is the matrix of Example 3, we can use the singular value decomposition obtained in that example:

$$A = U\Sigma V^* = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{6} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{pmatrix}^*.$$

By Theorem 6.29, we have

$$A^\dagger = V\Sigma^\dagger U^* = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}.$$

◆

Notice that the linear transformation T of Example 5 is L_A, where A is the matrix of Example 6, and that $\mathsf{T}^\dagger = \mathsf{L}_{A^\dagger}$.

The Pseudoinverse and Systems of Linear Equations

Let A be an $m \times n$ matrix with entries in F. Then for any $b \in \mathsf{F}^m$, the matrix equation $Ax = b$ is a system of linear equations, and so it either has no solutions, a unique solution, or infinitely many solutions. We know that the system has a unique solution for every $b \in \mathsf{F}^m$ if and only if A is invertible,

in which case the solution is given by $A^{-1}b$. Furthermore, if A is invertible, then $A^{-1} = A^\dagger$, and so the solution can be written as $x = A^\dagger b$. If, on the other hand, A is not invertible or the system $Ax = b$ is inconsistent, then $A^\dagger b$ still exists. We therefore pose the following question: In general, how is the vector $A^\dagger b$ related to the system of linear equations $Ax = b$?

In order to answer this question, we need the following lemma.

Lemma. *Let* V *and* W *be finite-dimensional inner product spaces, and let* $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ *be linear. Then*

(a) $\mathsf{T}^\dagger\mathsf{T}$ *is the orthogonal projection of* V *on* $\mathsf{N}(\mathsf{T})^\perp$.

(b) $\mathsf{T}\mathsf{T}^\dagger$ *is the orthogonal projection of* W *on* $\mathsf{R}(\mathsf{T})$.

Proof. As in the earlier discussion, we define $\mathsf{L}: \mathsf{N}(\mathsf{T})^\perp \to \mathsf{W}$ by $\mathsf{L}(x) = \mathsf{T}(x)$ for all $x \in \mathsf{N}(\mathsf{T})^\perp$. If $x \in \mathsf{N}(\mathsf{T})^\perp$, then $\mathsf{T}^\dagger\mathsf{T}(x) = \mathsf{L}^{-1}\mathsf{L}(x) = x$, and if $x \in \mathsf{N}(\mathsf{T})$, then $\mathsf{T}^\dagger\mathsf{T}(x) = \mathsf{T}^\dagger(0) = 0$. Consequently $\mathsf{T}^\dagger\mathsf{T}$ is the orthogonal projection of V on $\mathsf{N}(\mathsf{T})^\perp$. This proves (a).

The proof of (b) is similar and is left as an exercise. ∎

Theorem 6.30. *Consider the system of linear equations* $Ax = b$, *where* A *is an* $m \times n$ *matrix and* $b \in \mathsf{F}^m$. *If* $z = A^\dagger b$, *then* z *has the following properties.*

(a) *If* $Ax = b$ *is consistent, then* z *is the unique solution to the system having minimum norm. That is,* z *is a solution to the system, and if* y *is any solution to the system, then* $\|z\| \leq \|y\|$ *with equality if and only if* $z = y$.

(b) *If* $Ax = b$ *is inconsistent, then* z *is the unique best approximation to a solution having minimum norm. That is,* $\|Az - b\| \leq \|Ay - b\|$ *for any* $y \in \mathsf{F}^n$, *with equality if and only if* $Az = Ay$. *Furthermore, if* $Az = Ay$, *then* $\|z\| \leq \|y\|$ *with equality if and only if* $z = y$.

Proof. For convenience, let $\mathsf{T} = \mathsf{L}_A$.

(a) Suppose that $Ax = b$ is consistent, and let $z = A^\dagger b$. Observe that $b \in \mathsf{R}(\mathsf{T})$, and therefore $Az = AA^\dagger b = \mathsf{T}\mathsf{T}^\dagger(b) = b$ by part (b) of the lemma. Thus z is a solution to the system. Now suppose that y is any solution to the system. Then

$$\mathsf{T}^\dagger\mathsf{T}(y) = A^\dagger Ay = A^\dagger b = z,$$

and hence z is the orthogonal projection of y on $\mathsf{N}(\mathsf{T})^\perp$ by part (a) of the lemma. Therefore, by the corollary to Theorem 6.6 (p. 348), we have that $\|z\| \leq \|y\|$ with equality if and only if $z = y$.

(b) Suppose that $Ax = b$ is inconsistent. By the lemma, $Az = AA^\dagger b = \mathsf{T}\mathsf{T}^\dagger(b) = b$ is the orthogonal projection of b on $\mathsf{R}(\mathsf{T})$; therefore, by the corollary to Theorem 6.6 (p. 348), Az is the vector in $\mathsf{R}(\mathsf{T})$ nearest b. That is, if Ay is any other vector in $\mathsf{R}(\mathsf{T})$, then $\|Az - b\| \leq \|Ay - b\|$ with equality if and only if $Az = Ay$.

Finally, suppose that y is any vector in F^n such that $Az = Ay = c$. Then

$$A^\dagger c = A^\dagger Az = A^\dagger AA^\dagger b = A^\dagger b = z$$

by Exercise 23; hence we may apply part (a) of this theorem to the system $Ax = c$ to conclude that $\|z\| \leq \|y\|$ with equality if and only if $z = y$. ∎

Note that the vector $z = A^\dagger b$ in Theorem 6.30 is the vector x_0 described in Theorem 6.12 that arises in the least squares application on pages 358–361.

Example 7

Consider the linear systems

$$\begin{aligned} x_1 + x_2 - x_3 &= 1 \\ x_1 + x_2 - x_3 &= 1 \end{aligned} \qquad \text{and} \qquad \begin{aligned} x_1 + x_2 - x_3 &= 1 \\ x_1 + x_2 - x_3 &= 2. \end{aligned}$$

The first system has infinitely many solutions. Let $A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \end{pmatrix}$, the coefficient matrix of the system, and let $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. By Example 6,

$$A^\dagger = \frac{1}{6}\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix},$$

and therefore

$$z = A^\dagger b = \frac{1}{6}\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$

is the solution of minimal norm by Theorem 6.30(a).

The second system is obviously inconsistent. Let $b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. Thus, although

$$z = A^\dagger b = \frac{1}{6}\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$

is not a solution to the second system, it is the "best approximation" to a solution having minimum norm, as described in Theorem 6.30(b). ♦

EXERCISES

1. Label the following statements as true or false.

(a) The singular values of any linear operator on a finite-dimensional vector space are also eigenvalues of the operator.
(b) The singular values of any matrix A are the eigenvalues of A^*A.
(c) For any matrix A and any scalar c, if σ is a singular value of A, then $|c|\sigma$ is a singular value of cA.
(d) The singular values of any linear operator are nonnegative.
(e) If λ is an eigenvalue of a self-adjoint matrix A, then λ is a singular value of A.
(f) For any $m \times n$ matrix A and any $b \in \mathsf{F}^n$, the vector $A^\dagger b$ is a solution to $Ax = b$.
(g) The pseudoinverse of any linear operator exists even if the operator is not invertible.

2. Let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be a linear transformation of rank r, where V and W are finite-dimensional inner product spaces. In each of the following, find orthonormal bases $\{v_1, v_2, \ldots, v_n\}$ for V and $\{u_1, u_2, \ldots, u_m\}$ for W, and the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ of T such that $\mathsf{T}(v_i) = \sigma_i u_i$ for $1 \leq i \leq r$.

(a) $\mathsf{T}: \mathsf{R}^2 \to \mathsf{R}^3$ defined by $\mathsf{T}(x_1, x_2) = (x_1, x_1 + x_2, x_1 - x_2)$
(b) $\mathsf{T}: \mathsf{P}_2(R) \to \mathsf{P}_1(R)$, where $\mathsf{T}(f(x)) = f''(x)$, and the inner products are defined as in Example 1
(c) Let $\mathsf{V} = \mathsf{W} = \text{span}(\{1, \sin x, \cos x\})$ with the inner product defined by $\langle f, g \rangle = \int_0^{2\pi} f(t)g(t)\,dt$, and T is defined by $\mathsf{T}(f) = f' + 2f$
(d) $\mathsf{T}: \mathsf{C}^2 \to \mathsf{C}^2$ defined by $\mathsf{T}(z_1, z_2) = ((1-i)z_2, (1+i)z_1 + z_2)$

3. Find a singular value decomposition for each of the following matrices.

(a) $\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}$ (b) $\begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix}$ (c) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$ (e) $\begin{pmatrix} 1+i & 1 \\ 1-i & -i \end{pmatrix}$ (f) $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & -2 & 1 \\ 1 & -1 & 1 & 1 \end{pmatrix}$

4. Find a polar decomposition for each of the following matrices.

(a) $\begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}$ (b) $\begin{pmatrix} 20 & 4 & 0 \\ 0 & 0 & 1 \\ 4 & 20 & 0 \end{pmatrix}$

5. Find an explicit formula for each of the following expressions.

(a) $\mathsf{T}^\dagger(x_1, x_2, x_3)$, where T is the linear transformation of Exercise 2(a)
(b) $\mathsf{T}^\dagger(a + bx + cx^2)$, where T is the linear transformation of Exercise 2(b)

(c) $\mathsf{T}^\dagger(a + b\sin x + c\cos x)$, where T is the linear transformation of Exercise 2(c)

(d) $\mathsf{T}^\dagger(z_1, z_2)$, where T is the linear transformation of Exercise 2(d)

6. Use the results of Exercise 3 to find the pseudoinverse of each of the following matrices.

(a) $\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}$ **(b)** $\begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix}$ **(c)** $\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$ **(e)** $\begin{pmatrix} 1+i & 1 \\ 1-i & -i \end{pmatrix}$ **(f)** $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & -2 & 1 \\ 1 & -1 & 1 & 1 \end{pmatrix}$

7. For each of the given linear transformations $\mathsf{T}: \mathsf{V} \to \mathsf{W}$,

(i) Describe the subspace Z_1 of V such that $\mathsf{T}^\dagger\mathsf{T}$ is the orthogonal projection of V on Z_1.

(ii) Describe the subspace Z_2 of W such that $\mathsf{T}\mathsf{T}^\dagger$ is the orthogonal projection of W on Z_2.

(a) T is the linear transformation of Exercise 2(a)
(b) T is the linear transformation of Exercise 2(b)
(c) T is the linear transformation of Exercise 2(c)
(d) T is the linear transformation of Exercise 2(d)

8. For each of the given systems of linear equations,

(i) If the system is consistent, find the unique solution having minimum norm.

(ii) If the system is inconsistent, find the "best approximation to a solution" having minimum norm, as described in Theorem 6.30(b).

(Use your answers to parts (a) and (f) of Exercise 6.)

(a)
$$\begin{aligned} x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \\ -x_1 + -x_2 &= 0 \end{aligned}$$

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

9. Let V and W be finite-dimensional inner product spaces over F, and suppose that $\{v_1, v_2, \ldots, v_n\}$ and $\{u_1, u_2, \ldots, u_m\}$ are orthonormal bases for V and W, respectively. Let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be a linear transformation of rank r, and suppose that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are such that

$$\mathsf{T}(v_i) = \begin{cases} \sigma_i u_i & \text{if } 1 \leq i \leq r \\ 0 & \text{if } r < i. \end{cases}$$

(a) Prove that $\{u_1, u_2, \ldots, u_m\}$ is a set of eigenvectors of TT^* with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$, where

$$\lambda_i = \begin{cases} \sigma_i^2 & \text{if } 1 \le i \le r \\ 0 & \text{if } r < i. \end{cases}$$

(b) Let A be an $m \times n$ matrix with real or complex entries. Prove that the nonzero singular values of A are the positive square roots of the nonzero eigenvalues of AA^*, including repetitions.

(c) Prove that TT^* and T^*T have the same nonzero eigenvalues, including repetitions.

(d) State and prove a result for matrices analogous to (c).

10. Use Exercise 8 of Section 2.5 to obtain another proof of Theorem 6.27, the singular value decomposition theorem for matrices.

11. This exercise relates the singular values of a well-behaved linear operator or matrix to its eigenvalues.

(a) Let T be a normal linear operator on an n-dimensional inner product space with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Prove that the singular values of T are $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|$.

(b) State and prove a result for matrices analogous to (a).

12. Let A be a normal matrix with an orthonormal basis of eigenvectors $\beta = \{v_1, v_2, \ldots, v_n\}$ and corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Let V be the $n \times n$ matrix whose columns are the vectors in β. Prove that for each i there is a scalar θ_i of absolute value 1 such that if U is the $n \times n$ matrix with $\theta_i v_i$ as column i and Σ is the diagonal matrix such that $\Sigma_{ii} = |\lambda_i|$ for each i, then $U\Sigma V^*$ is a singular value decomposition of A.

13. Prove that if A is a positive semidefinite matrix, then the singular values of A are the same as the eigenvalues of A.

14. Prove that if A is a positive definite matrix and $A = U\Sigma V^*$ is a singular value decomposition of A, then $U = V$.

15. Let A be a square matrix with a polar decomposition $A = WP$.

(a) Prove that A is normal if and only if $WP^2 = P^2W$.

(b) Use (a) to prove that A is normal if and only if $WP = PW$.

16. Let A be a square matrix. Prove an alternate form of the polar decomposition for A: There exists a unitary matrix W and a positive semidefinite matrix P such that $A = PW$.

17. Let T and U be linear operators on R^2 defined for all $(x_1, x_2) \in \mathsf{R}^2$ by

$$\mathsf{T}(x_1, x_2) = (x_1, 0) \text{ and } \mathsf{U}(x_1, x_2) = (x_1 + x_2, 0).$$

(a) Prove that $(\mathsf{UT})^\dagger \neq \mathsf{T}^\dagger\mathsf{U}^\dagger$.
(b) Exhibit matrices A and B such that AB is defined, but $(AB)^\dagger \neq B^\dagger A^\dagger$.

18. Let A be an $m \times n$ matrix. Prove the following results.

(a) For any $m \times m$ unitary matrix G, $(GA)^\dagger = A^\dagger G^*$.
(b) For any $n \times n$ unitary matrix H, $(AH)^\dagger = H^* A^\dagger$.

19. Let A be a matrix with real or complex entries. Prove the following results.

(a) The nonzero singular values of A are the same as the nonzero singular values of A^*, which are the same as the nonzero singular values of A^t.
(b) $(A^\dagger)^* = (A^*)^\dagger$.
(c) $(A^\dagger)^t = (A^t)^\dagger$.

20. Let A be a square matrix such that $A^2 = O$. Prove that $(A^\dagger)^2 = O$.

21. Let V and W be finite-dimensional inner product spaces, and let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be linear. Prove the following results.

(a) $\mathsf{TT}^\dagger\mathsf{T} = \mathsf{T}$.
(b) $\mathsf{T}^\dagger\mathsf{TT}^\dagger = \mathsf{T}^\dagger$.
(c) Both $\mathsf{T}^\dagger\mathsf{T}$ and $\mathsf{TT}^\dagger$ are self-adjoint.

Visit `goo.gl/Dz3WQE` for a solution. The preceding three statements are called the **Penrose conditions**, and they characterize the pseudoinverse of a linear transformation as shown in Exercise 22.

22. Let V and W be finite-dimensional inner product spaces. Let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ and $\mathsf{U}: \mathsf{W} \to \mathsf{V}$ be linear transformations such that $\mathsf{TUT} = \mathsf{T}$, $\mathsf{UTU} = \mathsf{U}$, and both UT and TU are self-adjoint. Prove that $\mathsf{U} = \mathsf{T}^\dagger$.

23. State and prove a result for matrices that is analogous to the result of Exercise 21.

24. State and prove a result for matrices that is analogous to the result of Exercise 22.

25. Let V and W be finite-dimensional inner product spaces, and let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be linear. Prove the following results.

(a) If T is one-to-one, then $\mathsf{T}^*\mathsf{T}$ is invertible and $\mathsf{T}^\dagger = (\mathsf{T}^*\mathsf{T})^{-1}\mathsf{T}^*$.
(b) If T is onto, then TT^* is invertible and $\mathsf{T}^\dagger = \mathsf{T}^*(\mathsf{TT}^*)^{-1}$.

26. Let V and W be finite-dimensional inner product spaces with orthonormal bases β and γ, respectively, and let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be linear. Prove that $([\mathsf{T}]_\beta^\gamma)^\dagger = [\mathsf{T}^\dagger]_\gamma^\beta$.

27. Let V and W be finite-dimensional inner product spaces, and let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be a linear transformation. Prove part (b) of the lemma to Theorem 6.30: $\mathsf{TT}^\dagger$ is the orthogonal projection of W on $\mathsf{R}(\mathsf{T})$.

6.8* BILINEAR AND QUADRATIC FORMS

There is a certain class of scalar-valued functions of two variables defined on a vector space that arises in the study of such diverse subjects as geometry and multivariable calculus. This is the class of *bilinear forms*. We study the basic properties of this class with a special emphasis on symmetric bilinear forms, and we consider some of its applications to quadratic surfaces and multivariable calculus. In this section, F denotes any field that does not have characteristic two, as defined on page 549.

Bilinear Forms

Definition. *Let V be a vector space over a field F. A function H from the set $\mathsf{V} \times \mathsf{V}$ of ordered pairs of vectors to F is called a* **bilinear form** *on V if H is linear in each variable when the other variable is held fixed; that is, H is a bilinear form on V if*

(a) $H(ax_1 + x_2, y) = aH(x_1, y) + H(x_2, y)$ for all $x_1, x_2, y \in \mathsf{V}$ and $a \in F$

(b) $H(x, ay_1 + y_2) = aH(x, y_1) + H(x, y_2)$ for all $x, y_1, y_2 \in \mathsf{V}$ and $a \in F$.

We denote the set of all bilinear forms on V by $\mathcal{B}(\mathsf{V})$. Observe that an inner product on a vector space is a bilinear form if the underlying field is real, but not if the underlying field is complex.

Example 1

Define a function $H: \mathsf{R}^2 \times \mathsf{R}^2 \to R$ by

$$H\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right) = 2a_1b_1 + 3a_1b_2 + 4a_2b_1 - a_2b_2 \quad \text{for } \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \in \mathsf{R}^2.$$

We could verify directly that H is a bilinear form on R^2. However, it is more enlightening and less tedious to observe that if

$$A = \begin{pmatrix} 2 & 3 \\ 4 & -1 \end{pmatrix}, \quad x = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad \text{and} \quad y = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

then

$$H(x, y) = x^t A y.$$

The bilinearity of H now follows directly from the distributive property of matrix multiplication over matrix addition. ◆

The preceding bilinear form is a special case of the next example.

Example 2

Let $\mathsf{V} = \mathsf{F}^n$, where the vectors are considered as column vectors. For any $A \in \mathsf{M}_{n\times n}(F)$, define $H: \mathsf{V} \times \mathsf{V} \to F$ by

$$H(x, y) = x^t Ay \quad \text{for } x, y \in \mathsf{V}.$$

Notice that since x and y are $n \times 1$ matrices and A is an $n \times n$ matrix, $H(x, y)$ is a 1×1 matrix. We identify this matrix with its single entry. The bilinearity of H follows as in Example 1. For example, for $a \in F$ and $x_1, x_2, y \in \mathsf{V}$, we have

$$\begin{aligned} H(ax_1 + x_2, y) &= (ax_1 + x_2)^t Ay = (ax_1^t + x_2^t)Ay \\ &= ax_1^t Ay + x_2^t Ay \\ &= aH(x_1, y) + H(x_2, y). \end{aligned}$$ ◆

We list several properties possessed by all bilinear forms. Their proofs are left to the reader (see Exercise 2).

For any bilinear form H on a vector space V over a field F, the following properties hold.

1. If, for any $x \in \mathsf{V}$, the functions $\mathsf{L}_x, \mathsf{R}_x: \mathsf{V} \to F$ are defined by

$$\mathsf{L}_x(y) = H(x, y) \quad \text{and} \quad \mathsf{R}_x(y) = H(y, x) \quad \text{for all } y \in \mathsf{V},$$

   then L_x and R_x are linear.
2. $H(0, x) = H(x, 0) = 0$ for all $x \in \mathsf{V}$.
3. For all $x, y, z, w \in \mathsf{V}$,

$$H(x + y, z + w) = H(x, z) + H(x, w) + H(y, z) + H(y, w).$$

4. If $J: \mathsf{V} \times \mathsf{V} \to F$ is defined by $J(x, y) = H(y, x)$, then J is a bilinear form.

Definitions. *Let V be a vector space, let H_1 and H_2 be bilinear forms on V, and let a be a scalar. We define the* **sum** *$H_1 + H_2$ and the* ***scalar product*** *aH_1 by the equations*

$$(H_1 + H_2)(x, y) = H_1(x, y) + H_2(x, y)$$

and

$$(aH_1)(x, y) = a(H_1(x, y)) \quad \text{for all } x, y \in \mathsf{V}.$$

The following theorem is an immediate consequence of the definitions.

Theorem 6.31. *For any vector space* V*, the sum of two bilinear forms and the product of a scalar and a bilinear form on* V *are again bilinear forms on* V*. Furthermore,* $\mathcal{B}(\mathsf{V})$ *is a vector space with respect to these operations.*

Proof. Exercise. ∎

Let $\beta = \{v_1, v_2, \ldots, v_n\}$ be an ordered basis for an n-dimensional vector space V, and let $H \in \mathcal{B}(\mathsf{V})$. We can associate with H an $n \times n$ matrix A whose entry in row i and column j is defined by

$$A_{ij} = H(v_i, v_j) \qquad \text{for } i, j = 1, 2, \ldots, n.$$

Definition. *The matrix* A *above is called the* ***matrix representation*** *of* H *with respect to the ordered basis* β *and is denoted by* $\psi_\beta(H)$*.*

We can therefore regard ψ_β as a mapping from $\mathcal{B}(\mathsf{V})$ to $\mathsf{M}_{n\times n}(F)$, where F is the field of scalars for V, that takes a bilinear form H into its matrix representation $\psi_\beta(H)$. We first consider an example and then show that ψ_β is an isomorphism.

Example 3

Consider the bilinear form H of Example 1, and let

$$\beta = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\} \quad \text{and} \quad B = \psi_\beta(H).$$

Then

$$B_{11} = H\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) = 2 + 3 + 4 - 1 = 8,$$

$$B_{12} = H\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right) = 2 - 3 + 4 + 1 = 4,$$

$$B_{21} = H\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) = 2 + 3 - 4 + 1 = 2,$$

and

$$B_{22} = H\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right) = 2 - 3 - 4 - 1 = -6.$$

So

$$\psi_\beta(H) = \begin{pmatrix} 8 & 4 \\ 2 & -6 \end{pmatrix}.$$

If γ is the standard ordered basis for R^2, the reader can verify that

$$\psi_\gamma(H) = \begin{pmatrix} 2 & 3 \\ 4 & -1 \end{pmatrix}. \quad \blacklozenge$$

Theorem 6.32. *For any n-dimensional vector space V over F and any ordered basis β for V, $\psi_\beta\colon \mathcal{B}(\mathsf{V}) \to \mathsf{M}_{n\times n}(F)$ is an isomorphism.*

Proof. We leave the proof that ψ_β is linear to the reader.

To show that ψ_β is one-to-one, suppose that $\psi_\beta(H) = O$ for some $H \in \mathcal{B}(\mathsf{V})$. Fix $v_i \in \beta$, and recall the mapping $\mathsf{L}_{v_i}\colon \mathsf{V} \to F$, which is linear by property 1 on page 420. By hypothesis, $\mathsf{L}_{v_i}(v_j) = H(v_i, v_j) = 0$ for all $v_j \in \beta$. Hence L_{v_i} is the zero transformation from V to F. So

$$H(v_i, x) = \mathsf{L}_{v_i}(x) = 0 \quad \text{for all } x \in \mathsf{V} \text{ and } v_i \in \beta. \tag{7}$$

Next fix an arbitrary $y \in \mathsf{V}$, and recall the linear mapping $\mathsf{R}_y\colon \mathsf{V} \to F$ defined in property 1 on page 420. By (7), $\mathsf{R}_y(v_i) = H(v_i, y) = 0$ for all $v_i \in \beta$, and hence R_y is the zero transformation. So $H(x, y) = \mathsf{R}_y(x) = 0$ for all $x, y \in \mathsf{V}$. Thus H is the zero bilinear form, and therefore ψ_β is one-to-one.

To show that ψ_β is onto, consider any $A \in \mathsf{M}_{n\times n}(F)$. Recall the isomorphism $\phi_\beta\colon \mathsf{V} \to \mathsf{F}^n$ defined in Section 2.4. For $x \in \mathsf{V}$, we view $\phi_\beta(x) \in \mathsf{F}^n$ as a column vector. Let $H\colon \mathsf{V} \times \mathsf{V} \to F$ be the mapping defined by

$$H(x, y) = [\phi_\beta(x)]^t A[\phi_\beta(y)] \quad \text{for all } x, y \in \mathsf{V}.$$

A slight embellishment of the method of Example 2 can be used to prove that $H \in \mathcal{B}(\mathsf{V})$. We show that $\psi_\beta(H) = A$. Let $v_i, v_j \in \beta$. Then $\phi_\beta(v_i) = e_i$ and $\phi_\beta(v_j) = e_j$; hence, for any i and j,

$$H(v_i, v_j) = [\phi_\beta(v_i)]^t A[\phi_\beta(v_j)] = e_i^t A e_j = A_{ij}.$$

We conclude that $\psi_\beta(H) = A$ and ψ_β is onto. ∎

Corollary 1. *For any n-dimensional vector space V, $\mathcal{B}(\mathsf{V})$ has dimension n^2.*

Proof. Exercise. ∎

The following corollary is easily established by reviewing the proof of Theorem 6.32.

Corollary 2. *Let V be an n-dimensional vector space over F with ordered basis β. If $H \in \mathcal{B}(\mathsf{V})$ and $A \in \mathsf{M}_{n\times n}(F)$, then $\psi_\beta(H) = A$ if and only if $H(x, y) = [\phi_\beta(x)]^t A[\phi_\beta(y)]$ for all $x, y \in \mathsf{V}$.*

The following result is now an immediate consequence of Corollary 2.

Corollary 3. *Let F be a field, n be a positive integer, and β be the standard ordered basis for F^n. Then for any $H \in \mathcal{B}(\mathsf{F}^n)$, there exists a unique matrix $A \in \mathsf{M}_{n\times n}(F)$, namely, $A = \psi_\beta(H)$, such that*

$$H(x, y) = x^t A y \quad \text{for all } x, y \in \mathsf{F}^n.$$

Example 4

Define a function $H: \mathsf{R}^2 \times \mathsf{R}^2 \to R$ by

$$H\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right) = \det \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} = a_1 b_2 - a_2 b_1 \quad \text{for } \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \in \mathsf{R}^2.$$

It can be shown that H is a bilinear form. We find the matrix A in Corollary 3 such that $H(x, y) = x^t A y$ for all $x, y \in \mathsf{R}^2$.

Since $A_{ij} = H(e_i, e_j)$ for all i and j, we have

$$A_{11} = \det \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = 0 \qquad A_{12} = \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1,$$

$$A_{21} = \det \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -1 \quad \text{and} \quad A_{22} = \det \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} = 0.$$

Therefore $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. ◆

There is an analogy between bilinear forms and linear operators on finite-dimensional vector spaces in that both are associated with unique square matrices and the correspondences depend on the choice of an ordered basis for the vector space. As in the case of linear operators, one can pose the following question: How does the matrix corresponding to a fixed bilinear form change when the ordered basis is changed? As we have seen, the corresponding question for matrix representations of linear operators leads to the definition of the similarity relation on square matrices. In the case of bilinear forms, the corresponding question leads to another relation on square matrices, the *congruence* relation.

Definition. *Let $A, B \in \mathsf{M}_{n\times n}(F)$. Then B is said to be* **congruent** *to A if there exists an invertible matrix $Q \in \mathsf{M}_{n\times n}(F)$ such that $B = Q^t A Q$.*

Observe that the relation of congruence is an equivalence relation (see Exercise 12).

The next theorem relates congruence to the matrix representation of a bilinear form.

Theorem 6.33. *Let V be a finite-dimensional vector space with ordered bases $\beta = \{v_1, v_2, \ldots, v_n\}$ and $\gamma = \{w_1, w_2, \ldots, w_n\}$, and let Q be the change of coordinate matrix changing γ-coordinates into β-coordinates. Then, for any $H \in \mathcal{B}(\mathsf{V})$, we have $\psi_\gamma(H) = Q^t \psi_\beta(H) Q$. Therefore $\psi_\gamma(H)$ is congruent to $\psi_\beta(H)$.*

Proof. There are essentially two proofs of this theorem. One involves a direct computation, while the other follows immediately from a clever observation. We give the more direct proof here, leaving the other proof for the exercises (see Exercise 13).

Suppose that $A = \psi_\beta(H)$ and $B = \psi_\gamma(H)$. Then for $1 \le i, j \le n$,

$$w_i = \sum_{k=1}^{n} Q_{ki} v_k \quad \text{and} \quad w_j = \sum_{r=1}^{n} Q_{rj} v_r.$$

Thus

$$\begin{aligned} B_{ij} = H(w_i, w_j) &= H\left(\sum_{k=1}^{n} Q_{ki} v_k, w_j\right) \\ &= \sum_{k=1}^{n} Q_{ki} H(v_k, w_j) \\ &= \sum_{k=1}^{n} Q_{ki} H\left(v_k, \sum_{r=1}^{n} Q_{rj} v_r\right) \\ &= \sum_{k=1}^{n} Q_{ki} \sum_{r=1}^{n} Q_{rj} H(v_k, v_r) \\ &= \sum_{k=1}^{n} Q_{ki} \sum_{r=1}^{n} Q_{rj} A_{kr} \\ &= \sum_{k=1}^{n} Q_{ki} \sum_{r=1}^{n} A_{kr} Q_{rj} \\ &= \sum_{k=1}^{n} Q_{ki} (AQ)_{kj} \\ &= \sum_{k=1}^{n} Q_{ik}^t (AQ)_{kj} = (Q^t AQ)_{ij}. \end{aligned}$$

Hence $B = Q^t AQ$. ∎

The following result is the converse of Theorem 6.33.

Corollary. *Let V be an n-dimensional vector space with ordered basis β, and let H be a bilinear form on V. For any $n \times n$ matrix B, if B is congruent to $\psi_\beta(H)$, then there exists an ordered basis γ for V such that $\psi_\gamma(H) = B$. In this case, if Q is a matrix such that $B = Q^t \psi_\beta(H) Q$, then Q changes γ-coordinates into β-coordinates.*

Proof. Suppose that $B = Q^t \psi_\beta(H) Q$ for some invertible matrix Q and that $\beta = \{v_1, v_2, \ldots, v_n\}$. Let $\gamma = \{w_1, w_2, \ldots, w_n\}$, where

$$w_j = \sum_{i=1}^{n} Q_{ij} v_i \quad \text{for } 1 \le j \le n.$$

Since Q is invertible, γ is an ordered basis for V, and Q is the change of coordinate matrix that changes γ-coordinates into β-coordinates. Therefore, by Theorem 6.33,

$$B = Q^t \psi_\beta(H) Q = \psi_\gamma(H). \qquad \blacksquare$$

Symmetric Bilinear Forms

Like the diagonalization problem for linear operators, there is an analogous *diagonalization* problem for bilinear forms, namely, the problem of determining those bilinear forms for which there are diagonal matrix representations. As we will see, there is a close relationship between *diagonalizable* bilinear forms and those that are called *symmetric*.

Definition. *A bilinear form H on a vector space V is* ***symmetric*** *if $H(x, y) = H(y, x)$ for all $x, y \in \mathsf{V}$.*

As the name suggests, symmetric bilinear forms correspond to symmetric matrices.

Theorem 6.34. *Let H be a bilinear form on a finite-dimensional vector space V, and let β be an ordered basis for V. Then H is symmetric if and only if $\psi_\beta(H)$ is symmetric.*

Proof. Let $\beta = \{v_1, v_2, \ldots, v_n\}$ and $B = \psi_\beta(H)$.

First assume that H is symmetric. Then for $1 \le i, j \le n$,

$$B_{ij} = H(v_i, v_j) = H(v_j, v_i) = B_{ji},$$

and it follows that B is symmetric.

Conversely, suppose that B is symmetric. Let $J\colon \mathsf{V} \times \mathsf{V} \to F$, where F is the field of scalars for V, be the mapping defined by $J(x, y) = H(y, x)$ for all $x, y \in \mathsf{V}$. By property 4 on page 420, J is a bilinear form. Let $C = \psi_\beta(J)$. Then, for $1 \le i, j \le n$,

$$C_{ij} = J(v_i, v_j) = H(v_j, v_i) = B_{ji} = B_{ij}.$$

Thus $C = B$. Since ψ_β is one-to-one, we have $J = H$. Hence $H(y, x) = J(x, y) = H(x, y)$ for all $x, y \in \mathsf{V}$, and therefore H is symmetric. $\blacksquare$

Definition. *A bilinear form H on a finite-dimensional vector space V is called* ***diagonalizable*** *if there is an ordered basis β for V such that $\psi_\beta(H)$ is a diagonal matrix.*

Corollary. *Let H be a diagonalizable bilinear form on a finite-dimensional vector space V. Then H is symmetric.*

Proof. Suppose that H is diagonalizable. Then there is an ordered basis β for V such that $\psi_\beta(H) = D$ is a diagonal matrix. Trivially, D is a symmetric matrix, and hence, by Theorem 6.34, H is symmetric. ∎

Unfortunately, the converse is not true, as is illustrated by the following example.

Example 5

Let $F = Z_2$, $\mathsf{V} = \mathsf{F}^2$, and $H: \mathsf{V} \times \mathsf{V} \to F$ be the bilinear form defined by

$$H\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right) = a_1 b_2 + a_2 b_1.$$

Clearly H is symmetric. In fact, if β is the standard ordered basis for V, then

$$A = \psi_\beta(H) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

a symmetric matrix. We show that H is not diagonalizable.

By way of contradiction, suppose that H is diagonalizable. Then there is an ordered basis γ for V such that $B = \psi_\gamma(H)$ is a diagonal matrix. So by Theorem 6.33, there exists an invertible matrix Q such that $B = Q^t A Q$. Since Q is invertible, it follows that $\operatorname{rank}(B) = \operatorname{rank}(A) = 2$, and consequently the diagonal entries of B are nonzero. Since the only nonzero scalar of F is 1,

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Suppose that

$$Q = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Then

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = B = Q^t A Q$$

$$= \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} ac + ac & bc + ad \\ bc + ad & bd + bd \end{pmatrix}.$$

But $p + p = 0$ for all $p \in F$; hence $ac + ac = 0$. Thus, comparing the row 1, column 1 entries of the matrices in the equation above, we conclude that $1 = 0$, a contradiction. Therefore H is not diagonalizable. ◆

The bilinear form of Example 5 is an anomaly. Its failure to be diagonalizable is due to the fact that the scalar field Z_2 is of characteristic two. Recall from Appendix C that a field F is of **characteristic two** if $1 + 1 = 0$ in F. If F is not of characteristic two, then $1 + 1 = 2$ has a multiplicative inverse, which we denote by $1/2$.

Before proving the converse of the corollary to Theorem 6.34 for scalar fields that are not of characteristic two, we establish the following lemma.

Lemma. *Let H be a nonzero symmetric bilinear form on a vector space V over a field F not of characteristic two. Then there is a vector x in V such that $H(x, x) \neq 0$.*

Proof. Since H is nonzero, we can choose vectors $u, v \in \mathsf{V}$ such that $H(u, v) \neq 0$. If $H(u, u) \neq 0$ or $H(v, v) \neq 0$, there is nothing to prove. Otherwise, set $x = u + v$. Then

$$H(x, x) = H(u, u) + H(u, v) + H(v, u) + H(v, v) = 2H(u, v) \neq 0$$

because $2 \neq 0$ and $H(u, v) \neq 0$. ∎

Theorem 6.35. *Let V be a finite-dimensional vector space over a field F not of characteristic two. Then every symmetric bilinear form on V is diagonalizable.*

Proof. We use mathematical induction on $n = \dim(\mathsf{V})$. If $n = 1$, then every element of $\mathcal{B}(\mathsf{V})$ is diagonalizable. Now suppose that the theorem is valid for vector spaces of dimension less than n for some fixed integer $n > 1$, and suppose that $\dim(\mathsf{V}) = n$. If H is the zero bilinear form on V, then trivially H is diagonalizable; so suppose that H is a nonzero symmetric bilinear form on V. By the lemma, there exists a nonzero vector x in V such that $H(x, x) \neq 0$. Recall the function $\mathsf{L}_x : \mathsf{V} \to F$ defined by $\mathsf{L}_x(y) = H(x, y)$ for all $y \in \mathsf{V}$. By property 1 on page 420, L_x is linear. Furthermore, since $\mathsf{L}_x(x) = H(x, x) \neq 0$, L_x is nonzero. Consequently, $\text{rank}(\mathsf{L}_x) = 1$, and hence $\dim(\mathsf{N}(\mathsf{L}_x)) = n - 1$.

The restriction of H to $\mathsf{N}(\mathsf{L}_x)$ is obviously a symmetric bilinear form on a vector space of dimension $n - 1$. Thus, by the induction hypothesis, there exists an ordered basis $\{v_1, v_2, \ldots, v_{n-1}\}$ for $\mathsf{N}(\mathsf{L}_x)$ such that $H(v_i, v_j) = 0$ for $i \neq j$ $(1 \leq i, j \leq n - 1)$. Set $v_n = x$. Then $v_n \notin \mathsf{N}(\mathsf{L}_x)$, and so $\beta = \{v_1, v_2, \ldots, v_n\}$ is an ordered basis for V. In addition, $H(v_i, v_n) = H(v_n, v_i) = 0$ for $i = 1, 2, \ldots, n - 1$. We conclude that $\psi_\beta(H)$ is a diagonal matrix, and therefore H is diagonalizable. ∎

Corollary. *Let F be a field that is not of characteristic two. If $A \in \mathsf{M}_{n \times n}(F)$ is a symmetric matrix, then A is congruent to a diagonal matrix.*

Proof. Exercise. ∎

Diagonalization of Symmetric Matrices

Let A be a symmetric $n \times n$ matrix with entries from a field F not of characteristic two. By the corollary to Theorem 6.35, there are matrices $Q, D \in \mathsf{M}_{n \times n}(F)$ such that Q is invertible, D is diagonal, and $Q^t A Q = D$. We now give a method for computing Q and D. This method requires familiarity with elementary matrices and their properties, which the reader may wish to review in Section 3.1.

If E is an elementary $n \times n$ matrix, then AE can be obtained by performing an elementary column operation on A. By Exercise 21, $E^t A$ can be obtained by performing the same operation on the rows of A rather than on its columns. Thus $E^t AE$ can be obtained from A by performing an elementary operation on the columns of A and then performing the same operation on the rows of AE. (Note that the order of the operations can be reversed because of the associative property of matrix multiplication.) Suppose that Q is an invertible matrix and D is a diagonal matrix such that $Q^t AQ = D$. By Corollary 3 to Theorem 3.6 (p. 158), Q is a product of elementary matrices, say $Q = E_1 E_2 \cdots E_k$. Thus

$$D = Q^t AQ = E_k^t E_{k-1}^t \cdots E_1^t A E_1 E_2 \cdots E_k.$$

From the preceding equation, we conclude that *by means of several elementary column operations and the corresponding row operations, A can be transformed into a diagonal matrix D. Furthermore, if $E_1, E_2, \ldots, E_k$ are the elementary matrices corresponding to these elementary column operations indexed in the order performed, and if $Q = E_1 E_2 \cdots E_k$, then $Q^t AQ = D$.*

Example 6

Let A be the symmetric matrix in $\mathsf{M}_{3\times 3}(R)$ defined by

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

We use the procedure just described to find an invertible matrix Q and a diagonal matrix D such that $Q^t AQ = D$.

We begin by eliminating all of the nonzero entries in the first row and first column except for the entry in column 1 and row 1. To this end, we add the first column of A to the second column to produce a zero in row 1 and column 2. The elementary matrix that corresponds to this elementary column operation is

$$E_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We perform the corresponding elementary operation on the rows of AE_1 to obtain

$$E_1^t AE_1 = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 4 \\ 3 & 4 & 1 \end{pmatrix}.$$

We now use the first column of $E_1^t AE_1$ to eliminate the 3 in row 1 column 3, and follow this operation with the corresponding row operation. The corresponding elementary matrix E_2 and the result of the elementary operations

$E_2^t E_1^t A E_1 E_2$ are, respectively,

$$E_2 = \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad E_2^t E_1^t A E_1 E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 4 & -8 \end{pmatrix}.$$

Finally, we subtract 4 times the second column of $E_2^t E_1^t A E_1 E_2$ from the third column and follow this with the corresponding row operation. The corresponding elementary matrix E_3 and the result of the elementary operations $E_3^t E_2^t E_1^t A E_1 E_2 E_3$ are, respectively,

$$E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -4 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad E_3^t E_2^t E_1^t A E_1 E_2 E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -24 \end{pmatrix}.$$

Since we have obtained a diagonal matrix, the process is complete. So we let

$$Q = E_1 E_2 E_3 = \begin{pmatrix} 1 & 1 & -7 \\ 0 & 1 & -4 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -24 \end{pmatrix}$$

to obtain the desired diagonalization $Q^t A Q = D$. ◆

The reader should justify the following method for computing Q without recording each elementary matrix separately. The method is inspired by the algorithm for computing the inverse of a matrix developed in Section 3.2. We use a sequence of elementary column operations and corresponding row operations to change the $n \times 2n$ matrix $(A|I)$ into the form $(D|B)$, where D is a diagonal matrix and $B = Q^t$. It then follows that $D = Q^t A Q$.

Starting with the matrix A of the preceding example, this method produces the following sequence of matrices:

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & -1 & 3 & 1 & 0 & 0 \\ -1 & 2 & 1 & 0 & 1 & 0 \\ 3 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 3 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 & 1 & 0 \\ 3 & 4 & 1 & 0 & 0 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 3 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 3 & 4 & 1 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 3 & 4 & -8 & 0 & 0 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 1 & 0 \\ 0 & 4 & -8 & -3 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 4 & -24 & -3 & 0 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & -24 & -7 & -4 & 1 \end{array}\right) = (D|Q^t).$$

Therefore

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -24 \end{pmatrix}, \quad Q^t = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -7 & -4 & 1 \end{pmatrix}, \quad \text{and} \quad Q = \begin{pmatrix} 1 & 1 & -7 \\ 0 & 1 & -4 \\ 0 & 0 & 1 \end{pmatrix}.$$

Quadratic Forms

Associated with symmetric bilinear forms are functions called *quadratic forms.*

Definition. *Let* V *be a vector space over a field* F *not of characteristic two.*[2] *A function* $K : \mathsf{V} \to F$ *is called a* ***quadratic form on*** V *if there exists a symmetric bilinear form* $H \in \mathcal{B}(\mathsf{V})$ *such that*

$$K(x) = H(x, x) \quad \text{for all } x \in \mathsf{V}. \tag{8}$$

There is a one-to-one correspondence between symmetric bilinear forms and quadratic forms given by (8). In fact, if K is a quadratic form on a vector space V over a field F not of characteristic two, and $K(x) = H(x, x)$ for some symmetric bilinear form H on V, then we can recover H from K because

$$H(x, y) = \frac{1}{2}[K(x + y) - K(x) - K(y)] \tag{9}$$

(See Exercise 16.)

