# Chapter 8

# Orthogonality

In Section 5.3 we introduced the dot product in $\mathbb{R}^n$ and extended the basic geometric notions of length and distance. A set $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$ of nonzero vectors in $\mathbb{R}^n$ was called an **orthogonal set** if $\mathbf{f}_i \cdot \mathbf{f}_j = 0$ for all $i \neq j$, and it was proved that every orthogonal set is independent. In particular, it was observed that the expansion of a vector as a linear combination of orthogonal basis vectors is easy to obtain because formulas exist for the coefficients. Hence the orthogonal bases are the "nice" bases, and much of this chapter is devoted to extending results about bases to orthogonal bases. This leads to some very powerful methods and theorems. Our first task is to show that every subspace of $\mathbb{R}^n$ *has* an orthogonal basis.

## 8.1 Orthogonal Complements and Projections

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is linearly independent in a general vector space, and if $\mathbf{v}_{m+1}$ is not in $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$, then $\{\mathbf{v}_1, \ldots, \mathbf{v}_m, \mathbf{v}_{m+1}\}$ is independent (Lemma 6.4.1). Here is the analog for *orthogonal* sets in $\mathbb{R}^n$.

**Lemma 8.1.1: Orthogonal Lemma**

*Let $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$ be an orthogonal set in $\mathbb{R}^n$. Given $\mathbf{x}$ in $\mathbb{R}^n$, write*

$$\mathbf{f}_{m+1} = \mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

*Then:*

1. *$\mathbf{f}_{m+1} \cdot \mathbf{f}_k = 0$ for $k = 1, 2, \ldots, m$.*
2. *If $\mathbf{x}$ is not in $\operatorname{span}\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$, then $\mathbf{f}_{m+1} \neq \mathbf{0}$ and $\{\mathbf{f}_1, \ldots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ is an orthogonal set.*

**Proof.** For convenience, write $t_i = (\mathbf{x} \cdot \mathbf{f}_i)/\|\mathbf{f}_i\|^2$ for each $i$. Given $1 \leq k \leq m$:

$$\begin{aligned}
\mathbf{f}_{m+1} \cdot \mathbf{f}_k &= (\mathbf{x} - t_1\mathbf{f}_1 - \cdots - t_k\mathbf{f}_k - \cdots - t_m\mathbf{f}_m) \cdot \mathbf{f}_k \\
&= \mathbf{x} \cdot \mathbf{f}_k - t_1(\mathbf{f}_1 \cdot \mathbf{f}_k) - \cdots - t_k(\mathbf{f}_k \cdot \mathbf{f}_k) - \cdots - t_m(\mathbf{f}_m \cdot \mathbf{f}_k) \\
&= \mathbf{x} \cdot \mathbf{f}_k - t_k\|\mathbf{f}_k\|^2 \\
&= 0
\end{aligned}$$

This proves (1), and (2) follows because $\mathbf{f}_{m+1} \neq \mathbf{0}$ if $\mathbf{x}$ is not in $\operatorname{span}\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$. □

The orthogonal lemma has three important consequences for $\mathbb{R}^n$. The first is an extension for orthogonal sets of the fundamental fact that any independent set is part of a basis (Theorem 6.4.1).

**Theorem 8.1.1**

*Let $U$ be a subspace of $\mathbb{R}^n$.*

1. *Every orthogonal subset $\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ in $U$ is a subset of an orthogonal basis of $U$.*
2. *$U$ has an orthogonal basis.*

**Proof.**

1. If $\text{span}\{\mathbf{f}_1, \ldots, \mathbf{f}_m\} = U$, it is *already* a basis. Otherwise, there exists $\mathbf{x}$ in $U$ outside $\text{span}\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$. If $\mathbf{f}_{m+1}$ is as given in the orthogonal lemma, then $\mathbf{f}_{m+1}$ is in $U$ and $\{\mathbf{f}_1, \ldots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ is orthogonal. If $\text{span}\{\mathbf{f}_1, \ldots, \mathbf{f}_m, \mathbf{f}_{m+1}\} = U$, we are done. Otherwise, the process continues to create larger and larger orthogonal subsets of $U$. They are all independent by Theorem 5.3.5, so we have a basis when we reach a subset containing $\dim U$ vectors.

2. If $U = \{\mathbf{0}\}$, the empty basis is orthogonal. Otherwise, if $\mathbf{f} \neq \mathbf{0}$ is in $U$, then $\{\mathbf{f}\}$ is orthogonal, so (2) follows from (1). □

We can improve upon (2) of Theorem 8.1.1. In fact, the second consequence of the orthogonal lemma is a procedure by which *any* basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ of a subspace $U$ of $\mathbb{R}^n$ can be systematically modified to yield an orthogonal basis $\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ of $U$. The $\mathbf{f}_i$ are constructed one at a time from the $\mathbf{x}_i$.

To start the process, take $\mathbf{f}_1 = \mathbf{x}_1$. Then $\mathbf{x}_2$ is not in $\text{span}\{\mathbf{f}_1\}$ because $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent, so take

$$\mathbf{f}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1$$

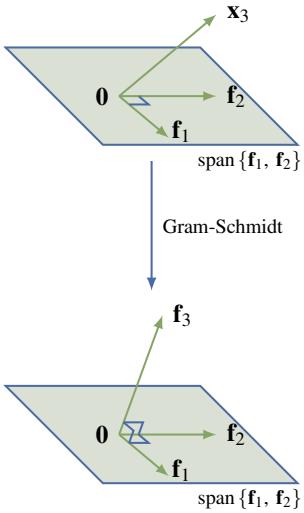
Thus $\{\mathbf{f}_1, \mathbf{f}_2\}$ is orthogonal by Lemma 8.1.1. Moreover, $\text{span}\{\mathbf{f}_1, \mathbf{f}_2\} = \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ (verify), so $\mathbf{x}_3$ is not in $\text{span}\{\mathbf{f}_1, \mathbf{f}_2\}$. Hence $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ is orthogonal where

$$\mathbf{f}_3 = \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2$$

Again, $\text{span}\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\} = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, so $\mathbf{x}_4$ is not in $\text{span}\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ and the process continues. At the $m$th iteration we construct an orthogonal set $\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ such that

$$\text{span}\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\} = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} = U$$

Hence $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$ is the desired orthogonal basis of $U$. The procedure can be summarized as follows.




**Theorem 8.1.2: Gram-Schmidt Orthogonalization Algorithm**[1]

*If $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ is any basis of a subspace $U$ of $\mathbb{R}^n$, construct $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m$ in $U$ successively as follows:*

$$\begin{aligned}
\mathbf{f}_1 &= \mathbf{x}_1 \\
\mathbf{f}_2 &= \mathbf{x}_2 - \tfrac{\mathbf{x}_2\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 \\
\mathbf{f}_3 &= \mathbf{x}_3 - \tfrac{\mathbf{x}_3\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \tfrac{\mathbf{x}_3\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 \\
&\vdots \\
\mathbf{f}_k &= \mathbf{x}_k - \tfrac{\mathbf{x}_k\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \tfrac{\mathbf{x}_k\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \tfrac{\mathbf{x}_k\cdot\mathbf{f}_{k-1}}{\|\mathbf{f}_{k-1}\|^2}\mathbf{f}_{k-1}
\end{aligned}$$

*for each $k = 2, 3, \ldots, m$. Then*

1. *$\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$ is an orthogonal basis of $U$.*
2. *$\operatorname{span}\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k\} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for each $k = 1, 2, \ldots, m$.*

The process (for $k = 3$) is depicted in the diagrams. Of course, the algorithm converts any basis of $\mathbb{R}^n$ itself into an orthogonal basis.