Example 7

The classic example of a quadratic form is the homogeneous second-degree polynomial of several variables. Given the variables $t_1, t_2, \ldots, t_n$ that take values in a field F not of characteristic two and given (not necessarily distinct) scalars a_{ij} $(1 \le i \le j \le n)$, define the polynomial

$$f(t_1, t_2, \ldots, t_n) = \sum_{i \le j} a_{ij} t_i t_j.$$

Any such polynomial is a quadratic form. In fact, if β is the standard ordered basis for F^n, then the symmetric bilinear form H corresponding to the quadratic form f has the matrix representation $\psi_\beta(H) = A$, where

$$A_{ij} = A_{ji} = \begin{cases} a_{ii} & \text{if } i = j \\ \frac{1}{2} a_{ij} & \text{if } i \ne j. \end{cases}$$

[2] Although it is possible to define quadratic forms over a field of characteristic 2, our definition can't be used in this case.

To see this, apply (9) to obtain $H(e_i, e_j) = A_{ij}$ from the quadratic form K, and verify that f is computable from H by (8) using f in place of K.

For example, given the polynomial

$$f(t_1, t_2, t_3) = 2t_1^2 - t_2^2 + 6t_1t_2 - 4t_2t_3$$

with real coefficients, let

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

Setting $H(x, y) = x^t Ay$ for all $x, y \in \mathsf{R}^3$, we see that

$$f(t_1, t_2, t_3) = (t_1, t_2, t_3)A \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} \quad \text{for } \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} \in \mathsf{R}^3. \quad \blacklozenge$$

Quadratic Forms on a Real Inner Product Space

Since symmetric matrices over R are *orthogonally diagonalizable* (see Theorem 6.20 p. 381), the theory of symmetric bilinear forms and quadratic forms on finite-dimensional vector spaces over R is especially nice. The following theorem and its corollary are useful.

Theorem 6.36. *Let V be a finite-dimensional real inner product space, and let H be a symmetric bilinear form on V. Then there exists an orthonormal basis β for V such that $\psi_\beta(H)$ is a diagonal matrix.*

Proof. Choose any orthonormal basis $\gamma = \{v_1, v_2, \ldots, v_n\}$ for V, and let $A = \psi_\gamma(H)$. Since A is symmetric, there exists an orthogonal matrix Q and a diagonal matrix D such that $D = Q^t AQ$ by Theorem 6.20. Let $\beta = \{w_1, w_2, \ldots, w_n\}$ be defined by

$$w_j = \sum_{i=1}^{n} Q_{ij} v_i \quad \text{for } 1 \le j \le n.$$

By Theorem 6.33, $\psi_\beta(H) = D$. Furthermore, since Q is orthogonal and γ is orthonormal, β is orthonormal by Exercise 30 of Section 6.5. ∎

Corollary. *Let K be a quadratic form on a finite-dimensional real inner product space V. There exists an orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$ for V and scalars $\lambda_1, \lambda_2, \ldots, \lambda_n$ (not necessarily distinct) such that if $x \in \mathsf{V}$ and*

$$x = \sum_{i=1}^{n} s_i v_i, \quad s_i \in R,$$

then

$$K(x) = \sum_{i=1}^{n} \lambda_i s_i^2.$$

In fact, if H is the symmetric bilinear form determined by K, then β can be chosen to be any orthonormal basis for V such that $\psi_\beta(H)$ is a diagonal matrix.

Proof. Let H be the symmetric bilinear form for which $K(x) = H(x, x)$ for all $x \in \mathsf{V}$. By Theorem 6.36, there exists an orthonormal basis $\beta = \{v_1, v_2, \ldots, v_n\}$ for V such that $\psi_\beta(H)$ is the diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

Let $x \in \mathsf{V}$, and suppose that $x = \sum_{i=1}^{n} s_i v_i$. Then

$$K(x) = H(x, x) = [\phi_\beta(x)]^t D[\phi_\beta(x)] = (s_1, s_2, \ldots, s_n) D \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} = \sum_{i=1}^{n} \lambda_i s_i^2. \quad \blacksquare$$

Example 8

For the homogeneous real polynomial of degree 2 defined by

$$f(t_1, t_2) = 5t_1^2 + 2t_2^2 + 4t_1 t_2, \tag{10}$$

we find an orthonormal basis $\gamma = \{x_1, x_2\}$ for R^2 and scalars λ_1 and λ_2 such that if

$$\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \in \mathsf{R}^2 \quad \text{and} \quad \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = s_1 x_1 + s_2 x_2,$$

then $f(t_1, t_2) = \lambda_1 s_1^2 + \lambda_2 s_2^2$. We can think of s_1 and s_2 as the coordinates of (t_1, t_2) relative to γ. Thus the polynomial $f(t_1, t_2)$, as an expression involving the coordinates of a point with respect to the standard ordered basis for R^2, is transformed into a new polynomial $g(s_1, s_2) = \lambda_1 s_1^2 + \lambda_2 s_2^2$ interpreted as an expression involving the coordinates of a point relative to the new ordered basis γ.

Let H denote the symmetric bilinear form corresponding to the quadratic form defined by (10), let β be the standard ordered basis for R^2, and let $A = \psi_\beta(H)$. Then

$$A = \psi_\beta(H) = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}.$$

Next, we find an orthogonal matrix Q such that Q^tAQ is a diagonal matrix. For this purpose, observe that $\lambda_1 = 6$ and $\lambda_2 = 1$ are the eigenvalues of A with corresponding orthonormal eigenvectors

$$v_1 = \frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix} \quad \text{and} \quad v_2 = \frac{1}{\sqrt{5}}\begin{pmatrix}1\\-2\end{pmatrix}.$$

Let $\gamma = \{v_1, v_2\}$. Then γ is an orthonormal basis for R^2 consisting of eigenvectors of A. Hence, setting

$$Q = \frac{1}{\sqrt{5}}\begin{pmatrix}2 & 1\\1 & -2\end{pmatrix},$$

we see that Q is an orthogonal matrix and

$$Q^tAQ = \begin{pmatrix}6 & 0\\0 & 1\end{pmatrix}.$$

Clearly Q is also a change of coordinate matrix. Consequently,

$$\psi_\gamma(H) = Q^t\psi_\beta(H)Q = Q^tAQ = \begin{pmatrix}6 & 0\\0 & 1\end{pmatrix}.$$

Thus by the corollary to Theorem 6.36,

$$K(x) = 6s_1^2 + s_2^2$$

for any $x = s_1v_1 + s_2v_2 \in \mathsf{R}^2$. So $g(s_1, s_2) = 6s_1^2 + s_2^2$. ◆

The next example illustrates how the theory of quadratic forms can be applied to the problem of describing quadratic surfaces in R^3.

Example 9

Let $\mathcal{S}$ be the surface in R^3 defined by the equation

$$2t_1^2 + 6t_1t_2 + 5t_2^2 - 2t_2t_3 + 2t_3^2 + 3t_1 - 2t_2 - t_3 + 14 = 0. \tag{11}$$

Then (11) describes the points of $\mathcal{S}$ in terms of their coordinates relative to β, the standard ordered basis for R^3. We find a new orthonormal basis γ for R^3 so that the equation describing the coordinates of $\mathcal{S}$ relative to γ is simpler than (11).

We begin with the observation that the terms of second degree on the left side of (11) add to form a quadratic form K on R^3:

$$K\begin{pmatrix}t_1\\t_2\\t_3\end{pmatrix} = 2t_1^2 + 6t_1t_2 + 5t_2^2 - 2t_2t_3 + 2t_3^2.$$

Next, we diagonalize K. Let H be the symmetric bilinear form corresponding to K, and let $A = \psi_\beta(H)$. Then

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

The characteristic polynomial of A is $(-1)(t-2)(t-7)t$; hence A has the eigenvalues $\lambda_1 = 2$, $\lambda_2 = 7$, and $\lambda_3 = 0$. Corresponding unit eigenvectors are

$$v_1 = \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \quad v_2 = \frac{1}{\sqrt{35}} \begin{pmatrix} 3 \\ 5 \\ -1 \end{pmatrix}, \quad \text{and} \quad v_3 = \frac{1}{\sqrt{14}} \begin{pmatrix} -3 \\ 2 \\ 1 \end{pmatrix}.$$

Set $\gamma = \{v_1, v_2, v_3\}$ and

$$Q = \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{35}} & \frac{-3}{\sqrt{14}} \\ 0 & \frac{5}{\sqrt{35}} & \frac{2}{\sqrt{14}} \\ \frac{3}{\sqrt{10}} & \frac{-1}{\sqrt{35}} & \frac{1}{\sqrt{14}} \end{pmatrix}.$$

As in Example 8, Q is a change of coordinate matrix changing γ-coordinates to β-coordinates, and

$$\psi_\gamma(H) = Q^t \psi_\beta(H) Q = Q^t A Q = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

By the corollary to Theorem 6.36, if $x = s_1 v_1 + s_2 v_2 + s_3 v_3$, then

$$K(x) = 2s_1^2 + 7s_2^2. \tag{12}$$

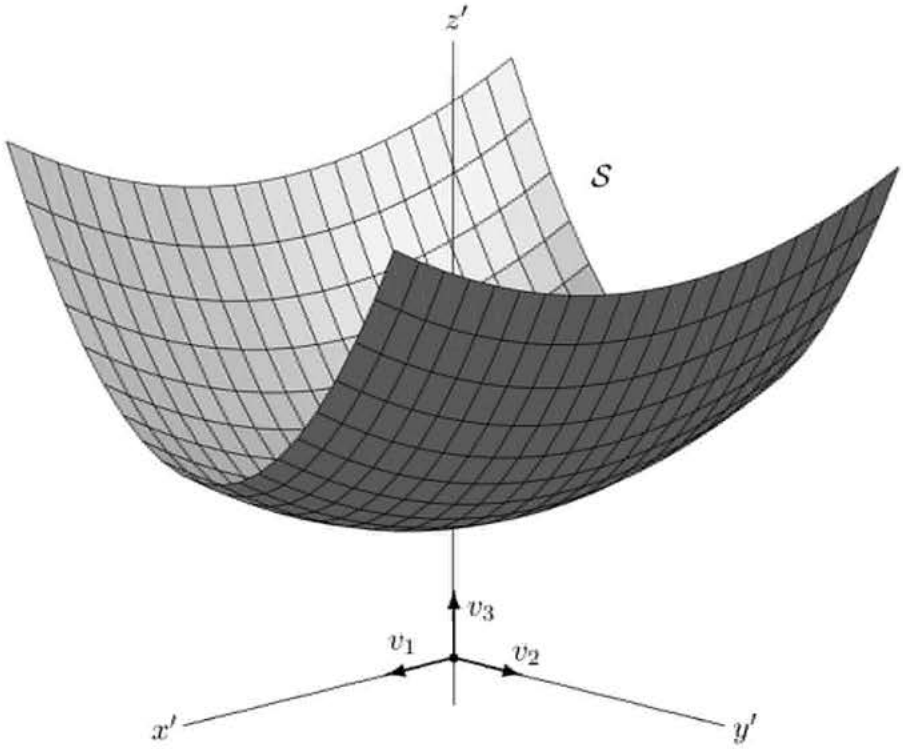
We are now ready to transform (11) into an equation involving coordinates relative to γ. Let $x = (t_1, t_2, t_3) \in \mathsf{R}^3$, and suppose that $x = s_1 v_1 + s_2 v_2 + s_3 v_3$. Then, by Theorem 2.22 (p. 112),

$$x = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} = Q \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix},$$

and therefore

$$t_1 = \frac{s_1}{\sqrt{10}} + \frac{3s_2}{\sqrt{35}} - \frac{3s_3}{\sqrt{14}},$$

$$t_2 = \frac{5s_2}{\sqrt{35}} + \frac{2s_3}{\sqrt{14}},$$



quadratic

Figure 6.7

and

$$t_3 = \frac{3s_1}{\sqrt{10}} - \frac{s_2}{\sqrt{35}} + \frac{s_3}{\sqrt{14}}.$$

Thus

$$3t_1 - 2t_2 - t_3 = -\frac{14s_3}{\sqrt{14}} = -\sqrt{14}s_3.$$

Combining (11), (12), and the preceding equation, we conclude that if $x \in \mathsf{R}^3$ and $x = s_1v_1 + s_2v_2 + s_3v_3$, then $x \in S$ if and only if

$$2s_1^2 + 7s_2^2 - \sqrt{14}s_3 + 14 = 0 \quad \text{or} \quad s_3 = \frac{\sqrt{14}}{7}s_1^2 + \frac{\sqrt{14}}{2}s_2^2 + \sqrt{14}.$$

Consequently, if we draw new axes x', y', and z' in the directions of v_1, v_2, and v_3, respectively, the graph of the equation, rewritten as

$$z' = \frac{\sqrt{14}}{7}(x')^2 + \frac{\sqrt{14}}{2}(y')^2 + \sqrt{14},$$

coincides with the surface $\mathcal{S}$. We recognize $\mathcal{S}$ to be an elliptic paraboloid.

Figure 6.7 is a sketch of the surface $\mathcal{S}$ drawn so that the vectors v_1, v_2 and v_3 are oriented to lie in the principal directions. For practical purposes, the scale of the z' axis has been adjusted so that the figure fits the page. ♦

The Second Derivative Test for Functions of Several Variables

We now consider an application of the theory of quadratic forms to multivariable calculus—the derivation of the second derivative test for local extrema of a function of several variables. We assume an acquaintance with the calculus of functions of several variables to the extent of Taylor's theorem. The reader is undoubtedly familiar with the one-variable version of Taylor's theorem. For a statement and proof of the multivariable version, consult, for example, *An Introduction to Analysis* 2d ed, by William R. Wade (Prentice Hall, Upper Saddle River, N.J., 2000).

Let $z = f(t_1, t_2, \ldots, t_n)$ be a fixed real-valued function of n real variables for which all third-order partial derivatives exist and are continuous. The function f is said to have a **local maximum** at a point $p \in \mathsf{R}^n$ if there exists a $\delta > 0$ such that $f(p) \geq f(x)$ whenever $\|x - p\| < \delta$. Likewise, f has a **local minimum** at $p \in \mathsf{R}^n$ if there exists a $\delta > 0$ such that $f(p) \leq f(x)$ whenever $\|x - p\| < \delta$. If f has either a local minimum or a local maximum at p, we say that f has a **local extremum** at p. A point $p \in \mathsf{R}^n$ is called a **critical point** of f if $\partial f(p)/\partial t_i = 0$ for $i = 1, 2, \ldots, n$. It is a well-known fact that if f has a local extremum at a point $p \in \mathsf{R}^n$, then p is a critical point of f. For, if f has a local extremum at $p = (p_1, p_2, \ldots, p_n)$, then for any $i = 1, 2, \ldots, n$ the function ϕ_i defined by $\phi_i(t) = f(p_1, p_2, \ldots, p_{i-1}, t, p_{i+1}, \ldots, p_n)$ has a local extremum at $t = p_i$. So, by an elementary single-variable argument,

$$\frac{\partial f(p)}{\partial t_i} = \frac{d\phi_i(p_i)}{dt} = 0.$$

Thus p is a critical point of f. But critical points are not necessarily local extrema.

The second-order partial derivatives of f at a critical point p can often be used to test for a local extremum at p. These partials determine a matrix $A(p)$ in which the row i, column j entry is

$$\frac{\partial^2 f(p)}{(\partial t_i)(\partial t_j)}.$$

This matrix is called the **Hessian matrix** of f at p. Note that if the third-order partial derivatives of f are continuous, then the mixed second-order partials of f at p are independent of the order in which they are taken, and hence $A(p)$ is a symmetric matrix. In this case, all of the eigenvalues of $A(p)$ are real.

Theorem 6.37 (The Second Derivative Test). *Let $f(t_1, t_2, \ldots, t_n)$ be a real-valued function in n real variables for which all third-order partial derivatives exist and are continuous. Let $p = (p_1, p_2, \ldots, p_n)$ be a critical point of f, and let $A(p)$ be the Hessian of f at p.*

(a) *If all eigenvalues of $A(p)$ are positive, then f has a local minimum at p.*
(b) *If all eigenvalues of $A(p)$ are negative, then f has a local maximum at p.*
(c) *If $A(p)$ has at least one positive and at least one negative eigenvalue, then f has no local extremum at p (p is called a* **saddle-point** *of f).*
(d) *If $\operatorname{rank}(A(p)) < n$ and $A(p)$ does not have both positive and negative eigenvalues, then the second derivative test is inconclusive.*

Proof. If $p \neq 0$, we may define a function $g\colon \mathsf{R}^n \to R$ by

$$g(t_1, t_2, \ldots, t_n) = f(t_1 + p_1, t_2 + p_2, \ldots, p_n + t_n) - f(p).$$

The following facts are easily verified.

1. The function f has a local maximum [minimum] at p if and only if g has a local maximum [minimum] at $0 = (0, 0, \ldots, 0)$.
2. The partial derivatives of g at 0 are equal to the corresponding partial derivatives of f at p.
3. 0 is a critical point of g.
4. $A_{ij}(p) = \dfrac{\partial^2 g(0)}{(\partial t_i)(\partial t_j)}$ for all i and j.

In view of these facts, we may assume without loss of generality that $p = 0$ and $f(p) = 0$.

Now we apply Taylor's theorem to f to obtain the first-order approximation of f around 0. We have

$$\begin{aligned} f(t_1, t_2, \ldots, t_n) &= f(0) + \sum_{i=1}^{n} \frac{\partial f(0)}{\partial t_i} t_i + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f(0)}{(\partial t_i)(\partial t_j)} t_i t_j + S(t_1, t_2, \ldots, t_n) \\ &= \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f(0)}{(\partial t_i)(\partial t_j)} t_i t_j + S(t_1, t_2, \ldots, t_n), \end{aligned} \tag{13}$$

where S is a real-valued function on R^n such that

$$\lim_{x \to 0} \frac{S(x)}{||x||^2} = \lim_{(t_1, t_2, \ldots, t_n) \to 0} \frac{S(t_1, t_2, \ldots, t_n)}{t_1^2 + t_2^2 + \cdots + t_n^2} = 0. \tag{14}$$

Let $K\colon \mathsf{R}^n \to R$ be the quadratic form defined by

$$K \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} = \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f(0)}{(\partial t_i)(\partial t_j)} t_i t_j, \tag{15}$$

H be the symmetric bilinear form corresponding to K, and β be the standard ordered basis for R^n. It is easy to verify that $\psi_\beta(H) = \frac{1}{2}A(p)$. Since $A(p)$ is symmetric, Theorem 6.20 (p. 381) implies that there exists an orthogonal matrix Q such that

$$Q^t A(p) Q = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

is a diagonal matrix whose diagonal entries are the eigenvalues of $A(p)$. Let $\gamma = \{v_1, v_2, \dots, v_n\}$ be the orthogonal basis for R^n whose ith vector is the ith column of Q. Then Q is the change of coordinate matrix changing γ-coordinates into β-coordinates, and by Theorem 6.33

$$\psi_\gamma(H) = Q^t \psi_\beta(H) Q = \frac{1}{2} Q^t A(p) Q = \begin{pmatrix} \frac{\lambda_1}{2} & 0 & \dots & 0 \\ 0 & \frac{\lambda_2}{2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{\lambda_n}{2} \end{pmatrix}.$$

Suppose that $A(p)$ is not the zero matrix. Then $A(p)$ has nonzero eigenvalues. Choose $\epsilon > 0$ such that $\epsilon < |\lambda_i|/2$ for all $\lambda_i \neq 0$. By (14), there exists $\delta > 0$ such that for any $x \in \mathsf{R}^n$ satisfying $0 < ||x|| < \delta$, we have $|S(x)| < \epsilon ||x||^2$. Consider any $x \in \mathsf{R}^n$ such that $0 < ||x|| < \delta$. Then, by (13) and (15),

$$|f(x) - K(x)| = |S(x)| < \epsilon ||x||^2,$$

and hence

$$K(x) - \epsilon ||x||^2 < f(x) < K(x) + \epsilon ||x||^2. \tag{16}$$

Suppose that $x = \sum_{i=1}^{n} s_i v_i$. Then

$$||x||^2 = \sum_{i=1}^{n} s_i^2 \quad \text{and} \quad K(x) = \frac{1}{2} \sum_{i=1}^{n} \lambda_i s_i^2.$$

Combining these equations with (16), we obtain

$$\sum_{i=1}^{n} \left(\frac{1}{2}\lambda_i - \epsilon \right) s_i^2 < f(x) < \sum_{i=1}^{n} \left(\frac{1}{2}\lambda_i + \epsilon \right) s_i^2. \tag{17}$$

Now suppose that all eigenvalues of $A(p)$ are positive. Then $\frac{1}{2}\lambda_i - \epsilon > 0$ for all i, and hence, by the left inequality in (17),

$$f(0) = 0 \leq \sum_{i=1}^{n} \left(\frac{1}{2}\lambda_i - \epsilon\right) s_i^2 < f(x).$$

Thus $f(0) \leq f(x)$ for $||x|| < \delta$, and so f has a local minimum at 0. By a similar argument using the right inequality in (17), we have that if all of the eigenvalues of $A(p)$ are negative, then f has a local maximum at 0. This establishes (a) and (b) of the theorem.

Next, suppose that $A(p)$ has both a positive and a negative eigenvalue, say, $\lambda_i > 0$ and $\lambda_j < 0$ for some i and j. Then $\frac{1}{2}\lambda_i - \epsilon > 0$ and $\frac{1}{2}\lambda_j + \epsilon < 0$. Let s be any real number such that $0 < |s| < \delta$. Substituting $x = sv_i$ and $x = sv_j$ into the left inequality and the right inequality of (17), respectively, we obtain

$$f(0) = 0 < (\tfrac{1}{2}\lambda_i - \epsilon)s^2 < f(sv_i) \quad \text{and} \quad f(sv_j) < (\tfrac{1}{2}\lambda_j + \epsilon)s^2 < 0 = f(0).$$

Thus f attains both positive and negative values arbitrarily close to 0; so f has neither a local maximum nor a local minimum at 0. This establishes (c).

To show that the second-derivative test is inconclusive under the conditions stated in (d), consider the functions

$$f(t_1, t_2) = t_1^2 - t_2^4 \quad \text{and} \quad g(t_1, t_2) = t_1^2 + t_2^4$$

at $p = 0$. In both cases, the function has a critical point at p, and

$$A(p) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

However, f does not have a local extremum at 0, whereas g has a local minimum at 0. ∎

Sylvester's Law of Inertia

Any two matrix representations of a bilinear form have the same rank because rank is preserved under congruence. We can therefore define the **rank** of a bilinear form to be the rank of any of its matrix representations. If a matrix representation is a diagonal matrix, then the rank is equal to the number of nonzero diagonal entries of the matrix.

We confine our analysis to symmetric bilinear forms on finite-dimensional real vector spaces. Each such form has a diagonal matrix representation in which the diagonal entries may be positive, negative, or zero. Although these entries are not unique, we show that the number of entries that are positive and the number that are negative are unique. That is, they are independent

of the choice of diagonal representation. This result is called *Sylvester's law of inertia*. We prove the law and apply it to describe the equivalence classes of congruent symmetric real matrices.

Theorem 6.38 (Sylvester's Law of Inertia). *Let H be a symmetric bilinear form on a finite-dimensional real vector space V. Then the number of positive diagonal entries and the number of negative diagonal entries in any diagonal matrix representation of H are each independent of the diagonal representation.*

Proof. Suppose that β and γ are ordered bases for V that determine diagonal representations of H. Without loss of generality, we may assume that β and γ are ordered so that on each diagonal the entries are in the order of positive, negative, and zero. It suffices to show that both representations have the same number of positive entries because the number of negative entries is equal to the difference between the rank and the number of positive entries. Let p and q be the number of positive diagonal entries in the matrix representations of H with respect to β and γ, respectively. We suppose that $p \neq q$ and arrive at a contradiction. Without loss of generality, assume that $p < q$. Let

$$\beta = \{v_1, v_2, \ldots, v_p, \ldots, v_r, \ldots, v_n\} \text{ and } \gamma = \{w_1, w_2, \ldots, w_q, \ldots, w_r, \ldots, w_n\},$$

where r is the rank of H and $n = \dim(\mathsf{V})$. Let $\mathsf{L}\colon \mathsf{V} \to \mathsf{R}^{p+r-q}$ be the mapping defined by

$$L(x) = (H(x, v_1), H(x, v_2), \ldots, H(x, v_p), H(x, w_{q+1}), \ldots, H(x, w_r)).$$

It is easily verified that L is linear and $\text{rank}(\mathsf{L}) \leq p + r - q$. Hence

$$\text{nullity}(\mathsf{L}) \geq n - (p + r - q) > n - r.$$

So there exists a nonzero vector v_0 such that $v_0 \notin \text{span}(\{v_{r+1}, v_{r+2}, \ldots, v_n\})$, but $v_0 \in \mathsf{N}(\mathsf{L})$. Since $v_0 \in \mathsf{N}(\mathsf{L})$, it follows that $H(v_0, v_i) = 0$ for $i \leq p$ and $H(v_0, w_i) = 0$ for $q < i \leq r$. Suppose that

$$v_0 = \sum_{j=1}^{n} a_j v_j = \sum_{j=1}^{n} b_j w_j.$$

For any $i \leq p$,

$$H(v_0, v_i) = H\left(\sum_{j=1}^{n} a_j v_j, v_i\right) = \sum_{j=1}^{n} a_j H(v_j, v_i) = a_i H(v_i, v_i).$$

But for $i \leq p$, we have $H(v_i, v_i) > 0$ and $H(v_0, v_i) = 0$, so that $a_i = 0$. Similarly, $b_i = 0$ for $q+1 \leq i \leq r$. Since v_0 is not in the span of $\{v_{r+1}, v_{r+2}, \ldots, v_n\}$, it follows that $a_i \neq 0$ for some $p < i \leq r$. Thus

$$H(v_0, v_0) = H\left(\sum_{j=1}^{n} a_j v_j, \sum_{i=1}^{n} a_i v_i\right) = \sum_{j=1}^{n} a_j^2 H(v_j, v_j) = \sum_{j=p+1}^{r} a_j^2 H(v_j, v_j) < 0.$$

Furthermore,

$$H(v_0, v_0) = H\left(\sum_{j=1}^{n} b_j w_j, \sum_{i=1}^{n} b_i w_i\right) = \sum_{j=1}^{n} b_j^2 H(w_j, w_j) = \sum_{j=p+1}^{r} b_j^2 H(w_j, w_j) \geq 0.$$

So $H(v_0, v_0) < 0$ and $H(v_0, v_0) \geq 0$, which is a contradiction. We conclude that $p = q$. ∎

Definitions. *The number of positive diagonal entries in a diagonal representation of a symmetric bilinear form on a real vector space is called the* ***index*** *of the form. The difference between the number of positive and the number of negative diagonal entries in a diagonal representation of a symmetric bilinear form is called the* ***signature*** *of the form. The three terms rank, index, and signature are called the* ***invariants*** *of the bilinear form because they are invariant with respect to matrix representations. These same terms apply to the associated quadratic form. Notice that the values of any two of these invariants determine the value of the third.*

Example 10

The bilinear form corresponding to the quadratic form K of Example 9 has a 3×3 diagonal matrix representation with diagonal entries of 2, 7, and 0. Therefore the rank, index, and signature of K are each 2. ♦

Example 11

The matrix representation of the bilinear form corresponding to the quadratic form $K(x, y) = x^2 - y^2$ on R^2 with respect to the standard ordered basis is the diagonal matrix with diagonal entries of 1 and -1. Therefore the rank of K is 2, the index of K is 1, and the signature of K is 0. ♦

Since the congruence relation is intimately associated with bilinear forms, we can apply Sylvester's law of inertia to study this relation on the set of real symmetric matrices. Let A be an $n \times n$ real symmetric matrix, and suppose that D and E are each diagonal matrices congruent to A. By Corollary 3 to Theorem 6.32, A is the matrix representation of the bilinear form H on R^n defined by $H(x, y) = x^t A y$ with respect to the standard ordered basis for R^n. Therefore Sylvester's law of inertia tells us that D and E have the same number of positive and negative diagonal entries. We can state this result as the matrix version of Sylvester's law.

Corollary 1 (Sylvester's Law of Inertia for Matrices). *Let A be a real symmetric matrix. Then the number of positive diagonal entries and the number of negative diagonal entries in any diagonal matrix congruent to A is independent of the choice of the diagonal matrix.*

Definitions. *Let A be a real symmetric matrix, and let D be a diagonal matrix that is congruent to A. The number of positive diagonal entries of D is called the* ***index*** *of A. The difference between the number of positive diagonal entries and the number of negative diagonal entries of D is called the* ***signature*** *of A. As before, the rank, index, and signature of a matrix are called the* ***invariants*** *of the matrix, and the values of any two of these invariants determine the value of the third.*

Any two of these invariants can be used to determine an equivalence class of congruent real symmetric matrices.

Corollary 2. *Two real symmetric $n \times n$ matrices are congruent if and only if they have the same invariants.*

Proof. If A and B are congruent $n \times n$ symmetric matrices, then they are both congruent to the same diagonal matrix, and it follows that they have the same invariants.

Conversely, suppose that A and B are $n \times n$ symmetric matrices with the same invariants. Let D and E be diagonal matrices congruent to A and B, respectively, chosen so that the diagonal entries are in the order of positive, negative, and zero. (Exercise 23 allows us to do this.) Since A and B have the same invariants, so do D and E. Let p and r denote the index and the rank, respectively, of both D and E. Let d_i denote the ith diagonal entry of D, and let Q be the $n \times n$ diagonal matrix whose ith diagonal entry q_i is given by

$$q_i = \begin{cases} \dfrac{1}{\sqrt{d_i}} & \text{if } 1 \le i \le p \\ \dfrac{1}{\sqrt{-d_i}} & \text{if } p < i \le r \\ 1 & \text{if } r < i. \end{cases}$$

Then $Q^t D Q = J_{pr}$, where

$$J_{pr} = \begin{pmatrix} I_p & O & O \\ O & -I_{r-p} & O \\ O & O & O \end{pmatrix}.$$

It follows that A is congruent to J_{pr}. Similarly, B is congruent to J_{pr}, and hence A is congruent to B. ∎

The matrix J_{pr} acts as a *canonical form* for the theory of real symmetric matrices. The next corollary, whose proof is contained in the proof of Corollary 2, describes the role of J_{pr}.

Corollary 3. *A real symmetric $n \times n$ matrix A has index p and rank r if and only if A is congruent to J_{pr} (as just defined).*

Example 12

Let

$$A = \begin{pmatrix} 1 & 1 & -3 \\ -1 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}, \quad \text{and} \quad C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 1 \end{pmatrix}.$$

We apply Corollary 2 to determine which pairs of the matrices A, B, and C are congruent.

The matrix A is the 3×3 matrix of Example 6, where it is shown that A is congruent to a diagonal matrix with diagonal entries 1, 1, and -24. Therefore A has rank 3 and index 2. Using the methods of Example 6 (it is not necessary to compute Q), it can be shown that B and C are congruent, respectively, to the diagonal matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -4 \end{pmatrix}.$$

It follows that both A and C have rank 3 and index 2, while B has rank 3 and index 1. We conclude that A and C are congruent but that B is congruent to neither A nor C. ♦

EXERCISES

1. Label the following statements as true or false.

- **(a)** Every quadratic form is a bilinear form.
- **(b)** If two matrices are congruent, they have the same eigenvalues.
- **(c)** Symmetric bilinear forms have symmetric matrix representations.
- **(d)** Any symmetric matrix is congruent to a diagonal matrix.
- **(e)** The sum of two symmetric bilinear forms is a symmetric bilinear form.
- **(f)** If two symmetric matrices over a field not of characteristic two have the the same characteristic polynomial, then they are matrix representations of the same bilinear form.
- **(g)** There exists a bilinear form H such that $H(x, y) \neq 0$ for all x and y.

(h) If V is a vector space of dimension n, then $\dim(\mathcal{B}(V)) = 2n$.
(i) Let H be a bilinear form on a finite-dimensional vector space V with $\dim(\mathsf{V}) > 1$. For any $x \in \mathsf{V}$, there exists $y \in \mathsf{V}$ such that $y \neq 0$, but $H(x, y) = 0$.
(j) If H is any bilinear form on a finite-dimensional real inner product space V, then there exists an ordered basis β for V such that $\psi_\beta(H)$ is a diagonal matrix.

2. Prove properties 1, 2, 3, and 4 on page 420.

3. **(a)** Prove that the sum of two bilinear forms is a bilinear form.
(b) Prove that the product of a scalar and a bilinear form is a bilinear form.
(c) Prove Theorem 6.31.

4. Determine which of the mappings that follow are bilinear forms. Justify your answers.

(a) Let $\mathsf{V} = \mathsf{C}[0, 1]$ be the space of continuous real-valued functions on the closed interval $[0, 1]$. For $f, g \in \mathsf{V}$, define

$$H(f, g) = \int_0^1 f(t)g(t)dt.$$

(b) Let V be a vector space over F, and let $J \in \mathcal{B}(\mathsf{V})$ be nonzero. Define $H : \mathsf{V} \times \mathsf{V} \to F$ by

$$H(x, y) = [J(x, y)]^2 \quad \text{for all } x, y \in \mathsf{V}.$$

(c) Define $H : R \times R \to R$ by $H(t_1, t_2) = t_1 + 2t_2$.
(d) Consider the vectors of R^2 as column vectors, and let $H : \mathsf{R}^2 \to R$ be the function defined by $H(x, y) = \det(x, y)$, the determinant of the 2×2 matrix with columns x and y.
(e) Let V be a real inner product space, and let $H : \mathsf{V} \times \mathsf{V} \to R$ be the function defined by $H(x, y) = \langle x, y \rangle$ for $x, y \in \mathsf{V}$.
(f) Let V be a complex inner product space, and let $H : \mathsf{V} \times \mathsf{V} \to C$ be the function defined by $H(x, y) = \langle x, y \rangle$ for $x, y \in \mathsf{V}$.

5. Verify that each of the given mappings is a bilinear form. Then compute its matrix representation with respect to the given ordered basis β.

(a) $H : \mathsf{R}^3 \times \mathsf{R}^3 \to R$, where

$$H\left(\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}\right) = a_1b_1 - 2a_1b_2 + a_2b_1 - a_3b_3$$

and

$$\beta = \left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

(b) Let $\mathsf{V} = \mathsf{M}_{2\times 2}(R)$ and

$$\beta = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

Define $H: \mathsf{V} \times \mathsf{V} \to R$ by $H(A, B) = \operatorname{tr}(A) \cdot \operatorname{tr}(B)$.

(c) Let $\beta = \{\cos t, \sin t, \cos 2t, \sin 2t\}$. Then β is an ordered basis for $\mathsf{V} = \operatorname{span}(\beta)$, a four-dimensional subspace of the space of all continuous functions on R. Let $H: \mathsf{V} \times \mathsf{V} \to R$ be the function defined by $H(f, g) = f'(0) \cdot g''(0)$.

6. Let $H: \mathsf{R}^2 \times \mathsf{R}^2 \to R$ be the function defined by

$$H\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right) = a_1 b_2 + a_2 b_1 \quad \text{for } \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \in \mathsf{R}^2.$$

(a) Prove that H is a bilinear form.

(b) Find the 2×2 matrix A such that $H(x, y) = x^t A y$ for all $x, y \in \mathsf{R}^2$.

7. Let V and W be vector spaces over the same field, and let $\mathsf{T}: \mathsf{V} \to \mathsf{W}$ be a linear transformation. For any $H \in \mathcal{B}(\mathsf{W})$, define $\widehat{\mathsf{T}}(H): \mathsf{V} \times \mathsf{V} \to F$ by $\widehat{\mathsf{T}}(H)(x, y) = H(\mathsf{T}(x), \mathsf{T}(y))$ for all $x, y \in \mathsf{V}$. Prove the following results.

(a) If $H \in \mathcal{B}(\mathsf{W})$, then $\widehat{\mathsf{T}}(H) \in \mathcal{B}(\mathsf{V})$.

(b) $\widehat{\mathsf{T}}: \mathcal{B}(\mathsf{W}) \to \mathcal{B}(\mathsf{V})$ is a linear transformation.

(c) If T is an isomorphism, then so is $\widehat{\mathsf{T}}$.

8. Assume the notation of Theorem 6.32.

(a) Prove that for any ordered basis β, ψ_β is linear.

(b) Let β be an ordered basis for an n-dimensional space V over F, and let $\phi_\beta: \mathsf{V} \to \mathsf{F}^n$ be the standard representation of V with respect to β. For $A \in \mathsf{M}_{n\times n}(F)$, define $H: \mathsf{V} \times \mathsf{V} \to F$ by $H(x, y) = [\phi_\beta(x)]^t A[\phi_\beta(y)]$. Prove that $H \in \mathcal{B}(\mathsf{V})$. Can you establish this as a corollary to Exercise 7?

(c) Prove the converse of (b): Let H be a bilinear form on V. If $A = \psi_\beta(H)$, then $H(x, y) = [\phi_\beta(x)]^t A[\phi_\beta(y)]$.

9. **(a)** Prove Corollary 1 to Theorem 6.32.

(b) For a finite-dimensional vector space V, describe a method for finding an ordered basis for $\mathcal{B}(\mathsf{V})$.

10. Prove Corollary 2 to Theorem 6.32.

11. Prove Corollary 3 to Theorem 6.32.

12. Prove that the relation of congruence is an equivalence relation.

13. Use Corollary 2 to Theorem 6.32 and Theorem 2.22(b) to obtain an alternate proof to Theorem 6.33.

14. Let V be a finite-dimensional vector space and $H \in \mathcal{B}(\mathsf{V})$. Prove that, for any ordered bases β and γ of V, $\operatorname{rank}(\psi_\beta(H)) = \operatorname{rank}(\psi_\gamma(H))$.

15. Prove the following results.

(a) Any square diagonal matrix is symmetric.
(b) Any matrix congruent to a diagonal matrix is symmetric.
(c) the corollary to Theorem 6.35

16. Let V be a vector space over a field F not of characteristic two, and let H be a symmetric bilinear form on V. Prove that if $K(x) = H(x, x)$ is the quadratic form associated with H, then, for all $x, y \in \mathsf{V}$,

$$H(x, y) = \frac{1}{2}[K(x + y) - K(x) - K(y)].$$

17. For each of the given quadratic forms K on a real inner product space V, find a symmetric bilinear form H such that $K(x) = H(x, x)$ for all $x \in \mathsf{V}$. Then find an orthonormal basis β for V such that $\psi_\beta(H)$ is a diagonal matrix.

(a) $K\colon \mathsf{R}^2 \to R$ defined by $K\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = -2t_1^2 + 4t_1t_2 + t_2^2$

(b) $K\colon \mathsf{R}^2 \to R$ defined by $K\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = 7t_1^2 - 8t_1t_2 + t_2^2$

(c) $K\colon \mathsf{R}^3 \to R$ defined by $K\begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} = 3t_1^2 + 3t_2^2 + 3t_3^2 - 2t_1t_3$

18. Let $\mathcal{S}$ be the set of all $(t_1, t_2, t_3) \in \mathsf{R}^3$ for which

$$3t_1^2 + 3t_2^2 + 3t_3^2 - 2t_1t_3 + 2\sqrt{2}(t_1 + t_3) + 1 = 0.$$

Find an orthonormal basis β for R^3 for which the equation relating the coordinates of points of $\mathcal{S}$ relative to β is simpler. Describe $\mathcal{S}$ geometrically.

19. Prove the following refinement of Theorem 6.37(d).

(a) If $0 < \text{rank}(A) < n$ and A has no negative eigenvalues, then f has no local maximum at p.
(b) If $0 < \text{rank}(A) < n$ and A has no positive eigenvalues, then f has no local minimum at p.

20. Prove the following variation of the second-derivative test for the case $n = 2$: Define

$$D = \left[\frac{\partial^2 f(p)}{\partial t_1^2}\right]\left[\frac{\partial^2 f(p)}{\partial t_2^2}\right] - \left[\frac{\partial^2 f(p)}{\partial t_1 \partial t_2}\right]^2.$$

(a) If $D > 0$ and $\partial^2 f(p)/\partial t_1^2 > 0$, then f has a local minimum at p.
(b) If $D > 0$ and $\partial^2 f(p)/\partial t_1^2 < 0$, then f has a local maximum at p.
(c) If $D < 0$, then f has no local extremum at p.
(d) If $D = 0$, then the test is inconclusive.

Hint: Observe that, as in Theorem 6.37, $D = \det(A) = \lambda_1\lambda_2$, where λ_1 and λ_2 are the eigenvalues of A.

21. Let A and E be in $\mathsf{M}_{n\times n}(F)$, with E an elementary matrix. In Section 3.1, it was shown that AE can be obtained from A by means of an elementary column operation. Prove that E^tA can be obtained by means of the same elementary operation performed on the rows rather than on the columns of A. *Hint:* Note that $E^tA = (A^tE)^t$.

22. For each of the following matrices A with entries from R, find a diagonal matrix D and an invertible matrix Q such that $Q^tAQ = D$.

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

Hint for (b): Use an elementary operation other than interchanging columns.

23. Prove that if the diagonal entries of a diagonal matrix are permuted, then the resulting diagonal matrix is congruent to the original one.

24. Let T be a linear operator on a real inner product space V, and define $H: \mathsf{V} \times \mathsf{V} \to R$ by $H(x, y) = \langle x, \mathsf{T}(y)\rangle$ for all $x, y \in \mathsf{V}$.

(a) Prove that H is a bilinear form.
(b) Prove that H is symmetric if and only if T is self-adjoint.
(c) What properties must T have for H to be an inner product on V?
(d) Explain why H may fail to be a bilinear form if V is a complex inner product space.

25. Prove the converse to Exercise 24(a): Let V be a finite-dimensional real inner product space, and let H be a bilinear form on V. Then there

exists a unique linear operator T on V such that $H(x, y) = \langle x, \mathsf{T}(y)\rangle$ for all $x, y \in \mathsf{V}$. *Hint:* Choose an orthonormal basis β for V, let $A = \psi_\beta(H)$, and let T be the linear operator on V such that $[\mathsf{T}]_\beta = A$. Visit `goo.gl/bGAfSy` for a solution.

26. Prove that the number of distinct equivalence classes of congruent $n \times n$ real symmetric matrices is

$$\frac{(n+1)(n+2)}{2}.$$

6.9* EINSTEIN'S SPECIAL THEORY OF RELATIVITY

As a consequence of physical experiments performed in the latter half of the nineteenth century (most notably the Michelson–Morley experiment of 1887), physicists concluded that *the results obtained in measuring the speed of light c are independent of the velocity of the instrument used to measure it.* For example, suppose that while on Earth, an experimenter measures the speed of light emitted from the sun and finds it to be 186,000 miles per second. Now suppose that the experimenter places the measuring equipment in a spaceship that leaves Earth traveling at 100,000 miles per second in a direction away from the sun. A repetition of the same experiment from the spaceship yields the same result: Light is traveling at 186,000 miles per second relative to the spaceship, rather than 86,000 miles per second as one might expect!

This revelation led to a new way of relating coordinate systems used to locate events in space–time. The result was Albert Einstein's *special theory of relativity*. In this section, we develop via a linear algebra viewpoint the essence of Einstein's theory.

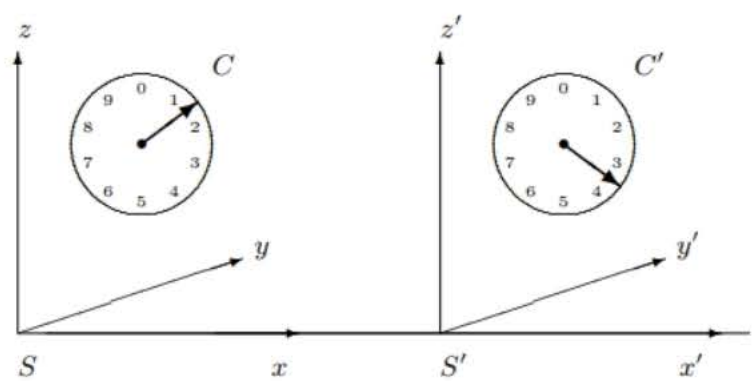


Figure 6.8

The basic problem is to compare two different inertial (nonaccelerating) coordinate systems S and S' that are in motion relative to each other under

the assumption that the speed of light is the same when measured in either system. We assume that S' moves at a constant velocity in relation to S as measured from S.

To simplify matters, we assume that the two coordinate systems have parallel axes and share the same x-axis, and that the motion of S' relative to S is along this common axis. (See Figure 6.8.)

We also suppose that there are two clocks, C and C', placed in space so that C is stationary relative to S and C' is stationary relative to S'. These clocks give readings that are real numbers in units of seconds. They are calibrated so that at the instant the origins of S and S' coincide, both give the reading of 0.

Given any event p (something whose position and time of occurrence can be described), we may assign a set of *space–time coordinates* to it. For example, if p is an event that occurs at position (x, y, z) relative to S and at time t as read on clock C, we can assign to p the set of coordinates

$$\begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix}.$$

This ordered 4-tuple is called the **space–time coordinates** of p relative to S and C. Likewise, p has a set of space–time coordinates

$$\begin{pmatrix} x' \\ y' \\ z' \\ t' \end{pmatrix}$$

relative to S' and C'.

Because motion is along a common x-axis, which lies in a common xy-plane, the third component of the space-time coordinates of p is always zero. Thus we consider only the first, second, and fourth coordinates of p, and write

$$\begin{pmatrix} x \\ y \\ t \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} x' \\ y' \\ t' \end{pmatrix}$$

to denote the space-time coordinates of an event p relative to S and S', respectively.

As we have mentioned, our unit of time is the second. Our measure of an object's velocity v is the ratio of its velocity (expressed in miles per second) to the speed of light expressed in the same units (which is approximately 186,000 miles per second). For example, if S' is moving at 18,600 miles per second relative to S and the speed of light, c, is 186,000 miles per second, the

velocity of S' relative to S, v, would have a value of $v = 0.1$. For this reason, the speed of light c has the value 1.

For a fixed velocity v, let $\mathsf{T}_v\colon \mathsf{R}^3 \to \mathsf{R}^3$ be the mapping defined by

$$\mathsf{T}_v \begin{pmatrix} x \\ y \\ t \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ t' \end{pmatrix},$$

where

$$\begin{pmatrix} x \\ y \\ t \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} x' \\ y' \\ t' \end{pmatrix}$$

are the space–time coordinates of the same event with respect to S and C and with respect to S' and C', respectively.

In what follows, we make four assumptions:

1. The origin of S' moves in the positive direction of the common x-axis relative to S at a constant velocity of $v > 0$.
2. The origin of S moves in the negative direction of the common x-axis relative to S' at the constant velocity of $-v < 0$.
3. T_v is a linear isomorphism.
4. The speed of any light beam, when measured in either S or S', using the clocks C in S and C' in S', is always $c = 1$.

Since motion is strictly along the x-axis and we assume that the y-axis is unaffected, we have that for any x, y, and t, there exist x' and t' such that

$$\mathsf{T}_v \begin{pmatrix} x \\ y \\ t \end{pmatrix} = \begin{pmatrix} x' \\ y \\ t' \end{pmatrix} \quad \text{and} \quad \mathsf{T}_v \begin{pmatrix} 0 \\ y \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ y \\ 0 \end{pmatrix}.$$

Our goal in this section is to calculate the matrix representation of T_v with respect to the standard basis for R^3.

Theorem 6.39. *Consider $\{e_1, e_2, e_3\}$, the standard ordered basis for R^3. Then*

(a) $\mathsf{T}_v(e_2) = e_2$.
(b) T_v *maps* $\operatorname{span}\{e_1, e_3\}$ *into itself.*
(c) T_v^* *maps* $\operatorname{span}\{e_1, e_3\}$ *into itself.*

Proof. Parts (a) and (b) follow immediately from the equations above.

For $i = 1$ and $i = 3$,

$$\langle \mathsf{T}_v^*(e_i), e_2 \rangle = \langle e_i, \mathsf{T}_v(e_2) \rangle = \langle e_i, e_2 \rangle = 0,$$

and hence (c) follows. ∎

Suppose that, at the instant the origins of S and S' coincide, a light flash is emitted from their common origin. The event of the light flash when measured either relative to S and C or relative to S' and C' has space–time coordinates

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Let P be the set of all events in the xy-plane whose space–time coordinates

$$\begin{pmatrix} x \\ y \\ t \end{pmatrix}$$

relative to S and C are such that the flash is observable in the common xy-plane at the point (x, y) (as measured relative to S) at the time t (as measured on C). Let us characterize P in terms of x, y, and t. Since the speed of light is 1, at any time $t \geq 0$ the light flash is observable from any point in the plane whose distance to the origin of S (as measured on S) is $t \cdot 1 = t$. These are precisely the points in the xy-plane with $x^2 + y^2 = t^2$, or $x^2 + y^2 - t^2 = 0$. Hence an event lies in P if and only if its space-time coordinates

$$\begin{pmatrix} x \\ y \\ t \end{pmatrix} \quad (t \geq 0)$$

relative to S and C satisfy the equation $x^2 + y^2 - t^2 = 0$. Since the speed of light when measured in either coordinate system is the same, we can characterize P in terms of the space–time coordinates relative to S' and C' similarly: An event lies in P if and only if, relative to S' and C', its space–time coordinates

$$\begin{pmatrix} x' \\ y \\ t' \end{pmatrix} \quad (t' \geq 0)$$

satisfy the equation $(x')^2 + y^2 - (t')^2 = 0$.

Let

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

Theorem 6.40. *If* $\langle \mathsf{L}_A(w), w \rangle = 0$ *for some* $w \in \mathsf{R}^3$, *then*

$$\langle \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w), w \rangle = 0.$$

Proof. Let

$$w = \begin{pmatrix} x \\ y \\ t \end{pmatrix} \in \mathsf{R}^3,$$

and suppose that $\langle \mathsf{L}_A(w), w\rangle = 0$.

CASE 1. $t \geq 0$. Since $\langle \mathsf{L}_A(w), w\rangle = x^2 + y^2 - t^2$, the vector w gives the coordinates of an event in P relative to S and C. Because

$$\begin{pmatrix} x' \\ y \\ t' \end{pmatrix}$$

are the space–time coordinates of the same event relative to S' and C', the discussion preceding Theorem 6.40 yields

$$(x')^2 + y^2 - (t')^2 = 0.$$

Thus $\langle \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w), w\rangle = \langle \mathsf{L}_A \mathsf{T}_v(w), \mathsf{T}_v(w)\rangle = (x')^2 + y^2 - (t')^2 = 0$, and the conclusion follows.

CASE 2. $t < 0$. The proof follows by applying case 1 to $-w$. ∎

We now proceed to deduce information about T_v. Let

$$w_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad w_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

Clearly $\{w_1, w_2\}$ is an orthogonal basis for the span of $\{e_1, e_3\}$. The next result tells us even more.

Theorem 6.41. *There exists a nonzero scalar a such that $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_1) = aw_2$ and $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_2) = aw_1$.*

Proof. Because $\langle \mathsf{L}_A(w_1), w_1\rangle = 0$, $\langle \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_1), w_1\rangle = 0$ by Theorem 6.40. Thus $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_1)$ is orthogonal to w_1. Since $\{w_1, w_2\}$ is an orthogonal basis for $\text{span}\{e_1, e_3\}$ and each of T_v^*, L_A, and T_v maps this span into itself, it follows that $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_1)$ must be a multiple of w_2, that is, $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_1) = aw_2$ for some scalar a. Since T_v and A are invertible, so is $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v$. Thus $a \neq 0$.

Similarly, there exists a nonzero scalar b such that $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_2) = bw_1$.

Finally, we show that $a = b$. Since $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_1) = aw_2$, we have

$$2a = \langle \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_1), w_2\rangle = \langle w_1, \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w_2)\rangle = \langle w_1, bw_1\rangle = 2b.$$

So $a = b$. ∎

Actually, $a = b = 1$, as we see in the following result.

For the rest of this section let $B_v = [\mathsf{T}_v]_\beta$, where β is the standard ordered basis for R^3.

Theorem 6.42. *Given $B_v = [\mathsf{T}_v]_\beta$, as defined above,*

(a) $B_v^* A B_v = A$.
(b) $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v = \mathsf{L}_A$.

Proof. Since

$$e_1 = \frac{1}{2}(w_1 + w_2) \quad \text{and} \quad e_3 = \frac{1}{2}(w_1 - w_2),$$

it follows from Theorem 6.41 that

$$\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(e_1) = a e_1 \quad \text{and} \quad \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(e_3) = -a e_3.$$

Furthermore, $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(e_2) = e_2$, and hence

$$B_v^* A B_v = \begin{pmatrix} a & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -a \end{pmatrix}.$$

Let

$$w = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

Then $\langle \mathsf{L}_A(w), w \rangle = 0$, and hence by Theorem 6.40

$$0 = \langle \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v(w), w \rangle = \langle B_v^* A B_v w, w \rangle = 1 - a.$$

Thus $a = 1$. As a consequence, $B_v^* A B_v = A$. This proves (a). Part (b) now follows. ∎

Now consider the situation 1 second after the origins of S and S' have coincided as measured by the clock C. Since the origin of S' is moving along the x-axis at a velocity v as measured in S, its space–time coordinates relative to S and C are

$$\begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix}.$$

Similarly, the space–time coordinates for the origin of S' relative to S' and C' must be

$$\begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix}$$

for some $t' > 0$. Thus we have

$$\mathsf{T}_v \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix} \quad \text{for some } t' > 0. \tag{18}$$

By Theorem 6.42

$$\left\langle \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix} \right\rangle = \left\langle \mathsf{L}_A \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix} \right\rangle = v^2 - 1. \tag{19}$$

But also

$$\begin{aligned} \left\langle \mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix} \right\rangle &= \left\langle \mathsf{L}_A \mathsf{T}_v \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix}, \mathsf{T}_v \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix} \right\rangle \\ &= \left\langle \mathsf{L}_A \begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix} \right\rangle = -(t')^2. \end{aligned} \tag{20}$$

Combining (19) and (20), we conclude that $v^2 - 1 = -(t')^2$, or

$$t' = \sqrt{1 - v^2}. \tag{21}$$

Thus, from (18) and (21), we obtain

$$\mathsf{T}_v \begin{pmatrix} v \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \sqrt{1 - v^2} \end{pmatrix}. \tag{22}$$

Next, recall that the origin of S moves in the negative direction of the x'-axis of S' at the constant velocity $-v < 0$ as measured from S'. Consequently, 1 second after the origins of S and S' have coincided as measured on clock C, there exists a time $t'' > 0$ as measured on clock C' such that

$$\mathsf{T}_v \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -vt'' \\ 0 \\ t'' \end{pmatrix}. \tag{23}$$

From (23), it follows in a manner similar to the derivation of (22) that

$$t'' = \frac{1}{\sqrt{1 - v^2}}. \tag{24}$$

Hence, from (23) and (24),

$$\mathsf{T}_v \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \dfrac{-v}{\sqrt{1 - v^2}} \\ 0 \\ \dfrac{1}{\sqrt{1 - v^2}} \end{pmatrix}. \tag{25}$$

The following result is now easily proved using (22), (25), and Theorem 6.39.

Theorem 6.43. *Let β be the standard ordered basis for R^3. Then*

$$[\mathsf{T}_v]_\beta = B_v = \begin{pmatrix} \frac{1}{\sqrt{1-v^2}} & 0 & \frac{-v}{\sqrt{1-v^2}} \\ 0 & 1 & 0 \\ \frac{-v}{\sqrt{1-v^2}} & 0 & \frac{1}{\sqrt{1-v^2}} \end{pmatrix}.$$

Time Contraction

A most curious and paradoxical conclusion follows if we accept Einstein's theory. Suppose that an astronaut leaves our solar system in a space vehicle traveling at a fixed velocity v as measured relative to our solar system. It follows from Einstein's theory that, at the end of time t as measured on Earth, the time that passes on the space vehicle is only $t\sqrt{1-v^2}$. To establish this result, consider the coordinate systems S and S' and clocks C and C' that we have been studying. Suppose that the origin of S' coincides with the space vehicle and the origin of S coincides with a point in the solar system (stationary relative to the sun) so that the origins of S and S' coincide and clocks C and C' read zero at the moment the astronaut embarks on the trip.

As viewed from S, the space–time coordinates of the vehicle at any time $t > 0$ as measured by C are

$$\begin{pmatrix} vt \\ 0 \\ t \end{pmatrix},$$

whereas, as viewed from S', the space–time coordinates of the vehicle at any time $t' > 0$ as measured by C' are

$$\begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix}.$$

But if two sets of space–time coordinates

$$\begin{pmatrix} vt \\ 0 \\ t \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix}$$

are to describe the same event, it must follow that

$$\mathsf{T}_v \begin{pmatrix} vt \\ 0 \\ t \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix}.$$

Thus

$$\begin{pmatrix} \frac{1}{\sqrt{1-v^2}} & 0 & \frac{-v}{\sqrt{1-v^2}} \\ 0 & 1 & 0 \\ \frac{-v}{\sqrt{1-v^2}} & 0 & \frac{1}{\sqrt{1-v^2}} \end{pmatrix} \begin{pmatrix} vt \\ 0 \\ t \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ t' \end{pmatrix}.$$

From the preceding equation, we obtain $\frac{-v^2t}{\sqrt{1-v^2}} + \frac{t}{\sqrt{1-v^2}} = t'$, or

$$t' = t\sqrt{1-v^2}. \tag{26}$$

This is the desired result.

A dramatic consequence of time contraction is that distances are contracted along the line of motion (see Exercise 7).

Let us make one additional point. Suppose that we choose units of distance and time commonly used in the study of motion, such as the mile, the kilometer, and the second. Recall that the velocity v we have been using is actually the ratio of the velocity using these units with the speed of light c, using the same units. For this reason, we can replace v in any of the equations given in this section with the ratio v/c ,where v and c are given using the same units of measurement. Thus, for example, given a set of units of distance and time, (26) becomes

$$t' = t\sqrt{1-\frac{v^2}{c^2}}.$$

EXERCISES

1. Complete the proof of Theorem 6.40 for the case $t < 0$.

2. For

$$w_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad w_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix},$$

show that

(a) $\{w_1, w_2\}$ is an orthogonal basis for $\text{span}(\{e_1, e_3\})$;

(b) $\text{span}(\{e_1, e_3\})$ is $\mathsf{T}_v^* \mathsf{L}_A \mathsf{T}_v$-invariant.

3. Derive (24), and prove that

$$\mathsf{T}_v \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{-v}{\sqrt{1-v^2}} \\ 0 \\ \frac{1}{\sqrt{1-v^2}} \end{pmatrix}. \tag{25}$$

Hint: Use a technique similar to the derivation of (22).

4. Consider three coordinate systems S, S', and S'' with the corresponding axes (x,x',x''; y,y',y''; and z,z',z'') parallel and such that the x-, x'-, and x''-axes coincide. Suppose that S' is moving past S at a velocity $v_1 > 0$ (as measured on S), S'' is moving past S' at a velocity $v_2 > 0$ (as measured on S'), and S'' is moving past S at a velocity $v_3 > 0$ (as measured on S), and that there are three clocks C, C', and C'' such that C is stationary relative to S, C' is stationary relative to S', and C'' is stationary relative to S''. Suppose that when measured on any of the three clocks, all the origins of S, S', and S'' coincide at time 0. Assuming that $\mathsf{T}_{v_3} = \mathsf{T}_{v_2}\mathsf{T}_{v_1}$ (i.e., $B_{v_3} = B_{v_2}B_{v_1}$), prove that

$$v_3 = \frac{v_1 + v_2}{1 + v_1 v_2}.$$

Note that substituting $v_2 = 1$ in this equation yields $v_3 = 1$. This tells us that the speed of light as measured in S or S' is the same. Why would we be surprised if this were not the case?

5. Compute $(B_v)^{-1}$. Show $(B_v)^{-1} = B_{(-v)}$. Conclude that if S' moves at a negative velocity v relative to S, then $[\mathsf{T}_v]_\beta = B_v$, where B_v is of the form given in Theorem 6.43. Visit `goo.gl/9gWNYu` for a solution.

6. Suppose that an astronaut left Earth in the year 2000 and traveled to a star 99 light years away from Earth at 99% of the speed of light and that upon reaching the star immediately turned around and returned to Earth at the same speed. Assuming Einstein's special theory of relativity, show that if the astronaut was 20 years old at the time of departure, then he or she would return to Earth at age 48.2 in the year 2200. Explain the use of Exercise 4 in solving this problem.

7. Recall the moving space vehicle considered in the study of time contraction. Suppose that the vehicle is moving toward a fixed star located on the x-axis of S at a distance b units from the origin of S. If the space vehicle moves toward the star at velocity v, Earthlings (who remain "almost" stationary relative to S) compute the time it takes for the vehicle to reach the star as $t = b/v$. Due to the phenomenon of time contraction, the astronaut perceives a time span of $t' = t\sqrt{1 - v^2} = (b/v)\sqrt{1 - v^2}$. A paradox appears in that the astronaut perceives a time span inconsistent with a distance of b and a velocity of v. The paradox is resolved by observing that the distance from the solar system to the star as measured by the astronaut is less than b.

Assuming that the coordinate systems S and S' and clocks C and C' are as in the discussion of time contraction, prove the following results.

(a) At time t (as measured on C), the space–time coordinates of the star relative to S and C are

$$\begin{pmatrix} b \\ 0 \\ t \end{pmatrix}.$$

(b) At time t (as measured on C), the space–time coordinates of the star relative to S' and C' are

$$\begin{pmatrix} \dfrac{b - vt}{\sqrt{1 - v^2}} \\ 0 \\ \dfrac{t - bv}{\sqrt{1 - v^2}} \end{pmatrix}.$$

(c) For

$$x' = \frac{b - tv}{\sqrt{1 - v^2}} \quad \text{and} \quad t' = \frac{t - bv}{\sqrt{1 - v^2}},$$

we have $x' = b\sqrt{1 - v^2} - t'v$.

This result may be interpreted to mean that at time t' as measured by the astronaut, the distance from the astronaut to the star, as measured by the astronaut, (see Figure 6.9) is

$$b\sqrt{1 - v^2} - t'v.$$

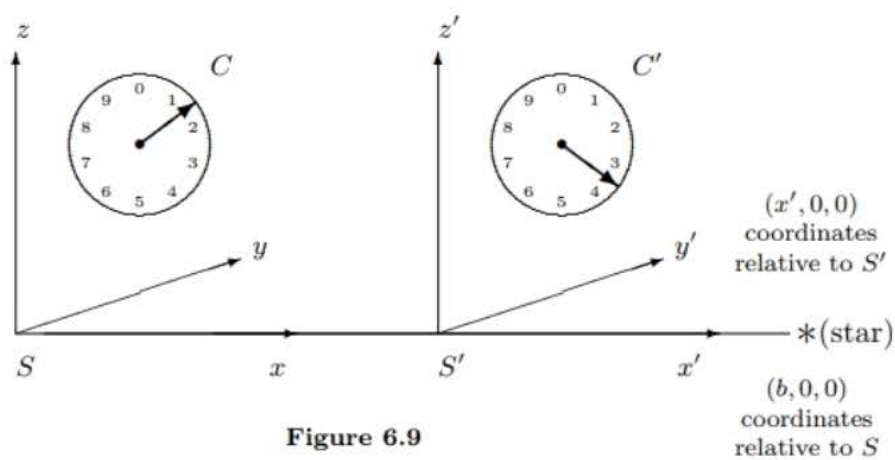


Figure 6.9

(d) Conclude from the preceding equation that

(1) the speed of the space vehicle relative to the star, as measured by the astronaut, is v;

(2) the distance from Earth to the star, as measured by the astronaut, is $b\sqrt{1-v^2}$.

Thus distances along the line of motion of the space vehicle appear to be contracted by a factor of $\sqrt{1-v^2}$.

6.10* CONDITIONING AND THE RAYLEIGH QUOTIENT

In Section 3.4, we studied specific techniques that allow us to solve systems of linear equations in the form $Ax = b$, where A is an $m \times n$ matrix and b is an $m \times 1$ vector. Such systems often arise in applications to the real world. The coefficients in the system are frequently obtained from experimental data, and, in many cases, both m and n are so large that a computer must be used in the calculation of the solution. Thus two types of errors must be considered. First, experimental errors arise in the collection of data since no instruments can provide completely accurate measurements. Second, computers introduce roundoff errors. One might intuitively feel that small relative changes in the coefficients of the system cause small relative errors in the solution. A system that has this property is called **well-conditioned**; otherwise, the system is called **ill-conditioned**.

We now consider several examples of these types of errors, concentrating primarily on changes in b rather than on changes in the entries of A. In addition, we assume that A is a square, complex (or real), invertible matrix since this is the case most frequently encountered in applications.

Example 1

Consider the system

$$\begin{aligned} x_1 + x_2 &= 5 \\ x_1 - x_2 &= 1. \end{aligned}$$

The solution to this system is

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix}.$$

Now suppose that we change the system somewhat and consider the new system

$$\begin{aligned} x_1 + x_2 &= 5 \\ x_1 - x_2 &= 1.0001. \end{aligned}$$

This modified system has the solution

$$\begin{pmatrix} 3.00005 \\ 1.99995 \end{pmatrix}.$$

We see that a change of 10^{-4} in one coefficient has caused a change of less than 10^{-4} in each coordinate of the new solution. More generally, the system

$$\begin{aligned} x_1 + x_2 &= 5 \\ x_1 - x_2 &= 1 + \delta \end{aligned}$$

has the solution

$$\begin{pmatrix} 3+\delta/2 \\ 2-\delta/2 \end{pmatrix}.$$

Hence small changes in b introduce small changes in the solution. Of course, we are really interested in *relative changes* since a change in the solution of, say, 10, is considered large if the original solution is of the order 10^{-2}, but small if the original solution is of the order 10^6.

We use the notation δb to denote the vector $b' - b$, where b is the vector in the original system and b' is the vector in the modified system. Thus we have

$$\delta b = \begin{pmatrix} 5 \\ 1+h \end{pmatrix} - \begin{pmatrix} 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ h \end{pmatrix}.$$

We now define the **relative change** in b to be the scalar $\|\delta b\|/\|b\|$, where $\|\cdot\|$ denotes the standard norm on C^n (or R^n); that is, $\|b\| = \sqrt{\langle b, b\rangle}$. Most of what follows, however, is true for any norm. Similar definitions hold for the **relative change** in x. In this example,

$$\frac{\|\delta b\|}{\|b\|} = \frac{|h|}{\sqrt{26}} \quad \text{and} \quad \frac{\|\delta x\|}{\|x\|} = \frac{\left\| \begin{pmatrix} 3+h/2 \\ 2-h/2 \end{pmatrix} - \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right\|}{\left\| \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right\|} = \frac{|h|}{\sqrt{26}}.$$

Thus the relative change in x equals, coincidentally, the relative change in b; so the system is well-conditioned. ◆

Example 2

Consider the system

$$\begin{aligned} x_1 &+ & x_2 &= 3 \\ x_1 &+ 1.00001x_2 &= 3.00001, \end{aligned}$$

which has

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

as its solution. The solution to the related system

$$\begin{aligned} x_1 &+ & x_2 &= 3 \\ x_1 &+ 1.00001x_2 &= 3.00001 + \delta \end{aligned}$$

is

$$\begin{pmatrix} 2-(10^5)h \\ 1+(10^5)h \end{pmatrix}.$$

Hence,

$$\frac{\|\delta x\|}{\|x\|} = 10^5\sqrt{2/5}\,|h| \geq 10^4|h|,$$

while

$$\frac{\|\delta b\|}{\|b\|} \approx \frac{|h|}{3\sqrt{2}}.$$

Thus the relative change in x is at least 10^4 times the relative change in b! This system is very ill-conditioned. Observe that the lines defined by the two equations are nearly coincident. So a small change in either line could greatly alter the point of intersection, that is, the solution to the system. ♦

To apply the full strength of the theory of self-adjoint matrices to the study of conditioning, we need the notion of the norm of a matrix. (See Exercises 26–30 of Section 6.1 for further results about norms.)

Definition. *Let A be a complex (or real) $n \times n$ matrix. Define the* ***norm*** *of A by*

$$\|A\|_E = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|},$$

where $x \in \mathsf{C}^n$ or $x \in \mathsf{R}^n$.

Intuitively, $\|A\|_E$ represents the maximum *magnification* of a vector by the matrix A. The question of whether or not this maximum exists, as well as the problem of how to compute it, can be answered by the use of the so-called *Rayleigh quotient.*

Definition. *Let B be an $n \times n$ self-adjoint matrix. The* ***Rayleigh quotient*** *for $x \neq 0$ is defined to be the scalar $R(x) = \langle Bx, x\rangle / \|x\|^2$.*

The following result characterizes the extreme values of the Rayleigh quotient of a self-adjoint matrix.

Theorem 6.44. *For a self-adjoint matrix $B \in \mathsf{M}_{n\times n}(F)$, we have that $\max_{x\neq 0} R(x)$ is the largest eigenvalue of B and $\min_{x\neq 0} R(x)$ is the smallest eigenvalue of B.*

Proof. By Theorems 6.19 (p. 381) and 6.20 (p. 381), we may choose an orthonormal basis $\{v_1, v_2, \ldots, v_n\}$ of eigenvectors of B such that $Bv_i = \lambda_i v_i$ $(1 \leq i \leq n)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. (Recall that by the lemma to Theorem 6.17, p. 370, the eigenvalues of B are real.) Now, for $x \in \mathsf{F}^n$, there exist scalars $a_1, a_2, \ldots, a_n$ such that

$$x = \sum_{i=1}^{n} a_i v_i.$$

Hence

$$R(x) = \frac{\langle Bx, x\rangle}{\|x\|^2} = \frac{\left\langle \sum_{i=1}^{n} a_i \lambda_i v_i, \sum_{j=1}^{n} a_j v_j \right\rangle}{\|x\|^2}$$

$$= \frac{\sum_{i=1}^{n} \lambda_i |a_i|^2}{\|x\|^2} \leq \frac{\lambda_1 \sum_{i=1}^{n} |a_i|^2}{\|x\|^2} = \frac{\lambda_1 \|x\|^2}{\|x\|^2} = \lambda_1.$$

It is easy to see that $R(v_1) = \lambda_1$, so we have demonstrated the first half of the theorem. The second half is proved similarly. ∎

Corollary 1. *For any square matrix A, $\|A\|_E$ is finite and, in fact, equals $\sqrt{\lambda}$, where λ is the largest eigenvalue of A^*A.*

Proof. Let B be the self-adjoint matrix A^*A, and let λ be the largest eigenvalue of B. Since, for $x \neq 0$,

$$0 \leq \frac{\|Ax\|^2}{\|x\|^2} = \frac{\langle Ax, Ax \rangle}{\|x\|^2} = \frac{\langle A^*Ax, x \rangle}{\|x\|^2} = \frac{\langle Bx, x \rangle}{\|x\|^2} = R(x),$$

it follows from Theorem 6.44 that $\|A\|_E^2 = \lambda$. ∎

Observe that the proof of Corollary 1 shows that all the eigenvalues of A^*A are nonnegative. For our next result, we need the following lemma.

Lemma. *For any square matrix A, λ is an eigenvalue of A^*A if and only if λ is an eigenvalue of AA^*.*

Proof. Let λ be an eigenvalue of A^*A. If $\lambda = 0$, then A^*A is not invertible. Hence A and A^* are not invertible, so that λ is also an eigenvalue of AA^*. The proof of the converse is similar.

Suppose now that $\lambda \neq 0$. Then there exists $x \neq 0$ such that $A^*Ax = \lambda x$. Apply A to both sides to obtain $(AA^*)(Ax) = \lambda(Ax)$. Since $Ax \neq 0$ (lest $\lambda x = 0$), we have that λ is an eigenvalue of AA^*. The proof of the converse is left as an exercise. ∎

Corollary 2. *Let A be an invertible matrix. Then $\|A^{-1}\|_E = 1/\sqrt{\lambda}$, where λ is the smallest eigenvalue of A^*A.*

Proof. Recall that λ is an eigenvalue of an invertible matrix if and only if λ^{-1} is an eigenvalue of its inverse.

Now let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A^*A, which by the lemma are the eigenvalues of AA^*. Then $\|A^{-1}\|_E^2$ equals the largest eigenvalue of $(A^{-1})^*A^{-1} = (AA^*)^{-1}$, which equals $1/\lambda_n$. ∎

For many applications, it is only the largest and smallest eigenvalues that are of interest. For example, in the case of vibration problems, the smallest eigenvalue represents the lowest frequency at which vibrations can occur.

We see the role of both of these eigenvalues in our study of conditioning.

Example 3

Let

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

Then

$$B = A^*A = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

The eigenvalues of B are 3, 3, and 0. Therefore $\|A\|_E = \sqrt{3}$. For any

$$x = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \neq 0,$$

we may compute $R(x)$ for the matrix B as

$$3 \geq R(x) = \frac{\langle Bx, x\rangle}{\|x\|^2} = \frac{2(a^2 + b^2 + c^2 - ab + ac + bc)}{a^2 + b^2 + c^2}. \quad \blacklozenge$$

Now that we know $\|A\|_E$ exists for every square matrix A, we can make use of the inequality $\|Ax\| \leq \|A\|_E \cdot \|x\|$, which holds for every x.

Assume in what follows that A is invertible, $b \neq 0$, and $Ax = b$. For a given δb, let δx be the vector that satisfies $A(x + \delta x) = b + \delta b$. Then $A(\delta x) = \delta b$, and so $\delta x = A^{-1}(\delta b)$. Hence

$$\|b\| = \|Ax\| \leq \|A\|_E \cdot \|x\| \quad \text{and} \quad \|\delta x\| = \|A^{-1}(\delta b)\| \leq \|A^{-1}\|_E \cdot \|\delta b\|.$$

Thus

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|x\|^2}{\|b\|/\|A\|_E} \leq \frac{\|A^{-1}\|_E \cdot \|\delta b\| \cdot \|A\|_E}{\|b\|} = \|A\|_E \cdot \|A^{-1}\|_E \cdot \left(\frac{\|\delta b\|}{\|b\|}\right).$$

Similarly (see Exercise 9),

$$\frac{1}{\|A\|_E \cdot \|A^{-1}\|_E} \left(\frac{\|\delta b\|}{\|b\|}\right) \leq \frac{\|\delta x\|}{\|x\|}.$$

The number $\|A\|_E \cdot \|A^{-1}\|_E$ is called the **condition number** of A and is denoted $\text{cond}(A)$. We summarize these results in the following theorem.

Theorem 6.45. *For the system $Ax = b$, where A is invertible and $b \neq 0$, the following statements are true.*

(a) *We have* $\dfrac{1}{\text{cond}(A)} \dfrac{\|\delta b\|}{\|b\|} \leq \dfrac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \dfrac{\|\delta b\|}{\|b\|}.$

(b) *If* λ_1 *and* λ_n *are the largest and smallest eigenvalues, respectively, of* A^*A, *then* $\text{cond}(A) = \sqrt{\lambda_1/\lambda_n}$.

Proof. Statement (a) follows from the previous inequalities, and (b) follows from Corollaries 1 and 2 to Theorem 6.44. ∎

It should be noted that the definition of $\text{cond}(A)$ depends on how the norm of A is defined. There are many reasonable ways of defining the norm of a matrix. In fact, the only property needed to establish Theorem 6.45(a) and the two displayed inequalities preceding it is that $\|Ax\| \leq \|A\|_E \cdot \|x\|$ for all x.

It is clear from Theorem 6.45(a) that $\text{cond}(A) \geq 1$. It is left as an exercise to prove that $\text{cond}(A) = 1$ if and only if A is a scalar multiple of a unitary or orthogonal matrix. Moreover, it can be shown with some work that equality can be obtained in (a) by an appropriate choice of b and δb.

We can see immediately from (a) that if $\text{cond}(A)$ is close to 1, then a small relative error in b forces a small relative error in x. If $\text{cond}(A)$ is large, however, then the relative error in x may be small even though the relative error in b is large, or the relative error in x may be large even though the relative error in b is small! In short, $\text{cond}(A)$ merely indicates the *potential* for large relative errors.

We have so far considered only errors in the vector b. If there is an error δA in the coefficient matrix of the system $Ax = b$, the situation is more complicated. For example, $A + \delta A$ may fail to be invertible. But under the appropriate assumptions, it can be shown that a bound for the relative error in x can be given in terms of $\text{cond}(A)$. For example, Charles Cullen (Charles G. Cullen, *An Introduction to Numerical Linear Algebra*, PWS Publishing Co., Boston 1994, p. 60) shows that if $A + \delta A$ is invertible, then

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \text{cond}(A) \frac{\|\delta A\|_E}{\|A\|_E}.$$

It should be mentioned that, in practice, one never computes $\text{cond}(A)$ from its definition, for it would be an unnecessary waste of time to compute A^{-1} merely to determine its norm. In fact, if a computer is used to find A^{-1}, the computed inverse of A in all likelihood only approximates A^{-1}, and the error in the computed inverse is affected by the size of $\text{cond}(A)$. So we are caught in a vicious circle! There are, however, some situations in which a usable approximation of $\text{cond}(A)$ can be found. Thus, in most cases, the estimate of the relative error in x is based on an estimate of $\text{cond}(A)$.

EXERCISES

1. Label the following statements as true or false.

(a) If $Ax = b$ is well-conditioned, then $\text{cond}(A)$ is small.
(b) If $\text{cond}(A)$ is large, then $Ax = b$ is ill-conditioned.
(c) If $\text{cond}(A)$ is small, then $Ax = b$ is well-conditioned.
(d) The norm of A equals the Rayleigh quotient.
(e) The norm of A always equals the largest eigenvalue of A.

2. Compute the norms of the following matrices.

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

3. Prove that if B is symmetric, then $\|B\|_E$ is the largest eigenvalue of B.

4. Let A and A^{-1} be as follows:

$$A = \begin{pmatrix} 6 & 13 & -17 \\ 13 & 29 & -38 \\ -17 & -38 & 50 \end{pmatrix} \quad \text{and} \quad A^{-1} = \begin{pmatrix} 6 & -4 & 1 \\ -4 & 11 & 7 \\ -1 & 7 & 5 \end{pmatrix}.$$

The eigenvalues of A are approximately 84.74, 0.2007, and 0.0588.

(a) Approximate $\|A\|_E$, $\|A^{-1}\|_E$, and $\text{cond}(A)$. (Note Exercise 3.)
(b) Suppose that we have vectors x and $\tilde{x}$ such that $Ax = b$ and $\|b - A\tilde{x}\| \leq 0.001$. Use (a) to determine upper bounds for $\|\tilde{x} - A^{-1}b\|$ (the absolute error) and $\|\tilde{x} - A^{-1}b\|/\|A^{-1}b\|$ (the relative error).

5. Suppose that x is the actual solution of $Ax = b$ and that a computer arrives at an approximate solution $\tilde{x}$. If $\text{cond}(A) = 100$, $\|b\| = 1$, and $\|b - A\tilde{x}\| = 0.1$, obtain upper and lower bounds for $\|x - \tilde{x}\|/\|x\|$.

6. Let

$$B = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Compute

$$R\begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}, \quad \|B\|_E, \quad \text{and} \quad \text{cond}(B).$$

7. Let B be a symmetric matrix. Prove that $\min_{x \neq 0} R(x)$ equals the smallest eigenvalue of B.

8. Prove that if λ is an eigenvalue of AA^*, then λ is an eigenvalue of A^*A. This completes the proof of the lemma to Corollary 2 to Theorem 6.44.

9. Prove that if A is an invertible matrix and $Ax = b$, then

$$\frac{1}{\|A\|_E \cdot \|A^{-1}\|_E}\left(\frac{\|\delta b\|}{\|b\|}\right) \leq \frac{\|\delta x\|}{\|x\|}.$$

10. Prove the left inequality of (a) in Theorem 6.45.

11. Prove that $\text{cond}(A) = 1$ if and only if A is a scalar multiple of a unitary or orthogonal matrix.

12. **(a)** Let A and B be square matrices that are unitarily equivalent. Prove that $\|A\|_E = \|B\|_E$.

(b) Let T be a linear operator on a finite-dimensional inner product space V. Define

$$\|\mathsf{T}\|_E = \max_{x \neq 0} \frac{\|\mathsf{T}(x)\|}{\|x\|}.$$

Prove that $\|\mathsf{T}\|_E = \|[\mathsf{T}]_\beta\|_E$, where β is any orthonormal basis for V.

(c) Let V be an infinite-dimensional inner product space with an orthonormal basis $\{v_1, v_2, \ldots\}$. Let T be the linear operator on V such that $\mathsf{T}(v_k) = kv_k$. Prove that $\|\mathsf{T}\|_E$ (defined in (b)) does not exist.

Visit `goo.gl/B8Uw33` for a solution.

The next exercise assumes the definitions of *singular value* and *pseudoinverse* and the results of Section 6.7.

13. Let A be an $n \times n$ matrix of rank r with the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. Prove each of the following results.

(a) $\|A\|_E = \sigma_1$.

(b) $\|A^\dagger\|_E = \dfrac{1}{\sigma_r}$.

(c) If A is invertible (and hence $r = n$), then $\text{cond}(A) = \dfrac{\sigma_1}{\sigma_n}$.

6.11* THE GEOMETRY OF ORTHOGONAL OPERATORS

By Theorem 6.22 (p. 383), any rigid motion on a finite-dimensional real inner product space is the composite of an orthogonal operator and a translation. Thus, to understand the geometry of rigid motions thoroughly, we must analyze the structure of orthogonal operators. In this section, we show that any orthogonal operator on a finite-dimensional real inner product space can be described in terms of rotations and reflections.

This material assumes familiarity with the results about direct sums developed at the end of Section 5.2 and with the definition of the determinant of a linear operator given in Section 5.1 as well as elementary properties of the determinant in Exercise 8 of Section 5.1.

We now extend our earlier definitions of *rotation* and *reflection* on R^2 to all 2-dimensional real inner product spaces.

Definitions. *Let* T *be a linear operator on a two-dimensional real inner product space* V.

We call T *a* ***rotation*** *if there exists an orthonormal basis* $\beta = \{x_1, x_2\}$ *for* V *and a real number* θ *such that*

$$\mathsf{T}(x_1) = (\cos\theta)x_1 + (\sin\theta)x_2 \quad \text{and} \quad \mathsf{T}(x_2) = (-\sin\theta)x_1 + (\cos\theta)x_2.$$

We call T *a* ***reflection*** *if there exists a one-dimensional subspace* W *of* V *such that* $\mathsf{T}(x) = -x$ *for all* $x \in \mathsf{W}$ *and* $\mathsf{T}(y) = y$ *for all* $y \in \mathsf{W}^\perp$. *In this context,* T *is called a* ***reflection of*** V ***about*** $\mathsf{W}^\perp$.

As a convenience, we define *rotations* and *reflections* on a 1-dimensional inner product space.

Definitions. *A linear operator* T *on a 1-dimensional inner product space* V *is called a* ***rotation*** *if* T *is the identity and a* ***reflection*** *if* $\mathsf{T}(x) = -x$ *for all* $x \in \mathsf{V}$.

Trivially, rotations and reflections on 1-dimensional inner product spaces are orthogonal operators. It should be noted that rotations and reflections on 2-dimensional real inner product spaces (or composites of these) are orthogonal operators (see Exercise 2).

Example 1

Some Typical Reflections

(a) Define $\mathsf{T}\colon \mathsf{R}^2 \to \mathsf{R}^2$ by $\mathsf{T}(a,b) = (-a,b)$, and let $\mathsf{W} = \operatorname{span}(\{e_1\})$. Then $\mathsf{T}(x) = -x$ for all $x \in \mathsf{W}$, and $\mathsf{T}(y) = y$ for all $y \in \mathsf{W}^\perp$. Thus T is a reflection of R^2 about $\mathsf{W}^\perp = \operatorname{span}(\{e_2\})$, the y-axis.

(b) Define $\mathsf{T}\colon \mathsf{R}^2 \to \mathsf{R}^2$ by $\mathsf{T}(a,b) = (b,a)$, and let $\mathsf{W} = \operatorname{span}(\{(1,1)\})$. Clearly $\mathsf{T}(w) = w$ for all $w \in \mathsf{W}$. Let $(a,b) \in \mathsf{W}^\perp$. Then (a,b) is orthogonal to $(1,1)$ and hence, $a + b = 0$. So $b = -a$. Thus $\mathsf{W}^\perp = \operatorname{span}(\{(1,-1)\})$. It follows that $\mathsf{T}(a,b) = (a,-a) = -(-a,a) = -(b,a)$. Hence T is the reflection of R^2 about W. ◆

The next theorem characterizes all orthogonal operators on a two-dimensional real inner product space V. The proof follows from Theorem 6.23 (p. 384) since all two-dimensional real inner product spaces are structurally identical. For a rigorous justification, apply Theorem 2.21 (p. 105), where β is an orthonormal basis for V. By Exercise 15 of Section 6.2, the resulting isomorphism $\phi_\beta\colon \mathsf{V} \to \mathsf{R}^2$ preserves inner products. (See Exercise 8.)

Theorem 6.46. *Let* T *be an orthogonal operator on a two-dimensional real inner product space* V. *Then* T *is either a rotation or a reflection. Furthermore,* T *is a rotation if and only if* $\det(\mathsf{T}) = 1$, *and* T *is a reflection if and only if* $\det(\mathsf{T}) = -1$.

A complete description of the reflections of R^2 is given in Section 6.5.

Corollary. *Let* V *be a two-dimensional real inner product space.*

(a) *The composite of a reflection and a rotation on* V *is a reflection on* V.
(b) *The composite of two reflections on* V *is a rotation on* V.
(c) *The product of two rotations on* V *is a rotation on* V.

Proof. If T_1 is a reflection on V and T_2 is a rotation on V, then by Theorem 6.46, $\det(\mathsf{T}_1) = 1$ and $\det(\mathsf{T}_2) = -1$. Let $\mathsf{T} = \mathsf{T}_2\mathsf{T}_1$ be the composite. Since T_2 and T_1 are orthogonal, so is T. Moreover, $\det(\mathsf{T}) = \det(\mathsf{T}_2) \cdot \det(\mathsf{T}_1) = -1$. Thus, by Theorem 6.46, T is a reflection. The proof for $\mathsf{T}_1\mathsf{T}_2$ is similar.

The proofs of (b) and (c) are similar to that of (a). ∎

We now study orthogonal operators on spaces of higher dimension.

Lemma. *If* T *is a linear operator on a nonzero finite-dimensional real vector space* V, *then there exists a* T*-invariant subspace* W *of* V *such that* $1 \leq \dim(\mathsf{W}) \leq 2$.