**Example 8.1.1**

Find an orthogonal basis of the row space of $A = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 3 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$.

**Solution.** Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ denote the rows of $A$ and observe that $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is linearly independent. Take $\mathbf{f}_1 = \mathbf{x}_1$. The algorithm gives

$$\mathbf{f}_2 = \mathbf{x}_2 - \tfrac{\mathbf{x}_2\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 = (3, 2, 0, 1) - \tfrac{4}{4}(1, 1, -1, -1) = (2, 1, 1, 2)$$

$$\mathbf{f}_3 = \mathbf{x}_3 - \tfrac{\mathbf{x}_3\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \tfrac{\mathbf{x}_3\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 = \mathbf{x}_3 - \tfrac{0}{4}\mathbf{f}_1 - \tfrac{3}{10}\mathbf{f}_2 = \tfrac{1}{10}(4, -3, 7, -6)$$

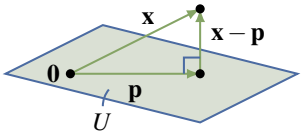
Hence $\{(1, 1, -1, -1), (2, 1, 1, 2), \frac{1}{10}(4, -3, 7, -6)\}$ is the orthogonal basis provided by the algorithm. In hand calculations it may be convenient to eliminate fractions (see the Remark below), so $\{(1, 1, -1, -1), (2, 1, 1, 2), (4, -3, 7, -6)\}$ is also an orthogonal basis for row $A$.

[1] Erhardt Schmidt (1876–1959) was a German mathematician who studied under the great David Hilbert and later developed the theory of Hilbert spaces. He first described the present algorithm in 1907. Jörgen Pederson Gram (1850–1916) was a Danish actuary.

***Remark***

Observe that the vector $\frac{\mathbf{x}\cdot\mathbf{f}_i}{\|\mathbf{f}_i\|^2}\mathbf{f}_i$ is unchanged if a nonzero scalar multiple of $\mathbf{f}_i$ is used in place of $\mathbf{f}_i$. Hence, if a newly constructed $\mathbf{f}_i$ is multiplied by a nonzero scalar at some stage of the Gram-Schmidt algorithm, the subsequent $\mathbf{f}$s will be unchanged. This is useful in actual calculations.

## Projections




Suppose a point $\mathbf{x}$ and a plane $U$ through the origin in $\mathbb{R}^3$ are given, and we want to find the point $\mathbf{p}$ in the plane that is closest to $\mathbf{x}$. Our geometric intuition assures us that such a point $\mathbf{p}$ exists. In fact (see the diagram), $\mathbf{p}$ must be chosen in such a way that $\mathbf{x}-\mathbf{p}$ is *perpendicular* to the plane.

Now we make two observations: first, the plane $U$ is a *subspace of* $\mathbb{R}^3$ (because $U$ contains the origin); and second, that the condition that $\mathbf{x}-\mathbf{p}$ is perpendicular to the plane $U$ means that $\mathbf{x}-\mathbf{p}$ is *orthogonal* to every vector in $U$. In these terms the whole discussion makes sense in $\mathbb{R}^n$. Furthermore, the orthogonal lemma provides exactly what is needed to find $\mathbf{p}$ in this more general setting.

**Definition 8.1 Orthogonal Complement of a Subspace of $\mathbb{R}^n$**

*If $U$ is a subspace of $\mathbb{R}^n$, define the* ***orthogonal complement*** *$U^\perp$ of $U$ (pronounced "$U$-perp") by*

$$U^\perp = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid \mathbf{x}\cdot\mathbf{y} = 0 \text{ for all } \mathbf{y} \text{ in } U\}$$

The following lemma collects some useful properties of the orthogonal complement; the proof of (1) and (2) is left as Exercise 8.1.6.

**Lemma 8.1.2**

*Let $U$ be a subspace of $\mathbb{R}^n$.*

1. *$U^\perp$ is a subspace of $\mathbb{R}^n$.*
2. *$\{\mathbf{0}\}^\perp = \mathbb{R}^n$ and $(\mathbb{R}^n)^\perp = \{\mathbf{0}\}$.*
3. *If $U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, then $U^\perp = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid \mathbf{x}\cdot\mathbf{x}_i = 0 \text{ for } i = 1, 2, \ldots, k\}$.*

**Proof.**

3. Let $U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$; we must show that $U^\perp = \{\mathbf{x} \mid \mathbf{x}\cdot\mathbf{x}_i = 0 \text{ for each } i\}$. If $\mathbf{x}$ is in $U^\perp$ then $\mathbf{x}\cdot\mathbf{x}_i = 0$ for all $i$ because each $\mathbf{x}_i$ is in $U$. Conversely, suppose that $\mathbf{x}\cdot\mathbf{x}_i = 0$ for all $i$; we must show that $\mathbf{x}$ is in $U^\perp$, that is, $\mathbf{x}\cdot\mathbf{y} = 0$ for each $\mathbf{y}$ in $U$. Write $\mathbf{y} = r_1\mathbf{x}_1 + r_2\mathbf{x}_2 + \cdots + r_k\mathbf{x}_k$, where each $r_i$ is in $\mathbb{R}$. Then, using Theorem 5.3.1,

$$\mathbf{x}\cdot\mathbf{y} = r_1(\mathbf{x}\cdot\mathbf{x}_1) + r_2(\mathbf{x}\cdot\mathbf{x}_2) + \cdots + r_k(\mathbf{x}\cdot\mathbf{x}_k) = r_1 0 + r_2 0 + \cdots + r_k 0 = 0$$

as required. □

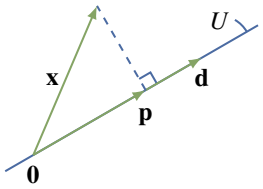
**Example 8.1.2**

Find $U^\perp$ if $U = \text{span}\{(1, -1, 2, 0), (1, 0, -2, 3)\}$ in $\mathbb{R}^4$.

**Solution.** By Lemma 8.1.2, $\mathbf{x} = (x, y, z, w)$ is in $U^\perp$ if and only if it is orthogonal to both $(1, -1, 2, 0)$ and $(1, 0, -2, 3)$; that is,

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

Gaussian elimination gives $U^\perp = \text{span}\{(2, 4, 1, 0), (3, 3, 0, -1)\}$.




Now consider vectors $\mathbf{x}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbb{R}^3$. The projection $\mathbf{p} = \text{proj}_{\mathbf{d}}\,\mathbf{x}$ of $\mathbf{x}$ on $\mathbf{d}$ was defined in Section 4.2 as in the diagram.

The following formula for $\mathbf{p}$ was derived in Theorem 4.2.4

$$\mathbf{p} = \text{proj}_{\mathbf{d}}\,\mathbf{x} = \left(\frac{\mathbf{x}\cdot\mathbf{d}}{\|\mathbf{d}\|^2}\right)\mathbf{d}$$

where it is shown that $\mathbf{x} - \mathbf{p}$ is orthogonal to $\mathbf{d}$. Now observe that the line $U = \mathbb{R}\mathbf{d} = \{t\mathbf{d} \mid t \in \mathbb{R}\}$ is a subspace of $\mathbb{R}^3$, that $\{\mathbf{d}\}$ is an orthogonal basis of $U$, and that $\mathbf{p} \in U$ and $\mathbf{x} - \mathbf{p} \in U^\perp$ (by Theorem 4.2.4).

In this form, this makes sense for any vector $\mathbf{x}$ in $\mathbb{R}^n$ and any subspace $U$ of $\mathbb{R}^n$, so we generalize it as follows. If $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$ is an orthogonal basis of $U$, we define the projection $\mathbf{p}$ of $\mathbf{x}$ on $U$ by the formula

$$\mathbf{p} = \left(\frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\right)\mathbf{f}_1 + \left(\frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\right)\mathbf{f}_2 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\right)\mathbf{f}_m \tag{8.1}$$

Then $\mathbf{p} \in U$ and (by the orthogonal lemma) $\mathbf{x} - \mathbf{p} \in U^\perp$, so it looks like we have a generalization of Theorem 4.2.4.

However there is a potential problem: the formula (8.1) for $\mathbf{p}$ must be shown to be independent of the choice of the orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$. To verify this, suppose that $\{\mathbf{f}'_1, \mathbf{f}'_2, \ldots, \mathbf{f}'_m\}$ is another orthogonal basis of $U$, and write

$$\mathbf{p}' = \left(\frac{\mathbf{x}\cdot\mathbf{f}'_1}{\|\mathbf{f}'_1\|^2}\right)\mathbf{f}'_1 + \left(\frac{\mathbf{x}\cdot\mathbf{f}'_2}{\|\mathbf{f}'_2\|^2}\right)\mathbf{f}'_2 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{f}'_m}{\|\mathbf{f}'_m\|^2}\right)\mathbf{f}'_m$$

As before, $\mathbf{p}' \in U$ and $\mathbf{x} - \mathbf{p}' \in U^\perp$, and we must show that $\mathbf{p}' = \mathbf{p}$. To see this, write the vector $\mathbf{p} - \mathbf{p}'$ as follows:

$$\mathbf{p} - \mathbf{p}' = (\mathbf{x} - \mathbf{p}') - (\mathbf{x} - \mathbf{p})$$

This vector is in $U$ (because $\mathbf{p}$ and $\mathbf{p}'$ are in $U$) and it is in $U^\perp$ (because $\mathbf{x} - \mathbf{p}'$ and $\mathbf{x} - \mathbf{p}$ are in $U^\perp$), and so it must be zero (it is orthogonal to itself!). This means $\mathbf{p}' = \mathbf{p}$ as desired.

Hence, the vector $\mathbf{p}$ in equation (8.1) depends only on $\mathbf{x}$ and the subspace $U$, and *not* on the choice of orthogonal basis $\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ of $U$ used to compute it. Thus, we are entitled to make the following definition:

**Definition 8.2 Projection onto a Subspace of $\mathbb{R}^n$**

*Let $U$ be a subspace of $\mathbb{R}^n$ with orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$. If $\mathbf{x}$ is in $\mathbb{R}^n$, the vector*

$$\text{proj}_U \mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 + \cdots + \frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

*is called the **orthogonal projection** of $\mathbf{x}$ on $U$. For the zero subspace $U = \{\mathbf{0}\}$, we define*

$$\text{proj}_{\{\mathbf{0}\}} \mathbf{x} = \mathbf{0}$$

The preceding discussion proves (1) of the following theorem.

**Theorem 8.1.3: Projection Theorem**

*If $U$ is a subspace of $\mathbb{R}^n$ and $\mathbf{x}$ is in $\mathbb{R}^n$, write $\mathbf{p} = \text{proj}_U \mathbf{x}$. Then:*

1. *$\mathbf{p}$ is in $U$ and $\mathbf{x} - \mathbf{p}$ is in $U^\perp$.*
2. *$\mathbf{p}$ is the vector in $U$ closest to $\mathbf{x}$ in the sense that*

$$\|\mathbf{x} - \mathbf{p}\| < \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{y} \in U,\ \mathbf{y} \neq \mathbf{p}$$

**Proof.**

1. This is proved in the preceding discussion (it is clear if $U = \{\mathbf{0}\}$).
2. Write $\mathbf{x} - \mathbf{y} = (\mathbf{x} - \mathbf{p}) + (\mathbf{p} - \mathbf{y})$. Then $\mathbf{p} - \mathbf{y}$ is in $U$ and so is orthogonal to $\mathbf{x} - \mathbf{p}$ by (1). Hence, the Pythagorean theorem gives

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{p}\|^2 + \|\mathbf{p} - \mathbf{y}\|^2 > \|\mathbf{x} - \mathbf{p}\|^2$$

because $\mathbf{p} - \mathbf{y} \neq \mathbf{0}$. This gives (2). □

**Example 8.1.3**

Let $U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ in $\mathbb{R}^4$ where $\mathbf{x}_1 = (1, 1, 0, 1)$ and $\mathbf{x}_2 = (0, 1, 1, 2)$. If $\mathbf{x} = (3, -1, 0, 2)$, find the vector in $U$ closest to $\mathbf{x}$ and express $\mathbf{x}$ as the sum of a vector in $U$ and a vector orthogonal to $U$.

**Solution.** $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent but not orthogonal. The Gram-Schmidt process gives an orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2\}$ of $U$ where $\mathbf{f}_1 = \mathbf{x}_1 = (1, 1, 0, 1)$ and

$$\mathbf{f}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 = \mathbf{x}_2 - \tfrac{3}{3}\mathbf{f}_1 = (-1, 0, 1, 1)$$

Hence, we can compute the projection using $\{\mathbf{f}_1, \mathbf{f}_2\}$:

$$\mathbf{p} = \text{proj}_U \mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 = \tfrac{4}{3}\mathbf{f}_1 + \tfrac{-1}{3}\mathbf{f}_2 = \tfrac{1}{3}\begin{bmatrix} 5 & 4 & -1 & 3 \end{bmatrix}$$

Thus, $\mathbf{p}$ is the vector in $U$ closest to $\mathbf{x}$, and $\mathbf{x}-\mathbf{p}=\frac{1}{3}(4,\ -7,\ 1,\ 3)$ is orthogonal to every vector in $U$. (This can be verified by checking that it is orthogonal to the generators $\mathbf{x}_1$ and $\mathbf{x}_2$ of $U$.) The required decomposition of $\mathbf{x}$ is thus

$$\mathbf{x}=\mathbf{p}+(\mathbf{x}-\mathbf{p})=\tfrac{1}{3}(5,\ 4,\ -1,\ 3)+\tfrac{1}{3}(4,\ -7,\ 1,\ 3)$$

**Example 8.1.4**

Find the point in the plane with equation $2x+y-z=0$ that is closest to the point $(2,\ -1,\ -3)$.