Proof. Fix an ordered basis $\beta = \{y_1, y_2, \ldots, y_n\}$ for V, and let $A = [\mathsf{T}]_\beta$. Let $\phi_\beta \colon \mathsf{V} \to \mathsf{R}^n$ be the linear transformation defined by $\phi_\beta(y_i) = e_i$ for $i = 1, 2, \ldots, n$. Then ϕ_β is an isomorphism, and, as we have seen in Section 2.4, the diagram in Figure 6.10 commutes, that is, $\mathsf{L}_A\phi_\beta = \phi_\beta\mathsf{T}$. As a consequence, it suffices to show that there exists an L_A-invariant subspace Z of R^n such that $1 \leq \dim(\mathsf{Z}) \leq 2$. If we then define $\mathsf{W} = \phi_\beta^{-1}(\mathsf{Z})$, it follows that W satisfies the conclusions of the lemma (see Exercise 12).

$$\begin{array}{ccc} \mathsf{V} & \xrightarrow{\mathsf{T}} & \mathsf{V} \\ \downarrow{\scriptstyle \phi_\beta} & & \downarrow{\scriptstyle \phi_\beta} \\ \mathsf{R}^n & \xrightarrow{\mathsf{L}_A} & \mathsf{R}^n \end{array}$$

Figure 6.10

The matrix A can be considered as an $n \times n$ matrix over C and, as such, can be used to define a linear operator U on C^n by $\mathsf{U}(v) = Av$. Since U is a linear operator on a finite-dimensional vector space over C, it has an

eigenvalue $\lambda \in C$. Let $x \in \mathsf{C}^n$ be an eigenvector corresponding to λ. We may write $\lambda = \lambda_1 + i\lambda_2$, where λ_1 and λ_2 are real, and

$$x = \begin{pmatrix} a_1 + ib_1 \\ a_2 + ib_2 \\ \vdots \\ a_n + ib_n \end{pmatrix},$$

where the a_i's and b_i's are real. Thus, setting

$$x_1 = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad \text{and} \quad x_2 = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix},$$

we have $x = x_1 + ix_2$, where x_1 and x_2 have real entries. Note that at least one of x_1 or x_2 is nonzero since $x \neq 0$. Hence

$$\mathsf{U}(x) = \lambda x = (\lambda_1 + i\lambda_2)(x_1 + ix_2) = (\lambda_1 x_1 - \lambda_2 x_2) + i(\lambda_1 x_2 + \lambda_2 x_1).$$

Similarly,

$$\mathsf{U}(x) = A(x_1 + ix_2) = Ax_1 + iAx_2.$$

Comparing the real and imaginary parts of these two expressions for $\mathsf{U}(x)$, we conclude that

$$Ax_1 = \lambda_1 x_1 - \lambda_2 x_2 \quad \text{and} \quad Ax_2 = \lambda_1 x_2 + \lambda_2 x_1.$$

Finally, let $\mathsf{Z} = \text{span}(\{x_1, x_2\})$, the span being taken as a subspace of R^n. Since $x_1 \neq 0$ or $x_2 \neq 0$, Z is a nonzero subspace. Thus $1 \leq \dim(\mathsf{Z}) \leq 2$, and the preceding pair of equations shows that Z is L_A-invariant. ∎

Theorem 6.47. *Let T be an orthogonal operator on a nonzero finite-dimensional real inner product space V. Then there exists a collection of pairwise orthogonal T-invariant subspaces $\{\mathsf{W}_1, \mathsf{W}_2, \ldots, \mathsf{W}_m\}$ of V such that*

(a) $1 \leq \dim(\mathsf{W}_i) \leq 2$ *for* $i = 1, 2, \ldots, m$.

(b) $\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 \oplus \cdots \oplus \mathsf{W}_m$.

Proof. The proof is by mathematical induction on $\dim(\mathsf{V})$. If $\dim(\mathsf{V}) = 1$, the result is obvious. So assume that the result is true whenever $\dim(\mathsf{V}) < n$ for some fixed integer $n > 1$.

Suppose $\dim(\mathsf{V}) = n$. By the lemma, there exists a T-invariant subspace W_1 of V such that $1 \leq \dim(\mathsf{W}) \leq 2$. If $\mathsf{W}_1 = \mathsf{V}$, the result is established. Otherwise, $\mathsf{W}_1^\perp \neq \{0\}$. By Exercise 13, $\mathsf{W}_1^\perp$ is T-invariant and the restriction

of T to $W_1^\perp$ is orthogonal. Since $\dim(W_1^\perp) < n$, we may apply the induction hypothesis to $T_{W_1^\perp}$ and conclude that there exists a collection of pairwise orthogonal T-invariant subspaces $\{W_1, W_2, \ldots, W_m\}$ of $W_1^\perp$ such that $1 \leq \dim(W_i) \leq 2$ for $i = 2, 3, \ldots, m$ and $W_1^\perp = W_2 \oplus W_3 \oplus \cdots \oplus W_m$. Thus $\{W_1, W_2, \ldots, W_m\}$ is pairwise orthogonal, and by Exercise 13(d) of Section 6.2,

$$V = W_1 \oplus W_1^\perp = W_1 \oplus W_2 \oplus \cdots \oplus W_m. \qquad \blacksquare$$

Applying Theorem 6.46 in the context of Theorem 6.47, we conclude that the restriction of T to W_i is either a rotation or a reflection for each $i = 1, 2, \ldots, m$. Thus, in some sense, T is composed of rotations and reflections. Unfortunately, very little can be said about the uniqueness of the decomposition of V in Theorem 6.47. For example, the W_i's, the number m of W_i's, and the number of W_i's for which T_{W_i} is a reflection are not unique. Although the number of W_i's for which T_{W_i} is a reflection is not unique, whether this number is even or odd is an intrinsic property of T. Moreover, we can always decompose V so that T_{W_i} is a reflection for at most one W_i. These facts are established in the following result.

Theorem 6.48. *Let $T, V, W_1, \ldots, W_m$ be as in Theorem 6.47.*

(a) *The number of W_i's for which T_{W_i} is a reflection is even or odd according to whether $\det(T) = 1$ or $\det(T) = -1$.*

(b) *It is always possible to decompose V as in Theorem 6.47 so that the number of W_i's for which T_{W_i} is a reflection is zero or one according to whether $\det(T) = 1$ or $\det(T) = -1$. Furthermore, if T_{W_i} is a reflection, then $\dim(W_i) = 1$.*

Proof. (a) Let r denote the number of W_i's in the decomposition for which T_{W_i} is a reflection. Then, by Exercise 14,

$$\det(T) = \det(T_{W_1}) \cdot \det(T_{W_2}) \cdot \cdots \cdot \det(T_{W_m}) = (-1)^r,$$

proving (a).

(b) Let $E = \{x \in V : T(x) = -x\}$; then E is a T-invariant subspace of V. Let $W = E^\perp$; then W is T-invariant. So by applying Theorem 6.47 to T_W, we obtain a collection of pairwise orthogonal T-invariant subspaces $\{W_1, W_2, \ldots, W_k\}$ of W such that $W = W_1 \oplus W_2 \oplus \cdots \oplus W_k$, and for $1 \leq i \leq k$, the dimension of each W_i is either 1 or 2. Observe that, for each $i = 1, 2, \ldots, k$, T_{W_i} is a rotation. For otherwise, if T_{W_i} is a reflection, there exists a nonzero $x \in W_i$ for which $T(x) = -x$. But then, $x \in W_i \cap E \subseteq E^\perp \cap E = \{0\}$, a contradiction. If $E = \{0\}$, the result follows. Otherwise, choose an orthonormal basis β for E containing p vectors ($p > 0$). It is possible to decompose β into a pairwise disjoint union $\beta = \beta_1 \cup \beta_2 \cup \cdots \cup \beta_r$ such that each β_i contains exactly two vectors for $i < r$, and β_r contains

two vectors if p is even and one vector if p is odd. For each $i = 1, 2, \ldots, r$, let $\mathsf{W}_{k+i} = \text{span}(\beta_i)$. Then, clearly, $\{\mathsf{W}_1, \mathsf{W}_2, \ldots, \mathsf{W}_k, \ldots, \mathsf{W}_{k+r}\}$ is pairwise orthogonal, and

$$\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 \oplus \cdots \oplus \mathsf{W}_k \oplus \cdots \oplus \mathsf{W}_{k+r}. \tag{27}$$

Moreover, if any β_i contains two vectors, then

$$\det(\mathsf{T}_{\mathsf{W}_{k+i}}) = \det([\mathsf{T}_{\mathsf{W}_{k+i}}]_{\beta_i}) = \det \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = 1.$$

So $\mathsf{T}_{\mathsf{W}_{k+i}}$ is a rotation, and hence $\mathsf{T}_{\mathsf{W}_j}$ is a rotation for $j < k + r$. If β_r consists of one vector, then $\dim(\mathsf{W}_{k+r}) = 1$ and

$$\det(\mathsf{T}_{\mathsf{W}_{k+r}}) = \det([\mathsf{T}_{\mathsf{W}_{k+r}}]_{\beta_r}) = \det(-1) = -1.$$

Thus $\mathsf{T}_{\mathsf{W}_{k+r}}$ is a reflection by Theorem 6.47, and we conclude that the decomposition in (27) satisfies (b). ∎

Example 2

Orthogonal Operators on a Three-Dimensional Real Inner Product Space

Let T be an orthogonal operator on a three-dimensional real inner product space V. Then, by Theorem 6.48(b), V can be decomposed into a direct sum of T-invariant orthogonal subspaces so that the restriction of T to each is either a rotation or a reflection, with at most one reflection. Let

$$\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 \oplus \cdots \oplus \mathsf{W}_m$$

be such a decomposition. Clearly, $m = 2$ or $m = 3$.

If $m = 2$, then $\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2$. Without loss of generality, suppose that $\dim(\mathsf{W}_1) = 1$ and $\dim(\mathsf{W}_2) = 2$. Thus $\mathsf{T}_{\mathsf{W}_1}$ is a reflection or the identity on W_1, and $\mathsf{T}_{\mathsf{W}_2}$ is a rotation.

If $m = 3$, then $\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 \oplus \mathsf{W}_3$ and $\dim(\mathsf{W}_i) = 1$ for all i. If $\mathsf{T}_{\mathsf{W}_i}$ is not a reflection, then it is the identity on W_i. If no $\mathsf{T}_{\mathsf{W}_i}$ is a reflection, then T is the identity operator. ♦

EXERCISES

1. Label the following statements as true or false. Assume that the underlying vector spaces are one or two-dimensional real inner product spaces.
   - **(a)** Any orthogonal operator is either a rotation or a reflection.
   - **(b)** The composite of any two rotations is a rotation.
   - **(c)** The identity operator is a rotation.

(d) The composite of two reflections is a reflection.
(e) Any orthogonal operator is a composite of rotations.
(f) For any orthogonal operator T, if $\det(\mathsf{T}) = -1$, then T is a reflection.
(g) Reflections always have eigenvalues.
(h) Rotations always have eigenvalues.
(i) If T is an operator on a 2-dimensional space V and W is a subspace of dimension 1 such that T is a reflection of V about $\mathsf{W}^{\perp}$, then W is the eigenspace of T corresponding to the eigenvalue $\lambda = -1$.
(j) The composite of an orthogonal operator and a translation is an orthogonal operator.

2. Prove that rotations, reflections, and composites of rotations and reflections are orthogonal operators.

3. Let

$$A = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

(a) Prove that L_A is a reflection.
(b) Find the subspace of R^2 on which L_A acts as the identity.
(c) Prove that L_{AB} and L_{BA} are rotations.

4. For any real number ϕ, let

$$A = \begin{pmatrix} \cos\phi & \sin\phi \\ \sin\phi & -\cos\phi \end{pmatrix}.$$

(a) Prove that L_A is a reflection.
(b) Find the axis in R^2 about which L_A reflects.

5. For any real number ϕ, define $\mathsf{T}_\phi = \mathsf{L}_A$, where

$$A = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}.$$

(a) Prove that any rotation on R^2 is of the form T_ϕ for some ϕ.
(b) Prove that $\mathsf{T}_\phi \mathsf{T}_\psi = \mathsf{T}_{(\phi+\psi)}$ for any $\phi, \psi \in R$.
(c) Deduce that any two rotations on R^2 commute.

6. Prove that if T is a rotation on a 2-dimensional inner product space, then $-\mathsf{T}$ is also a rotation.

7. Prove that if T is a reflection on a 2-dimensional inner product space, then T^2 is the identity operator.

8. Prove Theorem 6.46 using the hints preceding the statement of the theorem.

9. Prove that no orthogonal operator can be both a rotation and a reflection.

10. Prove that if V is a two-dimensional real inner product space, then the composite of two reflections on V is a rotation of V.

11. Let V be a one- or a two-dimensional real inner product space. Define $\mathsf{T}: \mathsf{V} \to \mathsf{V}$ by $\mathsf{T}(x) = -x$. Prove that T is a rotation if and only if $\dim(\mathsf{V}) = 2$.

12. Complete the proof of the lemma to Theorem 6.47 by showing that $\mathsf{W} = \phi_\beta^{-1}(\mathsf{Z})$ satisfies the required conditions.

13. Let T be an orthogonal [unitary] operator on a finite-dimensional real [complex] inner product space V. If W is a T-invariant subspace of V, prove the following results.

(a) T_W is an orthogonal [unitary] operator on W.

(b) $\mathsf{W}^\perp$ is a T-invariant subspace of V. *Hint:* Use the fact that T_W is one-to-one and onto to conclude that, for any $y \in \mathsf{W}$, $\mathsf{T}^*(y) = \mathsf{T}^{-1}(y) \in \mathsf{W}$.

(c) $\mathsf{T}_{\mathsf{W}^\perp}$ is an orthogonal [unitary] operator on W.

14. Let T be a linear operator on a finite-dimensional vector space V, where V is a direct sum of T-invariant subspaces, say, $\mathsf{V} = \mathsf{W}_1 \oplus \mathsf{W}_2 \oplus \cdots \oplus \mathsf{W}_k$. Prove that $\det(\mathsf{T}) = \det(\mathsf{T}_{\mathsf{W}_1}) \cdot \det(\mathsf{T}_{\mathsf{W}_2}) \cdot \cdots \cdot \det(\mathsf{T}_{\mathsf{W}_k})$.

15. Complete the proof of the corollary to Theorem 6.48.

16. Let V be a real inner product space of dimension 2. For any $x, y \in \mathsf{V}$ such that $x \neq y$ and $\|x\| = \|y\| = 1$, show that there exists a unique rotation T on V such that $\mathsf{T}(x) = y$. Visit `goo.gl/ahQT67` for a solution.

INDEX OF DEFINITIONS FOR CHAPTER 6

7

Canonical Forms

As we learned in Chapter 5, the advantage of a diagonalizable linear operator lies in the simplicity of its description. Such an operator has a diagonal matrix representation, or, equivalently, there is an ordered basis for the underlying vector space consisting of eigenvectors of the operator. However, not every linear operator is diagonalizable, even if its characteristic polynomial splits. Example 3 of Section 5.2 describes such an operator.

It is the purpose of this chapter to consider alternative matrix representations for nondiagonalizable operators. These representations are called *canonical forms*. There are different kinds of canonical forms, and their advantages and disadvantages depend on how they are applied. Every canonical form of an operator is obtained by an appropriate choice of an ordered basis. Naturally, the canonical forms of a linear operator are not diagonal matrices if the linear operator is not diagonalizable.

In this chapter, we treat two common canonical forms. The first of these, the *Jordan canonical form*, requires that the characteristic polynomial of the operator splits. This form is always available if the underlying field is algebraically closed, that is, if every polynomial with coefficients from the field splits. For example, the field of complex numbers is algebraically closed by the fundamental theorem of algebra (see Appendix D). The first two sections deal with this form. The *rational canonical form*, treated in Section 7.4, does not require such a factorization.

7.1 THE JORDAN CANONICAL FORM I

Let T be a linear operator on a finite-dimensional vector space V, and suppose that the characteristic polynomial of T splits. Recall from Section 5.2 that the diagonalizability of T depends on whether the union of ordered bases for the distinct eigenspaces of T is an ordered basis for V. So a lack of diagonalizability means that at least one eigenspace of T is too "small."

In this section, we extend the definition of eigenspace to *generalized eigenspace*. From these subspaces, we select ordered bases whose union is an ordered basis β for V such that

$$[\mathsf{T}]_\beta = \begin{pmatrix} A_1 & O & \cdots & O \\ O & A_2 & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & A_k \end{pmatrix},$$

where each O is a zero matrix, and each A_i is a square matrix of the form (λ) or

$$\begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}$$

for some eigenvalue λ of T. Such a matrix A_i is called a **Jordan block** corresponding to λ, and the matrix $[\mathsf{T}]_\beta$ is called a **Jordan canonical form** of T. We also say that the ordered basis β is a **Jordan canonical basis** for T. Observe that each Jordan block A_i is "almost" a diagonal matrix—in fact, $[\mathsf{T}]_\beta$ is a diagonal matrix if and only if each A_i is of the form (λ).

Example 1

Suppose that T is a linear operator on C^8, and $\beta = \{v_1, v_2, \ldots, v_8\}$ is an ordered basis for C^8 such that

$$J = [\mathsf{T}]_\beta = \left(\begin{array}{ccc|ccccc} 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

is a Jordan canonical form of T. Notice that the characteristic polynomial of T is $\det(J - tI) = (t-2)^4(t-3)^2t^2$, and hence the multiplicity of each eigenvalue is the number of times that the eigenvalue appears on the diagonal of J. Also observe that v_1, v_4, v_5, and v_7 are the only vectors in β that are eigenvectors of T. These are the vectors corresponding to the columns of J with no 1 above the diagonal entry. ◆

In Sections 7.1 and 7.2, we prove that every linear operator whose characteristic polynomial splits has a Jordan canonical form that is unique up to

the order of its Jordan blocks. Nevertheless, it is not the case that the Jordan canonical form is completely determined by the characteristic polynomial of the operator. For example, let T' be the linear operator on C^8 such that $[\mathsf{T}']_\beta = J'$, where β is the ordered basis in Example 1 and

$$J' = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then the characteristic polynomial of T' is also $(t-2)^4(t-3)^2t^2$. But the operator T' has the Jordan canonical form J', which is different from J, the Jordan canonical form of the linear operator T of Example 1.

Consider again the matrix J and the ordered basis β of Example 1. Notice that $\mathsf{T}(v_2) = v_1 + 2v_2$, and so $(\mathsf{T} - 2\mathsf{I})(v_2) = v_1$. Similarly, $(\mathsf{T} - 2\mathsf{I})(v_3) = v_2$. Since v_1 and v_4 are eigenvectors of T corresponding to $\lambda = 2$, it follows that $(\mathsf{T} - 2\mathsf{I})^3(v_i) = 0$ for $i = 1, 2, 3$, and 4. Similarly $(\mathsf{T} - 3\mathsf{I})^2(v_i) = 0$ for $i = 5, 6$, and $(\mathsf{T} - 0\mathsf{I})^2(v_i) = 0$ for $i = 7, 8$.

Because of the structure of each Jordan block in a Jordan canonical form, we can generalize these observations: *If v lies in a Jordan canonical basis for a linear operator T and is associated with a $k \times k$ Jordan block with diagonal entry λ, then $(\mathsf{T} - \lambda\mathsf{I})^k(v) = 0$.* Eigenvectors satisfy this condition for $k = 1$.

Definition. *Let T be a linear operator on a vector space V, and let λ be a scalar. A nonzero vector x in V is called a* ***generalized eigenvector of T corresponding to λ*** *if $(\mathsf{T} - \lambda\mathsf{I})^p(x) = 0$ for some positive integer p.*

Notice that if x is a generalized eigenvector of T corresponding to λ, and p is the smallest positive integer for which $(\mathsf{T} - \lambda\mathsf{I})^p(x) = 0$, then $(\mathsf{T} - \lambda\mathsf{I})^{p-1}(x)$ is an eigenvector of T corresponding to λ. Therefore λ is an eigenvalue of T.

In the context of Example 1, each vector in β is a generalized eigenvector of T. In fact, v_1, v_2, v_3, and v_4 correspond to the eigenvalue 2, v_5 and v_6 correspond to the eigenvalue 3, and v_7 and v_8 correspond to the eigenvalue 0.

Just as eigenvectors lie in eigenspaces, generalized eigenvectors lie in "generalized eigenspaces."

Definition. *Let T be a linear operator on a vector space V, and let λ be an eigenvalue of T. The* ***generalized eigenspace of T corresponding to λ****, denoted K_λ, is the subset of V defined by*

$$\mathsf{K}_\lambda = \{x \in \mathsf{V} : (\mathsf{T} - \lambda\mathsf{I})^p(x) = 0 \text{ for some positive integer } p\}.$$

Note that K_λ consists of the zero vector and all generalized eigenvectors corresponding to λ.

Recall that a subspace W of V is T-invariant for a linear operator T if $\mathsf{T}(\mathsf{W}) \subseteq \mathsf{W}$. In the development that follows, we assume the results of Exercises 3 and 4 of Section 5.4. In particular, for any polynomial $g(x)$, if W is T-invariant, then it is also $g(\mathsf{T})$-invariant. Furthermore, the range of a linear operator T is T-invariant.

Theorem 7.1. *Let T be a linear operator on a vetor space V, and let λ be an eigenvalue of T. Then*

(a) *K_λ is a T-invariant subspace of V containing E_λ (the eigenspace of T corresponding to λ).*

(b) *For any eigenvalue μ of T such that $\mu \neq \lambda$,*

$$\mathsf{K}_\lambda \bigcap \mathsf{K}_\mu = \{0\}.$$

(c) *For any scalar $\mu \neq \lambda$, the restriction of $\mathsf{T} - \mu\mathsf{I}$ to K_λ is one-to-one and onto.*

Proof. (a) Clearly, $0 \in \mathsf{K}_\lambda$. Suppose that x and y are in K_λ. Then there exist positive integers p and q such that

$$(\mathsf{T} - \lambda\mathsf{I})^p(x) = (\mathsf{T} - \lambda\mathsf{I})^q(y) = 0.$$

Therefore

$$\begin{aligned}(\mathsf{T} - \lambda\mathsf{I})^{p+q}(x+y) &= (\mathsf{T} - \lambda\mathsf{I})^{p+q}(x) + (\mathsf{T} - \lambda\mathsf{I})^{p+q}(y) \\ &= (\mathsf{T} - \lambda\mathsf{I})^q(0) + (\mathsf{T} - \lambda\mathsf{I})^p(0) \\ &= 0,\end{aligned}$$

and hence $x+y \in \mathsf{K}_\lambda$. The proof that K_λ is closed under scalar multiplication is straightforward.

To show that K_λ is T-invariant, consider any $x \in \mathsf{K}_\lambda$. Choose a positive integer p such that $(\mathsf{T} - \lambda\mathsf{I})^p(x) = 0$. Then

$$(\mathsf{T} - \lambda\mathsf{I})^p\mathsf{T}(x) = \mathsf{T}(\mathsf{T} - \lambda\mathsf{I})^p(x) = \mathsf{T}(0) = 0.$$

Therefore $\mathsf{T}(x) \in \mathsf{K}_\lambda$.

Finally, it is a simple observation that E_λ is contained in K_λ.

(b) Suppose that $w \in \mathsf{K}_\lambda \bigcap \mathsf{K}_\mu$. Then there exist positive integers p and q such that

$$(\mathsf{T} - \lambda\mathsf{I})^p(w) = (\mathsf{T} - \mu\mathsf{I})^q(w) = 0.$$

As polynomials in t, $(t-\lambda)^p$ and $(t-\mu)^q$ are relatively prime, and hence by Theorem E.2 in Appendix E there exist polymomials $q_1(t)$ and $q_2(t)$ such that

$$q_1(t)(t-\lambda)^p + q_2(t)(t-\mu)^q = 1.$$

It follows that

$$q_1(\mathsf{T})(\mathsf{T} - \lambda\mathsf{I})^p(w) + q_2(\mathsf{T})(\mathsf{T} - \mu\mathsf{I})^q(w) = w,$$

and hence $0 = w$.

(c) Let μ be any scalar such that $\mu \neq \lambda$. Since K_λ is T-invariant, it is also $(\mathsf{T} - \mu\mathsf{I})$-invariant. If $(\mathsf{T} - \mu\mathsf{I})(w) = 0$ for some $w \in \mathsf{K}_\lambda$, then $w \in \mathsf{E}_\mu \subset \mathsf{K}_\mu$. Hence $w = 0$ by (b), and we conclude that the restriction of $\mathsf{T} - \mu\mathsf{I}$ to K_λ is one-to-one.

We now show that the restriction of $\mathsf{T} - \mu\mathsf{I}$ to K_λ is onto. Let $x \in \mathsf{K}_\lambda$, p be the smallest positive integer for which $(\mathsf{T} - \lambda\mathsf{I})^p(x) = 0$, and

$$\mathsf{W} = \operatorname{span}\{x, (\mathsf{T} - \lambda\mathsf{I})(x), \ldots, (\mathsf{T} - \lambda\mathsf{I})^{p-1}(x)\}.$$

It is easily shown that W is a T-invariant subspace of K_λ, and hence it is also $(\mathsf{T} - \mu\mathsf{I})$-invariant. So $\mathsf{T} - \mu\mathsf{I}$ maps W to W. Since $\mathsf{T} - \mu\mathsf{I}$ is one-to-one on K_λ, it must be one-to-one on W. Thus, because W is finite-dimensional, $\mathsf{T} - \mu\mathsf{I}$ maps W onto W. So there exists a $y \in \mathsf{W}$ such that $(\mathsf{T} - \mu\mathsf{I})(y) = x$. Therefore $\mathsf{T} - \mu\mathsf{I}$ maps K_λ onto K_λ. ∎

Theorem 7.2. *Let T be a linear operator on a finite-dimensional vector space V such that the characteristic polynomial of T splits. Suppose that λ is an eigenvalue of T with multiplicity m. Then*

(a) $\dim(\mathsf{K}_\lambda) \leq m$.

(b) $\mathsf{K}_\lambda = \mathsf{N}((\mathsf{T} - \lambda\mathsf{I})^m)$.

Proof. (a) Let $\mathsf{W} = \mathsf{K}_\lambda$, and let $h(t)$ be the characteristic polynomial of T_W. By Theorem 5.20 (p. 312), $h(t)$ divides the characteristic polynomial of T, and by Theorem 7.1(b), λ is the only eigenvalue of T_W. Hence $h(t) = (-1)^d(t-\lambda)^d$, where $d = \dim(\mathsf{W})$, and $d \leq m$.

(b) Clearly $\mathsf{N}((\mathsf{T} - \lambda\mathsf{I})^m) \subseteq \mathsf{K}_\lambda$. The characteristic polynomial of T is of the form $f(t) = (t - \lambda)^m g(t)$, where $g(t)$ is a product of powers of the form $(t - \mu)$ for eigenvalues $\mu \neq \lambda$. By Theorem 7.1, $g(\mathsf{T})$ is one-to one on K_λ. Since K_λ is finite-dimensional, it is also onto. Let $x \in \mathsf{K}_\lambda$. Then there exists $y \in \mathsf{K}_\lambda$ such that $g(\mathsf{T})(y) = x$. Hence

$$(\mathsf{T} - \lambda\mathsf{I})^m(x) = (\mathsf{T} - \lambda\mathsf{I})^m g(\mathsf{T})(y) = f(\mathsf{T})(y) = 0$$

by the Cayley-Hamilton Theorem. Therefore $x \in \mathsf{N}((\mathsf{T} - \lambda\mathsf{I})^m)$, and thus $\mathsf{K}_\lambda \subseteq \mathsf{N}((\mathsf{T} - \lambda\mathsf{I})^m)$. ∎

Theorem 7.3. *Let T be a linear operator on a finite-dimensional vector space V such that the characteristic polynomial of T splits, and let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T. Then, for every $x \in \mathsf{V}$, there exist unique vectors $v_i \in \mathsf{K}_{\lambda_i}$, for $i = 1, 2, \ldots, k$, such that*

$$x = v_1 + v_2 + \cdots + v_k.$$

Proof. Suppose the characteristic polynomial of T is

$$f(t) = (t - \lambda_1)^{n_1}(t - \lambda_2)^{n_2} \dots (t - \lambda_k)^{n_k}.$$

For $j = 1, 2, \dots, k$, let

$$f_j(t) = \prod_{\substack{i=1 \\ i \neq j}}^{k} (t - \lambda_i)^{n_i}.$$

First, we prove the existence of the vectors v_i. Because the polynomials $f_j(t)$ are relatively prime for $j = 1, 2, \dots, k$, by Theorem E.2 there exist polynomials $q_j(t)$ for $j = 1, 2, \dots, k$ such that

$$q_1(t)f_1(t) + q_2(t)f_2(t) + \cdots + q_k(t)f_k(t) = 1.$$

It follows that

$$v = q_1(\mathsf{T})f_1(\mathsf{T})(v) + q_2(\mathsf{T})f_2(\mathsf{T})(v) + \cdots + q_k(\mathsf{T})f_k(\mathsf{T})(v).$$

For $i = 1, 2, \dots, k$, let $v_i = q_i(\mathsf{T})f_i(\mathsf{T})(v) = f_i(\mathsf{T})q_i(\mathsf{T})(v)$. Then for each i,

$$(\mathsf{T} - \lambda_i \mathsf{I})^{n_i}(v_i) = f(\mathsf{T})(q_i(\mathsf{T})(v)) = 0$$

by the Cayley-Hamilton theorem. It follows that $v_i \in \mathsf{K}_{\lambda_i}$ for $i = 1, 2, \dots, k$.

Now we prove the uniqueness of the vectors v_i. Since each K_{λ_j} is T-invariant, each K_{λ_j} is invariant under the operator $f_j(\mathsf{T})$ for $j = 1, 2, \dots, k$. Furthermore, $f_j(\mathsf{T})$ is one-to-one on K_{λ_j} by Theorem 7.1(c) and maps K_{λ_i} to $\{0\}$ for $i \neq j$ by Theorem 7.2(b).

Suppose that

$$v = \sum_{i=1}^{k} v_i = \sum_{i=1}^{k} w_i,$$

where $v_i, w_i \in \mathsf{K}_{\lambda_i}$ for $i = 1, 2, \dots, k$. Then for $j = 1, 2, \dots, k$,

$$f_j(\mathsf{T})(v) = \sum_{i=1}^{k} f_j(\mathsf{T})(v_i) = \sum_{i=1}^{k} f_j(\mathsf{T})(w_i),$$

and hence $f_j(\mathsf{T})(v_j) = f_j(\mathsf{T})(w_j)$ since $f_j(\mathsf{T})(v_i) = f_j(\mathsf{T})(w_i) = 0$ for $i \neq j$. Because $f_j(\mathsf{T})$ is one-to-one on K_{λ_j}, it follows that $v_j = w_j$. ∎

The next result extends Theorem 5.8(b) (p. 267) to all linear operators whose characteristic polynomials split. In this case, the eigenspaces are replaced by generalized eigenspaces.

Theorem 7.4. *Let T be a linear operator on a finite-dimensional vector space V whose characteristic polynomial $(t - \lambda_1)^{m_1}(t - \lambda_2)^{m_2} \dots (t - \lambda_k)^{m_k}$ splits. For $i = 1, 2, \dots, k$, let β_i be an ordered basis for K_{λ_i}. Then*

(a) $\beta_i \cap \beta_j = \varnothing$ *for* $i \neq j$.
(b) $\beta = \beta_1 \cup \beta_2 \cup \cdots \cup \beta_k$ *is an ordered basis for* V.
(c) $\dim(\mathsf{K}_{\lambda_i}) = m_i$ *for all* i.

Proof. (a) This is a direct consequence of Theorem 7.1(b).

(b) First, we prove that β is linearly independent. Consider a linear combination of vectors in β equal to 0. For each i, let v_i be the sum of the terms of this linear combination involving the vectors in β_i. Then

$$v_1 + v_2 + \cdots + v_k = 0 = 0 + 0 + \cdots + 0,$$

and hence $v_i = 0$ for all i by Theorem 7.3. Since each β_i is linearly independent, it follows that all of the coefficients in the linear combination are zeros. Hence β is linearly independent.

We now prove that β is a spanning set for V. Consider any vector $v \in \mathsf{V}$. By Theorem 7.3, for $i = 1, 2, \ldots, k$ there exist vectors $v_i \in \mathsf{K}_{\lambda_i}$ such that $v = v_1 + v_2 + \cdots + v_k$. Since each v_i is a linear combination of the vectors in β_i, v is a linear combination of the vectors in β. Hence β spans V.

(c) Since $\dim(\mathsf{V})$ is equal to the degree of the characteristic polynomial of T, it follows that $\dim(\mathsf{V}) = \sum_{i=1}^{k} m_i$. By (b), $\dim(\mathsf{V}) = \sum_{i=1}^{k} \dim(\mathsf{K}_{\lambda_i})$. Hence $\sum_{i=1}^{m}(m_i - \dim(\mathsf{K}_{\lambda_i})) = 0$. By Theorem 7.2(a), $m_i - \dim(\mathsf{K}_{\lambda_i}) \geq 0$ for all i. Consequently, $m_i - \dim(\mathsf{K}_{\lambda_i}) = 0$ for all i, which is the desired result. ∎

Corollary. *Let* T *be a linear operator on a finite-dimensional vector space* V *such that the characteristic polynomial of* T *splits. Then* T *is diagonalizable if and only if* $\mathsf{E}_\lambda = \mathsf{K}_\lambda$ *for every eigenvalue* λ *of* T.

Proof. Combining Theorems 7.4 and 5.8(a) (p. 267), we see that T is diagonalizable if and only if $\dim(\mathsf{E}_\lambda) = \dim(\mathsf{K}_\lambda)$ for each eigenvalue λ of T. But $\mathsf{E}_\lambda \subseteq \mathsf{K}_\lambda$, and hence these subspaces have the same dimension if and only if they are equal. ∎

We now focus our attention on the problem of selecting suitable bases for the generalized eigenspaces of a linear operator so that we may use Theorem 7.4 to obtain a Jordan canonical basis for the operator. For this purpose, we consider again the basis β of Example 1. We have seen that the first four vectors of β lie in the generalized eigenspace K_2. Observe that the vectors in β that determine the first Jordan block of J are of the form

$$\{v_1, v_2, v_3\} = \{(\mathsf{T} - 2\mathsf{I})^2(v_3), (\mathsf{T} - 2\mathsf{I})(v_3), v_3\}.$$

Furthermore, observe that $(\mathsf{T} - 2\mathsf{I})^3(v_3) = 0$. The relation between these vectors is the key to finding Jordan canonical bases. This leads to the following definitions.

Definitions. *Let* T *be a linear operator on a vector space* V*, and let* x *be a generalized eigenvector of* T *corresponding to the eigenvalue* λ*. Suppose that* p *is the smallest positive integer for which* $(\mathsf{T} - \lambda\mathsf{I})^p(x) = 0$*. Then the ordered set*

$$\{(\mathsf{T} - \lambda\mathsf{I})^{p-1}(x), (\mathsf{T} - \lambda\mathsf{I})^{p-2}(x), \ldots, (\mathsf{T} - \lambda\mathsf{I})(x), x\}$$

is called a ***cycle of generalized eigenvectors*** *of* T *corresponding to* λ*. The vectors* $(\mathsf{T} - \lambda\mathsf{I})^{p-1}(x)$ *and* x *are called the* ***initial vector*** *and the* ***end vector*** *of the cycle, respectively. We say that the* ***length*** *of the cycle is* p*.*

Notice that the initial vector of a cycle of generalized eigenvectors of a linear operator T is the only eigenvector of T in the cycle. Also observe that if x is an eigenvector of T corresponding to the eigenvalue λ, then the set $\{x\}$ is a cycle of generalized eigenvectors of T corresponding to λ of length 1.

In Example 1, the subsets $\beta_1 = \{v_1, v_2, v_3\}$, $\beta_2 = \{v_4\}$, $\beta_3 = \{v_5, v_6\}$, and $\beta_4 = \{v_7, v_8\}$ are the cycles of generalized eigenvectors of T that occur in β. Notice that β is a disjoint union of these cycles. Furthermore, setting $\mathsf{W}_i = \text{span}(\beta_i)$ for $i = 1, 2, 3, 4$, we see that β_i is a basis for W_i and $[\mathsf{T}_{\mathsf{W}_i}]_{\beta_i}$ is the ith Jordan block of the Jordan canonical form of T. This is precisely the condition that is required for a Jordan canonical basis.

Theorem 7.5. *Let* T *be a linear operator on a finite-dimensional vector space* V *whose characteristic polynomial splits, and suppose that* β *is a basis for* V *such that* β *is a disjoint union of cycles of generalized eigenvectors of* T*. Then the following statements are true.*

(a) *For each cycle* γ *of generalized eigenvectors contained in* β*,* $\mathsf{W} = \text{span}(\gamma)$ *is* T*-invariant, and* $[\mathsf{T}_{\mathsf{W}}]_\gamma$ *is a Jordan block.*

(b) β *is a Jordan canonical basis for* V*.*

Proof. (a) Suppose that γ corresponds to the eigenvalue λ, γ has length p, and x is the end vector of γ. Then $\gamma = \{v_1, v_2, \ldots, v_p\}$, where

$$v_i = (\mathsf{T} - \lambda\mathsf{I})^{p-i}(x) \text{ for } i < p \quad \text{and} \quad v_p = x.$$

So

$$(\mathsf{T} - \lambda\mathsf{I})(v_1) = (\mathsf{T} - \lambda\mathsf{I})^p(x) = 0,$$

and hence $\mathsf{T}(v_1) = \lambda v_1$. For $i > 1$,

$$(\mathsf{T} - \lambda\mathsf{I})(v_i) = (\mathsf{T} - \lambda\mathsf{I})^{p-(i-1)}(x) = v_{i-1}.$$

This shows that $\mathsf{T} - \lambda\mathsf{I}$ maps W into itself, and so T does also. Thus, by the preceding equations, we see that $[\mathsf{T}_{\mathsf{W}}]_\gamma$ is a Jordan block.

For (b), simply repeat the arguments of (a) for each cycle in β in order to obtain $[\mathsf{T}]_\beta$. We leave the details as an exercise. ∎

Theorem 7.7. *Let* T *be a linear operator on a finite-dimensional vector space* V*, and let* λ *be an eigenvalue of* T*. Then* K_λ *has an ordered basis consisting of a union of disjoint cycles of generalized eigenvectors corresponding to* λ*.*

Proof. The proof is by mathematical induction on $n = \dim(\mathsf{K}_\lambda)$. The result is clear for $n = 1$. So suppose that for some integer $n > 1$ the result is true whenever $\dim(\mathsf{K}_\lambda) < n$, and assume that $\dim(\mathsf{K}_\lambda) = n$. Let U denote the restriction of $\mathsf{T} - \lambda\mathsf{I}$ to K_λ. Then $\mathsf{R}(\mathsf{U})$ is a subspace of K_λ of lesser dimension, and $\mathsf{R}(\mathsf{U})$ is the space of generalized eigenvectors corresponding to λ for the restriction of T to $\mathsf{R}(\mathsf{U})$. Therefore, by the induction hypothesis, there exist disjoint cycles $\gamma_1, \gamma_2, \ldots, \gamma_q$ of generalized eigenvectors of this restriction, and hence of T itself, corresponding to λ for which $\gamma = \bigcup_{i=1}^{q} \gamma_i$ is a basis for $\mathsf{R}(\mathsf{U})$. For $i = 1, 2, \ldots, q$, the end vector of γ_i is the image under U of a vector $v_i \in \mathsf{K}_\lambda$, and so we can extend each γ_i to a larger cycle $\tilde{\gamma}_i = \gamma_i \cup \{v_i\}$ of generalized eigenvectors of T corresponding to λ. For $i = 1, 2, \ldots, q$, let w_i be the initial vector of $\tilde{\gamma}_i$ (and hence of γ_i). Since $\{w_1, w_2, \ldots, w_q\}$ is a linearly independent subset of E_λ, this set can be extended to a basis $\{w_1, w_2, \ldots, w_q, u_1, u_2, \ldots, u_s\}$ for E_λ. Then $\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_q, \{u_1\}, \{u_2\}, \ldots, \{u_s\}$ are disjoint cycles of generalized eigenvectors of T corresponding to λ such that the initial vectors of these cycles are linearly independent. Therefore their union $\tilde{\gamma}$ is a linearly independent subset of K_λ by Theorem 7.6.

Finally, we show that $\tilde{\gamma}$ is a basis for K_λ. Suppose that γ consists of $r = \operatorname{rank}(\mathsf{U})$ vectors. Then $\tilde{\gamma}$ consists of $r + q + s$ vectors. Furthermore, since $\{w_1, w_2, \ldots, w_q, u_1, u_2, \ldots, u_s\}$ is a basis for $\mathsf{E}_\lambda = \mathsf{N}(\mathsf{U})$, it follows that $\operatorname{nullity}(\mathsf{U}) = q + s$. Therefore

$$\dim(\mathsf{K}_\lambda) = \operatorname{rank}(\mathsf{U}) + \operatorname{nullity}(\mathsf{U}) = r + q + s.$$

So $\tilde{\gamma}$ is a linearly independent subset of K_λ containing $\dim(\mathsf{K}_\lambda)$ vectors, and thus $\tilde{\gamma}$ is a basis for K_λ. ∎

Corollary 1. *Let* T *be a linear operator on a finite-dimensional vector space* V *whose characteristic polynomial splits. Then* T *has a Jordan canonical form.*

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T. By Theorem 7.7, for each i there is an ordered basis β_i consisting of a disjoint union of cycles of generalized eigenvectors corresponding to λ_i. Let $\beta = \beta_1 \cup \beta_2 \cup \cdots \cup \beta_k$. Then, by Theorem 7.5(b), β is a Jordan canonical basis for V. ∎

The Jordan canonical form also can be studied from the viewpoint of matrices.

Definition. *Let $A \in \mathsf{M}_{n\times n}(F)$ be such that the characteristic polynomial of A (and hence of L_A) splits. Then the* ***Jordan canonical form*** *of A is defined to be the Jordan canonical form of the linear operator L_A on F^n.*

The next result is an immediate consequence of this definition and Corollary 1.

Corollary 2. *Let A be an $n \times n$ matrix whose characteristic polynomial splits. Then A has a Jordan canonical form J, and A is similar to J.*

Proof. Exercise. ∎

We can now compute the Jordan canonical forms of matrices and linear operators in some simple cases, as is illustrated in the next two examples. The tools necessary for computing the Jordan canonical forms in general are developed in the next section.

Example 2

Let

$$A = \begin{pmatrix} 3 & 1 & -2 \\ -1 & 0 & 5 \\ -1 & -1 & 4 \end{pmatrix} \in \mathsf{M}_{3\times 3}(R).$$

To find the Jordan canonical form for A, we need to find a Jordan canonical basis for $\mathsf{T} = \mathsf{L}_A$.

The characteristic polynomial of A is

$$f(t) = \det(A - tI) = -(t-3)(t-2)^2.$$

Hence $\lambda_1 = 3$ and $\lambda_2 = 2$ are the eigenvalues of A with multiplicities 1 and 2, respectively. By Theorem 7.4, $\dim(\mathsf{K}_{\lambda_1}) = 1$, and $\dim(\mathsf{K}_{\lambda_2}) = 2$. By Theorem 7.2, $\mathsf{K}_{\lambda_1} = \mathsf{N}(\mathsf{T}-3\mathsf{I})$, and $\mathsf{K}_{\lambda_2} = \mathsf{N}((\mathsf{T}-2\mathsf{I})^2)$. Since $\mathsf{E}_{\lambda_1} = \mathsf{N}(\mathsf{T}-3\mathsf{I})$, we have that $\mathsf{E}_{\lambda_1} = \mathsf{K}_{\lambda_1}$. Observe that $(-1, 2, 1)$ is an eigenvector of T corresponding to $\lambda_1 = 3$; therefore

$$\beta_1 = \left\{ \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} \right\}$$

is a basis for K_{λ_1}.