**Solution.** We write $\mathbb{R}^3$ as rows. The plane is the subspace $U$ whose points $(x,\ y,\ z)$ satisfy $z=2x+y$. Hence

$$U=\{(s,\ t,\ 2s+t)\mid s,\ t \text{ in } \mathbb{R}\}=\operatorname{span}\{(0,\ 1,\ 1),\ (1,\ 0,\ 2)\}$$

The Gram-Schmidt process produces an orthogonal basis $\{\mathbf{f}_1,\ \mathbf{f}_2\}$ of $U$ where $\mathbf{f}_1=(0,\ 1,\ 1)$ and $\mathbf{f}_2=(1,\ -1,\ 1)$. Hence, the vector in $U$ closest to $\mathbf{x}=(2,\ -1,\ -3)$ is

$$\operatorname{proj}_U\mathbf{x}=\tfrac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1+\tfrac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2=-2\mathbf{f}_1+0\mathbf{f}_2=(0,\ -2,\ -2)$$

Thus, the point in $U$ closest to $(2,\ -1,\ -3)$ is $(0,\ -2,\ -2)$.

The next theorem shows that projection on a subspace of $\mathbb{R}^n$ is actually a linear operator $\mathbb{R}^n\to\mathbb{R}^n$.

**Theorem 8.1.4**

*Let $U$ be a fixed subspace of $\mathbb{R}^n$. If we define $T:\mathbb{R}^n\to\mathbb{R}^n$ by*

$$T(\mathbf{x})=\operatorname{proj}_U\mathbf{x}\quad \textit{for all } \mathbf{x} \textit{ in } \mathbb{R}^n$$

1. *$T$ is a linear operator.*
2. *$\operatorname{im}T=U$ and $\ker T=U^\perp$.*
3. *$\dim U+\dim U^\perp=n$.*

**Proof.** If $U=\{\mathbf{0}\}$, then $U^\perp=\mathbb{R}^n$, and so $T(\mathbf{x})=\operatorname{proj}_{\{\mathbf{0}\}}\mathbf{x}=\mathbf{0}$ for all $\mathbf{x}$. Thus $T=0$ is the zero (linear) operator, so (1), (2), and (3) hold. Hence assume that $U\neq\{\mathbf{0}\}$.

1. If $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \ldots,\ \mathbf{f}_m\}$ is an orthonormal basis of $U$, then

$$T(\mathbf{x})=(\mathbf{x}\cdot\mathbf{f}_1)\mathbf{f}_1+(\mathbf{x}\cdot\mathbf{f}_2)\mathbf{f}_2+\cdots+(\mathbf{x}\cdot\mathbf{f}_m)\mathbf{f}_m\quad\text{for all }\mathbf{x}\text{ in }\mathbb{R}^n \tag{8.2}$$

by the definition of the projection. Thus $T$ is linear because

$$(\mathbf{x}+\mathbf{y})\cdot\mathbf{f}_i=\mathbf{x}\cdot\mathbf{f}_i+\mathbf{y}\cdot\mathbf{f}_i\quad\text{and}\quad(r\mathbf{x})\cdot\mathbf{f}_i=r(\mathbf{x}\cdot\mathbf{f}_i)\quad\text{for each } i$$

2. We have $\text{im } T \subseteq U$ by (8.2) because each $\mathbf{f}_i$ is in $U$. But if $\mathbf{x}$ is in $U$, then $\mathbf{x} = T(\mathbf{x})$ by (8.2) and the expansion theorem applied to the space $U$. This shows that $U \subseteq \text{im } T$, so $\text{im } T = U$.

   Now suppose that $\mathbf{x}$ is in $U^\perp$. Then $\mathbf{x} \cdot \mathbf{f}_i = 0$ for each $i$ (again because each $\mathbf{f}_i$ is in $U$) so $\mathbf{x}$ is in $\ker T$ by (8.2). Hence $U^\perp \subseteq \ker T$. On the other hand, Theorem 8.1.3 shows that $\mathbf{x} - T(\mathbf{x})$ is in $U^\perp$ for all $\mathbf{x}$ in $\mathbb{R}^n$, and it follows that $\ker T \subseteq U^\perp$. Hence $\ker T = U^\perp$, proving (2).

3. This follows from (1), (2), and the dimension theorem (Theorem 7.2.4). □

## Exercises for 8.1

**Exercise 8.1.1** In each case, use the Gram-Schmidt algorithm to convert the given basis $B$ of $V$ into an orthogonal basis.

a. $V = \mathbb{R}^2$, $B = \{(1, -1), (2, 1)\}$

b. $V = \mathbb{R}^2$, $B = \{(2, 1), (1, 2)\}$

c. $V = \mathbb{R}^3$, $B = \{(1, -1, 1), (1, 0, 1), (1, 1, 2)\}$

d. $V = \mathbb{R}^3$, $B = \{(0, 1, 1), (1, 1, 1), (1, -2, 2)\}$

**Exercise 8.1.2** In each case, write $\mathbf{x}$ as the sum of a vector in $U$ and a vector in $U^\perp$.

a. $\mathbf{x} = (1, 5, 7)$, $U = \text{span}\{(1, -2, 3), (-1, 1, 1)\}$

b. $\mathbf{x} = (2, 1, 6)$, $U = \text{span}\{(3, -1, 2), (2, 0, -3)\}$

c. $\mathbf{x} = (3, 1, 5, 9)$,
$U = \text{span}\{(1, 0, 1, 1), (0, 1, -1, 1), (-2, 0, 1, 1)\}$

d. $\mathbf{x} = (2, 0, 1, 6)$,
$U = \text{span}\{(1, 1, 1, 1), (1, 1, -1, -1), (1, -1, 1, -1)\}$

e. $\mathbf{x} = (a, b, c, d)$,
$U = \text{span}\{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0)\}$

f. $\mathbf{x} = (a, b, c, d)$,
$U = \text{span}\{(1, -1, 2, 0), (-1, 1, 1, 1)\}$

**Exercise 8.1.3** Let $\mathbf{x} = (1, -2, 1, 6)$ in $\mathbb{R}^4$, and let $U = \text{span}\{(2, 1, 3, -4), (1, 2, 0, 1)\}$.

a. Compute $\text{proj}_U \mathbf{x}$.

b. Show that $\{(1, 0, 2, -3), (4, 7, 1, 2)\}$ is another orthogonal basis of $U$.

c. Use the basis in part (b) to compute $\text{proj}_U \mathbf{x}$.

**Exercise 8.1.4** In each case, use the Gram-Schmidt algorithm to find an orthogonal basis of the subspace $U$, and find the vector in $U$ closest to $\mathbf{x}$.

a. $U = \text{span}\{(1, 1, 1), (0, 1, 1)\}$, $\mathbf{x} = (-1, 2, 1)$

b. $U = \text{span}\{(1, -1, 0), (-1, 0, 1)\}$, $\mathbf{x} = (2, 1, 0)$

c. $U = \text{span}\{(1, 0, 1, 0), (1, 1, 1, 0), (1, 1, 0, 0)\}$,
$\mathbf{x} = (2, 0, -1, 3)$

d. $U = \text{span}\{(1, -1, 0, 1), (1, 1, 0, 0), (1, 1, 0, 1)\}$,
$\mathbf{x} = (2, 0, 3, 1)$

**Exercise 8.1.5** Let $U = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$, $\mathbf{v}_i$ in $\mathbb{R}^n$, and let $A$ be the $k \times n$ matrix with the $\mathbf{v}_i$ as rows.

a. Show that $U^\perp = \{\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n, A\mathbf{x}^T = \mathbf{0}\}$.

b. Use part (a) to find $U^\perp$ if
$U = \text{span}\{(1, -1, 2, 1), (1, 0, -1, 1)\}$.

**Exercise 8.1.6**

a. Prove part 1 of Lemma 8.1.2.

b. Prove part 2 of Lemma 8.1.2.

**Exercise 8.1.7** Let $U$ be a subspace of $\mathbb{R}^n$. If $\mathbf{x}$ in $\mathbb{R}^n$ can be written in any way at all as $\mathbf{x} = \mathbf{p} + \mathbf{q}$ with $\mathbf{p}$ in $U$ and $\mathbf{q}$ in $U^\perp$, show that necessarily $\mathbf{p} = \text{proj}_U \mathbf{x}$.

**Exercise 8.1.8** Let $U$ be a subspace of $\mathbb{R}^n$ and let $\mathbf{x}$ be a vector in $\mathbb{R}^n$. Using Exercise 8.1.7, or otherwise, show that $\mathbf{x}$ is in $U$ if and only if $\mathbf{x} = \text{proj}_U \mathbf{x}$.

**Exercise 8.1.9** Let $U$ be a subspace of $\mathbb{R}^n$.

a. Show that $U^\perp = \mathbb{R}^n$ if and only if $U = \{\mathbf{0}\}$.

b. Show that $U^\perp = \{\mathbf{0}\}$ if and only if $U = \mathbb{R}^n$.

**Exercise 8.1.10** If $U$ is a subspace of $\mathbb{R}^n$, show that $\text{proj}_U \mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$ in $U$.

**Exercise 8.1.11** If $U$ is a subspace of $\mathbb{R}^n$, show that $\mathbf{x} = \text{proj}_U \mathbf{x} + \text{proj}_{U^\perp} \mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

**Exercise 8.1.12** If $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ is an orthogonal basis of $\mathbb{R}^n$ and $U = \text{span}\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$, show that $U^\perp = \text{span}\{\mathbf{f}_{m+1}, \ldots, \mathbf{f}_n\}$.

**Exercise 8.1.13** If $U$ is a subspace of $\mathbb{R}^n$, show that $U^{\perp\perp} = U$. [*Hint*: Show that $U \subseteq U^{\perp\perp}$, then use Theorem 8.1.4 (3) twice.]

**Exercise 8.1.14** If $U$ is a subspace of $\mathbb{R}^n$, show how to find an $n \times n$ matrix $A$ such that $U = \{\mathbf{x} \mid A\mathbf{x} = \mathbf{0}\}$. [*Hint*: Exercise 8.1.13.]

**Exercise 8.1.15** Write $\mathbb{R}^n$ as rows. If $A$ is an $n \times n$ matrix, write its null space as $\text{null } A = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid A\mathbf{x}^T = \mathbf{0}\}$. Show that:

a. $\text{null } A = (\text{row } A)^\perp$; b. $\text{null } A^T = (\text{col } A)^\perp$.

**Exercise 8.1.16** If $U$ and $W$ are subspaces, show that $(U+W)^\perp = U^\perp \cap W^\perp$. [See Exercise 5.1.22.]

**Exercise 8.1.17** Think of $\mathbb{R}^n$ as consisting of rows.

a. Let $E$ be an $n \times n$ matrix, and let $U = \{\mathbf{x}E \mid \mathbf{x} \text{ in } \mathbb{R}^n\}$. Show that the following are equivalent.

   i. $E^2 = E = E^T$ ($E$ is a **projection matrix**).

   ii. $(\mathbf{x} - \mathbf{x}E) \cdot (\mathbf{y}E) = 0$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.

   iii. $\text{proj}_U \mathbf{x} = \mathbf{x}E$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

   [*Hint*: For (ii) implies (iii): Write $\mathbf{x} = \mathbf{x}E + (\mathbf{x} - \mathbf{x}E)$ and use the uniqueness argument preceding the definition of $\text{proj}_U \mathbf{x}$. For (iii) implies (ii): $\mathbf{x} - \mathbf{x}E$ is in $U^\perp$ for all $\mathbf{x}$ in $\mathbb{R}^n$.]

b. If $E$ is a projection matrix, show that $I - E$ is also a projection matrix.

c. If $EF = 0 = FE$ and $E$ and $F$ are projection matrices, show that $E + F$ is also a projection matrix.

d. If $A$ is $m \times n$ and $AA^T$ is invertible, show that $E = A^T(AA^T)^{-1}A$ is a projection matrix.

**Exercise 8.1.18** Let $A$ be an $n \times n$ matrix of rank $r$. Show that there is an invertible $n \times n$ matrix $U$ such that $UA$ is a row-echelon matrix with the property that the first $r$ rows are orthogonal. [*Hint*: Let $R$ be the row-echelon form of $A$, and use the Gram-Schmidt process on the nonzero rows of $R$ from the bottom up. Use Lemma 2.4.1.]

**Exercise 8.1.19** Let $A$ be an $(n-1) \times n$ matrix with rows $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}$ and let $A_i$ denote the $(n-1) \times (n-1)$ matrix obtained from $A$ by deleting column $i$. Define the vector $\mathbf{y}$ in $\mathbb{R}^n$ by

$$\mathbf{y} = \begin{bmatrix} \det A_1 & -\det A_2 & \det A_3 & \cdots & (-1)^{n+1} \det A_n \end{bmatrix}$$

Show that:

a. $\mathbf{x}_i \cdot \mathbf{y} = 0$ for all $i = 1, 2, \ldots, n-1$. [*Hint*: Write $B_i = \begin{bmatrix} x_i \\ A \end{bmatrix}$ and show that $\det B_i = 0$.]

b. $\mathbf{y} \neq \mathbf{0}$ if and only if $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}\}$ is linearly independent. [*Hint*: If some $\det A_i \neq 0$, the rows of $A_i$ are linearly independent. Conversely, if the $\mathbf{x}_i$ are independent, consider $A = UR$ where $R$ is in reduced row-echelon form.]

c. If $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}\}$ is linearly independent, use Theorem 8.1.3(3) to show that all solutions to the system of $n-1$ homogeneous equations

$$A\mathbf{x}^T = \mathbf{0}$$

are given by $t\mathbf{y}$, $t$ a parameter.

## 8.2 Orthogonal Diagonalization

Recall (Theorem 5.5.3) that an $n \times n$ matrix $A$ is diagonalizable if and only if it has $n$ linearly independent eigenvectors. Moreover, the matrix $P$ with these eigenvectors as columns is a diagonalizing matrix for $A$, that is

$$P^{-1}AP \text{ is diagonal.}$$

As we have seen, the really nice bases of $\mathbb{R}^n$ are the orthogonal ones, so a natural question is: which $n \times n$ matrices have an *orthogonal* basis of eigenvectors? These turn out to be precisely the symmetric matrices, and this is the main result of this section.