Since $\dim(\mathsf{K}_{\lambda_2}) = 2$ and a generalized eigenspace has a basis consisting of a union of cycles, this basis is either a union of two cycles of length 1 or a single cycle of length 2. Since the rank of $A - 2I$ is 2 it follows that $\dim(\mathsf{E}_{\lambda_2}) = 1$. Hence the former is not the case. It can easily be shown that $w = (1, -3, -1)$ is an eigenvector of A corresponding to λ_2. This vector can serve as the initial

vector of the cycle. Any solution to the equation $(A - 2I)v = w$ will serve as the end vector. For example, we can take $v = (1, -2, 0)$. Thus

$$\beta_2 = \{(A - 2I)v, v\} = \left\{ \begin{pmatrix} 1 \\ -3 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix} \right\}$$

is a cycle of generalized eigenvectors for $\lambda_2 = 2$. Finally, we take the union of these two bases to obtain

$$\beta = \beta_1 \cup \beta_2 = \left\{ \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -3 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix} \right\},$$

which is a Jordan canonical basis for A. Therefore,

$$J = [\mathsf{T}]_\beta = \left(\begin{array}{c|cc} 3 & 0 & 0 \\ \hline 0 & 2 & 1 \\ 0 & 0 & 2 \end{array}\right)$$

is a Jordan canonical form for A. Notice that A is similar to J. In fact, $J = Q^{-1}AQ$, where Q is the matrix whose columns are the vectors in β. ◆

Example 3

Let T be the linear operator on $\mathsf{P}_2(R)$ defined by $\mathsf{T}(g(x)) = -g(x) - g'(x)$. We find a Jordan canonical form of T and a Jordan canonical basis for T.

Let β be the standard ordered basis for $\mathsf{P}_2(R)$. Then

$$[\mathsf{T}]_\beta = \begin{pmatrix} -1 & -1 & 0 \\ 0 & -1 & -2 \\ 0 & 0 & -1 \end{pmatrix},$$

which has the characteristic polynomial $f(t) = -(t + 1)^3$. Thus $\lambda = -1$ is the only eigenvalue of T, and hence $\mathsf{K}_\lambda = \mathsf{P}_2(R)$ by Theorem 7.4. So β is a basis for K_λ. Now

$$\dim(\mathsf{E}_\lambda) = 3 - \operatorname{rank}(A + I) = 3 - \operatorname{rank} \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix} = 3 - 2 = 1.$$

Therefore a basis for K_λ cannot be a union of two or three cycles because the initial vector of each cycle is an eigenvector, and there do not exist two or more linearly independent eigenvectors. So the desired basis must consist

of a single cycle of length 3. If γ is such a cycle, then γ determines a single Jordan block

$$[\mathsf{T}]_\gamma = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix},$$

which is a Jordan canonical form of T.

The end vector $h(x)$ of such a cycle must satisfy $(\mathsf{T}+\mathsf{I})^2(h(x)) \neq 0$. In any basis for K_λ, there must be a vector that satisfies this condition, or else no vector in K_λ satisfies this condition, contrary to our reasoning. Testing the vectors in β, we see that $h(x) = x^2$ is acceptable. Therefore

$$\gamma = \{(\mathsf{T}+\mathsf{I})^2(x^2), (\mathsf{T}+\mathsf{I})(x^2), x^2\} = \{2, -2x, x^2\}$$

is a Jordan canonical basis for T. ◆

In the next section, we develop a computational approach for finding a Jordan canonical form and a Jordan canonical basis. In the process, we prove that Jordan canonical forms are unique up to the order of the Jordan blocks.

Let T be a linear operator on a finite-dimensional vector space V, and suppose that the characteristic polynomial of T splits. By Theorem 5.10 (p. 277), T is diagonalizable if and only if V is the direct sum of the eigenspaces of T. If T is diagonalizable, then the eigenspaces and the generalized eigenspaces coincide.

For those familiar with the material on direct sums in Section 5.2, we conclude by stating a generalization of Theorem 5.10 for nondiagonalizable operators.

Theorem 7.8. *Let T be a linear operator on a finite-dimensional vector space V whose characteristic polynomial splits. Then V is the direct sum of the generalized eigenspaces of T.*

Proof. Exercise. ∎

EXERCISES

1. Label the following statements as true or false.

- **(a)** Eigenvectors of a linear operator T are also generalized eigenvectors of T.
- **(b)** It is possible for a generalized eigenvector of a linear operator T to correspond to a scalar that is not an eigenvalue of T.
- **(c)** Any linear operator on a finite-dimensional vector space has a Jordan canonical form.
- **(d)** A cycle of generalized eigenvectors is linearly independent.

(e) There is exactly one cycle of generalized eigenvectors corresponding to each eigenvalue of a linear operator on a finite-dimensional vector space.

(f) Let T be a linear operator on a finite-dimensional vector space whose characteristic polynomial splits, and let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T. If, for each i, β_i is a basis for K_{λ_i}, then $\beta_1 \cup \beta_2 \cup \cdots \cup \beta_k$ is a Jordan canonical basis for T.

(g) For any Jordan block J, the operator L_J has Jordan canonical form J.

(h) Let T be a linear operator on an n-dimensional vector space whose characteristic polynomial splits. Then, for any eigenvalue λ of T, $\mathsf{K}_\lambda = \mathsf{N}((\mathsf{T} - \lambda\mathsf{I})^n)$.

2. For each matrix A, find a basis for each generalized eigenspace of L_A consisting of a union of disjoint cycles of generalized eigenvectors. Then find a Jordan canonical form J of A.

(a) $A = \begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$ **(b)** $A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$

(c) $A = \begin{pmatrix} 11 & -4 & -5 \\ 21 & -8 & -11 \\ 3 & -1 & 0 \end{pmatrix}$ **(d)** $A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 1 & -1 & 3 \end{pmatrix}$

3. For each linear operator T, find a basis for each generalized eigenspace of T consisting of a union of disjoint cycles of generalized eigenvectors. Then find a Jordan canonical form J of T.

(a) Define T on $\mathsf{P}_2(R)$ by $\mathsf{T}(f(x)) = 2f(x) - f'(x)$.

(b) V is the real vector space of functions spanned by the set of real-valued functions $\{1, t, t^2, e^t, te^t\}$, and T is the linear operator on V defined by $\mathsf{T}(f) = f'$.

(c) T is the linear operator on $\mathsf{M}_{2\times 2}(R)$ defined for all $A \in \mathsf{M}_{2\times 2}(R)$ by $\mathsf{T}(A) = BA$, where $B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. 1

(d) $\mathsf{T}(A) = 2A + A^t$ for all $A \in \mathsf{M}_{2\times 2}(R)$.

4.† Let T be a linear operator on a vector space V, and let γ be a cycle of generalized eigenvectors that corresponds to the eigenvalue λ. Prove that span(γ) is a T-invariant subspace of V. Visit `goo.gl/Lw4ahY` for a solution.

5. Let $\gamma_1, \gamma_2, \ldots, \gamma_p$ be cycles of generalized eigenvectors of a linear operator T corresponding to an eigenvalue λ. Prove that if the initial eigenvectors are distinct, then the cycles are disjoint.

6. Let $T: V \to W$ be a linear transformation. Prove the following results.

(a) $N(T) = N(-T)$.

(b) $N(T^k) = N((-T)^k)$..

(c) If $V = W$ (so that T is a linear operator on V) and λ is an eigenvalue of T, then for any positive integer k

$$N((T - \lambda I_V)^k) = N((\lambda I_V - T)^k).$$

7. Let U be a linear operator on a finite-dimensional vector space V. Prove the following results.

(a) $N(U) \subseteq N(U^2) \subseteq \cdots \subseteq N(U^k) \subseteq N(U^{k+1}) \subseteq \cdots$.

(b) If $\text{rank}(U^m) = \text{rank}(U^{m+1})$ for some positive integer m, then $\text{rank}(U^m) = \text{rank}(U^k)$ for any positive integer $k \geq m$.

(c) If $\text{rank}(U^m) = \text{rank}(U^{m+1})$ for some positive integer m, then $N(U^m) = N(U^k)$ for any positive integer $k \geq m$.

(d) Let T be a linear operator on V, and let λ be an eigenvalue of T. Prove that if $\text{rank}((T-\lambda I)^m) = \text{rank}((T-\lambda I)^{m+1})$ for some integer m, then $K_\lambda = N((T - \lambda I)^m)$.

(e) *Second Test for Diagonalizability.* Let T be a linear operator on V whose characteristic polynomial splits, and let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T. Then T is diagonalizable if and only if $\text{rank}(T - \lambda_i I) = \text{rank}((T - \lambda_i I)^2)$ for $1 \leq i \leq k$.

(f) Use (e) to obtain a simpler proof of Exercise 24 of Section 5.4: If T is a diagonalizable linear operator on a finite-dimensional vector space V and W is a T-invariant subspace of V, then T_W is diagonalizable.

8. Let T be a linear operator on a finite-dimensional vector space V whose characteristic polynomial splits and has Jordan canonical form J. Prove that for any nonzero scalar c, cJ is a Jordan canonical form for cT.

9. Let T be a linear operator on a finite-dimensional vector space V whose characteristic polynomial splits.

(a) Prove Theorem 7.5(b).

(b) Suppose that β is a Jordan canonical basis for T, and let λ be an eigenvalue of T. Let $\beta' = \beta \cap K_\lambda$. Prove that β' is a basis for K_λ.

10. Let T be a linear operator on a finite-dimensional vector space whose characteristic polynomial splits, and let λ be an eigenvalue of T.

(a) Suppose that γ is a basis for K_λ consisting of the union of q disjoint cycles of generalized eigenvectors. Prove that $q \leq \dim(E_\lambda)$.

(b) Let β be a Jordan canonical basis for T, and suppose that $J = [T]_\beta$ has q Jordan blocks with λ in the diagonal positions. Prove that $q \leq \dim(E_\lambda)$.

11. Prove Corollary 2 to Theorem 7.7.

12. Let T be a linear operator on a finite-dimensional vector space V, and let λ be an eigenvalue of T with corresponding eigenspace and generalized eigenspace E_λ and K_λ, respectively. Let U be an invertible linear operator on V that commutes with T (i.e., $\mathsf{TU} = \mathsf{UT}$). Prove that $\mathsf{U}(\mathsf{E}_\lambda) = \mathsf{E}_\lambda$ and $\mathsf{U}(\mathsf{K}_\lambda) = \mathsf{K}_\lambda$.

Exercises 13 and 14 are concerned with direct sums of matrices, defined in Section 5.4 on page 318.

13. Prove Theorem 7.8.

14. Let T be a linear operator on a finite-dimensional vector space V such that the characteristic polynomial of T splits, and let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T. For each i, let J_i be the Jordan canonical form of the restriction of T to K_{λ_i}. Prove that

$$J = J_1 \oplus J_2 \oplus \cdots \oplus J_k$$

is the Jordan canonical form of J.

7.2 THE JORDAN CANONICAL FORM II

For the purposes of this section, we fix a linear operator T on an n-dimensional vector space V such that the characteristic polynomial of T splits. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T.

By Theorem 7.7 (p. 484), each generalized eigenspace K_{λ_i} contains an ordered basis β_i consisting of a union of disjoint cycles of generalized eigenvectors corresponding to λ_i. So by Theorems 7.4(b) (p. 480) and 7.5 (p. 482), the union $\beta = \bigcup_{i=1}^{k} \beta_i$ is a Jordan canonical basis for T. For each i, let T_i be the restriction of T to K_{λ_i}, and let $A_i = [\mathsf{T}_i]_{\beta_i}$. Then A_i is the Jordan canonical form of T_i, and

$$J = [\mathsf{T}]_\beta = \begin{pmatrix} A_1 & O & \cdots & O \\ O & A_2 & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & A_k \end{pmatrix}$$

is the Jordan canonical form of T. In this matrix, each O is a zero matrix of appropriate size.

In this section, we compute the matrices A_i and the bases β_i, thereby computing J and β as well. While developing a method for finding J, it becomes evident that in some sense the matrices A_i are unique.

To aid in formulating the uniqueness theorem for J, we adopt the following convention: The basis β_i for K_{λ_i} will henceforth be ordered in such a way that the cycles appear in order of decreasing length. That is, if β_i is a disjoint union of cycles $\gamma_1, \gamma_2, \ldots, \gamma_{n_i}$ and if the length of the cycle γ_j is p_j, we index the cycles so that $p_1 \geq p_2 \geq \cdots \geq p_{n_i}$. This ordering of the cycles limits the possible orderings of vectors in β_i, which in turn determines the matrix A_i. It is in this sense that A_i is unique. It then follows that the Jordan canonical form for T is unique up to an ordering of the eigenvalues of T. As we will see, there is no uniqueness theorem for the bases β_i or for β. However, we show that for each i, the number n_i of cycles that form β_i, and the length p_j $(j = 1, 2, \ldots, n_i)$ of each cycle, is completely determined by T.

Example 1

To illustrate the discussion above, suppose that, for some i, the ordered basis β_i for K_{λ_i} is the union of four cycles $\beta_i = \gamma_1 \cup \gamma_2 \cup \gamma_3 \cup \gamma_4$ with respective lengths $p_1 = 3$, $p_2 = 3$, $p_3 = 2$, and $p_4 = 1$. Then

$$A_i = \left(\begin{array}{ccc|ccc|cc|c} \lambda_i & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_i & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_i & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_i & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i \end{array}\right). \quad \blacklozenge$$

To help us visualize each of the matrices A_i and ordered bases β_i, we use an array of dots called a **dot diagram** of T_i, where T_i is the restriction of T to K_{λ_i}. Suppose that β_i is a disjoint union of cycles of generalized eigenvectors $\gamma_1, \gamma_2, \ldots, \gamma_{n_i}$ with lengths $p_1 \geq p_2 \geq \cdots \geq p_{n_i}$, respectively. The dot diagram of T_i contains one dot for each vector in β_i, and the dots are configured according to the following rules.

1. The array consists of n_i columns (one column for each cycle).
2. Counting from left to right, the jth column consists of the p_j dots that correspond to the vectors of γ_j starting with the initial vector at the top and continuing down to the end vector.

Denote the end vectors of the cycles by $v_1, v_2, \ldots, v_{n_i}$. In the following dot diagram of T_i, each dot is labeled with the name of the vector in β_i to

which it corresponds.

$$\begin{array}{llll}
\bullet\,(\mathsf{T}-\lambda_i\mathsf{I})^{p_1-1}(v_1) & \bullet\,(\mathsf{T}-\lambda_i\mathsf{I})^{p_2-1}(v_2) & \cdots & \bullet\,(\mathsf{T}-\lambda_i\mathsf{I})^{p_{n_i}-1}(v_{n_i}) \\
\bullet\,(\mathsf{T}-\lambda_i\mathsf{I})^{p_1-2}(v_1) & \bullet\,(\mathsf{T}-\lambda_i\mathsf{I})^{p_2-2}(v_2) & \cdots & \bullet\,(\mathsf{T}-\lambda_i\mathsf{I})^{p_{n_i}-2}(v_{n_i}) \\
 & & & \vdots \\
 & \vdots & & \bullet\,(\mathsf{T}-\lambda_i\mathsf{I})(v_{n_i}) \\
\vdots & & & \bullet\,v_{n_i} \\
 & \bullet\,(\mathsf{T}-\lambda_i\mathsf{I})(v_2) & & \\
 & \bullet\,v_2 & & \\
\bullet\,(\mathsf{T}-\lambda_i\mathsf{I})(v_1) & & & \\
\bullet\,v_1 & & &
\end{array}$$

Notice that the dot diagram of T_i has n_i columns (one for each cycle) and p_1 rows. Since $p_1 \geq p_2 \geq \cdots \geq p_{n_i}$, the columns of the dot diagram become shorter (or at least not longer) as we move from left to right.

Now let r_j denote the number of dots in the jth row of the dot diagram. Observe that $r_1 \geq r_2 \geq \cdots \geq r_{p_1}$. Furthermore, the diagram can be reconstructed from the values of the r_i's. The proofs of these facts, which are combinatorial in nature, are treated in Exercise 9.

In Example 1, with $n_i = 4$, $p_1 = p_2 = 3$, $p_3 = 2$, and $p_4 = 1$, the dot diagram of T_i is as follows:

$$\begin{array}{cccc}
\bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \\
\bullet & \bullet & &
\end{array}$$

Here $r_1 = 4$, $r_2 = 3$, and $r_3 = 2$.

We now devise a method for computing the dot diagram of T_i using the ranks of linear operators determined by T and λ_i. It will follow that the dot diagram is completely determined by T, from which it will follow that it is unique. On the other hand, β_i is not unique. For example, see Exercise 8. (It is for this reason that we associate the dot diagram with T_i rather than with β_i.)

To determine the dot diagram of T_i, we devise a method for computing each r_j, the number of dots in the jth row of the dot diagram, using only T and λ_i. The next three results give us the required method. To facilitate our arguments, we fix a basis β_i for K_{λ_i} so that β_i is a disjoint union of n_i cycles of generalized eigenvectors with lengths $p_1 \geq p_2 \geq \cdots \geq p_{n_i}$.

Theorem 7.9. *For any positive integer r, the vectors in β_i that are associated with the dots in the first r rows of the dot diagram of T_i constitute a basis for $\mathsf{N}((\mathsf{T}-\lambda_i\mathsf{I})^r)$. Hence the number of dots in the first r rows of the dot diagram equals* $\text{nullity}((\mathsf{T}-\lambda_i\mathsf{I})^r)$.

Proof. Clearly, $\mathsf{N}((\mathsf{T}-\lambda_i\mathsf{I})^r) \subseteq \mathsf{K}_{\lambda_i}$, and K_{λ_i} is invariant under $(\mathsf{T}-\lambda_i\mathsf{I})^r$. Let U denote the restriction of $(\mathsf{T}-\lambda_i\mathsf{I})^r$ to K_{λ_i}. By the preceding remarks, $\mathsf{N}((\mathsf{T}-\lambda_i\mathsf{I})^r) = \mathsf{N}(\mathsf{U})$, and hence it suffices to establish the theorem for U. Now define

$$S_1 = \{x \in \beta_i : \mathsf{U}(x) = 0\} \quad \text{and} \quad S_2 = \{x \in \beta_i : \mathsf{U}(x) \neq 0\}.$$

Let a and b denote the number of vectors in S_1 and S_2, respectively, and let $m_i = \dim(\mathsf{K}_{\lambda_i})$. Then $a + b = m_i$. For any $x \in \beta_i$, $x \in S_1$ if and only if x is one of the first r vectors of a cycle, and this is true if and only if x corresponds to a dot in the first r rows of the dot diagram. Hence a is the number of dots in the first r rows of the dot diagram. For any $x \in S_2$, the effect of applying U to x is to move the dot corresponding to x exactly r places up its column to another dot. It follows that U maps S_2 in a one-to-one fashion into β_i. Thus $\{\mathsf{U}(x) : x \in S_2\}$ is a basis for $\mathsf{R}(\mathsf{U})$ consisting of b vectors. Hence $\text{rank}(\mathsf{U}) = b$, and so $\text{nullity}(\mathsf{U}) = m_i - b = a$. But S_1 is a linearly independent subset of $\mathsf{N}(\mathsf{U})$ consisting of a vectors; therefore S_1 is a basis for $\mathsf{N}(\mathsf{U})$. ∎

In the case that $r = 1$, Theorem 7.9 yields the following corollary.

Corollary. *The dimension of E_{λ_i} is n_i. Hence in a Jordan canonical form of T, the number of Jordan blocks corresponding to λ_i equals the dimension of E_{λ_i}.*

Proof. Exercise. ∎

We are now able to devise a method for describing the dot diagram in terms of the ranks of operators.

Theorem 7.10. *Let r_j denote the number of dots in the jth row of the dot diagram of T_i, the restriction of T to K_{λ_i}. Then the following statements are true.*

(a) $r_1 = \dim(\mathsf{V}) - \text{rank}(\mathsf{T}-\lambda_i\mathsf{I})$.

(b) $r_j = \text{rank}((\mathsf{T}-\lambda_i\mathsf{I})^{j-1}) - \text{rank}((\mathsf{T}-\lambda_i\mathsf{I})^j)$ *if* $j > 1$.

Proof. By Theorem 7.9, for $j = 1, 2, \ldots, p-1$, we have

$$\begin{aligned} r_1 + r_2 + \cdots + r_j &= \text{nullity}((\mathsf{T}-\lambda_i\mathsf{I})^j) \\ &= \dim(\mathsf{V}) - \text{rank}((\mathsf{T}-\lambda_i\mathsf{I})^j). \end{aligned}$$

Hence

$$r_1 = \dim(\mathsf{V}) - \text{rank}(\mathsf{T}-\lambda_i\mathsf{I}),$$

and for $j > 1$,

$$r_j = (r_1 + r_2 + \cdots + r_j) - (r_1 + r_2 + \cdots + r_{j-1})$$

$$= [\dim(\mathsf{V}) - \operatorname{rank}((\mathsf{T} - \lambda_i \mathsf{I})^j)] - [\dim(\mathsf{V}) - \operatorname{rank}((\mathsf{T} - \lambda_i \mathsf{I})^{j-1})]$$
$$= \operatorname{rank}((\mathsf{T} - \lambda_i \mathsf{I})^{j-1}) - \operatorname{rank}((\mathsf{T} - \lambda_i \mathsf{I})^j). \quad \blacksquare$$

Theorem 7.10 shows that the dot diagram of T_i is completely determined by T and λ_i. Hence we have proved the following result.

Corollary. *For any eigenvalue λ_i of T, the dot diagram of T_i is unique. Thus, subject to the convention that the cycles of generalized eigenvectors for the bases of each generalized eigenspace are listed in order of decreasing length, the Jordan canonical form of a linear operator or a matrix is unique up to the ordering of the eigenvalues.*

We apply these results to find the Jordan canonical forms of two matrices and a linear operator.

Example 2

Let

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

We find the Jordan canonical form of A and a Jordan canonical basis for the linear operator $\mathsf{T} = \mathsf{L}_A$. The characteristic polynomial of A is

$$\det(A - tI) = (t-2)^3(t-3).$$

Thus A has two distinct eigenvalues, $\lambda_1 = 2$ and $\lambda_2 = 3$, with multiplicities 3 and 1, respectively. Let T_1 and T_2 be the restrictions of L_A to the generalized eigenspaces K_{λ_1} and K_{λ_2}, respectively.

Suppose that β_1 is a Jordan canonical basis for T_1. Since λ_1 has multiplicity 3, it follows that $\dim(\mathsf{K}_{\lambda_1}) = 3$ by Theorem 7.4(c) (p. 480); hence the dot diagram of T_1 has three dots. As we did earlier, let r_j denote the number of dots in the jth row of this dot diagram. Then, by Theorem 7.10,

$$r_1 = 4 - \operatorname{rank}(A - 2I) = 4 - \operatorname{rank}\begin{pmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} = 4 - 2 = 2,$$

and

$$r_2 = \operatorname{rank}(A - 2I) - \operatorname{rank}((A - 2I)^2) = 2 - 1 = 1.$$

(Actually, the computation of r_2 is unnecessary in this case because $r_1 = 2$ and the dot diagram only contains three dots.) Hence the dot diagram associated with β_1 is

So

$$A_1 = [\mathsf{T}_1]_{\beta_1} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Since $\lambda_2 = 3$ has multiplicity 1, it follows that $\dim(\mathsf{K}_{\lambda_2}) = 1$, and consequently any basis β_2 for K_{λ_2} consists of a single eigenvector corresponding to $\lambda_2 = 3$. Therefore

$$A_2 = [\mathsf{T}_2]_{\beta_2} = (3).$$

Setting $\beta = \beta_1 \cup \beta_2$, we have

$$J = [\mathsf{L}_A]_\beta = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix},$$

and so J is the Jordan canonical form of A.

We now find a Jordan canonical basis for $\mathsf{T} = \mathsf{L}_A$. We begin by determining a Jordan canonical basis β_1 for T_1. Since the dot diagram of T_1 has two columns, each corresponding to a cycle of generalized eigenvectors, there are two such cycles. Let v_1 and v_2 denote the end vectors of the first and second cycles, respectively. We reprint below the dot diagram with the dots labeled with the names of the vectors to which they correspond.

$$\begin{array}{ll} \bullet\, (\mathsf{T} - 2\mathsf{I})(v_1) & \bullet\, v_2 \\ \bullet\, v_1 & \end{array}$$

From this diagram we see that $v_1 \in \mathsf{N}((\mathsf{T} - 2\mathsf{I})^2)$ but $v_1 \notin \mathsf{N}(\mathsf{T} - 2\mathsf{I})$. Now

$$A - 2I = \begin{pmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad (A - 2I)^2 = \begin{pmatrix} 0 & -2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 1 \end{pmatrix}.$$

It is easily seen that

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 2 \end{pmatrix} \right\}$$

is a basis for $\mathsf{N}((\mathsf{T} - 2\mathsf{I})^2) = \mathsf{K}_{\lambda_1}$. Of these three basis vectors, the last two do not belong to $\mathsf{N}(\mathsf{T} - 2\mathsf{I})$, and hence we select one of these for v_1. Suppose that we choose

$$v_1 = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}.$$

Then

$$(\mathsf{T}-2\mathsf{I})(v_1)=(A-2I)(v_1)=\begin{pmatrix}0&-1&0&1\\0&1&-1&0\\0&1&-1&0\\0&-1&0&1\end{pmatrix}\begin{pmatrix}0\\1\\2\\0\end{pmatrix}=\begin{pmatrix}-1\\-1\\-1\\-1\end{pmatrix}.$$

Now simply choose v_2 to be a vector in E_{λ_1} that is linearly independent of $(\mathsf{T}-2\mathsf{I})(v_1)$; for example, select

$$v_2=\begin{pmatrix}1\\0\\0\\0\end{pmatrix}.$$

Thus we have associated the Jordan canonical basis

$$\beta_1=\left\{\begin{pmatrix}-1\\-1\\-1\\-1\end{pmatrix},\begin{pmatrix}0\\1\\2\\0\end{pmatrix},\begin{pmatrix}1\\0\\0\\0\end{pmatrix}\right\}$$

with the dot diagram in the following manner.

$$\bullet\begin{pmatrix}-1\\-1\\-1\\-1\end{pmatrix}\quad\bullet\begin{pmatrix}1\\0\\0\\0\end{pmatrix}$$

$$\bullet\begin{pmatrix}0\\1\\2\\0\end{pmatrix}$$

By Theorem 7.6 (p. 483), the linear independence of β_1 is guaranteed since v_2 was chosen to be linearly independent of $(\mathsf{T}-2\mathsf{I})(v_1)$.

Since $\lambda_2=3$ has multiplicity 1, $\dim(\mathsf{K}_{\lambda_2})=\dim(\mathsf{E}_{\lambda_2})=1$. Hence any eigenvector of L_A corresponding to $\lambda_2=3$ constitutes an appropriate basis β_2. For example,

$$\beta_2=\left\{\begin{pmatrix}1\\0\\0\\1\end{pmatrix}\right\}.$$

Thus

$$\beta=\beta_1\cup\beta_2=\left\{\begin{pmatrix}-1\\-1\\-1\\-1\end{pmatrix},\begin{pmatrix}0\\1\\2\\0\end{pmatrix},\begin{pmatrix}1\\0\\0\\0\end{pmatrix},\begin{pmatrix}1\\0\\0\\1\end{pmatrix}\right\}$$

is a Jordan canonical basis for L_A.

Notice that if

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

then $J = Q^{-1}AQ$. ◆

Example 3

Let

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

We find the Jordan canonical form J of A, a Jordan canonical basis for L_A, and a matrix Q such that $J = Q^{-1}AQ$.

The characteristic polynomial of A is $\det(A - tI) = (t-2)^2(t-4)^2$. Let $\mathsf{T} = \mathsf{L}_A$, $\lambda_1 = 2$, and $\lambda_2 = 4$, and let T_i be the restriction of L_A to K_{λ_i} for $i = 1, 2$.

We begin by computing the dot diagram of T_1. Let r_1 denote the number of dots in the first row of this diagram. Then

$$r_1 = 4 - \operatorname{rank}(A - 2I) = 4 - 2 = 2;$$

hence the dot diagram of T_1 is as follows.

• •

Therefore

$$A_1 = [\mathsf{T}_1]_{\beta_1} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

where β_1 is any basis corresponding to the dots. In this case, β_1 is an arbitrary basis for $\mathsf{E}_{\lambda_1} = \mathsf{N}(\mathsf{T} - 2\mathsf{I})$, for example,

$$\beta_1 = \left\{ \begin{pmatrix} 2 \\ 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix} \right\}.$$

Next we compute the dot diagram of T_2. Since $\operatorname{rank}(A - 4I) = 3$, there is only $4 - 3 = 1$ dot in the first row of the diagram. Since $\lambda_2 = 4$ has

multiplicity 2, we have $\dim(\mathsf{K}_{\lambda_2}) = 2$, and hence this dot diagram has the following form:

Thus

$$A_2 = [\mathsf{T}_2]_{\beta_2} = \begin{pmatrix} 4 & 1 \\ 0 & 4 \end{pmatrix},$$

where β_2 is any basis for K_{λ_2} corresponding to the dots. In this case, β_2 is a cycle of length 2. The end vector of this cycle is a vector $v \in \mathsf{K}_{\lambda_2} = \mathsf{N}((\mathsf{T} - 4\mathsf{I})^2)$ such that $v \notin \mathsf{N}(\mathsf{T} - 4\mathsf{I})$. One way of finding such a vector was used to select the vector v_1 in Example 2. In this example, we illustrate another method. A simple calculation shows that a basis for the null space of $\mathsf{L}_A - 4\mathsf{I}$ is

$$\left\{ \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

Choose v to be any solution to the system of linear equations

$$(A - 4I)x = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

for example,

$$v = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 0 \end{pmatrix}.$$

Thus

$$\beta_2 = \{(\mathsf{L}_A - 4\mathsf{I})(v), v\} = \left\{ \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \\ 0 \end{pmatrix} \right\}.$$

Therefore

$$\beta = \beta_1 \cup \beta_2 = \left\{ \begin{pmatrix} 2 \\ 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \\ 0 \end{pmatrix} \right\}$$

is a Jordan canonical basis for L_A. The corresponding Jordan canonical form is given by

$$J = [\mathsf{L}_A]_\beta = \begin{pmatrix} A_1 & O \\ O & A_2 \end{pmatrix} = \left(\begin{array}{cc|cc} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{array}\right).$$

Finally, we define Q to be the matrix whose columns are the vectors of β listed in the same order, namely,

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

Then $J = Q^{-1}AQ$. ♦

Example 4

Let V be the vector space of polynomial functions in two real variables x and y of degree at most 2. Then V is a vector space over R and $\alpha = \{1, x, y, x^2, y^2, xy\}$ is an ordered basis for V. Let T be the linear operator on V defined by

$$\mathsf{T}(f(x,y)) = \frac{\partial}{\partial x} f(x,y).$$

For example, if $f(x,y) = x + 2x^2 - 3xy + y$, then

$$\mathsf{T}(f(x,y)) = \frac{\partial}{\partial x}(x + 2x^2 - 3xy + y) = 1 + 4x - 3y.$$

We find the Jordan canonical form and a Jordan canonical basis for T.

Let $A = [\mathsf{T}]_\alpha$. Then

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

and hence the characteristic polynomial of T is

$$\det(A - tI) = \det \begin{pmatrix} -t & 1 & 0 & 0 & 0 & 0 \\ 0 & -t & 0 & 2 & 0 & 0 \\ 0 & 0 & -t & 0 & 0 & 1 \\ 0 & 0 & 0 & -t & 0 & 0 \\ 0 & 0 & 0 & 0 & -t & 0 \\ 0 & 0 & 0 & 0 & 0 & -t \end{pmatrix} = t^6.$$

Thus $\lambda = 0$ is the only eigenvalue of T, and $\mathsf{K}_\lambda = \mathsf{V}$. For each j, let r_j denote the number of dots in the jth row of the dot diagram of T. By Theorem 7.10,

$$r_1 = 6 - \text{rank}(A) = 6 - 3 = 3,$$

and since

$$A^2 = \begin{pmatrix} 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$r_2 = \text{rank}(A) - \text{rank}(A^2) = 3 - 1 = 2$.

Because there are a total of six dots in the dot diagram and $r_1 = 3$ and $r_2 = 2$, it follows that $r_3 = 1$. So the dot diagram of T is

We conclude that the Jordan canonical form of T is

$$J = \left(\begin{array}{ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

We now find a Jordan canonical basis for T. Since the first column of the dot diagram of T consists of three dots, we must find a polynomial $f_1(x, y)$ such that $\dfrac{\partial^2}{\partial x^2} f_1(x, y) \neq 0$. Examining the basis $\alpha = \{1, x, y, x^2, y^2, xy\}$ for $\mathsf{K}_\lambda = \mathsf{V}$, we see that x^2 is a suitable candidate. Setting $f_1(x, y) = x^2$, we see that

$$(\mathsf{T} - \lambda \mathsf{I})(f_1(x, y)) = \mathsf{T}(f_1(x, y)) = \frac{\partial}{\partial x}(x^2) = 2x$$

and

$$(\mathsf{T} - \lambda \mathsf{I})^2(f_1(x, y)) = \mathsf{T}^2(f_1(x, y)) = \frac{\partial^2}{\partial x^2}(x^2) = 2.$$

Likewise, since the second column of the dot diagram consists of two dots, we must find a polynomial $f_2(x, y)$ such that

$$\frac{\partial}{\partial x}(f_2(x, y)) \neq 0, \quad \text{but} \quad \frac{\partial^2}{\partial x^2}(f_2(x, y)) = 0.$$

Since our choice must be linearly independent of the polynomials already chosen for the first cycle, the only choice in α that satisfies these constraints is xy. So we set $f_2(x, y) = xy$. Thus

$$(\mathsf{T} - \lambda \mathsf{I})(f_2(x, y)) = \mathsf{T}(f_2(x, y)) = \frac{\partial}{\partial x}(xy) = y.$$

Finally, the third column of the dot diagram consists of a single polynomial that lies in the null space of T. The only remaining polynomial in α is y^2, and it is suitable here. So set $f_3(x, y) = y^2$. Therefore we have identified polynomials with the dots in the dot diagram as follows.

$$\begin{array}{lll} \bullet\, 2 & \bullet\, y & \bullet\, y^2 \\ \bullet\, 2x & \bullet\, xy & \\ \bullet\, x^2 & & \end{array}$$

Thus $\beta = \{2, 2x, x^2, y, xy, y^2\}$ is a Jordan canonical basis for T. ♦

In the three preceding examples, we relied on our ingenuity and the context of the problem to find Jordan canonical bases. The reader can do the same in the exercises. We are successful in these cases because the dimensions of the generalized eigenspaces under consideration are small. We do not attempt, however, to develop a general algorithm for computing Jordan canonical bases, although one could be devised by following the steps in the proof of the existence of such a basis (Theorem 7.7 p. 484).

The following result may be thought of as a corollary to Theorem 7.10.

Theorem 7.11. *Let A and B be $n \times n$ matrices, each having Jordan canonical forms computed according to the conventions of this section. Then A and B are similar if and only if they have (up to an ordering of their eigenvalues) the same Jordan canonical form.*

Proof. If A and B have the same Jordan canonical form J, then A and B are each similar to J and hence are similar to each other.

Conversely, suppose that A and B are similar. Then A and B have the same eigenvalues. Let J_A and J_B denote the Jordan canonical forms of A and B, respectively, with the same ordering of their eigenvalues. Then A is similar to both J_A and J_B, and therefore, by the corollary to Theorem 2.23 (p. 115), J_A and J_B are matrix representations of L_A. Hence J_A and J_B are Jordan canonical forms of L_A. Thus $J_A = J_B$ by the corollary to Theorem 7.10. ∎

Example 5

We determine which of the matrices

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

$$C = \begin{pmatrix} 0 & -1 & -1 \\ -3 & -1 & -2 \\ 7 & 5 & 6 \end{pmatrix}, \quad \text{and} \quad D = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

are similar. Observe that A, B, and C have the same characteristic polynomial $-(t-1)(t-2)^2$, whereas D has $-t(t-1)(t-2)$ as its characteristic polynomial. Because similar matrices have the same characteristic polynomials, D cannot be similar to A, B, or C. Let J_A, J_B, and J_C be the Jordan canonical forms of A, B, and C, respectively, using the ordering 1, 2 for their common eigenvalues. Then (see Exercise 4)

$$J_A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}, \quad J_B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad \text{and} \quad J_C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

Since $J_A = J_C$, A is similar to C. Since J_B is different from J_A and J_C, B is similar to neither A nor C. ♦

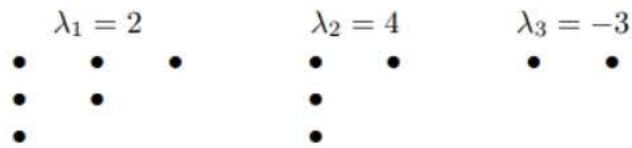
The reader should observe that any diagonal matrix is a Jordan canonical form. Thus a linear operator T on a finite-dimensional vector space V *is diagonalizable if and only if its Jordan canonical form is a diagonal matrix. Hence* T *is diagonalizable if and only if the Jordan canonical basis for* T *consists of eigenvectors of* T. Similar statements can be made about matrices. Thus, of the matrices A, B, and C in Example 5, A and C are not diagonalizable because their Jordan canonical forms are not diagonal matrices.

EXERCISES

1. Label the following statements as true or false. Assume that the characteristic polynomial of the matrix or linear operator splits.

- **(a)** The Jordan canonical form of a diagonal matrix is the matrix itself.
- **(b)** Let T be a linear operator on a finite-dimensional vector space V that has a Jordan canonical form J. If β is any basis for V, then the Jordan canonical form of $[\mathsf{T}]_\beta$ is J.
- **(c)** Linear operators having the same characteristic polynomial are similar.
- **(d)** Matrices having the same Jordan canonical form are similar.
- **(e)** Every matrix is similar to its Jordan canonical form.
- **(f)** Every linear operator with the characteristic polynomial $(-1)^n(t-\lambda)^n$ has the same Jordan canonical form.
- **(g)** Every linear operator on a finite-dimensional vector space has a unique Jordan canonical basis.
- **(h)** The dot diagrams of a linear operator on a finite-dimensional vector space are unique.

2. Let T be a linear operator on a finite-dimensional vector space V such that the characteristic polynomial of T splits. Suppose that $\lambda_1 = 2$, $\lambda_2 = 4$, and $\lambda_3 = -3$ are the distinct eigenvalues of T and that the dot diagrams for the restriction of T to K_{λ_i} $(i = 1, 2, 3)$ are as follows:



Find the Jordan canonical form J of T.

3. Let T be a linear operator on a finite-dimensional vector space V with Jordan canonical form

$$\left(\begin{array}{ccc|cc|cc} 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{array}\right).$$

(a) Find the characteristic polynomial of T.
(b) Find the dot diagram corresponding to each eigenvalue of T.
(c) For which eigenvalues λ_i, if any, does $\mathsf{E}_{\lambda_i} = \mathsf{K}_{\lambda_i}$?
(d) For each eigenvalue λ_i, find the smallest positive integer p_i for which $\mathsf{K}_{\lambda_i} = \mathsf{N}((\mathsf{T} - \lambda_i \mathsf{I})^{p_i})$.
(e) Compute the following numbers for each i, where U_i denotes the restriction of $\mathsf{T} - \lambda_i \mathsf{I}$ to K_{λ_i}.
(i) $\text{rank}(\mathsf{U}_i)$
(ii) $\text{rank}(\mathsf{U}_i^2)$
(iii) $\text{nullity}(\mathsf{U}_i)$
(iv) $\text{nullity}(\mathsf{U}_i^2)$

4. For each of the matrices A that follow, find a Jordan canonical form J and an invertible matrix Q such that $J = Q^{-1}AQ$. Notice that the matrices in (a), (b), and (c) are those used in Example 5.

(a) $A = \begin{pmatrix} -3 & 3 & -2 \\ -7 & 6 & -3 \\ 1 & -1 & 2 \end{pmatrix}$ **(b)** $A = \begin{pmatrix} 0 & 1 & -1 \\ -4 & 4 & -2 \\ -2 & 1 & 1 \end{pmatrix}$

(c) $A = \begin{pmatrix} 0 & -1 & -1 \\ -3 & -1 & -2 \\ 7 & 5 & 6 \end{pmatrix}$ **(d)** $A = \begin{pmatrix} 0 & -3 & 1 & 2 \\ -2 & 1 & -1 & 2 \\ -2 & 1 & -1 & 2 \\ -2 & -3 & 1 & 4 \end{pmatrix}$

5. For each linear operator T, find a Jordan canonical form J of T and a Jordan canonical basis β for T.

(a) V is the real vector space of functions spanned by the set of real-valued functions $\{e^t, te^t, t^2e^t, e^{2t}\}$, and T is the linear operator on V defined by $\mathsf{T}(f) = f'$.

(b) T is the linear operator on $\mathsf{P}_3(R)$ defined by $\mathsf{T}(f(x)) = xf''(x)$.

(c) T is the linear operator on $\mathsf{P}_3(R)$ defined by $\mathsf{T}(f(x)) = f''(x) + 2f(x)$.

(d) T is the linear operator on $\mathsf{M}_{2\times 2}(R)$ defined by

$$\mathsf{T}(A) = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix} \cdot A - A^t.$$

(e) T is the linear operator on $\mathsf{M}_{2\times 2}(R)$ defined by

$$\mathsf{T}(A) = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix} \cdot (A - A^t).$$

(f) V is the vector space of polynomial functions in two real variables x and y of degree at most 2, as defined in Example 4, and T is the linear operator on V defined by

$$\mathsf{T}(f(x,y)) = \frac{\partial}{\partial x} f(x,y) + \frac{\partial}{\partial y} f(x,y).$$

6. Let A be an $n \times n$ matrix whose characteristic polynomial splits. Prove that A and A^t have the same Jordan canonical form, and conclude that A and A^t are similar. *Hint:* For any eigenvalue λ of A and A^t and any positive integer r, show that $\operatorname{rank}((A - \lambda I)^r) = \operatorname{rank}((A^t - \lambda I)^r)$.

7. Let A be an $n \times n$ matrix whose characteristic polynomial splits, γ be a cycle of generalized eigenvectors corresponding to an eigenvalue λ, and W be the subspace spanned by γ. Define γ' to be the ordered set obtained from γ by reversing the order of the vectors in γ.

(a) Prove that $[\mathsf{T}_\mathsf{W}]_{\gamma'} = ([\mathsf{T}_\mathsf{W}]_\gamma)^t$.

(b) Let J be the Jordan canonical form of A. Use (a) to prove that J and J^t are similar.

(c) Use (b) to prove that A and A^t are similar.

8. Let T be a linear operator on a finite-dimensional vector space, and suppose that the characteristic polynomial of T splits. Let β be a Jordan canonical basis for T.

(a) Prove that for any nonzero scalar c, $\{cx : x \in \beta\}$ is a Jordan canonical basis for T.