Before proceeding, recall that an orthogonal set of vectors is called *orthonormal* if $\|\mathbf{v}\| = 1$ for each vector $\mathbf{v}$ in the set, and that any orthogonal set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ can be "*normalized*", that is converted into an orthonormal set $\{\frac{1}{\|\mathbf{v}_1\|}\mathbf{v}_1, \frac{1}{\|\mathbf{v}_2\|}\mathbf{v}_2, \ldots, \frac{1}{\|\mathbf{v}_k\|}\mathbf{v}_k\}$. In particular, if a matrix $A$ has $n$ orthogonal eigenvectors, they can (by normalizing) be taken to be orthonormal. The corresponding diagonalizing matrix $P$ has orthonormal columns, and such matrices are very easy to invert.

**Theorem 8.2.1**

*The following conditions are equivalent for an $n \times n$ matrix $P$.*

1. *$P$ is invertible and $P^{-1} = P^T$.*
2. *The rows of $P$ are orthonormal.*
3. *The columns of $P$ are orthonormal.*

**Proof.** First recall that condition (1) is equivalent to $PP^T = I$ by Corollary 2.4.1 of Theorem 2.4.5. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ denote the rows of $P$. Then $\mathbf{x}_j^T$ is the $j$th column of $P^T$, so the $(i, j)$-entry of $PP^T$ is $\mathbf{x}_i \cdot \mathbf{x}_j$. Thus $PP^T = I$ means that $\mathbf{x}_i \cdot \mathbf{x}_j = 0$ if $i \neq j$ and $\mathbf{x}_i \cdot \mathbf{x}_j = 1$ if $i = j$. Hence condition (1) is equivalent to (2). The proof of the equivalence of (1) and (3) is similar. □

**Definition 8.3 Orthogonal Matrices**

*An $n \times n$ matrix $P$ is called an **orthogonal matrix**[2] if it satisfies one (and hence all) of the conditions in Theorem 8.2.1.*

**Example 8.2.1**

The rotation matrix $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ is orthogonal for any angle $\theta$.

These orthogonal matrices have the virtue that they are easy to invert—simply take the transpose. But they have many other important properties as well. If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear operator, we will prove (Theorem 10.4.3) that $T$ is distance preserving if and only if its matrix is orthogonal. In particular, the matrices of rotations and reflections about the origin in $\mathbb{R}^2$ and $\mathbb{R}^3$ are all orthogonal (see Example 8.2.1).

[2] In view of (2) and (3) of Theorem 8.2.1, orthonormal matrix might be a better name. But orthogonal matrix is standard.

It is not enough that the rows of a matrix $A$ are merely orthogonal for $A$ to be an orthogonal matrix. Here is an example.

**Example 8.2.2**

The matrix $\begin{bmatrix} 2 & 1 & 1 \\ -1 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix}$ has orthogonal rows but the columns are not orthogonal. However, if the rows are normalized, the resulting matrix $\begin{bmatrix} \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ is orthogonal (so the columns are now orthonormal as the reader can verify).

**Example 8.2.3**

If $P$ and $Q$ are orthogonal matrices, then $PQ$ is also orthogonal, as is $P^{-1} = P^T$.

**Solution.** $P$ and $Q$ are invertible, so $PQ$ is also invertible and

$$(PQ)^{-1} = Q^{-1}P^{-1} = Q^T P^T = (PQ)^T$$

Hence $PQ$ is orthogonal. Similarly,

$$(P^{-1})^{-1} = P = (P^T)^T = (P^{-1})^T$$

shows that $P^{-1}$ is orthogonal.

**Definition 8.4 Orthogonally Diagonalizable Matrices**

*An $n \times n$ matrix $A$ is said to be **orthogonally diagonalizable** when an orthogonal matrix $P$ can be found such that $P^{-1}AP = P^TAP$ is diagonal.*

This condition turns out to characterize the symmetric matrices.

**Theorem 8.2.2: Principal Axes Theorem**

*The following conditions are equivalent for an $n \times n$ matrix $A$.*

1. *$A$ has an orthonormal set of $n$ eigenvectors.*
2. *$A$ is orthogonally diagonalizable.*
3. *$A$ is symmetric.*

**Proof.** (1) $\Leftrightarrow$ (2). Given (1), let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be orthonormal eigenvectors of $A$. Then $P = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}$ is orthogonal, and $P^{-1}AP$ is diagonal by Theorem 3.3.4. This proves (2). Conversely, given (2) let $P^{-1}AP$ be diagonal where $P$ is orthogonal. If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are the columns of $P$ then $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ that consists of eigenvectors of $A$ by Theorem 3.3.4. This proves (1).

(2) $\Rightarrow$ (3). If $P^TAP = D$ is diagonal, where $P^{-1} = P^T$, then $A = PDP^T$. But $D^T = D$, so this gives $A^T = P^{TT}D^TP^T = PDP^T = A$.

(3) $\Rightarrow$ (2). If $A$ is an $n \times n$ symmetric matrix, we proceed by induction on $n$. If $n = 1$, $A$ is already diagonal. If $n > 1$, assume that (3) $\Rightarrow$ (2) for $(n-1) \times (n-1)$ symmetric matrices. By Theorem 5.5.7 let $\lambda_1$ be a (real) eigenvalue of $A$, and let $A\mathbf{x}_1 = \lambda_1\mathbf{x}_1$, where $\|\mathbf{x}_1\| = 1$. Use the Gram-Schmidt algorithm to find an orthonormal basis $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ for $\mathbb{R}^n$. Let $P_1 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}$, so $P_1$ is an orthogonal matrix and $P_1^TAP_1 = \begin{bmatrix} \lambda_1 & B \\ 0 & A_1 \end{bmatrix}$ in block form by Lemma 5.5.2. But $P_1^TAP_1$ is symmetric ($A$ is), so it follows that $B = 0$ and $A_1$ is symmetric. Then, by induction, there exists an $(n-1) \times (n-1)$ orthogonal matrix $Q$ such that $Q^TA_1Q = D_1$ is diagonal. Observe that $P_2 = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}$ is orthogonal, and compute:

$$\begin{aligned}(P_1P_2)^TA(P_1P_2) &= P_2^T(P_1^TAP_1)P_2 \\ &= \begin{bmatrix} 1 & 0 \\ 0 & Q^T \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & 0 \\ 0 & D_1 \end{bmatrix}\end{aligned}$$

is diagonal. Because $P_1P_2$ is orthogonal, this proves (2). □

A set of orthonormal eigenvectors of a symmetric matrix $A$ is called a set of **principal axes** for $A$. The name comes from geometry, and this is discussed in Section 8.9. Because the eigenvalues of a (real) symmetric matrix are real, Theorem 8.2.2 is also called the **real spectral theorem**, and the set of distinct eigenvalues is called the **spectrum** of the matrix. In full generality, the spectral theorem is a similar result for matrices with complex entries (Theorem 8.7.8).

**Example 8.2.4**

Find an orthogonal matrix $P$ such that $P^{-1}AP$ is diagonal, where $A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ -1 & 2 & 5 \end{bmatrix}$.

**Solution.** The characteristic polynomial of $A$ is (adding twice row 1 to row 2):

$$c_A(x) = \det\begin{bmatrix} x-1 & 0 & 1 \\ 0 & x-1 & -2 \\ 1 & -2 & x-5 \end{bmatrix} = x(x-1)(x-6)$$

Thus the eigenvalues are $\lambda = 0$, 1, and 6, and corresponding eigenvectors are

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} -1 \\ 2 \\ 5 \end{bmatrix}$$

respectively. Moreover, by what appears to be remarkably good luck, these eigenvectors are *orthogonal*. We have $\|\mathbf{x}_1\|^2 = 6$, $\|\mathbf{x}_2\|^2 = 5$, and $\|\mathbf{x}_3\|^2 = 30$, so

$$P = \begin{bmatrix} \frac{1}{\sqrt{6}}\mathbf{x}_1 & \frac{1}{\sqrt{5}}\mathbf{x}_2 & \frac{1}{\sqrt{30}}\mathbf{x}_3 \end{bmatrix} = \frac{1}{\sqrt{30}} \begin{bmatrix} \sqrt{5} & 2\sqrt{6} & -1 \\ -2\sqrt{5} & \sqrt{6} & 2 \\ \sqrt{5} & 0 & 5 \end{bmatrix}$$

is an orthogonal matrix. Thus $P^{-1} = P^T$ and

$$P^T A P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \end{bmatrix}$$

by the diagonalization algorithm.

Actually, the fact that the eigenvectors in Example 8.2.4 are orthogonal is no coincidence. Theorem 5.5.4 guarantees they are linearly independent (they correspond to distinct eigenvalues); the fact that the matrix is *symmetric* implies that they are orthogonal. To prove this we need the following useful fact about symmetric matrices.

**Theorem 8.2.3**

*If $A$ is an $n \times n$ symmetric matrix, then*

$$(A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$$

*for all columns $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.*[3]

**Proof.** Recall that $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y}$ for all columns $\mathbf{x}$ and $\mathbf{y}$. Because $A^T = A$, we get

$$(A\mathbf{x}) \cdot \mathbf{y} = (A\mathbf{x})^T\mathbf{y} = \mathbf{x}^T A^T\mathbf{y} = \mathbf{x}^T A\mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$$

□

**Theorem 8.2.4**

*If $A$ is a symmetric matrix, then eigenvectors of $A$ corresponding to distinct eigenvalues are orthogonal.*

**Proof.** Let $A\mathbf{x} = \lambda\mathbf{x}$ and $A\mathbf{y} = \mu\mathbf{y}$, where $\lambda \neq \mu$. Using Theorem 8.2.3, we compute

$$\lambda(\mathbf{x} \cdot \mathbf{y}) = (\lambda\mathbf{x}) \cdot \mathbf{y} = (A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y}) = \mathbf{x} \cdot (\mu\mathbf{y}) = \mu(\mathbf{x} \cdot \mathbf{y})$$

Hence $(\lambda - \mu)(\mathbf{x} \cdot \mathbf{y}) = 0$, and so $\mathbf{x} \cdot \mathbf{y} = 0$ because $\lambda \neq \mu$. □

Now the procedure for diagonalizing a symmetric $n \times n$ matrix is clear. Find the distinct eigenvalues (all real by Theorem 5.5.7) and find orthonormal bases for each eigenspace (the Gram-Schmidt algorithm

[3] The converse also holds (Exercise 8.2.15).

may be needed). Then the set of all these basis vectors is orthonormal (by Theorem 8.2.4) and contains $n$ vectors. Here is an example.

**Example 8.2.5**

Orthogonally diagonalize the symmetric matrix $A = \begin{bmatrix} 8 & -2 & 2 \\ -2 & 5 & 4 \\ 2 & 4 & 5 \end{bmatrix}$.

**Solution.** The characteristic polynomial is

$$c_A(x) = \det \begin{bmatrix} x-8 & 2 & -2 \\ 2 & x-5 & -4 \\ -2 & -4 & x-5 \end{bmatrix} = x(x-9)^2$$

Hence the distinct eigenvalues are 0 and 9 of multiplicities 1 and 2, respectively, so $\dim(E_0) = 1$ and $\dim(E_9) = 2$ by Theorem 5.5.6 ($A$ is diagonalizable, being symmetric). Gaussian elimination gives

$$E_0(A) = \text{span}\{\mathbf{x}_1\}, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}, \quad \text{and} \quad E_9(A) = \text{span}\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \right\}$$

The eigenvectors in $E_9$ are both orthogonal to $\mathbf{x}_1$ as Theorem 8.2.4 guarantees, but not to each other. However, the Gram-Schmidt process yields an orthogonal basis

$$\{\mathbf{x}_2, \mathbf{x}_3\} \text{ of } E_9(A) \quad \text{where} \quad \mathbf{x}_2 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} \text{ and } \mathbf{x}_3 = \begin{bmatrix} 2 \\ 4 \\ 5 \end{bmatrix}$$

Normalizing gives orthonormal vectors $\{\frac{1}{3}\mathbf{x}_1, \frac{1}{\sqrt{5}}\mathbf{x}_2, \frac{1}{3\sqrt{5}}\mathbf{x}_3\}$, so

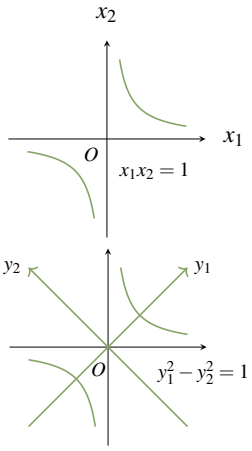
$$P = \begin{bmatrix} \frac{1}{3}\mathbf{x}_1 & \frac{1}{\sqrt{5}}\mathbf{x}_2 & \frac{1}{3\sqrt{5}}\mathbf{x}_3 \end{bmatrix} = \frac{1}{3\sqrt{5}} \begin{bmatrix} \sqrt{5} & -6 & 2 \\ 2\sqrt{5} & 3 & 4 \\ -2\sqrt{5} & 0 & 5 \end{bmatrix}$$

is an orthogonal matrix such that $P^{-1}AP$ is diagonal.

It is worth noting that other, more convenient, diagonalizing matrices $P$ exist. For example, $\mathbf{y}_2 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$ and $\mathbf{y}_3 = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix}$ lie in $E_9(A)$ and they are orthogonal. Moreover, they both have norm 3 (as does $\mathbf{x}_1$), so

$$Q = \begin{bmatrix} \frac{1}{3}\mathbf{x}_1 & \frac{1}{3}\mathbf{y}_2 & \frac{1}{3}\mathbf{y}_3 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{bmatrix}$$

is a nicer orthogonal matrix with the property that $Q^{-1}AQ$ is diagonal.




If $A$ is symmetric and a set of orthogonal eigenvectors of $A$ is given, the eigenvectors are called principal axes of $A$. The name comes from geometry. An expression $q = ax_1^2 + bx_1x_2 + cx_2^2$ is called a **quadratic form** in the variables $x_1$ and $x_2$, and the graph of the equation $q = 1$ is called a **conic** in these variables. For example, if $q = x_1x_2$, the graph of $q = 1$ is given in the first diagram.

But if we introduce new variables $y_1$ and $y_2$ by setting $x_1 = y_1 + y_2$ and $x_2 = y_1 - y_2$, then $q$ becomes $q = y_1^2 - y_2^2$, a diagonal form with no cross term $y_1y_2$ (see the second diagram). Because of this, the $y_1$ and $y_2$ axes are called the principal axes for the conic (hence the name). Orthogonal diagonalization provides a systematic method for finding principal axes. Here is an illustration.