(b) Suppose that γ is one of the cycles of generalized eigenvectors that forms β, and suppose that γ corresponds to the eigenvalue λ and has length greater than 1. Let x be the end vector of γ, and let y be a nonzero vector in E_λ. Let γ' be the ordered set obtained from γ by replacing x by $x + y$. Prove that γ' is a cycle of generalized eigenvectors corresponding to λ, and that if γ' replaces γ in the union that defines β, then the new union is also a Jordan canonical basis for T.

(c) Apply (b) to obtain a Jordan canonical basis for L_A, where A is the matrix given in Example 2, that is different from the basis given in the example.

9. Suppose that a dot diagram has k columns and m rows with p_j dots in column j and r_i dots in row i. Prove the following results.

(a) $m = p_1$ and $k = r_1$.

(b) $p_j = \max\{i : r_i \geq j\}$ for $1 \leq j \leq k$ and $r_i = \max\{j : p_j \geq i\}$ for $1 \leq i \leq m$. *Hint:* Use mathematical induction on m.

(c) $r_1 \geq r_2 \geq \cdots \geq r_m$.

(d) Deduce that the number of dots in each column of a dot diagram is completely determined by the number of dots in the rows.

10. Let T be a linear operator whose characteristic polynomial splits, and let λ be an eigenvalue of T.

(a) Prove that $\dim(\mathsf{K}_\lambda)$ is the sum of the lengths of all the blocks corresponding to λ in the Jordan canonical form of T.

(b) Deduce that $\mathsf{E}_\lambda = \mathsf{K}_\lambda$ if and only if all the Jordan blocks corresponding to λ are 1×1 matrices.

The following definitions are used in Exercises 11–19.

Definitions. *A linear operator T on a vector space V is called* ***nilpotent*** *if $\mathsf{T}^p = \mathsf{T}_0$ for some positive integer p. An $n \times n$ matrix A is called* ***nilpotent*** *if $A^p = O$ for some positive integer p.*

11. Let T be a linear operator on a finite-dimensional vector space V, and let β be an ordered basis for V. Prove that T is nilpotent if and only if $[\mathsf{T}]_\beta$ is nilpotent.

12. Prove that any square upper triangular matrix with each diagonal entry equal to zero is nilpotent.

13. Let T be a nilpotent operator on an n-dimensional vector space V, and suppose that p is the smallest positive integer for which $\mathsf{T}^p = \mathsf{T}_0$. Prove the following results.

(a) $\mathsf{N}(\mathsf{T}^i) \subseteq \mathsf{N}(\mathsf{T}^{i+1})$ for every positive integer i.

(b) There is a sequence of ordered bases $\beta_1, \beta_2, \ldots, \beta_p$ such that β_i is a basis for $\mathsf{N}(\mathsf{T}^i)$ and β_{i+1} contains β_i for $1 \le i \le p-1$.

(c) Let $\beta = \beta_p$ be the ordered basis for $\mathsf{N}(\mathsf{T}^p) = \mathsf{V}$ in (b). Then $[\mathsf{T}]_\beta$ is an upper triangular matrix with each diagonal entry equal to zero.

(d) The characteristic polynomial of T is $(-1)^n t^n$. Hence the characteristic polynomial of T splits, and 0 is the only eigenvalue of T.

14. Prove the converse of Exercise 13(d): If T is a linear operator on an n-dimensional vector space V and $(-1)^n t^n$ is the characteristic polynomial of T, then T is nilpotent.

15. Give an example of a linear operator T on a finite-dimensional vector space over the field of real numbers such that T is not nilpotent, but zero is the only eigenvalue of T. Characterize all such operators. Visit `goo.gl/nDjsWm` for a solution.

16. Let T be a nilpotent linear operator on a finite-dimensional vector space V. Recall from Exercise 13 that $\lambda = 0$ is the only eigenvalue of T, and hence $\mathsf{V} = \mathsf{K}_\lambda$. Let β be a Jordan canonical basis for T. Prove that for any positive integer i, if we delete from β the vectors corresponding to the last i dots in each column of a dot diagram of β, the resulting set is a basis for $\mathsf{R}(\mathsf{T}^i)$. (If a column of the dot diagram contains fewer than i dots, all the vectors associated with that column are removed from β.)

17. Let T be a linear operator on a finite-dimensional vector space V such that the characteristic polynomial of T splits, and let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T. For each i, let v_i denote the unique vector in K_{λ_i} such that $x = v_1 + v_2 + \cdots + v_k$. (This unique representation is guaranteed by Theorem 7.3 (p. 479).) Define a mapping $\mathsf{S}: \mathsf{V} \to \mathsf{V}$ by

$$\mathsf{S}(x) = \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_k v_k.$$

(a) Prove that S is a diagonalizable linear operator on V.

(b) Let $\mathsf{U} = \mathsf{T} - \mathsf{S}$. Prove that U is nilpotent and commutes with S, that is, $\mathsf{SU} = \mathsf{US}$.

18. Let T be a linear operator on a finite-dimensional vector space V over C, and let J be the Jordan canonical form of T. Let D be the diagonal matrix whose diagonal entries are the diagonal entries of J, and let $M = J - D$. Prove the following results.

(a) M is nilpotent.

(b) $MD = DM$.

(c) If p is the smallest positive integer for which $M^p = O$, then, for any positive integer $r < p$,

$$J^r = D^r + rD^{r-1}M + \frac{r(r-1)}{2!}D^{r-2}M^2 + \cdots + rDM^{r-1} + M^r,$$

and, for any positive integer $r \geq p$,

$$J^r = D^r + rD^{r-1}M + \frac{r(r-1)}{2!}D^{r-2}M^2 + \cdots + \frac{r!}{(r-p+1)!(p-1)!}D^{r-p+1}M^{p-1}.$$

19. Let $F = C$ and

$$J = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}$$

be the $m \times m$ Jordan block corresponding to λ, and let $N = J - \lambda I_m$. Prove the following results:

(a) $N^m = O$, and for $1 \leq r < m$,

$$N_{ij}^r = \begin{cases} 1 & \text{if } j = i + r \\ 0 & \text{otherwise.} \end{cases}$$

(b) For any integer $r \geq m$,

$$J^r = \begin{pmatrix} \lambda^r & r\lambda^{r-1} & \frac{r(r-1)}{2!}\lambda^{r-2} & \cdots & \frac{r(r-1)\cdots(r-m+2)}{(m-1)!}\lambda^{r-m+1} \\ 0 & \lambda^r & r\lambda^{r-1} & \cdots & \frac{r(r-1)\cdots(r-m+3)}{(m-2)!}\lambda^{r-m+2} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda^r \end{pmatrix}.$$

(c) $\lim_{r\to\infty} J^r$ exists if and only if one of the following holds:

(i) $|\lambda| < 1$.

(ii) $\lambda = 1$ and $m = 1$.

(Note that $\lim_{r\to\infty} \lambda^r$ exists under these conditions. See the discussion preceding Theorem 5.12 on page 284.) Furthermore, $\lim_{r\to\infty} J^r$ is the zero matrix if condition (i) holds and is the 1×1 matrix (1) if condition (ii) holds.

(d) Prove Theorem 5.12 on page 284.

The following definition is used in Exercises 20 and 21.

Definition. *For any $A \in \mathsf{M}_{n\times n}(C)$, define the norm of A by*

$$\|A\|_m = \max\{|A_{ij}| : 1 \le i, j \le n\}.$$

20. Let $A, B \in \mathsf{M}_{n\times n}(C)$. Prove the following results.

(a) $\|A\|_m \ge 0$.

(b) $\|A\|_m = 0$ if and only if $A = O$.

(c) $\|cA\|_m = |c| \cdot \|A\|_m$ for any scalar c.

(d) $\|A + B\|_m \le \|A\|_m + \|B\|_m$.

(e) $\|AB\|_m \le n\|A\|_m\|B\|_m$.

21. Let $A \in \mathsf{M}_{n\times n}(C)$ be a transition matrix. (See Section 5.3.) Since C is an algebraically closed field, A has a Jordan canonical form J to which A is similar. Let P be an invertible matrix such that $P^{-1}AP = J$. Prove the following results.

(a) $\|A^k\|_m \le 1$ for every positive integer k.

(b) There exists a positive number c such that $\|J^k\|_m \le c$ for every positive integer k.

(c) Each Jordan block of J corresponding to the eigenvalue $\lambda = 1$ is a 1×1 matrix.

(d) $\lim_{k\to\infty} A^k$ exists if and only if 1 is the only eigenvalue of A with absolute value 1.

(e) Theorem 5.19(a), using (c) and Theorem 5.18.

The next exercise requires knowledge of absolutely convergent series as well as the definition of e^A for a matrix A. (See page 310.)

22. Use Exercise 20(d) to prove that e^A exists for every $A \in \mathsf{M}_{n\times n}(C)$.

23. Let $x' = Ax$ be a system of n linear differential equations, where x is an n-tuple of differentiable functions $x_1(t), x_2(t), \ldots, x_n(t)$ of the real variable t, and A is an $n \times n$ coefficient matrix as in Exercise 16 of Section 5.2. In contrast to that exercise, however, do not assume that A is diagonalizable, but assume that the characteristic polynomial of A splits. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of A.

(a) Prove that if u is the end vector of a cycle of generalized eigenvectors of L_A of length p and u corresponds to the eigenvalue λ_i, then for any polynomial $f(t)$ of degree less than p, the function

$$e^{\lambda_i t}[f(t)(A - \lambda_i I)^{p-1} + f'(t)(A - \lambda_i I)^{p-2} + \cdots + f^{(p-1)}(t)]u$$

is a solution to the system $x' = Ax$.

(b) Prove that the general solution to $x' = Ax$ is a sum of the functions of the form given in (a), where the vectors u are the end vectors of the distinct cycles that constitute a fixed Jordan canonical basis for L_A.

24. Use Exercise 23 to find the general solution to each of the following systems of linear equations, where x, y, and z are real-valued differentiable functions of the real variable t.

(a)
$$\begin{aligned} x' &= 2x + y \\ y' &= 2y - z \\ z' &= 3z \end{aligned}$$

(b)
$$\begin{aligned} x' &= 2x + y \\ y' &= 2y + z \\ z' &= 2z \end{aligned}$$

7.3 THE MINIMAL POLYNOMIAL

The Cayley–Hamilton theorem (Theorem 5.22 p. 315) tells us that for any linear operator T on an n-dimensional vector space, there is a polynomial $f(t)$ of degree n such that $f(\mathsf{T}) = \mathsf{T}_0$, namely, the characteristic polynomial of T. Hence there is a polynomial of least degree with this property, and this degree is at most n. If $g(t)$ is such a polynomial, we can divide $g(t)$ by its leading coefficient to obtain another polynomial $p(t)$ of the same degree with leading coefficient 1, that is, $p(t)$ is a *monic* polynomial. (See Appendix E.)

Definition. *Let T be a linear operator on a finite-dimensional vector space. A polynomial $p(t)$ is called a* ***minimal polynomial*** *of T if $p(t)$ is a monic polynomial of least positive degree for which $p(\mathsf{T}) = \mathsf{T}_0$.*

The preceding discussion shows that every linear operator on a finite-dimensional vector space has a minimal polynomial. The next result shows that it is unique, and hence we can speak of *the* minimal polynomial of T.

Theorem 7.12. *Let $p(t)$ be a minimal polynomial of a linear operator T on a finite-dimensional vector space V.*

(a) *For any polynomial $g(t)$, if $g(\mathsf{T}) = \mathsf{T}_0$, then $p(t)$ divides $g(t)$. In particular, $p(t)$ divides the characteristic polynomial of T.*

(b) *The minimal polynomial of T is unique.*

Proof. (a) Let $g(t)$ be a polynomial for which $g(\mathsf{T}) = \mathsf{T}_0$. By the division algorithm for polynomials (Theorem E.1 of Appendix E, p. 555), there exist polynomials $q(t)$ and $r(t)$ such that

$$g(t) = q(t)p(t) + r(t), \tag{1}$$

where $r(t)$ has degree less than the degree of $p(t)$. Substituting T into (1) and using that $g(\mathsf{T}) = p(\mathsf{T}) = \mathsf{T}_0$, we have $r(\mathsf{T}) = \mathsf{T}_0$. Since $r(t)$ has degree less than $p(t)$ and $p(t)$ is the minimal polynomial of T, $r(t)$ must be the zero polynomial. Thus (1) simplifies to $g(t) = q(t)p(t)$, proving (a).

(b) Suppose that $p_1(t)$ and $p_2(t)$ are each minimal polynomials of T. Then $p_1(t)$ divides $p_2(t)$ by (a). Since $p_1(t)$ and $p_2(t)$ have the same degree, we have that $p_2(t) = cp_1(t)$ for some nonzero scalar c. Because $p_1(t)$ and $p_2(t)$ are monic, $c = 1$; hence $p_1(t) = p_2(t)$. ∎

The minimal polynomial of a linear operator has an obvious analog for a matrix.

Definition. *Let $A \in \mathsf{M}_{n\times n}(F)$. The* ***minimal polynomial*** *$p(t)$ of A is the monic polynomial of least positive degree for which $p(A) = O$.*

The following results are now immediate.

Theorem 7.13. *Let T be a linear operator on a finite-dimensional vector space V, and let β be an ordered basis for V. Then the minimal polynomial of T is the same as the minimal polynomial of $[\mathsf{T}]_\beta$.*

Proof. Exercise. ∎

Corollary. *For any $A \in \mathsf{M}_{n\times n}(F)$, the minimal polynomial of A is the same as the minimal polynomial of L_A.*

Proof. Exercise. ∎

In view of the preceding theorem and corollary, Theorem 7.12 and all subsequent theorems in this section that are stated for operators are also valid for matrices.

For the remainder of this section, we study primarily minimal polynomials of operators (and hence matrices) whose characteristic polynomials split. A more general treatment of minimal polynomials is given in Section 7.4.

Theorem 7.14. *Let T be a linear operator on a finite-dimensional vector space V, and let $p(t)$ be the minimal polynomial of T. A scalar λ is an eigenvalue of T if and only if $p(\lambda) = 0$. Hence the characteristic polynomial and the minimal polynomial of T have the same zeros.*

Proof. Let $f(t)$ be the characteristic polynomial of T. Since $p(t)$ divides $f(t)$, there exists a polynomial $q(t)$ such that $f(t) = q(t)p(t)$. If λ is a zero of $p(t)$, then

$$f(\lambda) = q(\lambda)p(\lambda) = q(\lambda)\cdot 0 = 0.$$

So λ is a zero of $f(t)$; that is, λ is an eigenvalue of T.

Conversely, suppose that λ is an eigenvalue of T, and let $x \in \mathsf{V}$ be an eigenvector corresponding to λ. By Exercise 22 of Section 5.1, we have

$$0 = \mathsf{T}_0(x) = p(\mathsf{T})(x) = p(\lambda)x.$$

Since $x \neq 0$, it follows that $p(\lambda) = 0$, and so λ is a zero of $p(t)$. ∎

The following corollary is immediate.

Corollary. *Let* T *be a linear operator on a finite-dimensional vector space* V *with minimal polynomial* $p(t)$ *and characteristic polynomial* $f(t)$. *Suppose that* $f(t)$ *factors as*

$$f(t) = (\lambda_1 - t)^{n_1}(\lambda_2 - t)^{n_2} \cdots (\lambda_k - t)^{n_k},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ *are the distinct eigenvalues of* T. *Then there exist integers* $m_1, m_2, \ldots, m_k$ *such that* $1 \le m_i \le n_i$ *for all* i *and*

$$p(t) = (t - \lambda_1)^{m_1}(t - \lambda_2)^{m_2} \cdots (t - \lambda_k)^{m_k}.$$

Example 1

We compute the minimal polynomial of the matrix

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

Since A has the characteristic polynomial

$$f(t) = \det \begin{pmatrix} 3-t & -1 & 0 \\ 0 & 2-t & 0 \\ 1 & -1 & 2-t \end{pmatrix} = -(t-2)^2(t-3),$$

the minimal polynomial of A must be either $(t-2)(t-3)$ or $(t-2)^2(t-3)$ by the corollary to Theorem 7.14. Substituting A into $p(t) = (t-2)(t-3)$, we find that $p(A) = O$; hence $p(t)$ is the minimal polynomial of A. ◆

Example 2

Let T be the linear operator on R^2 defined by

$$\mathsf{T}(a, b) = (2a + 5b, 6a + b)$$

and β be the standard ordered basis for R^2. Then

$$[\mathsf{T}]_\beta = \begin{pmatrix} 2 & 5 \\ 6 & 1 \end{pmatrix},$$

and hence the characteristic polynomial of T is

$$f(t) = \det \begin{pmatrix} 2-t & 5 \\ 6 & 1-t \end{pmatrix} = (t-7)(t+4).$$

Thus the minimal polynomial of T is also $(t-7)(t+4)$. ◆

Example 3

Let D be the linear operator on $\mathsf{P}_2(R)$ defined by $\mathsf{D}(g(x)) = g'(x)$, the derivative of $g(x)$. We compute the minimal polynomial of T. Let β be the standard ordered basis for $\mathsf{P}_2(R)$. Then

$$[\mathsf{D}]_\beta = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix},$$

and it follows that the characteristic polynomial of D is $-t^3$. So by the corollary to Theorem 7.14, the minimal polynomial of D is t, t^2, or t^3. Since $\mathsf{D}^2(x^2) = 2 \neq 0$, it follows that $\mathsf{D}^2 \neq \mathsf{T}_0$; hence the minimal polynomial of D must be t^3. ◆

In Example 3, it is easily verified that $\mathsf{P}_2(R)$ is a D-cyclic subspace (of itself). Here the minimal and characteristic polynomials are of the same degree. This is no coincidence.

Theorem 7.15. *Let T be a linear operator on an n-dimensional vector space V such that V is a T-cyclic subspace of itself. Then the characteristic polynomial $f(t)$ and the minimal polynomial $p(t)$ have the same degree, and hence $f(t) = (-1)^n p(t)$.*

Proof. Since V is a T-cyclic space, there exists an $x \in \mathsf{V}$ such that

$$\beta = \{x, \mathsf{T}(x), \ldots, \mathsf{T}^{n-1}(x)\}$$

is a basis for V (Theorem 5.21 p. 314). Let

$$g(t) = a_0 + a_1 t + \cdots + a_k t^k,$$

be a polynomial of degree $k < n$. Then $a_k \neq 0$ and

$$g(\mathsf{T})(x) = a_0 x + a_1 \mathsf{T}(x) + \cdots + a_k \mathsf{T}^k(x),$$

and so $g(\mathsf{T})(x)$ is a linear combination of the vectors of β having at least one nonzero coefficient, namely, a_k. Since β is linearly independent, it follows that $g(\mathsf{T})(x) \neq 0$; hence $g(\mathsf{T}) \neq \mathsf{T}_0$. Therefore the minimal polynomial of T has degree n, which is also the degree of the characteristic polynomial of T. ∎

Theorem 7.15 gives a condition under which the degree of the minimal polynomial of an operator is as large as possible. We now investigate the other extreme. By Theorem 7.14, the degree of the minimal polynomial of an operator must be greater than or equal to the number of distinct eigenvalues of the operator. The next result shows that the operators for which the degree of the minimal polynomial is as small as possible are precisely the diagonalizable operators.

Theorem 7.16. *Let* T *be a linear operator on a finite-dimensional vector space* V. *Then* T *is diagonalizable if and only if the minimal polynomial of* T *is of the form*

$$p(t) = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_k),$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ *are the distinct eigenvalues of* T.

Proof. Suppose that T is diagonalizable. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T, and define

$$p(t) = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_k).$$

By Theorem 7.14, $p(t)$ divides the minimal polynomial of T. Let $\beta = \{v_1, v_2, \ldots, v_n\}$ be a basis for V consisting of eigenvectors of T, and consider any $v_i \in \beta$. Then $(\mathsf{T} - \lambda_j \mathsf{I})(v_i) = 0$ for some eigenvalue λ_j. Since $t - \lambda_j$ divides $p(t)$, there is a polynomial $q_j(t)$ such that $p(t) = q_j(t)(t - \lambda_j)$. Hence

$$p(\mathsf{T})(v_i) = q_j(\mathsf{T})(\mathsf{T} - \lambda_j \mathsf{I})(v_i) = 0.$$

It follows that $p(\mathsf{T}) = \mathsf{T}_0$, since $p(\mathsf{T})$ takes each vector in a basis for V into 0. Therefore $p(t)$ is the minimal polynomial of T.

Conversely, suppose that there are distinct scalars $\lambda_1, \lambda_2, \ldots, \lambda_k$ such that the minimal polynomial $p(t)$ of T factors as

$$p(t) = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_k).$$

By Theorem 7.14, the λ_i's are eigenvalues of T. We apply mathematical induction on $n = \dim(\mathsf{V})$. Clearly T is diagonalizable for $n = 1$. Now assume that T is diagonalizable whenever $\dim(\mathsf{V}) < n$ for some $n > 1$, and let $\dim(\mathsf{V}) = n$ and $\mathsf{W} = \mathsf{R}(\mathsf{T} - \lambda_k \mathsf{I})$. Obviously $\mathsf{W} \neq \mathsf{V}$, because λ_k is an eigenvalue of T. If $\mathsf{W} = \{0\}$, then $\mathsf{T} = \lambda_k \mathsf{I}$, which is clearly diagonalizable. So suppose that $0 < \dim(\mathsf{W}) < n$. Then W is T-invariant, and for any $x \in \mathsf{W}$,

$$(\mathsf{T} - \lambda_1 \mathsf{I})(\mathsf{T} - \lambda_2 \mathsf{I}) \cdots (\mathsf{T} - \lambda_{k-1} \mathsf{I})(x) = 0.$$

It follows that the minimal polynomial of T_W divides the polynomial $(t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_{k-1})$. Hence by the induction hypothesis, T_W is diagonalizable. Furthermore, λ_k is not an eigenvalue of T_W by Theorem 7.14. Therefore $\mathsf{W} \cap \mathsf{N}(\mathsf{T} - \lambda_k \mathsf{I}) = \{0\}$. Now let $\beta_1 = \{v_1, v_2, \ldots, v_m\}$ be a basis for W consisting of eigenvectors of T_W (and hence of T), and let $\beta_2 = \{w_1, w_2, \ldots, w_p\}$ be a basis for $\mathsf{N}(\mathsf{T} - \lambda_k \mathsf{I})$, the eigenspace of T corresponding to λ_k. Then β_1 and β_2 are disjoint by the previous comment. Moreover, $m + p = n$ by the dimension theorem applied to $\mathsf{T} - \lambda_k \mathsf{I}$. We show that $\beta = \beta_1 \cup \beta_2$ is linearly independent. Consider scalars $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_p$ such that

$$a_1 v_1 + a_2 v_2 + \cdots + a_m v_m + b_1 w_1 + b_2 w_2 + \cdots + b_p w_p = 0.$$

Let

$$x = \sum_{i=1}^{m} a_i v_i \qquad \text{and} \qquad y = \sum_{i=1}^{p} b_i w_i.$$

Then $x \in \mathsf{W}$, $y \in \mathsf{N}(\mathsf{T} - \lambda_k \mathsf{I})$, and $x + y = 0$. It follows that $x = -y \in \mathsf{W} \cap \mathsf{N}(\mathsf{T} - \lambda_k \mathsf{I})$, and therefore $x = 0$. Since β_1 is linearly independent, we have that $a_1 = a_2 = \cdots = a_m = 0$. Similarly, $b_1 = b_2 = \cdots = b_p = 0$, and we conclude that β is a linearly independent subset of V consisting of n eigenvectors. It follows that β is a basis for V consisting of eigenvectors of T, and consequently T is diagonalizable. ∎

In addition to diagonalizable operators, there are methods for determining the minimal polynomial of any linear operator on a finite-dimensional vector space. In the case that the characteristic polynomial of the operator splits, the minimal polynomial can be described using the Jordan canonical form of the operator. (See Exercise 13.) In the case that the characteristic polynomial does not split, the minimal polynomial can be described using the *rational canonical form*, which we study in the next section. (See Exercise 7 of Section 7.4.)

Example 4

We determine all matrices $A \in \mathsf{M}_{2\times 2}(R)$ for which $A^2 - 3A + 2I = O$. Let $g(t) = t^2 - 3t + 2 = (t-1)(t-2)$. Since $g(A) = O$, the minimal polynomial $p(t)$ of A divides $g(t)$. Hence the only possible candidates for $p(t)$ are $t-1$, $t-2$, and $(t-1)(t-2)$. If $p(t) = t-1$ or $p(t) = t-2$, then $A = I$ or $A = 2I$, respectively. If $p(t) = (t-1)(t-2)$, then A is diagonalizable with eigenvalues 1 and 2, and hence A is similar to

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

that is, $A = QDQ^{-1}$ for some invertible matrix Q. ◆

Example 5

Let $A \in \mathsf{M}_{n\times n}(R)$ satisfy $A^3 = A$. We show that A is diagonalizable. Let $g(t) = t^3 - t = t(t+1)(t-1)$. Then $g(A) = O$, and hence the minimal polynomial $p(t)$ of A divides $g(t)$. Since $g(t)$ has no repeated factors, neither does $p(t)$. Thus A is diagonalizable by Theorem 7.16. ◆

Example 6

In Example 3, we saw that the minimal polynomial of the differential operator D on $\mathsf{P}_2(R)$ is t^3. Hence, by Theorem 7.16, D is not diagonalizable. ◆

EXERCISES

1. Label the following statements as true or false. Assume that all vector spaces are finite-dimensional.

(a) Every linear operator T has a polynomial $p(t)$ of largest degree for which $p(\mathsf{T}) = \mathsf{T}_0$.
(b) Every linear operator has a unique minimal polynomial.
(c) The characteristic polynomial of a linear operator divides the minimal polynomial of that operator.
(d) The minimal and the characteristic polynomials of any diagonalizable operator are equal.
(e) Let T be a linear operator on an n-dimensional vector space V, $p(t)$ be the minimal polynomial of T, and $f(t)$ be the characteristic polynomial of T. Suppose that $f(t)$ splits. Then $f(t)$ divides $[p(t)]^n$.
(f) The minimal polynomial of a linear operator always has the same degree as the characteristic polynomial of the operator.
(g) A linear operator is diagonalizable if its minimal polynomial splits.
(h) Let T be a linear operator on a vector space V such that V is a T-cyclic subspace of itself. Then the degree of the minimal polynomial of T equals $\dim(\mathsf{V})$.
(i) Let T be a linear operator on a vector space V such that T has n distinct eigenvalues, where $n = \dim(\mathsf{V})$. Then the degree of the minimal polynomial of T equals n.

2. Find the minimal polynomial of each of the following matrices.

(a) $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ **(b)** $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

(c) $\begin{pmatrix} 4 & -14 & 5 \\ 1 & -4 & 2 \\ 1 & -6 & 4 \end{pmatrix}$ **(d)** $\begin{pmatrix} 3 & 0 & 1 \\ 2 & 2 & 2 \\ -1 & 0 & 1 \end{pmatrix}$

3. For each linear operator T on V, find the minimal polynomial of T.

(a) $\mathsf{V} = \mathsf{R}^2$ and $\mathsf{T}(a, b) = (a + b, a - b)$
(b) $\mathsf{V} = \mathsf{P}_2(R)$ and $\mathsf{T}(g(x)) = g'(x) + 2g(x)$
(c) $\mathsf{V} = \mathsf{P}_2(R)$ and $\mathsf{T}(f(x)) = -xf''(x) + f'(x) + 2f(x)$
(d) $\mathsf{V} = \mathsf{M}_{n\times n}(R)$ and $\mathsf{T}(A) = A^t$. *Hint:* Note that $\mathsf{T}^2 = \mathsf{I}$.

4. Determine which of the matrices and operators in Exercises 2 and 3 are diagonalizable.

5. Describe all linear operators T on R^2 such that T is diagonalizable and $\mathsf{T}^3 - 2\mathsf{T}^2 + \mathsf{T} = \mathsf{T}_0$.

6. Prove Theorem 7.13 and its corollary.

7. Prove the corollary to Theorem 7.14.

8. Let T be a linear operator on a finite-dimensional vector space, and let $p(t)$ be the minimal polynomial of T. Prove the following results.

(a) T is invertible if and only if $p(0) \neq 0$.
(b) If T is invertible and $p(t) = t^m + a_{m-1}t^{m-1} + \cdots + a_1 t + a_0$, then

$$\mathsf{T}^{-1} = -\frac{1}{a_0}\left(\mathsf{T}^{m-1} + a_{m-1}\mathsf{T}^{m-2} + \cdots + a_2\mathsf{T} + a_1\mathsf{I}\right).$$

9. Let T be a diagonalizable linear operator on a finite-dimensional vector space V. Prove that V is a T-cyclic subspace if and only if each of the eigenspaces of T is one-dimensional.

10. Let T be a linear operator on a finite-dimensional vector space V, and suppose that W is a T-invariant subspace of V. Prove that the minimal polynomial of T_W divides the minimal polynomial of T.

11. Let $g(t)$ be the auxiliary polynomial associated with a homogeneous linear differential equation with constant coefficients (as defined in Section 2.7), and let V denote the solution space of this differential equation. Prove the following results.

(a) V is a D-invariant subspace, where D is the differentiation operator on C^∞.
(b) The minimal polynomial of D_V (the restriction of D to V) is $g(t)$.
(c) If the degree of $g(t)$ is n, then the characteristic polynomial of D_V is $(-1)^n g(t)$.

Hint: Use Theorem 2.32 (p. 135) for (b) and (c).

12. Let D be the differentiation operator on $\mathsf{P}(R)$, the space of polynomials over R. Prove that there exists no polynomial $g(t)$ for which $g(\mathsf{D}) = \mathsf{T}_0$. Hence D has no minimal polynomial.

13. Let T be a linear operator on a finite-dimensional vector space, and suppose that the characteristic polynomial of T splits. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of T, and for each i let p_i be the number of rows in the largest Jordan block corresponding to λ_i in a Jordan canonical form of T. Prove that the minimal polynomial of T is

$$(t - \lambda_1)^{p_1}(t - \lambda_2)^{p_2} \cdots (t - \lambda_k)^{p_k}.$$

The following exercise requires knowledge of direct sums (see Section 5.2).

14. Let T be a linear operator on a finite-dimensional vector space V, and let W_1 and W_2 be T-invariant subspaces of V such that $V = W_1 \oplus W_2$. Suppose that $p_1(t)$ and $p_2(t)$ are the minimal polynomials of T_{W_1} and T_{W_2}, respectively. Either prove that the minimal polynomial $f(t)$ of T always equals $p_1(t)p_2(t)$ or give an example in which $f(t) \neq p_1(t)p_2(t)$.

Exercise 15 uses the following definition.

Definition. *Let T be a linear operator on a finite-dimensional vector space V, and let x be a nonzero vector in V. The polynomial $p(t)$ is called a* **T-annihilator** *of x if $p(t)$ is a monic polynomial of least degree for which $p(T)(x) = 0$.*

15.† Let T be a linear operator on a finite-dimensional vector space V, and let x be a nonzero vector in V. Prove the following results.

(a) The vector x has a unique T-annihilator.
(b) The T-annihilator of x divides any polynomial $g(t)$ for which $g(T) = T_0$.
(c) If $p(t)$ is the T-annihilator of x and W is the T-cyclic subspace generated by x, then $p(t)$ is the minimal polynomial of T_W, and $\dim(W)$ equals the degree of $p(t)$.
(d) The degree of the T-annihilator of x is 1 if and only if x is an eigenvector of T.

Visit `goo.gl/8KD6Gw` for a solution.

16. Let T be a linear operator on a finite-dimensional vector space V, and let W_1 be a T-invariant subspace of V. Let $x \in V$ such that $x \notin W_1$. Prove the following results.

(a) There exists a unique monic polynomial $g_1(t)$ of least positive degree such that $g_1(T)(x) \in W_1$.
(b) If $h(t)$ is a polynomial for which $h(T)(x) \in W_1$, then $g_1(t)$ divides $h(t)$.
(c) $g_1(t)$ divides the minimal and the characteristic polynomials of T.
(d) Let W_2 be a T-invariant subspace of V such that $W_2 \subseteq W_1$, and let $g_2(t)$ be the unique monic polynomial of least degree such that $g_2(T)(x) \in W_2$. Then $g_1(t)$ divides $g_2(t)$.

7.4* THE RATIONAL CANONICAL FORM

Until now we have used eigenvalues, eigenvectors, and generalized eigenvectors in our analysis of linear operators with characteristic polynomials that

split. In general, characteristic polynomials need not split, and indeed, operators need not have eigenvalues! However, the unique factorization theorem for polynomials (see page 562) guarantees that the characteristic polynomial $f(t)$ of any linear operator T on an n-dimensional vector space factors uniquely as

$$f(t) = (-1)^n(\phi_1(t))^{n_1}(\phi_2(t))^{n_2}\cdots(\phi_k(t))^{n_k},$$

where the $\phi_i(t)$'s $(1 \le i \le k)$ are distinct irreducible monic polynomials and the n_i's are positive integers. In the case that $f(t)$ splits, each irreducible monic polynomial factor is of the form $\phi_i(t) = t - \lambda_i$, where λ_i is an eigenvalue of T, and there is a one-to-one correspondence between eigenvalues of T and the irreducible monic factors of the characteristic polynomial. In general, eigenvalues need not exist, but the irreducible monic factors always exist. In this section, we establish structure theorems based on the irreducible monic factors of the characteristic polynomial instead of eigenvalues.

In this context, the following definition is the appropriate replacement for eigenspace and generalized eigenspace.

Definition. *Let T be a linear operator on a finite-dimensional vector space V with characteristic polynomial*

$$f(t) = (-1)^n(\phi_1(t))^{n_1}(\phi_2(t))^{n_2}\cdots(\phi_k(t))^{n_k},$$

where the $\phi_i(t)$'s $(1 \le i \le k)$ are distinct irreducible monic polynomials and the n_i's are positive integers. For $1 \le i \le k$, we define the subset K_{ϕ_i} of V by

$$\mathsf{K}_{\phi_i} = \{x \in \mathsf{V} \colon (\phi_i(\mathsf{T}))^p(x) = 0 \text{ for some positive integer } p\}.$$

We show that each K_{ϕ_i} is a nonzero T-invariant subspace of V. Note that if $\phi_i(t) = t - \lambda$ is of degree one, then K_{ϕ_i} is the generalized eigenspace of T corresponding to the eigenvalue λ.

Having obtained suitable generalizations of the related concepts of eigenvalue and eigenspace, our next task is to describe a canonical form of a linear operator suitable to this context. The one that we study is called the *rational canonical form*. Since a canonical form is a description of a matrix representation of a linear operator, it can be defined by specifying the form of the ordered bases allowed for these representations.

Here the bases of interest naturally arise from the generators of certain cyclic subspaces. For this reason, the reader should recall the definition of a T-cyclic subspace generated by a vector and Theorem 5.21 (p. 314). We briefly review this concept and introduce some new notation and terminology.

Let T be a linear operator on a finite-dimensional vector space V, and let x be a nonzero vector in V. We use the notation C_x for the T-cyclic subspace generated by x. Recall (Theorem 5.21) that if $\dim(\mathsf{C}_x) = k$, then the set

$$\{x, \mathsf{T}(x), \mathsf{T}^2(x), \ldots, \mathsf{T}^{k-1}(x)\}$$

is an ordered basis for C_x. To distinguish this basis from all other ordered bases for C_x, we call it the **T-cyclic basis generated by x** and denote it by β_x. Let A be the matrix representation of the restriction of T to C_x in the ordered basis β_x. Recall from the proof of Theorem 5.21 that

$$A = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{k-1} \end{pmatrix},$$

where

$$a_0 x + a_1 \mathsf{T}(x) + \cdots + a_{k-1}\mathsf{T}^{k-1}(x) + \mathsf{T}^k(x) = 0.$$

Furthermore, the characteristic polynomial of A is given by

$$\det(A - tI) = (-1)^k(a_0 + a_1 t + \cdots + a_{k-1}t^{k-1} + t^k).$$

The matrix A is called the **companion matrix** of the monic polynomial $h(t) = a_0 + a_1 t + \cdots + a_{k-1}t^{k-1} + t^k$. Every monic polynomial has a companion matrix, and the characteristic polynomial of the companion matrix of a monic polynomial $g(t)$ of degree k is equal to $(-1)^k g(t)$. (See Exercise 19 of Section 5.4.) By Theorem 7.15 (p. 512), the monic polynomial $h(t)$ is also the minimal polynomial of A. Since A is the matrix representation of the restriction of T to C_x, $h(t)$ is also the minimal polynomial of this restriction. By Exercise 15 of Section 7.3, $h(t)$ is also the T-annihilator of x.

It is the object of this section to prove that for every linear operator T on a finite-dimensional vector space V, there exists an ordered basis β for V such that the matrix representation $[\mathsf{T}]_\beta$ is of the form

$$\begin{pmatrix} C_1 & O & \cdots & O \\ O & C_2 & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & C_r \end{pmatrix},$$

where each C_i is the companion matrix of a polynomial $(\phi(t))^m$ such that $\phi(t)$ is a monic irreducible divisor of the characteristic polynomial of T and m is a positive integer. A matrix representation of this kind is called a **rational canonical form** of T. We call the accompanying basis a **rational canonical basis** for T.

The next theorem is a simple consequence of the following lemma, which relies on the concept of T-annihilator, introduced in the Exercises of Section 7.3.

Lemma. *Let* T *be a linear operator on a finite-dimensional vector space* V*, let* x *be a nonzero vector in* V*, and suppose that the* T*-annihilator of* x *is of the form* $(\phi(t))^p$ *for some irreducible monic polynomial* $\phi(t)$*. Then* $\phi(t)$ *divides the minimal polynomial of* T*, and* $x \in \mathsf{K}_\phi$*.*

Proof. By Exercise 15(b) of Section 7.3, $(\phi(t))^p$ divides the minimal polynomial of T. Therefore $\phi(t)$ divides the minimal polynomial of T. Furthermore, $x \in \mathsf{K}_\phi$ by the definition of K_ϕ. ∎

Theorem 7.17. *Let* T *be a linear operator on a finite-dimensional vector space* V*, and let* β *be an ordered basis for* V*. Then* β *is a rational canonical basis for* T *if and only if* β *is the disjoint union of* T*-cyclic bases* β_{v_i}*, where each* v_i *lies in* K_ϕ *for some irreducible monic divisor* $\phi(t)$ *of the characteristic polynomial of* T*.*

Proof. Exercise. ∎

Example 1

Suppose that T is a linear operator on R^8 and

$$\beta = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$$

is a rational canonical basis for T such that

$$C = [\mathsf{T}]_\beta = \left(\begin{array}{cc|cccc|cc} 0 & -3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right)$$

is a rational canonical form of T. In this case, the submatrices C_1, C_2, and C_3 are the companion matrices of the polynomials $\phi_1(t)$, $(\phi_2(t))^2$, and $\phi_2(t)$, respectively, where

$$\phi_1(t) = t^2 - t + 3 \quad \text{and} \quad \phi_2(t) = t^2 + 1.$$

In the context of Theorem 7.17, β is the disjoint union of the T-cyclic bases; that is,

$$\begin{aligned} \beta &= \beta_{v_1} \cup \beta_{v_3} \cup \beta_{v_7} \\ &= \{v_1, v_2\} \cup \{v_3, v_4, v_5, v_6\} \cup \{v_7, v_8\}. \end{aligned}$$

By Exercise 39 of Section 5.4, the characteristic polynomial $f(t)$ of T is the product of the characteristic polynomials of the companion matrices:

$$f(t) = \phi_1(t)(\phi_2(t))^2\phi_2(t) = \phi_1(t)(\phi_2(t))^3. \quad \blacklozenge$$

The rational canonical form C of the operator T in Example 1 is constructed from matrices of the form C_i, each of which is the companion matrix of some power of a monic irreducible divisor of the characteristic polynomial of T. Furthermore, each such divisor is used in this way at least once.

In the course of showing that every linear operator T on a finite dimensional vector space has a rational canonical form C, we show that the companion matrices C_i that constitute C are always constructed from powers of the monic irreducible divisors of the characteristic polynomial of T. A key role in our analysis is played by the subspaces K_ϕ, where $\phi(t)$ is an irreducible monic divisor of the minimal polynomial of T. Since the minimal polynomial of an operator divides the characteristic polynomial of the operator, every irreducible divisor of the former is also an irreducible divisor of the latter. We eventually show that the converse is also true; that is, the minimal polynomial and the characteristic polynomial have the same irreducible divisors.

We begin with a result that lists several properties of irreducible divisors of the minimal polynomial. The reader is advised to review the definition of T-annihilator and the accompanying Exercise 15 of Section 7.3.

Theorem 7.18. *Let T be a linear operator on a finite-dimensional vector space V, and suppose that*

$$p(t) = (\phi_1(t))^{m_1}(\phi_2(t))^{m_2}\cdots(\phi_k(t))^{m_k}$$

is the minimal polynomial of T, where the $\phi_i(t)$'s $(1 \le i \le k)$ are the distinct irreducible monic factors of $p(t)$ and the m_i's are positive integers. Then the following statements are true.

(a) *K_{ϕ_i} is a nonzero T-invariant subspace of V for each i.*

(b) *If x is a nonzero vector in some K_{ϕ_i}, then the T-annihilator of x is of the form $(\phi_i(t))^p$ for some integer p.*

(c) *$\mathsf{K}_{\phi_i} \cap \mathsf{K}_{\phi_j} = \{0\}$ for $i \neq j$.*

(d) *K_{ϕ_i} is invariant under $\phi_j(\mathsf{T})$ for $i \neq j$, and the restriction of $\phi_j(\mathsf{T})$ to K_{ϕ_i} is one-to-one and onto.*

(e) *$\mathsf{K}_{\phi_i} = \mathsf{N}((\phi_i(\mathsf{T}))^{m_i})$ for each i.*

Proof. If $k = 1$, then (a), (b), and (e) are obvious, while (c) and (d) are vacuously true. Now suppose that $k > 1$.

(a) The proof that K_{ϕ_i} is a T-invariant subspace of V is left as an exercise. Let $f_i(t)$ be the polynomial obtained from $p(t)$ by omitting the factor $(\phi_i(t))^{m_i}$. To prove that K_{ϕ_i} is nonzero, first observe that $f_i(t)$ is a proper divisor of $p(t)$; therefore there exists a vector $z \in \mathsf{V}$ such that $x = f_i(\mathsf{T})(z) \neq 0$. Then $x \in \mathsf{K}_{\phi_i}$ because

$$(\phi_i(\mathsf{T}))^{m_i}(x) = (\phi_i(\mathsf{T}))^{m_i} f_i(\mathsf{T})(z) = p(\mathsf{T})(z) = 0.$$

(b) Assume the hypothesis. Then $(\phi_i(\mathsf{T}))^q(x) = 0$ for some positive integer q. Hence the T-annihilator of x divides $(\phi_i(t))^q$ by Exercise 15(b) of Section 7.3, and the result follows.

(c) Assume $i \neq j$. Let $x \in \mathsf{K}_{\phi_i} \cap \mathsf{K}_{\phi_j}$, and suppose that $x \neq 0$. By (b), the T-annihilator of x is a power of both $\phi_i(t)$ and $\phi_j(t)$. But this is impossible because $\phi_i(t)$ and $\phi_j(t)$ are relatively prime (see Appendix E). We conclude that $x = 0$.

(d) Assume $i \neq j$. Since K_{ϕ_i} is T-invariant, it is also $\phi_j(\mathsf{T})$-invariant. Suppose that $\phi_j(\mathsf{T})(x) = 0$ for some $x \in \mathsf{K}_{\phi_i}$. Then $x \in \mathsf{K}_{\phi_i} \cap \mathsf{K}_{\phi_j} = \{0\}$ by (c). Therefore the restriction of $\phi_j(\mathsf{T})$ to K_{ϕ_i} is one-to-one. Since V is finite-dimensional, this restriction is also onto.

(e) Suppose that $1 \leq i \leq k$. Clearly, $\mathsf{N}((\phi_i(\mathsf{T}))^{m_i}) \subseteq \mathsf{K}_{\phi_i}$. Let $f_i(t)$ be the polynomial defined in (a). Since $f_i(t)$ is a product of polynomials of the form $\phi_j(t)$ for $j \neq i$, we have by (d) that the restriction of $f_i(\mathsf{T})$ to K_{ϕ_i} is onto. Let $x \in \mathsf{K}_{\phi_i}$. Then there exists $y \in \mathsf{K}_{\phi_i}$ such that $f_i(\mathsf{T})(y) = x$. Therefore

$$((\phi_i(\mathsf{T}))^{m_i})(x) = ((\phi_i(\mathsf{T}))^{m_i})f_i(\mathsf{T})(y) = p(\mathsf{T})(y) = 0,$$

and hence $x \in \mathsf{N}((\phi_i(\mathsf{T}))^{m_i})$. Thus $\mathsf{K}_{\phi_i} = \mathsf{N}((\phi_i(\mathsf{T}))^{m_i})$. ∎

Since a rational canonical basis for an operator T is obtained from a union of T-cyclic bases, we need to know when such a union is linearly independent. The next major result, Theorem 7.19, reduces this problem to the study of T-cyclic bases within K_ϕ, where $\phi(t)$ is an irreducible monic divisor of the minimal polynomial of T. We begin with the following lemma.

Lemma. *Let T be a linear operator on a finite-dimensional vector space V, and suppose that*

$$p(t) = (\phi_1(t))^{m_1}(\phi_2(t))^{m_2} \cdots (\phi_k(t))^{m_k}$$

is the minimal polynomial of T, where the ϕ_i's $(1 \leq i \leq k)$ are the distinct irreducible monic factors of $p(t)$ and the m_i's are positive integers. For $1 \leq i \leq k$, let $v_i \in \mathsf{K}_{\phi_i}$ be such that

$$v_1 + v_2 + \cdots + v_k = 0. \tag{2}$$

Then $v_i = 0$ for all i.

Proof. The result is trivial if $k = 1$, so suppose that $k > 1$. Consider any i. Let $f_i(t)$ be the polynomial obtained from $p(t)$ by omitting the factor $(\phi_i(t))^{m_i}$. As a consequence of Theorem 7.18, $f_i(\mathsf{T})$ is one-to-one on K_{ϕ_i}, and $f_i(\mathsf{T})(v_j) = 0$ for $i \neq j$. Thus, applying $f_i(\mathsf{T})$ to (2), we obtain $f_i(\mathsf{T})(v_i) = 0$, from which it follows that $v_i = 0$. ∎

Theorem 7.19. *Let T be a linear operator on a finite-dimensional vector space V, and suppose that*

$$p(t) = (\phi_1(t))^{m_1}(\phi_2(t))^{m_2} \cdots (\phi_k(t))^{m_k}$$

is the minimal polynomial of T, *where the* ϕ_i*'s* $(1 \le i \le k)$ *are the distinct irreducible monic factors of* $p(t)$ *and the* m_i*'s are positive integers. For* $1 \le i \le k$, *let* S_i *be a linearly independent subset of* K_{ϕ_i}. *Then*

(a) $S_i \cap S_j = \varnothing$ *for* $i \neq j$

(b) $S_1 \cup S_2 \cup \cdots \cup S_k$ *is linearly independent.*

Proof. If $k = 1$, then (a) is vacuously true and (b) is obvious. Now suppose that $k > 1$. Then (a) follows immediately from Theorem 7.18(c). Furthermore, the proof of (b) is identical to the proof of Theorem 5.5 (p. 261) with the eigenvectors replaced by the generalized eigenvectors. ∎

In view of Theorem 7.19, we can focus on bases of individual spaces of the form K_ϕ, where $\phi(t)$ is an irreducible monic divisor of the minimal polynomial of T. The next several results give us ways to construct bases for these spaces that are unions of T-cyclic bases. These results serve the dual purposes of leading to the existence theorem for the rational canonical form and of providing methods for constructing rational canonical bases.

For Theorems 7.20 and 7.21 and the latter's corollary, we fix a linear operator T on a finite-dimensional vector space V and an irreducible monic divisor $\phi(t)$ of the minimal polynomial of T.

Theorem 7.20. *Let* $v_1, v_2, \ldots, v_k$ *be distinct vectors in* K_ϕ *such that*

$$S_1 = \beta_{v_1} \cup \beta_{v_2} \cup \cdots \cup \beta_{v_k}$$

is linearly independent. For each i, *suppose there exists* $w_i \in \mathsf{V}$ *such that* $\phi(\mathsf{T})(w_i) = v_i$. *Then*

$$S_2 = \beta_{w_1} \cup \beta_{w_2} \cup \cdots \cup \beta_{w_k}$$

is also linearly independent.

Proof. Consider any linear combination of vectors in S_2 that sums to zero, say,

$$\sum_{i=1}^{k} \sum_{j=0}^{n_i} a_{ij} \mathsf{T}^j(w_i) = 0. \tag{3}$$

For each i, let $f_i(t)$ be the polynomial defined by

$$f_i(t) = \sum_{j=0}^{n_i} a_{ij} t^j.$$

Then (3) can be rewritten as

$$\sum_{i=1}^{k} f_i(\mathsf{T})(w_i) = 0. \tag{4}$$

Apply $\phi(\mathsf{T})$ to both sides of (4) to obtain

$$\sum_{i=1}^{k} \phi(\mathsf{T}) f_i(\mathsf{T})(w_i) = \sum_{i=1}^{k} f_i(\mathsf{T})\phi(\mathsf{T})(w_i) = \sum_{i=1}^{k} f_i(\mathsf{T})(v_i) = 0.$$

This last sum can be rewritten as a linear combination of the vectors in S_1 so that each $f_i(\mathsf{T})(v_i)$ is a linear combination of the vectors in β_{v_i}. Since S_1 is linearly independent, it follows that

$$f_i(\mathsf{T})(v_i) = 0 \quad \text{for all } i.$$

Therefore the T-annihilator of v_i divides $f_i(t)$ for all i. (See Exercise 15 of Section 7.3.) By Theorem 7.18(b), $\phi(t)$ divides the T-annihilator of v_i, and hence $\phi(t)$ divides $f_i(t)$ for all i. Thus, for each i, there exists a polynomial $g_i(t)$ such that $f_i(t) = g_i(t)\phi(t)$. So (4) becomes

$$\sum_{i=1}^{k} g_i(\mathsf{T})\phi(\mathsf{T})(w_i) = \sum_{i=1}^{k} g_i(\mathsf{T})(v_i) = 0.$$

Again, linear independence of S_1 requires that

$$f_i(\mathsf{T})(w_i) = g_i(\mathsf{T})(v_i) = 0 \quad \text{for all } i.$$

But $f_i(\mathsf{T})(w_i)$ is the result of grouping the terms of the linear combination in (3) that arise from the linearly independent set β_{w_i}. We conclude that for each i, $a_{ij} = 0$ for all j. Therefore S_2 is linearly independent. ∎

We now show that K_ϕ has a basis consisting of a union of T-cycles.

Lemma. *Let W be a T-invariant subspace of K_ϕ, and let β be a basis for W. Then the following statements are true.*

(a) *Suppose that $x \in \mathsf{N}(\phi(\mathsf{T}))$, but $x \notin \mathsf{W}$. Then $\beta \cup \beta_x$ is linearly independent.*

(b) *For some $w_1, w_2, \ldots, w_s$ in $\mathsf{N}(\phi(\mathsf{T}))$, β can be extended to the linearly independent set*

$$\beta' = \beta \cup \beta_{w_1} \cup \beta_{w_2} \cup \cdots \cup \beta_{w_s},$$

whose span contains $\mathsf{N}(\phi(\mathsf{T}))$.

Proof. (a) Let $\beta = \{v_1, v_2, \ldots, v_k\}$, and suppose that

$$\sum_{i=1}^{k} a_i v_i + z = 0 \qquad \text{and} \qquad z = \sum_{j=0}^{d-1} b_j \mathsf{T}^j(x),$$

where d is the degree of $\phi(t)$. Then $z \in \mathsf{C}_x \cap \mathsf{W}$, and hence $\mathsf{C}_z \subseteq \mathsf{C}_x \cap \mathsf{W}$. Suppose that $z \neq 0$. Then z has $\phi(t)$ as its T-annihilator, and therefore

$$d = \dim(\mathsf{C}_z) \leq \dim(\mathsf{C}_x \cap \mathsf{W}) \leq \dim(\mathsf{C}_x) = d.$$

It follows that $\mathsf{C}_x \cap \mathsf{W} = \mathsf{C}_x$, and consequently $x \in \mathsf{W}$, contrary to hypothesis. Therefore $z = 0$, from which it follows that $b_j = 0$ for all j. Since β is linearly independent, it follows that $a_i = 0$ for all i. Thus $\beta \cup \beta_x$ is linearly independent.

(b) Suppose that W does not contain $\mathsf{N}(\phi(\mathsf{T}))$. Choose a vector $w_1 \in \mathsf{N}(\phi(\mathsf{T}))$ that is not in W. By (a), $\beta_1 = \beta \cup \beta_{w_1}$ is linearly independent. Let $\mathsf{W}_1 = \text{span}(\beta_1)$. If W_1 does not contain $\mathsf{N}(\phi(\mathsf{T}))$, choose a vector w_2 in $\mathsf{N}(\phi(\mathsf{T}))$, but not in W_1, so that $\beta_2 = \beta_1 \cup \beta_{w_2} = \beta \cup \beta_{w_1} \cup \beta_{w_2}$ is linearly independent. Continuing this process, we eventually obtain vectors $w_1, w_2, \ldots, w_s$ in $\mathsf{N}(\phi(\mathsf{T}))$ such that the union

$$\beta' = \beta \cup \beta_{w_1} \cup \beta_{w_2} \cup \cdots \cup \beta_{w_s}$$

is a linearly independent set whose span contains $\mathsf{N}(\phi(\mathsf{T}))$. ∎

Theorem 7.21. *If the minimal polynomial of* T *is of the form* $p(t) = (\phi(t))^m$, *then there exists a rational canonical basis for* T.

Proof. The proof is by mathematical induction on m. Suppose that $m = 1$. Apply (b) of the lemma to $\mathsf{W} = \{0\}$ to obtain a linearly independent subset of V of the form $\beta_{v_1} \cup \beta_{v_2} \cup \cdots \cup \beta_{v_k}$, whose span contains $\mathsf{N}(\phi(\mathsf{T}))$. Since $\mathsf{V} = \mathsf{N}(\phi(\mathsf{T}))$, this set is a rational canonical basis for V.

Now suppose that, for some integer $m > 1$, the result is valid whenever the minimal polynomial of T is of the form $(\phi(t))^k$, where $k < m$, and assume that the minimal polynomial of T is $p(t) = (\phi(t))^m$. Let $r = \text{rank}(\phi(\mathsf{T}))$. Then $\mathsf{R}(\phi(\mathsf{T}))$ is a T-invariant subspace of V, and the restriction of T to this subspace has $(\phi(t))^{m-1}$ as its minimal polynomial. Therefore we may apply the induction hypothesis to obtain a rational canonical basis for the restriction of T to $\mathsf{R}(\mathsf{T})$. Suppose that $v_1, v_2, \ldots, v_k$ are the generating vectors of the T-cyclic bases that constitute this rational canonical basis. For each i, choose w_i in V such that $v_i = \phi(\mathsf{T})(w_i)$. By Theorem 7.20, the union β of the sets β_{w_i} is linearly independent. Let $\mathsf{W} = \text{span}(\beta)$. Then W contains $\mathsf{R}(\phi(\mathsf{T}))$. Apply (b) of the lemma and adjoin additional T-cyclic bases $\beta_{w_{k+1}}, \beta_{w_{k+2}}, \ldots, \beta_{w_s}$ to β, if necessary, where w_i is in $\mathsf{N}(\phi(\mathsf{T}))$ for $i \geq k$, to obtain a linearly independent set

$$\beta' = \beta_{w_1} \cup \beta_{w_2} \cup \cdots \cup \beta_{w_k} \cup \cdots \cup \beta_{w_s}$$

whose span W' contains both W and $\mathsf{N}(\phi(\mathsf{T}))$.

We show that $W' = V$. Let U denote the restriction of $\phi(T)$ to W', which is $\phi(T)$-invariant. By the way in which W' was obtained from $R(\phi(T))$, it follows that $R(U) = R(\phi(T))$ and $N(U) = N(\phi(T))$. Therefore

$$\begin{aligned}\dim(W') &= \text{rank}(U) + \text{nullity}(U)\\ &= \text{rank}(\phi(T)) + \text{nullity}(\phi(T))\\ &= \dim(V).\end{aligned}$$

Thus $W' = V$, and β' is a rational canonical basis for T. ∎

Corollary. *K_ϕ has a basis consisting of the union of T-cyclic bases.*

Proof. Apply Theorem 7.21 to the restriction of T to K_ϕ. ∎

We are now ready to study the general case.

Theorem 7.22. *Every linear operator on a finite-dimensional vector space has a rational canonical basis and, hence, a rational canonical form.*

Proof. Let T be a linear operator on a finite-dimensional vector space V, and let $p(t) = (\phi_1(t))^{m_1}(\phi_2(t))^{m_2}\cdots(\phi_k(t))^{m_k}$ be the minimal polynomial of T, where the $\phi_i(t)$'s are the distinct irreducible monic factors of $p(t)$ and $m_i > 0$ for all i. The proof is by mathematical induction on k. The case $k = 1$ is proved in Theorem 7.21.

Suppose that the result is valid whenever the minimal polynomial contains fewer than k distinct irreducible factors for some $k > 1$, and suppose that $p(t)$ contains k distinct factors. Let U be the restriction of T to the T-invariant subspace $W = R((\phi_k(T))^{m_k})$, and let $q(t)$ be the minimal polynomial of U. Then $q(t)$ divides $p(t)$ by Exercise 10 of Section 7.3. Furthermore, $\phi_k(t)$ does not divide $q(t)$. For otherwise, there would exist a nonzero vector $x \in W$ such that $\phi_k(U)(x) = 0$ and a vector $y \in V$ such that $x = (\phi_k(T))^{m_k}(y)$. It follows that $(\phi_k(T))^{m_k+1}(y) = 0$, and hence $y \in K_{\phi_k}$ and $x = (\phi_k(T))^{m_k}(y) = 0$ by Theorem 7.18(e), a contradiction. Thus $q(t)$ contains fewer than k distinct irreducible divisors. So by the induction hypothesis, U has a rational canonical basis β_1 consisting of a union of U-cyclic bases (and hence T-cyclic bases) of vectors from some of the subspaces K_{ϕ_i}, $1 \le i \le k-1$. By the corollary to Theorem 7.21, K_{ϕ_k} has a basis β_2 consisting of a union of T-cyclic bases. By Theorem 7.19, β_1 and β_2 are disjoint, and $\beta = \beta_1 \cup \beta_2$ is linearly independent. Let s denote the number of vectors in β. Then

$$\begin{aligned}s &= \dim(R((\phi_k(T))^{m_k})) + \dim(K_{\phi_k})\\ &= \text{rank}((\phi_k(T))^{m_k}) + \text{nullity}((\phi_k(T))^{m_k})\\ &= n.\end{aligned}$$

We conclude that β is a basis for V. Therefore β is a rational canonical basis, and T has a rational canonical form. ∎

In our study of the rational canonical form, we relied on the minimal polynomial. We are now able to relate the rational canonical form to the characteristic polynomial.

Theorem 7.23. *Let T be a linear operator on an n-dimensional vector space V with characteristic polynomial*

$$f(t) = (-1)^n(\phi_1(t))^{n_1}(\phi_2(t))^{n_2}\cdots(\phi_k(t))^{n_k},$$

where the $\phi_i(t)$'s $(1 \le i \le k)$ are distinct irreducible monic polynomials and the n_i's are positive integers. Then the following statements are true.

(a) *$\phi_1(t), \phi_2(t), \ldots, \phi_k(t)$ are the irreducible monic factors of the minimal polynomial.*
(b) *For each i, $\dim(\mathsf{K}_{\phi_i}) = d_i n_i$, where d_i is the degree of $\phi_i(t)$.*
(c) *If β is a rational canonical basis for T, then $\beta_i = \beta \cap \mathsf{K}_{\phi_i}$ is a basis for K_{ϕ_i} for each i.*
(d) *If γ_i is a basis for K_{ϕ_i} for each i, then $\gamma = \gamma_1 \cup \gamma_2 \cup \cdots \cup \gamma_k$ is a basis for V. In particular, if each γ_i is a disjoint union of T-cyclic bases, then γ is a rational canonical basis for T.*

Proof. (a) By Theorem 7.22, T has a rational canonical form C. By Exercise 39 of Section 5.4, the characteristic polynomial of C, and hence of T, is the product of the characteristic polynomials of the companion matrices that compose C. Therefore each irreducible monic divisor $\phi_i(t)$ of $f(t)$ divides the characteristic polynomial of at least one of the companion matrices, and hence for some integer p, $(\phi_i(t))^p$ is the T-annihilator of a nonzero vector of V. We conclude that $(\phi_i(t))^p$, and so $\phi_i(t)$, divides the minimal polynomial of T. Conversely, if $\phi(t)$ is an irreducible monic polynomial that divides the minimal polynomial of T, then $\phi(t)$ divides the characteristic polynomial of T because the minimal polynomial divides the characteristic polynomial.

(b), (c), and (d) Let $C = [\mathsf{T}]_\beta$, which is a rational canonical form of T. Consider any i $(1 \le i \le k)$. Since $f(t)$ is the product of the characteristic polynomials of the companion matrices that compose C, we may multiply those characteristic polynomials that arise from the T-cyclic bases in β_i to obtain the factor $(\phi_i(t))^{n_i}$ of $f(t)$. Since this polynomial has degree $n_i d_i$, and the union of these bases is a linearly independent subset β_i of K_{ϕ_i}, we have

$$n_i d_i \le \dim(\mathsf{K}_{\phi_i}).$$

Furthermore, $n = \sum_{i=1}^{k} d_i n_i$, because this sum is equal to the degree of $f(t)$. Now let s denote the number of vectors in γ. By Theorem 7.19, γ is linearly independent, and therefore

$$n = \sum_{i=1}^{k} d_i n_i \le \sum_{i=1}^{k} \dim(\mathsf{K}_{\phi_i}) = s \le n.$$

Hence $n = s$, and $d_i n_i = \dim(\mathsf{K}_{\phi_i})$ for all i. It follows that γ is a basis for V and β_i is a basis for K_{ϕ_i} for each i. ∎

Uniqueness of the Rational Canonical Form

Having shown that a rational canonical form exists, we are now in a position to ask about the extent to which it is unique. Certainly, the rational canonical form of a linear operator T can be modified by permuting the T-cyclic bases that constitute the corresponding rational canonical basis. This has the effect of permuting the companion matrices that make up the rational canonical form. As in the case of the Jordan canonical form, we show that except for these permutations, the rational canonical form is unique, although the rational canonical bases are not.

To simplify this task, we adopt the convention of ordering every rational canonical basis so that all the T-cyclic bases associated with the same irreducible monic divisor of the characteristic polynomial are grouped together. Furthermore, within each such grouping, we arrange the T-cyclic bases in decreasing order of size. Our task is to show that, subject to this order, the rational canonical form of a linear operator is unique up to the arrangement of the irreducible monic divisors.

As in the case of the Jordan canonical form, we introduce arrays of dots from which we can reconstruct the rational canonical form. For the Jordan canonical form, we devised a dot diagram for each eigenvalue of the given operator. In the case of the rational canonical form, we define a dot diagram for each irreducible monic divisor of the characteristic polynomial of the given operator. A proof that the resulting dot diagrams are completely determined by the operator is also a proof that the rational canonical form is unique.

In what follows, T is a linear operator on a finite-dimensional vector space with rational canonical basis β; $\phi(t)$ is an irreducible monic divisor of the characteristic polynomial of T; $\beta_{v_1}, \beta_{v_2}, \ldots, \beta_{v_k}$ are the T-cyclic bases of β that are contained in K_ϕ; and d is the degree of $\phi(t)$. For each j, let $(\phi(t))^{p_j}$ be the annihilator of v_j. This polynomial has degree dp_j; therefore, by Exercise 15 of Section 7.3, β_{v_j} contains dp_j vectors. Furthermore, $p_1 \geq p_2 \geq \cdots \geq p_k$ since the T-cyclic bases are arranged in decreasing order of size. We define the **dot diagram** of $\phi(t)$ to be the array consisting of k columns of dots with p_j dots in the jth column, arranged so that the jth column begins at the top and terminates after p_j dots. For example, if $k = 3$, $p_1 = 4$, $p_2 = 2$, and $p_3 = 2$, then the dot diagram is

Although each column of a dot diagram corresponds to a T-cyclic basis

β_{v_i} in K_ϕ, there are fewer dots in the column than there are vectors in the basis.

Example 2

Recall the linear operator T of Example 1 with the rational canonical basis β and the rational canonical form $C = [\mathsf{T}]_\beta$. Since there are two irreducible monic divisors of the characteristic polynomial of T, $\phi_1(t) = t^2 - t + 3$ and $\phi_2(t) = t^2 + 1$, there are two dot diagrams to consider. Because $\phi_1(t)$ is the T-annihilator of v_1 and β_{v_1} is a basis for K_{ϕ_1}, the dot diagram for $\phi_1(t)$ consists of a single dot. The other two T-cyclic bases, β_{v_3} and β_{v_7}, lie in K_{ϕ_2}. Since v_3 has T-annihilator $(\phi_2(t))^2$ and v_7 has T-annihilator $\phi_2(t)$, in the dot diagram of $\phi_2(t)$ we have $p_1 = 2$ and $p_2 = 1$. These diagrams are as follows:

Dot diagram for $\phi_1(t)$ Dot diagram for $\phi_2(t)$ ♦

In practice, we obtain the rational canonical form of a linear operator from the information provided by dot diagrams. This is illustrated in the next example.

Example 3

Let T be a linear operator on a finite-dimensional vector space over R, and suppose that the irreducible monic divisors of the characteristic polynomial of T are

$$\phi_1(t) = t - 1, \quad \phi_2(t) = t^2 + 2, \quad \text{and} \quad \phi_3(t) = t^2 + t + 1.$$

Suppose, furthermore, that the dot diagrams associated with these divisors are as follows:

Diagram for $\phi_1(t)$ Diagram for $\phi_2(t)$ Diagram for $\phi_3(t)$

Since the dot diagram for $\phi_1(t)$ has two columns, it contributes two companion matrices to the rational canonical form. The first column has two dots, and therefore corresponds to the 2×2 companion matrix of $(\phi_1(t))^2 = (t-1)^2$. The second column, with only one dot, corresponds to the 1×1 companion matrix of $\phi_1(t) = t - 1$. These two companion matrices are given by

$$C_1 = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad C_2 = (1).$$

The dot diagram for $\phi_2(t) = t^2 + 2$ consists of two columns. each containing a single dot; hence this diagram contributes two copies of the 2×2 companion

matrix for $\phi_2(t)$, namely,

$$C_3 = C_4 = \begin{pmatrix} 0 & -2 \\ 1 & 0 \end{pmatrix}.$$

The dot diagram for $\phi_3(t) = t^2 + t + 1$ consists of a single column with a single dot contributing the single 2×2 companion matrix

$$C_5 = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

Therefore the rational canonical form of T is the 9×9 matrix

$$C = \begin{pmatrix} C_1 & O & O & O & O \\ O & C_2 & O & O & O \\ O & O & C_3 & O & O \\ O & O & O & C_4 & O \\ O & O & O & O & C_5 \end{pmatrix}$$

$$= \left(\begin{array}{cc|ccccccc} 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{array}\right). \quad \blacklozenge$$

We return to the general problem of finding dot diagrams. As we did before, we fix a linear operator T on a finite-dimensional vector space and an irreducible monic divisor $\phi(t)$ of the characteristic polynomial of T. Let U denote the restriction of the linear operator $\phi(\mathsf{T})$ to K_ϕ. By Theorem 7.18(d), $\mathsf{U}^q = \mathsf{T}_0$ for some positive integer q. Consequently, by Exercise 12 of Section 7.2, the characteristic polynomial of U is $(-1)^m t^m$, where $m = \dim(\mathsf{K}_\phi)$. Therefore K_ϕ is the generalized eigenspace of U corresponding to $\lambda = 0$, and U has a Jordan canonical form. The dot diagram associated with the Jordan canonical form of U gives us a key to understanding the dot diagram of T that is associated with $\phi(t)$. We now relate the two diagrams.

Let β be a rational canonical basis for T, and $\beta_{v_1}, \beta_{v_2}, \ldots, \beta_{v_k}$ be the T-cyclic bases of β that are contained in K_ϕ. Consider one of these T-cyclic bases β_{v_j}, and suppose again that the T-annihilator of v_j is $(\phi(t))^{p_j}$. Then β_{v_j} consists of dp_j vectors in β. For $0 \le i < d$, let γ_i be the cycle of generalized eigenvectors of U corresponding to $\lambda = 0$ with end vector $\mathsf{T}^i(v_j)$,

where $\mathsf{T}^0(v_j) = b_j$. Then

$$\gamma_i = \{(\phi(\mathsf{T}))^{p_j-1}\mathsf{T}^i(v_j), (\phi(\mathsf{T}))^{p_j-2}\mathsf{T}^i(v_j), \ldots, (\phi(\mathsf{T}))\mathsf{T}^i(v_j), \mathsf{T}^i(v_j)\}.$$

By Theorem 7.1 (p. 478), γ_i is a linearly independent subset of C_{v_i}. Now let

$$\alpha_j = \gamma_0 \cup \gamma_1 \cup \cdots \cup \gamma_{d-1}.$$

Notice that α_j contains dp_j vectors.

Lemma 1. *α_j is an ordered basis for C_{v_j}.*

Proof. The key to this proof is Theorem 7.4 (p. 480). Since α_j is the union of cycles of generalized eigenvectors of U corresponding to $\lambda = 0$, it suffices to show that the set of initial vectors of these cycles

$$\{(\phi(\mathsf{T}))^{p_j-1}(v_j), (\phi(\mathsf{T}))^{p_j-1}\mathsf{T}(v_j), \ldots, (\phi(\mathsf{T}))^{p_j-1}\mathsf{T}^{d-1}(v_j)\}$$

is linearly independent. Consider any linear combination of these vectors

$$a_0(\phi(\mathsf{T}))^{p_j-1}(v_j) + a_1(\phi(\mathsf{T}))^{p_j-1}\mathsf{T}(v_j) + \cdots + a_{d-1}(\phi(\mathsf{T}))^{p_j-1}\mathsf{T}^{d-1}(v_j),$$

where not all of the coefficients are zero. Let $g(t)$ be the polynomial defined by $g(t) = a_0 + a_1 t + \cdots + a_{d-1}t^{d-1}$. Then $g(t)$ is a nonzero polynomial of degree less than d, and hence $(\phi(t))^{p_j-1}g(t)$ is a nonzero polynomial with degree less than dp_j. Since $(\phi(t))^{p_j}$ is the T-annihilator of v_j, it follows that $(\phi(\mathsf{T}))^{p_j-1}g(\mathsf{T})(v_j) \neq 0$. Therefore the set of initial vectors is linearly independent. So by Theorem 7.4, α_j is linearly independent, and the γ_i's are disjoint. Consequently, α_j consists of dp_j linearly independent vectors in C_{v_j}, which has dimension dp_j. We conclude that α_j is a basis for C_{v_j}. ∎

Thus we may replace β_{v_j} by α_j as a basis for C_{v_j}. We do this for each j to obtain a subset $\alpha = \alpha_1 \cup \alpha_2 \cup \cdots \cup \alpha_k$ of K_ϕ.

Lemma 2. *α is a Jordan canonical basis for K_ϕ.*

Proof. Since $\beta_{v_1} \cup \beta_{v_2} \cup \cdots \cup \beta_{v_k}$ is a basis for K_ϕ, and since $\text{span}(\alpha_i) = \text{span}(\beta_{v_i}) = \mathsf{C}_{v_i}$, Exercise 9 implies that α is a basis for K_ϕ. Because α is a union of cycles of generalized eigenvectors of U, we conclude that α is a Jordan canonical basis. ∎

We are now in a position to relate the dot diagram of T corresponding to $\phi(t)$ to the dot diagram of U, bearing in mind that in the first case we are considering a rational canonical form and in the second case we are considering a Jordan canonical form. For convenience, we designate the first diagram D_1, and the second diagram D_2. For each j, the presence of the T-cyclic basis β_{x_j} results in a column of p_j dots in D_1. By Lemma 1, this basis is

replaced by the union α_j of d cycles of generalized eigenvectors of U, each of length p_j, which becomes part of the Jordan canonical basis for U. In effect, α_j determines d columns each containing p_j dots in D_2. So each column in D_1 determines d columns in D_2 of the same length, and all columns in D_2 are obtained in this way. Alternatively, each row in D_2 has d times as many dots as the corresponding row in D_1. Since Theorem 7.10 (p. 493) gives us the number of dots in any row of D_2, we may divide the appropriate expression in this theorem by d to obtain the number of dots in the corresponding row of D_1. Thus we have the following result.

Theorem 7.24. *Let T be a linear operator on a finite-dimensional vector space V, let $\phi(t)$ be an irreducible monic divisor of the characteristic polynomial of T of degree d, and let r_i denote the number of dots in the ith row of the dot diagram for $\phi(t)$ with respect to a rational canonical basis for T. Then*

(a) $r_1 = \dfrac{1}{d}[\dim(\mathsf{V}) - \operatorname{rank}(\phi(\mathsf{T}))]$;

(b) $r_i = \dfrac{1}{d}[\operatorname{rank}((\phi(\mathsf{T}))^{i-1}) - \operatorname{rank}((\phi(\mathsf{T}))^i)]$ *for $i > 1$.*

Thus the dot diagrams associated with a rational canonical form of an operator are completely determined by the operator. Since the rational canonical form is completely determined by its dot diagrams, we have the following uniqueness condition.

Corollary. *Under the conventions described earlier, the rational canonical form of a linear operator is unique up to the arrangement of the irreducible monic divisors of the characteristic polynomial.*

Since the rational canonical form of a linear operator is unique, the polynomials corresponding to the companion matrices that determine this form are also unique. These polynomials, which are powers of the irreducible monic divisors, are called the **elementary divisors** of the linear operator. Since a companion matrix may occur more than once in a rational canonical form, the same is true for the elementary divisors. We call the number of such occurrences the **multiplicity** of the elementary divisor.

Conversely, the elementary divisors and their multiplicities determine the companion matrices and, therefore, the rational canonical form of a linear operator.

Example 4

Let

$$\beta = \{e^x \cos 2x, e^x \sin 2x, xe^x \cos 2x, xe^x \sin 2x\}$$

be viewed as a subset of $\mathcal{F}(R, R)$, the space of all real-valued functions defined on R, and let $\mathsf{V} = \operatorname{span}(\beta)$. Then V is a four-dimensional subspace of $\mathcal{F}(R, R)$,

and β is an ordered basis for V. Let D be the linear operator on V defined by $\mathsf{D}(y) = y'$, the derivative of y, and let $A = [\mathsf{D}]_\beta$. Then

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

and the characteristic polynomial of D, and hence of A, is

$$f(t) = (t^2 - 2t + 5)^2.$$

Thus $\phi(t) = t^2 - 2t + 5$ is the only irreducible monic divisor of $f(t)$. Since $\phi(t)$ has degree 2 and V is four-dimensional, the dot diagram for $\phi(t)$ contains only two dots. Therefore the dot diagram is determined by r_1, the number of dots in the first row. Because ranks are preserved under matrix representations, we can use A in place of D in the formula given in Theorem 7.24. Now

$$\phi(A) = \begin{pmatrix} 0 & 0 & 0 & 4 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and so

$$r_1 = \tfrac{1}{2}[4 - \operatorname{rank}(\phi(A))] = \tfrac{1}{2}[4 - 2] = 1.$$

It follows that the second dot lies in the second row, and the dot diagram is as follows:

•
•

Hence V is a D-cyclic space generated by a single function with D-annihilator $(\phi(t))^2$. Furthermore, its rational canonical form is given by the companion matrix of $(\phi(t))^2 = t^4 - 4t^3 + 14t^2 - 20t + 25$, which is

$$\begin{pmatrix} 0 & 0 & 0 & -25 \\ 1 & 0 & 0 & 20 \\ 0 & 1 & 0 & -14 \\ 0 & 0 & 1 & 4 \end{pmatrix}.$$

Thus $(\phi(t))^2$ is the only elementary divisor of D, and it has multiplicity 1. For the cyclic generator, it suffices to find a function g in V for which $\phi(\mathsf{D})(g) \neq 0$. Since $\phi(A)(e_3) \neq 0$, it follows that $\phi(\mathsf{D})(xe^x \cos 2x) \neq 0$; therefore $g(x) = xe^x \cos 2x$ can be chosen as the cyclic generator. Hence

$$\beta_g = \{xe^x \cos 2x, \mathsf{D}(xe^x \cos 2x), \mathsf{D}^2(xe^x \cos 2x), \mathsf{D}^3(xe^x \cos 2x)\}$$

is a rational canonical basis for D. Notice that the function h defined by $h(x) = xe^x \sin 2x$ can be chosen in place of g. This shows that the rational canonical basis is not unique. ♦

It is convenient to refer to the rational canonical form and elementary divisors of a matrix, which are defined in the obvious way.

Definitions. *Let $A \in \mathsf{M}_{n\times n}(F)$. The* ***rational canonical form*** *of A is defined to be the rational canonical form of L_A. Likewise, for A, the* ***elementary divisors*** *and their* ***multiplicities*** *are the same as those of L_A.*

Let A be an $n \times n$ matrix, let C be a rational canonical form of A, and let β be the appropriate rational canonical basis for L_A. Then $C = [\mathsf{L}_A]_\beta$, and therefore A is similar to C. In fact, if Q is the matrix whose columns are the vectors of β in the same order, then $Q^{-1}AQ = C$.

Example 5

For the following real matrix A, we find the rational canonical form C of A and a matrix Q such that $Q^{-1}AQ = C$.

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

The characteristic polynomial of A is $f(t) = -(t^2+2)^2(t-2)$; therefore $\phi_1(t) = t^2+2$ and $\phi_2(t) = t-2$ are the distinct irreducible monic divisors of $f(t)$. By Theorem 7.23, $\dim(\mathsf{K}_{\phi_1}) = 4$ and $\dim(\mathsf{K}_{\phi_2}) = 1$. Since the degree of $\phi_1(t)$ is 2, the total number of dots in the dot diagram of $\phi_1(t)$ is $4/2 = 2$, and the number of dots r_1 in the first row is given by

$$\begin{aligned} r_1 &= \tfrac{1}{2}[\dim(\mathsf{R}^5) - \operatorname{rank}(\phi_1(A))] \\ &= \tfrac{1}{2}[5 - \operatorname{rank}(A^2 + 2I)] \\ &= \tfrac{1}{2}[5-1] = 2. \end{aligned}$$

Thus the dot diagram of $\phi_1(t)$ is

$$\bullet \quad \bullet$$

and each column contributes the companion matrix

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

for $\phi_1(t) = t^2+2$ to the rational canonical form C. Consequently $\phi_1(t)$ is an elementary divisor with multiplicity 2. Since $\dim(\mathsf{K}_{\phi_2}) = 1$, the dot diagram of $\phi_2(t) = t-2$ consists of a single dot, which contributes the 1×1 matrix

(2). Hence $\phi_2(t)$ is an elementary divisor with multiplicity 1. Therefore the rational canonical form C is

$$C = \left(\begin{array}{cc|cc|c} 0 & -2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 2 \end{array}\right).$$

We can infer from the dot diagram of $\phi_1(t)$ that if β is a rational canonical basis for L_A, then $\beta \cap \mathsf{K}_{\phi_1}$ is the union of two cyclic bases β_{v_1} and β_{v_2}, where v_1 and v_2 each have annihilator $\phi_1(t)$. It follows that both v_1 and v_2 lie in $\mathsf{N}(\phi_1(\mathsf{L}_A))$. It can be shown that

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 1 \end{pmatrix} \right\}$$

is a basis for $\mathsf{N}(\phi_1(\mathsf{L}_A))$. Setting $v_1 = e_1$, we see that

$$Av_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Next choose v_2 in $\mathsf{K}_{\phi_1} = \mathsf{N}(\phi(\mathsf{L}_A))$, but not in the span of $\beta_{v_1} = \{v_1, Av_1\}$. For example, $v_2 = e_2$. Then it can be seen that

$$Av_2 = \begin{pmatrix} 2 \\ -2 \\ 0 \\ -2 \\ -4 \end{pmatrix},$$

and $\beta_{v_1} \cup \beta_{v_2}$ is a basis for K_{ϕ_1}.

Since the dot diagram of $\phi_2(t) = t-2$ consists of a single dot, any nonzero vector in K_{ϕ_2} is an eigenvector of A corresponding to the eigenvalue $\lambda = 2$. For example, choose

$$v_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 2 \end{pmatrix}.$$

By Theorem 7.23, $\beta = \{v_1, Av_1, v_2, Av_2, v_3\}$ is a rational canonical basis for L_A. So setting

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

we have $Q^{-1}AQ = C$. ◆

Example 6

For the following matrix A, we find the rational canonical form C and a matrix Q such that $Q^{-1}AQ = C$.

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

Since the characteristic polynomial of A is $f(t) = (t-2)^4$, the only irreducible monic divisor of $f(t)$ is $\phi(t) = t - 2$, and so $\mathsf{K}_\phi = \mathsf{R}^4$. In this case, $\phi(t)$ has degree 1; hence in applying Theorem 7.24 to compute the dot diagram for $\phi(t)$, we obtain

$$r_1 = 4 - \text{rank}(\phi(A)) = 4 - 2 = 2,$$
$$r_2 = \text{rank}(\phi(A)) - \text{rank}((\phi(A))^2) = 2 - 1 = 1,$$

and

$$r_3 = \text{rank}((\phi(A))^2) - \text{rank}((\phi(A))^3) = 1 - 0 = 1,$$

where r_i is the number of dots in the ith row of the dot diagram. Since there are $\dim(\mathsf{R}^4) = 4$ dots in the diagram, we may terminate these computations with r_3. Thus the dot diagram for A is

Since $(t-2)^3$ has the companion matrix

$$\begin{pmatrix} 0 & 0 & 8 \\ 1 & 0 & -12 \\ 0 & 1 & 6 \end{pmatrix}$$

and $(t-2)$ has the companion matrix $\begin{pmatrix}2\end{pmatrix}$, the rational canonical form of A is given by

$$C = \left(\begin{array}{ccc|c} 0 & 0 & 8 & 0 \\ 1 & 0 & -12 & 0 \\ 0 & 1 & 6 & 0 \\ \hline 0 & 0 & 0 & 2 \end{array}\right).$$

Next we find a rational canonical basis for L_A. The preceding dot diagram indicates that there are two vectors v_1 and v_2 in R^4 with annihilators $(\phi(t))^3$ and $\phi(t)$, respectively, and such that

$$\beta = \beta_{v_1} \cup \beta_{v_2} = \{v_1, Av_1, A^2v_1, v_2\}$$

is a rational canonical basis for L_A. Furthermore, $v_1 \notin \mathsf{N}((\mathsf{L}_A - 2\mathsf{I})^2)$, and $v_2 \in \mathsf{N}(\mathsf{L}_A - 2\mathsf{I})$. It can easily be shown that

$$\mathsf{N}(\mathsf{L}_A - 2\mathsf{I}) = \operatorname{span}(\{e_1, e_4\})$$

and

$$\mathsf{N}((\mathsf{L}_A - 2\mathsf{I})^2) = \operatorname{span}(\{e_1, e_2, e_4\}).$$

The standard vector e_3 meets the criteria for v_1; so we set $v_1 = e_3$. It follows that

$$Av_1 = \begin{pmatrix}0\\1\\2\\0\end{pmatrix} \quad \text{and} \quad A^2v_1 = \begin{pmatrix}1\\4\\4\\0\end{pmatrix}.$$

Next we choose a vector $v_2 \in \mathsf{N}(\mathsf{L}_A - 2\mathsf{I})$ that is not in the span of β_{v_1}. Clearly, $v_2 = e_4$ satisfies this condition. Thus

$$\left\{\begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\1\\2\\0\end{pmatrix}, \begin{pmatrix}1\\4\\4\\0\end{pmatrix}, \begin{pmatrix}0\\0\\0\\1\end{pmatrix}\right\}$$

is a rational canonical basis for L_A.

Finally, let Q be the matrix whose columns are the vectors of β in the same order:

$$Q = \begin{pmatrix}0 & 0 & 1 & 0\\ 0 & 1 & 4 & 0\\ 1 & 2 & 4 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}.$$

Then $C = Q^{-1}AQ$. ♦

Direct Sums*

The next theorem is a simple consequence of Theorem 7.23.

Theorem 7.25 (Primary Decomposition Theorem). *Let* T *be a linear operator on an* n*-dimensional vector space* V *with characteristic polynomial*

$$f(t) = (-1)^n(\phi_1(t))^{n_1}(\phi_2(t))^{n_2}\cdots(\phi_k(t))^{n_k},$$

where the $\phi_i(t)$*'s* $(1 \le i \le k)$ *are distinct irreducible monic polynomials and the* n_i*'s are positive integers. Then the following statements are true.*

(a) $\mathsf{V} = \mathsf{K}_{\phi_1} \oplus \mathsf{K}_{\phi_2} \oplus \cdots \oplus \mathsf{K}_{\phi_k}$.

(b) *If* T_i $(1 \le i \le k)$ *is the restriction of* T *to* K_{ϕ_i} *and* C_i *is the rational canonical form of* T_i*, then* $C_1 \oplus C_2 \oplus \cdots \oplus C_k$ *is the rational canonical form of* T.

Proof. Exercise. ∎

The next theorem is a simple consequence of Theorem 7.17.

Theorem 7.26. *Let* T *be a linear operator on a finite-dimensional vector space* V*. Then* V *is a direct sum of* T*-cyclic subspaces* C_{v_i}*, where each* v_i *lies in* K_ϕ *for some irreducible monic divisor* $\phi(t)$ *of the characteristic polynomial of* T.

Proof. Exercise. ∎

EXERCISES

1. Label the following statements as true or false.

(a) Every rational canonical basis for a linear operator T is the union of T-cyclic bases.

(b) If a basis is the union of T-cyclic bases for a linear operator T, then it is a rational canonical basis for T.

(c) There exist square matrices having no rational canonical form.

(d) A square matrix is similar to its rational canonical form.

(e) For any linear operator T on a finite-dimensional vector space, any irreducible factor of the characteristic polynomial of T divides the minimal polynomial of T.

(f) Let $\phi(t)$ be an irreducible monic divisor of the characteristic polynomial of a linear operator T. The dots in the diagram used to compute the rational canonical form of the restriction of T to K_ϕ are in one-to-one correspondence with the vectors in a basis for K_ϕ.

(g) If a matrix has a Jordan canonical form, then its Jordan canonical form and rational canonical form are similar.

2. For each of the following matrices $A \in \mathsf{M}_{n\times n}(F)$, find the rational canonical form C of A and a matrix $Q \in \mathsf{M}_{n\times n}(F)$ such that $Q^{-1}AQ = C$.

(a) $A = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix}$ $F = R$

(b) $A = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$ $F = R$

(c) $A = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$ $F = C$

(d) $A = \begin{pmatrix} 0 & -7 & 14 & -6 \\ 1 & -4 & 6 & -3 \\ 0 & -4 & 9 & -4 \\ 0 & -4 & 11 & -5 \end{pmatrix}$ $F = R$

(e) $A = \begin{pmatrix} 0 & -4 & 12 & -7 \\ 1 & -1 & 3 & -3 \\ 0 & -1 & 6 & -4 \\ 0 & -1 & 8 & -5 \end{pmatrix}$ $F = R$

3. For each of the following linear operators T, find the elementary divisors, the rational canonical form C, and a rational canonical basis β.

(a) T is the linear operator on $\mathsf{P}_3(R)$ defined by

$$\mathsf{T}(f(x)) = f(0)x - f'(1).$$

(b) Let $S = \{\sin x, \cos x, x\sin x, x\cos x\}$, a subset of $\mathcal{F}(R, R)$, and let $\mathsf{V} = \operatorname{span}(S)$. Define T to be the linear operator on V such that

$$\mathsf{T}(f) = f'.$$

(c) T is the linear operator on $\mathsf{M}_{2\times 2}(R)$ defined by

$$\mathsf{T}(A) = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix} \cdot A.$$

(d) Let $S = \{\sin x \sin y, \sin x \cos y, \cos x \sin y, \cos x \cos y\}$, a subset of $\mathcal{F}(R \times R, R)$, and let $\mathsf{V} = \operatorname{span}(S)$. Define T to be the linear operator on V such that

$$\mathsf{T}(f)(x, y) = \frac{\partial f(x, y)}{\partial x} + \frac{\partial f(x, y)}{\partial y}.$$

4. Let T be a linear operator on a finite-dimensional vector space V with minimal polynomial $(\phi(t))^m$ for some positive integer m.

(a) Prove that $\mathsf{R}(\phi(\mathsf{T})) \subseteq \mathsf{N}((\phi(\mathsf{T}))^{m-1})$.

(b) Give an example to show that the subspaces in (a) need not be equal.

(c) Prove that the minimal polynomial of the restriction of T to $\mathsf{R}(\phi(\mathsf{T}))$ equals $(\phi(t))^{m-1}$.

5. Let T be a linear operator on a finite-dimensional vector space. Prove that the rational canonical form of T is a diagonal matrix if and only if T is diagonalizable. Visit `goo.gl/tK8pru` for a solution.

6. Let T be a linear operator on a finite-dimensional vector space V with characteristic polynomial $f(t) = (-1)^n \phi_1(t)\phi_2(t)$, where $\phi_1(t)$ and $\phi_2(t)$ are distinct irreducible monic polynomials and $n = \dim(\mathsf{V})$.

(a) Prove that there exist $v_1, v_2 \in \mathsf{V}$ such that v_1 has T-annihilator $\phi_1(t)$, v_2 has T-annihilator $\phi_2(t)$, and $\beta_{v_1} \cup \beta_{v_2}$ is a basis for V.

(b) Prove that there is a vector $v_3 \in \mathsf{V}$ with T-annihilator $\phi_1(t)\phi_2(t)$ such that β_{v_3} is a basis for V.

(c) Describe the difference between the matrix representation of T with respect to $\beta_{v_1} \cup \beta_{v_2}$ and the matrix representation of T with respect to β_{v_3}.

Thus, to assure the uniqueness of the rational canonical form, we require that the generators of the T-cyclic bases that constitute a rational canonical basis have T-annihilators equal to powers of irreducible monic factors of the characteristic polynomial of T.

7. Let T be a linear operator on a finite-dimensional vector space with minimal polynomial

$$f(t) = (\phi_1(t))^{m_1}(\phi_2(t))^{m_2} \cdots (\phi_k(t))^{m_k},$$

where the $\phi_i(t)$'s are distinct irreducible monic factors of $f(t)$. Prove that for each i, m_i is the number of entries in the first column of the dot diagram for $\phi_i(t)$.

8. Let T be a linear operator on a finite-dimensional vector space V. Prove that for any irreducible polynomial $\phi(t)$, if $\phi(\mathsf{T})$ is not one-to-one, then $\phi(t)$ divides the characteristic polynomial of T. *Hint:* Apply Exercise 15 of Section 7.3.

9. Let V be a vector space and $\beta_1, \beta_2, \ldots, \beta_k$ be disjoint subsets of V whose union is a basis for V. Now suppose that $\gamma_1, \gamma_2, \ldots, \gamma_k$ are linearly independent subsets of V such that $\text{span}(\gamma_i) = \text{span}(\beta_i)$ for all i. Prove that $\gamma_1 \cup \gamma_2 \cup \cdots \cup \gamma_k$ is also a basis for V.

10. Let T be a linear operator on a finite-dimensional vector space, and suppose that $\phi(t)$ is an irreducible monic factor of the characteristic polynomial of T. Prove that if $\phi(t)$ is the T-annihilator of vectors x and y, then $x \in \mathsf{C}_y$ if and only if $\mathsf{C}_x = \mathsf{C}_y$.

Exercises 11 and 12 are concerned with direct sums.

11. Prove Theorem 7.25.

12. Prove Theorem 7.26.

INDEX OF DEFINITIONS FOR CHAPTER 7

Appendices

APPENDIX A SETS

A **set**[1] is a collection of objects, called **elements** of the set. If x is an element of the set A, then we write $x \in A$; otherwise, we write $x \notin A$. For example, if Z is the set of integers, then $3 \in Z$ and $\frac{1}{2} \notin Z$.

One set that appears frequently is the set of real numbers, which we denote by R throughout this text.

Two sets A and B are called **equal**, written $A = B$, if they contain exactly the same elements. Sets may be described in one of two ways:

1. By listing the elements of the set between set braces { }.
2. By describing the elements of the set in terms of some characteristic property.

For example, the set consisting of the elements 1, 2, 3, and 4 can be written as $\{1, 2, 3, 4\}$ or as

$$\{x\colon x \text{ is a positive integer less than } 5\}.$$

Note that the order in which the elements of a set are listed is immaterial, and listing an element in a set more than once does not change the set. Hence

$$\{1, 2, 3, 4\} = \{3, 1, 2, 4\} = \{1, 3, 1, 4, 2\}.$$

Example 1

Let A denote the set of real numbers strictly between 1 and 2. Then A may be written as

$$A = \{x \in R\colon 1 < x < 2\}. \quad \blacklozenge$$

A set B is called a **subset** of a set A, written $B \subseteq A$ or $A \supseteq B$, if every element of B is an element of A. For example, $\{1, 2, 6\} \subseteq \{2, 8, 7, 6, 1\}$. If $B \subseteq A$, and $B \neq A$, then B is called a **proper subset** of A. Observe that $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$, a fact that is often used to prove that two sets are equal.

The **empty set**, denoted by $\varnothing$, is the set containing no elements. The empty set is a subset of every set.

[1]It is impossible to give a formal definition of the word *set* in terms of simpler concepts. In this appendix we explain the conceptual ideas behind sets and the facts about them that we will use in this book.

Sets may be combined to form other sets in two basic ways. The **union** of two sets A and B, denoted $A \cup B$, is the set of elements that are in A, or B, or both; that is,

$$A \cup B = \{x: x \in A \text{ or } x \in B\}.$$

The **intersection** of two sets A and B, denoted $A \cap B$, is the set of elements that are in both A and B; that is,

$$A \cap B = \{x: x \in A \text{ and } x \in B\}.$$

Two sets are called **disjoint** if their intersection equals the empty set.

Example 2

Let $A = \{1, 3, 5\}$ and $B = \{1, 5, 7, 8\}$. Then

$$A \cup B = \{1, 3, 5, 7, 8\} \quad \text{and} \quad A \cap B = \{1, 5\}.$$

Likewise, if $X = \{1, 2, 8\}$ and $Y = \{3, 4, 5\}$, then

$$X \cup Y = \{1, 2, 3, 4, 5, 8\} \quad \text{and} \quad X \cap Y = \varnothing.$$

Thus X and Y are disjoint sets. ◆

The union and intersection of more than two sets can be defined analogously. Specifically, if $A_1, A_2, \ldots, A_n$ are sets, then the union and intersections of these sets are defined, respectively, by

$$\bigcup_{i=1}^{n} A_i = \{x: x \in A_i \text{ for some } i = 1, 2, \ldots, n\}$$

and

$$\bigcap_{i=1}^{n} A_i = \{x: x \in A_i \text{ for all } i = 1, 2, \ldots, n\}.$$

In this situation, we have one set A_i for each $i \in \{1, 2, \ldots, n\}$. It can be convenient to regard these sets A_i as being *indexed* by the set $I = \{1, 2, \ldots, n\}$. When doing so, I is called an **index set** for the collection $\{A_i : i \in I\}$.

The use of an index set is especially useful with an infinite collection of sets, as in Example 3 below. If Λ is an index set and $\{A_\alpha: \alpha \in \Lambda\}$ is a collection of sets, the union and intersection of these sets are defined, respectively, by

$$\bigcup_{\alpha \in \Lambda} A_\alpha = \{x: x \in A_\alpha \text{ for some } \alpha \in \Lambda\}$$

and

$$\bigcap_{\alpha \in \Lambda} A_\alpha = \{x: x \in A_\alpha \text{ for all } \alpha \in \Lambda\}.$$

Example 3

Let $\Lambda = \{\alpha \in R\colon \alpha > 1\}$, and let

$$A_\alpha = \left\{x \in R\colon \frac{-1}{\alpha} \le x \le 1 + \alpha\right\}$$

for each $\alpha \in \Lambda$. Then

$$\bigcup_{\alpha\in\Lambda} A_\alpha = \{x \in R\colon x > -1\} \quad \text{and} \quad \bigcap_{\alpha\in\Lambda} A_\alpha = \{x \in R\colon 0 \le x \le 2\}. \quad \blacklozenge$$

By a relation on a set A, we mean a rule for determining whether or not, for any elements x and y in A, x stands in a given relationship to y. To make this concept precise, we define a **relation** on A to be any set S of ordered pairs of elements of A, and say that the elements x and y of A satisfy the relation S if and only if $(x, y) \in S$. On the set of real numbers, for instance, "is equal to," "is less than," and "is greater than or equal to" are familiar relations. If S is a relation on a set A, we often write $x \sim y$ in place of $(x, y) \in S$.

A relation S on a set A is called an **equivalence relation** on A if the following three conditions hold:

1. For each $x \in A$, $x \sim x$ *(reflexivity).*
2. If $x \sim y$, then $y \sim x$ *(symmetry).*
3. If $x \sim y$ and $y \sim z$, then $x \sim z$ *(transitivity).*

For example, if we define $x \sim y$ to mean that $x - y$ is divisible by a fixed integer n, then $\sim$ is an equivalence relation on the set of integers.

APPENDIX B FUNCTIONS

If A and B are sets, then a **function** f from A to B, written $f\colon A \to B$, is a rule that associates to each element x in A a unique element denoted $f(x)$ in B. The element $f(x)$ is called the **image** of x (under f), and x is called a **preimage** of $f(x)$ (under f). If $f\colon A \to B$, then A is called the **domain** of f, B is called the **codomain** of f, and the set $\{f(x)\colon x \in A\}$ is called the **range** of f. Note that the range of f is a subset of B. If $S \subseteq A$, we denote by $f(S)$ the set $\{f(x)\colon x \in S\}$ of all images of elements of S. Likewise, if $T \subseteq B$, we denote by $f^{-1}(T)$ the set $\{x \in A\colon f(x) \in T\}$ of all preimages of elements in T. Finally, two functions $f\colon A \to B$ and $g\colon A \to B$ are **equal**, written $f = g$, if $f(x) = g(x)$ for all $x \in A$.

Example 1

Suppose that $A = [-10, 10]$. Let $f\colon A \to R$ be the function that assigns to each element x in A the element $x^2 + 1$ in R; that is, f is defined by

$f(x) = x^2 + 1$. Then A is the domain of f, R is the codomain of f, and $[1, 101]$ is the range of f. Since $f(2) = 5$, the image of 2 is 5, and 2 is a preimage of 5. Notice that -2 is another preimage of 5. Moreover, if $S = [1, 2]$ and $T = [82, 101]$, then $f(S) = [2, 5]$ and $f^{-1}(T) = [-10, -9] \cup [9, 10]$. ♦

As Example 1 shows, the preimage of an element in the range need not be unique. Functions such that each element of the range has a unique preimage are called **one-to-one**; that is $f: A \to B$ is one-to-one if $f(x) = f(y)$ implies $x = y$ or, equivalently, if $x \neq y$ implies $f(x) \neq f(y)$.

If $f: A \to B$ is a function with range B, that is, if $f(A) = B$, then f is called **onto**. So f is onto if and only if the range of f equals the codomain of f.

Let $f: A \to B$ be a function and $S \subseteq A$. Then a function $f_S: S \to B$, called the **restriction** of f to S, can be formed by defining $f_S(x) = f(x)$ for each $x \in S$.

The next example illustrates these concepts.

Example 2

Let $f: [-1, 1] \to [0, 1]$ be defined by $f(x) = x^2$. This function is onto, but not one-to-one since $f(-1) = f(1) = 1$. Note that if $S = [0, 1]$, then f_S is both onto and one-to-one. Finally, if $T = [\frac{1}{2}, 1]$, then f_T is one-to-one, but not onto. ♦

Let A, B, and C be sets and $f: A \to B$ and $g: B \to C$ be functions. By following f with g, we obtain a function $g \circ f: A \to C$ called the **composite** of g and f. Thus $(g \circ f)(x) = g(f(x))$ for all $x \in A$. For example, let $A = B = C = R$, $f(x) = \sin x$, and $g(x) = x^2 + 3$. Then $(g \circ f)(x) = (g(f(x)) = \sin^2 x + 3$, whereas $(f \circ g)(x) = f(g(x)) = \sin(x^2 + 3)$. Hence, $g \circ f \neq f \circ g$. Functional composition is associative, however; that is, if $h: C \to D$ is another function, then $h \circ (g \circ f) = (h \circ g) \circ f$.

A function $f: A \to B$ is said to be **invertible** if there exists a function $g: B \to A$ such that $(f \circ g)(y) = y$ for all $y \in B$ and $(g \circ f)(x) = x$ for all $x \in A$. If such a function g exists, then it is unique and is called the **inverse** of f. We denote the inverse of f (when it exists) by f^{-1}. It can be shown that f is invertible if and only if f is both one-to-one and onto.

Example 3

The function $f: R \to R$ defined by $f(x) = 3x + 1$ is one-to-one and onto; hence f is invertible. The inverse of f is the function $f^{-1}: R \to R$ defined by $f^{-1}(x) = (x - 1)/3$. ♦

The following facts about invertible functions are easily proved.

1. If $f: A \to B$ is invertible, then f^{-1} is invertible, and $(f^{-1})^{-1} = f$.

2. If $f: A \to B$ and $g: B \to C$ are invertible, then $g \circ f$ is invertible, and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

APPENDIX C FIELDS

The set of real numbers is an example of an algebraic structure called a *field*. Basically, a field is a set in which four operations (called addition, multiplication, subtraction, and division) can be defined so that, with the exception of division by zero, the sum, product, difference, and quotient of any two elements in the set is an element of the set. More precisely, a field is defined as follows.

Definitions. *A field F is a set on which two operations $+$ and $\cdot$ (called* ***addition*** *and* ***multiplication****, respectively) are defined so that, for each pair of elements x, y in F, there are unique elements in F, denoted $x+y$ and $x \cdot y$, and such that the following conditions hold for all elements a, b, c in F.*

(F 1) $a+b=b+a$ and $a \cdot b = b \cdot a$
(commutativity of addition and multiplication)

(F 2) $(a+b)+c = a+(b+c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
(associativity of addition and multiplication)

(F 3) *There exist distinct elements* 0 *and* 1 *in F such that*

$$0+a=a \quad \text{and} \quad 1 \cdot a = a$$

(existence of identity elements for addition and multiplication)

(F 4) *For each element a in F and each nonzero element b in F, there exist elements c and d in F such that*

$$a+c=0 \quad \text{and} \quad b \cdot d = 1$$

(existence of inverses for addition and multiplication)

(F 5) $a \cdot (b+c) = a \cdot b + a \cdot c$
(distributivity of multiplication over addition)

The elements $x+y$ and $x \cdot y$ are called the ***sum*** *and* ***product****, respectively, of x and y. The elements* 0 *(read "***zero***") and* 1 *(read "***one***") mentioned in* (F 3) *are called* ***identity elements*** *for addition and multiplication, respectively, and the elements c and d referred to in* (F 4) *are called an* ***additive inverse*** *for a and a* ***multiplicative inverse*** *for b, respectively.*

Example 1

The set of real numbers R with the usual definitions of addition and multiplication is a field. ♦

Example 2

The set of rational numbers with the usual definitions of addition and multiplication is a field. ◆

Example 3

The set of all real numbers of the form $a + b\sqrt{2}$, where a and b are rational numbers, with addition and multiplication as in R is a field. ◆

Example 4

The field Z_2 consists of two elements 0 and 1 with the operations of addition and multiplication defined by the equations

$$0+0=0, \quad 0+1=1+0=1, \quad 1+1=0,$$
$$0\cdot 0=0, \quad 0\cdot 1=1\cdot 0=0, \quad \text{and} \quad 1\cdot 1=1. \quad ◆$$

Example 5

Neither the set of positive integers nor the set of integers with the usual definitions of addition and multiplication is a field, for in either case (F 4) does not hold. ◆

The identity and inverse elements guaranteed by (F 3) and (F 4) are unique; this is a consequence of the following theorem.

Theorem C.1 (Cancellation Laws). *For arbitrary elements a, b, and c in a field, the following statements are true.*

(a) *If $a + b = c + b$, then $a = c$.*

(b) *If $a \cdot b = c \cdot b$ and $b \neq 0$, then $a = c$.*

Proof. (a) The proof of (a) is left as an exercise.

(b) If $b \neq 0$, then (F 4) guarantees the existence of an element d in the field such that $b \cdot d = 1$. Multiply both sides of the equality $a \cdot b = c \cdot b$ by d to obtain $(a \cdot b) \cdot d = (c \cdot b) \cdot d$. Consider the left side of this equality: By (F 2) and (F 3), we have

$$(a \cdot b) \cdot d = a \cdot (b \cdot d) = a \cdot 1 = a.$$

Similarly, the right side of the equality reduces to c. Thus $a = c$. ∎

Corollary. *The elements 0 and 1 mentioned in* (F 3), *and the elements c and d mentioned in* (F 4), *are unique.*

Proof. Suppose that $0' \in F$ satisfies $0' + a = a$ for each $a \in F$. Since $0 + a = a$ for each $a \in F$, we have $0' + a = 0 + a$ for each $a \in F$. Thus $0' = 0$ by Theorem C.1.

The proofs of the remaining parts are similar. ∎

Thus each element b in a field has a unique additive inverse and, if $b \neq 0$, a unique multiplicative inverse. (It is shown in the corollary to Theorem C.2 that 0 has no multiplicative inverse.) The additive inverse and the multiplicative inverse of b are denoted by $-b$ and b^{-1}, respectively. Note that $-(-b) = b$ and $(b^{-1})^{-1} = b$.

Subtraction and *division* can be defined in terms of addition and multiplication by using the additive and multiplicative inverses. Specifically, subtraction of b is defined to be addition of $-b$ and division by $b \neq 0$ is defined to be multiplication by b^{-1}; that is,

$$a - b = a + (-b) \quad \text{and} \quad \frac{a}{b} = a \cdot b^{-1}.$$

In particular, the symbol $\frac{1}{b}$ denotes b^{-1}. Division by zero is undefined, but, with this exception, the sum, product, difference, and quotient of any two elements of a field are defined.

Many of the familiar properties of multiplication of real numbers are true in any field, as the next theorem shows.

Theorem C.2. *Let a and b be arbitrary elements of a field. Then each of the following statements is true.*

(a) $a \cdot 0 = 0$.
(b) $(-a) \cdot b = a \cdot (-b) = -(a \cdot b)$.
(c) $(-a) \cdot (-b) = a \cdot b$.

Proof. (a) Since $0 + 0 = 0$, (F 5) shows that

$$0 + a \cdot 0 = a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0.$$

Thus $0 = a \cdot 0$ by Theorem C.1.

(b) By definition, $-(a \cdot b)$ is the unique element of F with the property $a \cdot b + [-(a \cdot b)] = 0$. So in order to prove that $(-a) \cdot b = -(a \cdot b)$, it suffices to show that $a \cdot b + (-a) \cdot b = 0$. But $-a$ is the element of F such that $a + (-a) = 0$; so

$$a \cdot b + (-a) \cdot b = [a + (-a)] \cdot b = 0 \cdot b = b \cdot 0 = 0$$

by (F 5) and (a). Thus $(-a) \cdot b = -(a \cdot b)$. The proof that $a \cdot (-b) = -(a \cdot b)$ is similar.

(c) By applying (b) twice, we find that

$$(-a) \cdot (-b) = -[a \cdot (-b)] = -[-(a \cdot b)] = a \cdot b. \qquad \blacksquare$$

Corollary. *The additive identity of a field has no multiplicative inverse.*

In an arbitrary field F, it may happen that a sum $1+1+\cdots+1$ (p summands) equals 0 for some positive integer p. For example, in the field Z_2 (defined in Example 4), $1+1=0$. In this case, the smallest positive integer p for which a sum of p 1's equals 0 is called the **characteristic** of F; if no such positive integer exists, then F is said to have **characteristic zero**. Thus Z_2 has characteristic two, and R has characteristic zero. Observe that if F is a field of characteristic $p \neq 0$, then $x+x+\cdots+x$ (p summands) equals 0 for all $x \in F$. In a field having nonzero characteristic (especially characteristic two), many unexpected problems arise. For this reason, some of the results about vector spaces stated in this book require that the field over which the vector space is defined be of characteristic zero (or, at least, of some characteristic other than two).

Finally, note that in other sections of this book, the product of two elements a and b in a field is usually denoted ab rather than $a \cdot b$.

APPENDIX D COMPLEX NUMBERS

For the purposes of algebra, the field of real numbers is not sufficient, for there are polynomials of nonzero degree with real coefficients that have no zeros in the field of real numbers (for example, x^2+1). It is often desirable to have a field in which any polynomial of nonzero degree with coefficients from that field has a zero in that field. It is possible to "enlarge" the field of real numbers to obtain such a field.

Definitions. *A **complex number** is an expression of the form $z = a+bi$, where a and b are real numbers called the **real part** and the **imaginary part** of z, respectively.*

*The **sum** and **product** of two complex numbers $z = a+bi$ and $w = c+di$ (where a, b, c, and d are real numbers) are defined, respectively, as follows:*

$$z + w = (a + bi) + (c + di) = (a + c) + (b + d)i$$

and

$$zw = (a + bi)(c + di) = (ac - bd) + (bc + ad)i.$$

Example 1

The sum and product of $z = 3 - 5i$ and $w = 9 + 7i$ are, respectively,

$$z + w = (3 - 5i) + (9 + 7i) = (3 + 9) + [(-5) + 7]i = 12 + 2i$$

and

$$zw = (3 - 5i)(9 + 7i) = [3 \cdot 9 - (-5) \cdot 7] + [(-5) \cdot 9 + 3 \cdot 7]i = 62 - 24i. \quad \blacklozenge$$

Any real number c may be regarded as a complex number by identifying c with the complex number $c + 0i$, that is, treating them as identical. Observe that this correspondence preserves sums and products; that is,

$$(c + 0i) + (d + 0i) = (c + d) + 0i \quad \text{and} \quad (c + 0i)(d + 0i) = cd + 0i.$$

Any complex number of the form $bi = 0 + bi$, where b is a nonzero real number, is called **imaginary**. The product of two imaginary numbers is real since

$$(bi)(di) = (0 + bi)(0 + di) = (0 - bd) + (b \cdot 0 + 0 \cdot d)i = -bd.$$

In particular, for $i = 0 + 1i$, we have $i \cdot i = -1$.

The observation that $i^2 = i \cdot i = -1$ provides an easy way to remember the definition of multiplication of complex numbers: simply multiply two complex numbers as you would any two algebraic expressions, and replace i^2 by -1. Example 2 illustrates this technique.

Example 2

The product of $-5 + 2i$ and $1 - 3i$ is

$$\begin{aligned}(-5 + 2i)(1 - 3i) &= -5(1 - 3i) + 2i(1 - 3i) \\ &= -5 + 15i + 2i - 6i^2 \\ &= -5 + 15i + 2i - 6(-1) \\ &= 1 + 17i. \quad \blacklozenge\end{aligned}$$

The real number 0, regarded as a complex number, is an additive identity element for the complex numbers since

$$(a + bi) + 0 = (a + bi) + (0 + 0i) = (a + 0) + (b + 0)i = a + bi.$$

Likewise the real number 1, regarded as a complex number, is a multiplicative identity element for the set of complex numbers since

$$(a + bi) \cdot 1 = (a + bi)(1 + 0i) = (a \cdot 1 - b \cdot 0) + (b \cdot 1 + a \cdot 0)i = a + bi.$$

Every complex number $a + bi$ has an additive inverse, namely $(-a) + (-b)i$. But also each complex number except 0 has a multiplicative inverse. In fact,

$$(a + bi)^{-1} = \left(\frac{a}{a^2 + b^2}\right) - \left(\frac{b}{a^2 + b^2}\right) i.$$

In view of the preceding statements, the following result is not surprising.

Theorem D.1. *The set of complex numbers with the operations of addition and multiplication previously defined is a field.*

Proof. Exercise. ∎

Definition. *The* ***(complex) conjugate*** *of a complex number $a + bi$ is the complex number $a - bi$. We denote the conjugate of the complex number z by $\overline{z}$.*

Example 3

The conjugates of $-3 + 2i$, $4 - 7i$, and 6 are, respectively,

$$\overline{-3 + 2i} = -3 - 2i, \quad \overline{4 - 7i} = 4 + 7i, \quad \text{and} \quad \overline{6} = \overline{6 + 0i} = 6 - 0i = 6. \quad \blacklozenge$$

The next theorem contains some important properties of the conjugate of a complex number.

Theorem D.2. *Let z and w be complex numbers. Then the following statements are true.*

(a) $\overline{\overline{z}} = z$.
(b) $\overline{(z + w)} = \overline{z} + \overline{w}$.
(c) $\overline{zw} = \overline{z} \cdot \overline{w}$.
(d) $\overline{\left(\frac{z}{w}\right)} = \frac{\overline{z}}{\overline{w}}$ *if* $w \neq 0$.
(e) z *is a real number if and only if* $\overline{z} = z$.

Proof. We leave the proofs of (a), (d), and (e) to the reader.
(b) Let $z = a + bi$ and $w = c + di$, where $a, b, c, d \in R$. Then

$$\begin{aligned} \overline{(z + w)} &= \overline{(a + c) + (b + d)i} = (a + c) - (b + d)i \\ &= (a - bi) + (c - di) = \overline{z} + \overline{w}. \end{aligned}$$

(c) For z and w, we have

$$\begin{aligned} \overline{zw} &= \overline{(a + bi)(c + di)} = \overline{(ac - bd) + (ad + bc)i} \\ &= (ac - bd) - (ad + bc)i = (a - bi)(c - di) = \overline{z} \cdot \overline{w}. \end{aligned}$$

∎

For any complex number $z = a + bi$, $z\overline{z}$ is real and nonnegative, for

$$z\overline{z} = (a + bi)(a - bi) = a^2 + b^2.$$

This fact can be used to define the absolute value of a complex number.

Definition. *Let $z = a + bi$, where $a, b \in R$. The* ***absolute value*** *(or* ***modulus****) of z is the real number $\sqrt{a^2 + b^2}$. We denote the absolute value of z by $|z|$.*

Observe that $z\overline{z} = |z|^2$. The fact that the product of a complex number and its conjugate is real provides an easy method for determining the quotient of two complex numbers; for if $c + di \neq 0$, then

$$\frac{a+bi}{c+di} = \frac{a+bi}{c+di} \cdot \frac{c-di}{c-di} = \frac{(ac+bd)+(bc-ad)i}{c^2+d^2} = \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}i.$$

Example 4

To illustrate this procedure, we compute the quotient $(1 + 4i)/(3 - 2i)$:

$$\frac{1+4i}{3-2i} = \frac{1+4i}{3-2i} \cdot \frac{3+2i}{3+2i} = \frac{-5+14i}{9+4} = -\frac{5}{13} + \frac{14}{13}i. \quad \blacklozenge$$

The absolute value of a complex number has the familiar properties of the absolute value of a real number, as the following result shows.

Theorem D.3. *Let z and w denote any two complex numbers. Then the following statements are true.*

(a) $|zw| = |z| \cdot |w|$.

(b) $\left|\frac{z}{w}\right| = \frac{|z|}{|w|}$ *if* $w \neq 0$.

(c) $|z + w| \leq |z| + |w|$.

(d) $|z| - |w| \leq |z + w|$.

Proof. (a) By Theorem D.2, we have

$$|zw|^2 = (zw)\overline{(zw)} = (zw)(\overline{z} \cdot \overline{w}) = (z\overline{z})(w\overline{w}) = |z|^2|w|^2,$$

proving (a).

(b) For the proof of (b), apply (a) to the product $\left(\frac{z}{w}\right) w$.

(c) For any complex number $x = a + bi$, where $a, b \in R$, observe that

$$x + \overline{x} = (a + bi) + (a - bi) = 2a \leq 2\sqrt{a^2 + b^2} = 2|x|.$$

Thus $x + \overline{x}$ is real and satisfies the inequality $x + \overline{x} \leq 2|x|$. Taking $x = w\overline{z}$, we have, by Theorem D.2 and (a),

$$w\overline{z} + \overline{w}z \leq 2|w\overline{z}| = 2|w||\overline{z}| = 2|z||w|.$$

Using Theorem D.2 again gives

$$\begin{aligned} |z + w|^2 &= (z + w)\overline{(z + w)} = (z + w)(\overline{z} + \overline{w}) = z\overline{z} + w\overline{z} + z\overline{w} + w\overline{w} \\ &\leq |z|^2 + 2|z||w| + |w|^2 = (|z| + |w|)^2. \end{aligned}$$

By taking square roots, we obtain (c).

(d) From (a) and (c), it follows that

$$|z| = |(z+w) - w| \leq |z+w| + |-w| = |z+w| + |w|.$$

So

$$|z| - |w| \leq |z+w|,$$

proving (d). ∎

It is interesting as well as useful that complex numbers have both a geometric and an algebraic representation. Suppose that $z = a + bi$, where a and b are real numbers. We may represent z as a vector in the complex plane (see Figure D.1(a)). Notice that, as in R^2, there are two axes, the **real axis** and the **imaginary axis**. The real and imaginary parts of z are the first and second coordinates, and the absolute value of z gives the length of the vector z. It is clear that addition of complex numbers may be represented as in R^2 using the parallelogram law.

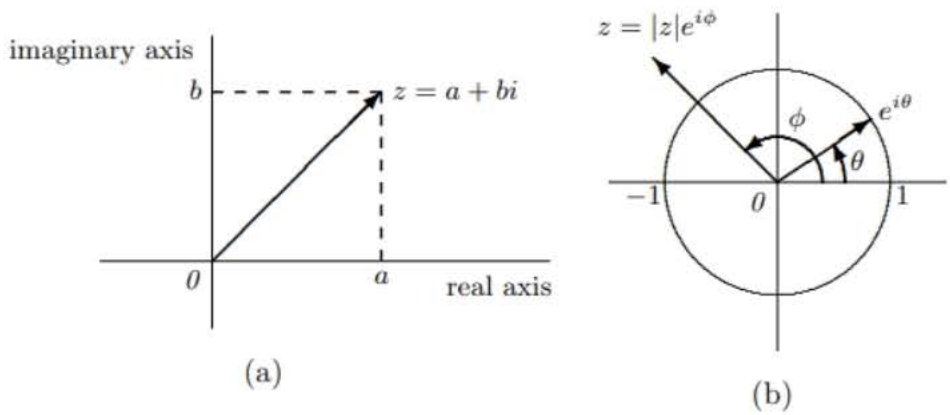


Figure D.1

In Section 2.7 (p.133), we introduce Euler's formula. The special case $e^{i\theta} = \cos\theta + i\sin\theta$ is of particular interest. Because of the geometry we have introduced, we may represent the vector $e^{i\theta}$ as in Figure D.1(b); that is, $e^{i\theta}$ is the unit vector that makes an angle θ with the positive real axis. From this figure, we see that any nonzero complex number z may be depicted as a multiple of a unit vector, namely, $z = |z|e^{i\phi}$, where ϕ is the angle that the vector z makes with the positive real axis. Thus multiplication, as well as addition, has a simple geometric interpretation: If $z = |z|e^{i\theta}$ and $w = |w|e^{i\omega}$ are two nonzero complex numbers, then from the properties established in Section 2.7 and Theorem D.3, we have

$$zw = |z|e^{i\theta} \cdot |w|e^{i\omega} = |zw|e^{i(\theta+\omega)}.$$

So zw is the vector whose length is the product of the lengths of z and w, and makes the angle $\theta + \omega$ with the positive real axis.

Our motivation for enlarging the set of real numbers to the set of complex numbers is to obtain a field such that every polynomial with nonzero degree having coefficients in that field has a zero. Our next result guarantees that the field of complex numbers has this property.

Theorem D.4 (The Fundamental Theorem of Algebra). *Suppose that* $p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ *is a polynomial in* $\mathsf{P}(C)$ *of degree* $n \geq 1$. *Then* $p(z)$ *has a zero.*

The following proof is based on one in the book *Principles of Mathematical Analysis* 3d., by Walter Rudin (McGraw-Hill Higher Education, New York, 1976).

Proof. We want to find z_0 in C such that $p(z_0) = 0$. Let m be the greatest lower bound of $\{|p(z)| : z \in C\}$. For $|z| = s > 0$, we have

$$\begin{aligned} |p(z)| &= |a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0| \\ &\geq |a_n||z|^n - |a_{n-1}||z|^{n-1} - \cdots - |a_0| \\ &= |a_n| s^n - |a_{n-1}| s^{n-1} - \cdots - |a_0| \\ &= s^n [|a_n| - |a_{n-1}| s^{-1} - \cdots - |a_0| s^{-n}]. \end{aligned}$$

Because the last expression approaches infinity as s approaches infinity, we may choose a closed disk D about the origin such that $|p(z)| > m + 1$ if z is not in D. It follows that m is the greatest lower bound of $\{|p(z)| : z \in D\}$. Because D is closed and bounded and $p(z)$ is continuous, there exists z_0 in D such that $|p(z_0)| = m$. We want to show that $m = 0$. We argue by contradiction.

Assume that $m \neq 0$. Let $q(z) = \dfrac{p(z + z_0)}{p(z_0)}$. Then $q(z)$ is a polynomial of degree n, $q(0) = 1$, and $|q(z)| \geq 1$ for all z in C. So we may write

$$q(z) = 1 + b_k z^k + b_{k+1} z^{k+1} + \cdots + b_n z^n,$$

where $b_k \neq 0$. Because $-\dfrac{|b_k|}{b_k}$ has modulus one, we may pick a real number θ such that $e^{ik\theta} = -\dfrac{|b_k|}{b_k}$, or $e^{ik\theta} b_k = -|b_k|$. For any $r > 0$, we have

$$\begin{aligned} q(re^{i\theta}) &= 1 + b_k r^k e^{ik\theta} + b_{k+1} r^{k+1} e^{i(k+1)\theta} + \cdots + b_n r^n e^{in\theta} \\ &= 1 - |b_k| r^k + b_{k+1} r^{k+1} e^{i(k+1)\theta} + \cdots + b_n r^n e^{in\theta}. \end{aligned}$$

Choose r small enough so that $1 - |b_k| r^k > 0$. Then

$$|q(re^{i\theta})| \leq 1 - |b_k| r^k + |b_{k+1}| r^{k+1} + \cdots + |b_n| r^n$$

$$= 1 - r^k[|b_k| - |b_{k+1}|r - \cdots - |b_n|r^{n-k}].$$

Now choose r even smaller, if necessary, so that the expression within the brackets is positive. We obtain that $|q(re^{i\theta})| < 1$. But this is a contradiction. ∎

The following important corollary is a consequence of Theorem D.4 and the division algorithm for polynomials (Theorem E.1).

Corollary. *If $p(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$ is a polynomial of degree $n \geq 1$ with complex coefficients, then there exist complex numbers $c_1, c_2, \ldots, c_n$ (not necessarily distinct) such that*

$$p(z) = a_n(z - c_1)(z - c_2)\cdots(z - c_n).$$

Proof. Exercise. ∎

A field is called **algebraically closed** if it has the property that every polynomial of positive degree with coefficients from that field factors as a product of polynomials of degree 1. Thus the preceding corollary asserts that the field of complex numbers is algebraically closed.

APPENDIX E POLYNOMIALS

In this appendix, we discuss some useful properties of the polynomials with coefficients from a field. For the definition of a polynomial, refer to Section 1.2. Throughout this appendix, we assume that all polynomials have coefficients from a fixed field F.

Definition. *A polynomial $f(x)$ **divides** a polynomial $g(x)$ if there exists a polynomial $q(x)$ such that $g(x) = f(x)q(x)$.*

Our first result shows that the familiar long division process for polynomials with real coefficients is valid for polynomials with coefficients from an arbitrary field.

Theorem E.1 (The Division Algorithm for Polynomials). *Let $f(x)$ be a polynomial of degree n, and let $g(x)$ be a polynomial of degree $m \geq 0$. Then there exist unique polynomials $q(x)$ and $r(x)$ such that*

$$f(x) = q(x)g(x) + r(x), \tag{1}$$

where the degree of $r(x)$ is less than m.

Proof. We begin by establishing the existence of $q(x)$ and $r(x)$ that satisfy (1). If $n < m$, take $q(x) = 0$ and $r(x) = f(x)$ to satisfy (1). Otherwise, if $0 \leq m \leq n$, we apply mathematical induction on n. First suppose that

$n = 0$. Then $m = 0$, and it follows that $f(x)$ and $g(x)$ are nonzero constants. Hence we may take $q(x) = f(x)/g(x)$ and $r(x) = 0$ to satisfy (1).

Now suppose that the result is valid for all polynomials with degree less than n for some fixed $n > 0$, and assume that $f(x)$ has degree n. Suppose that

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

and

$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0,$$

and let $h(x)$ be the polynomial defined by

$$h(x) = f(x) - a_n b_m^{-1} x^{n-m} g(x). \tag{2}$$

Then $h(x)$ is a polynomial of degree $k < n$, and therefore we may apply the induction hypothesis or the case where $k < m$ (whichever is relevant) to obtain polynomials $q_1(x)$ and $r(x)$ such that $r(x)$ has degree less than m and

$$h(x) = q_1(x)g(x) + r(x). \tag{3}$$

Combining (2) and (3) and solving for $f(x)$ gives us $f(x) = q(x)g(x) + r(x)$ with $q(x) = a_n b_m^{-1} x^{n-m} + q_1(x)$, which establishes the existence of $q(x)$ and $r(x)$ by mathematical induction.

We now show the uniqueness of $q(x)$ and $r(x)$. Suppose that $q_1(x)$, $q_2(x)$, $r_1(x)$, and $r_2(x)$ exist such that $r_1(x)$ and $r_2(x)$ each has degree less than m and

$$f(x) = q_1(x)g(x) + r_1(x) = q_2(x)g(x) + r_2(x).$$

Then

$$[q_1(x) - q_2(x)]\, g(x) = r_2(x) - r_1(x). \tag{4}$$

The right side of (4) is a polynomial of degree less than m. Since $g(x)$ has degree m, it must follow that $q_1(x) - q_2(x)$ is the zero polynomial. Hence $q_1(x) = q_2(x)$, and thus $r_1(x) = r_2(x)$ by (4). ∎

In the context of Theorem E.1, we call $q(x)$ and $r(x)$ the **quotient** and **remainder**, respectively, for the division of $f(x)$ by $g(x)$. For example, suppose that F is the field of complex numbers. Then the quotient and remainder for the division of

$$f(x) = (3+i)x^5 - (1-i)x^4 + 6x^3 + (-6+2i)x^2 + (2+i)x + 1$$

by

$$g(x) = (3+i)x^2 - 2ix + 4$$

are, respectively,

$$q(x) = x^3 + ix^2 - 2 \quad \text{and} \quad r(x) = (2-3i)x + 9.$$

Corollary 1. *Let $f(x)$ be a polynomial of positive degree, and let $a \in F$. Then $f(a) = 0$ if and only if $x - a$ divides $f(x)$.*

Proof. Suppose that $x - a$ divides $f(x)$. Then there exists a polynomial $q(x)$ such that $f(x) = (x - a)q(x)$. Thus $f(a) = (a - a)q(a) = 0 \cdot q(a) = 0$.

Conversely, suppose that $f(a) = 0$. By the division algorithm, there exist polynomials $q(x)$ and $r(x)$ such that $r(x)$ has degree less than 1 and

$$f(x) = q(x)(x - a) + r(x).$$

Substituting a for x in the equation above, we obtain $r(a) = 0$. Since $r(x)$ has degree less than 1, it must be the constant polynomial $r(x) = 0$. Thus $f(x) = q(x)(x - a)$. ∎

For any polynomial $f(x)$ with coefficients from a field F, an element $a \in F$ is called a **zero** of $f(x)$ if $f(a) = 0$. With this terminology, the preceding corollary states that a is a zero of $f(x)$ if and only if $x - a$ divides $f(x)$.

Corollary 2. *Any polynomial of degree $n \geq 1$ has at most n distinct zeros.*

Proof. The proof is by mathematical induction on n. The result is obvious if $n = 1$. Now suppose that the result is true for some positive integer n, and let $f(x)$ be a polynomial of degree $n + 1$. If $f(x)$ has no zeros, then there is nothing to prove. Otherwise, if a is a zero of $f(x)$, then by Corollary 1 we may write $f(x) = (x - a)q(x)$ for some polynomial $q(x)$. Note that $q(x)$ must be of degree n; therefore, by the induction hypothesis, $q(x)$ can have at most n distinct zeros. Since any zero of $f(x)$ distinct from a is also a zero of $q(x)$, it follows that $f(x)$ can have at most $n + 1$ distinct zeros. ∎

Polynomials having no common divisors arise naturally in the study of *canonical forms*. (See Chapter 7.)

Definition. *Polynomials $f_1(x), f_2(x), \ldots, f_n(x)$ are called* ***relatively prime*** *if no polynomial of positive degree divides all of them.*

For example, the polynomials with real coefficients $f(x) = x^2(x - 1)$ and $h(x) = (x - 1)(x - 2)$ are not relatively prime because $x - 1$ divides each of them. On the other hand, consider $f(x)$ and $g(x) = (x - 2)(x - 3)$, which do not appear to have common factors. Could other factorizations of $f(x)$ and $g(x)$ reveal a hidden common factor? We will soon see (Theorem E.9) that the preceding factors are the only ones. Thus $f(x)$ and $g(x)$ are relatively prime because they have no common factors of positive degree.