**Example 8.2.6**

Find principal axes for the quadratic form $q = x_1^2 - 4x_1x_2 + x_2^2$.

**Solution.** In order to utilize diagonalization, we first express $q$ in matrix form. Observe that

$$q = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -4 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The matrix here is not symmetric, but we can remedy that by writing

$$q = x_1^2 - 2x_1x_2 - 2x_2x_1 + x_2^2$$

Then we have

$$q = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}^T A \mathbf{x}$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $A = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$ is symmetric. The eigenvalues of $A$ are $\lambda_1 = 3$ and $\lambda_2 = -1$, with corresponding (orthogonal) eigenvectors $\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Since $\|\mathbf{x}_1\| = \|\mathbf{x}_2\| = \sqrt{2}$, so

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \text{ is orthogonal and } P^T A P = D = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$$

Now define new variables $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{y}$ by $\mathbf{y} = P^T\mathbf{x}$, equivalently $\mathbf{x} = P\mathbf{y}$ (since $P^{-1} = P^T$). Hence

$$y_1 = \frac{1}{\sqrt{2}}(x_1 - x_2) \quad \text{and} \quad y_2 = \frac{1}{\sqrt{2}}(x_1 + x_2)$$

In terms of $y_1$ and $y_2$, $q$ takes the form

$$q = \mathbf{x}^T A \mathbf{x} = (P\mathbf{y})^T A (P\mathbf{y}) = \mathbf{y}^T (P^T A P)\mathbf{y} = \mathbf{y}^T D \mathbf{y} = 3y_1^2 - y_2^2$$

Note that $\mathbf{y} = P^T\mathbf{x}$ is obtained from $\mathbf{x}$ by a counterclockwise rotation of $\frac{\pi}{4}$ (see Theorem 2.4.6).

Observe that the quadratic form $q$ in Example 8.2.6 can be diagonalized in other ways. For example

$$q = x_1^2 - 4x_1x_2 + x_2^2 = z_1^2 - \tfrac{1}{3}z_2^2$$

where $z_1 = x_1 - 2x_2$ and $z_2 = 3x_2$. We examine this more carefully in Section 8.9.

If we are willing to replace "diagonal" by "upper triangular" in the principal axes theorem, we can weaken the requirement that $A$ is symmetric to insisting only that $A$ has real eigenvalues.

**Theorem 8.2.5: Triangulation Theorem**

*If $A$ is an $n \times n$ matrix with $n$ real eigenvalues, an orthogonal matrix $P$ exists such that $P^TAP$ is upper triangular.*[4]

**Proof.** We modify the proof of Theorem 8.2.2. If $A\mathbf{x}_1 = \lambda_1\mathbf{x}_1$ where $\|\mathbf{x}_1\| = 1$, let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be an orthonormal basis of $\mathbb{R}^n$, and let $P_1 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}$. Then $P_1$ is orthogonal and $P_1^TAP_1 = \begin{bmatrix} \lambda_1 & B \\ 0 & A_1 \end{bmatrix}$ in block form. By induction, let $Q^TA_1Q = T_1$ be upper triangular where $Q$ is of size $(n-1) \times (n-1)$ and orthogonal. Then $P_2 = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}$ is orthogonal, so $P = P_1P_2$ is also orthogonal and $P^TAP = \begin{bmatrix} \lambda_1 & BQ \\ 0 & T_1 \end{bmatrix}$ is upper triangular. □

The proof of Theorem 8.2.5 gives no way to construct the matrix $P$. However, an algorithm will be given in Section 11.1 where an improved version of Theorem 8.2.5 is presented. In a different direction, a version of Theorem 8.2.5 holds for an arbitrary matrix with complex entries (Schur's theorem in Section 8.7).

As for a diagonal matrix, the eigenvalues of an upper triangular matrix are displayed along the main diagonal. Because $A$ and $P^TAP$ have the same determinant and trace whenever $P$ is orthogonal, Theorem 8.2.5 gives:

**Corollary 8.2.1**

*If $A$ is an $n \times n$ matrix with real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (possibly not all distinct), then* $\det A = \lambda_1\lambda_2\ldots\lambda_n$ *and* $\operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n$.

This corollary remains true even if the eigenvalues are not real (using Schur's theorem).

## Exercises for 8.2

**Exercise 8.2.1** Normalize the rows to make each of the following matrices orthogonal.

a. $A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ b. $A = \begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix}$

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

d. $A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$, $(a, b) \neq (0, 0)$

e. $A = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 2 \end{bmatrix}$

[4] There is also a lower triangular version.

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

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

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

**Exercise 8.2.2** If $P$ is a triangular orthogonal matrix, show that $P$ is diagonal and that all diagonal entries are 1 or $-1$.

**Exercise 8.2.3** If $P$ is orthogonal, show that $kP$ is orthogonal if and only if $k = 1$ or $k = -1$.

**Exercise 8.2.4** If the first two rows of an orthogonal matrix are $(\frac{1}{3}, \frac{2}{3}, \frac{2}{3})$ and $(\frac{2}{3}, \frac{1}{3}, \frac{-2}{3})$, find all possible third rows.

**Exercise 8.2.5** For each matrix $A$, find an orthogonal matrix $P$ such that $P^{-1}AP$ is diagonal.

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

c. $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 5 \end{bmatrix}$ d. $A = \begin{bmatrix} 3 & 0 & 7 \\ 0 & 5 & 0 \\ 7 & 0 & 3 \end{bmatrix}$

e. $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ f. $A = \begin{bmatrix} 5 & -2 & -4 \\ -2 & 8 & -2 \\ -4 & -2 & 5 \end{bmatrix}$

g. $A = \begin{bmatrix} 5 & 3 & 0 & 0 \\ 3 & 5 & 0 & 0 \\ 0 & 0 & 7 & 1 \\ 0 & 0 & 1 & 7 \end{bmatrix}$

h. $A = \begin{bmatrix} 3 & 5 & -1 & 1 \\ 5 & 3 & 1 & -1 \\ -1 & 1 & 3 & 5 \\ 1 & -1 & 5 & 3 \end{bmatrix}$

**Exercise 8.2.6** Consider $A = \begin{bmatrix} 0 & a & 0 \\ a & 0 & c \\ 0 & c & 0 \end{bmatrix}$ where one of $a$, $c \neq 0$. Show that $c_A(x) = x(x-k)(x+k)$, where $k = \sqrt{a^2 + c^2}$ and find an orthogonal matrix $P$ such that $P^{-1}AP$ is diagonal.

**Exercise 8.2.7** Consider $A = \begin{bmatrix} 0 & 0 & a \\ 0 & b & 0 \\ a & 0 & 0 \end{bmatrix}$. Show that $c_A(x) = (x-b)(x-a)(x+a)$ and find an orthogonal matrix $P$ such that $P^{-1}AP$ is diagonal.

**Exercise 8.2.8** Given $A = \begin{bmatrix} b & a \\ a & b \end{bmatrix}$, show that $c_A(x) = (x-a-b)(x+a-b)$ and find an orthogonal matrix $P$ such that $P^{-1}AP$ is diagonal.

**Exercise 8.2.9** Consider $A = \begin{bmatrix} b & 0 & a \\ 0 & b & 0 \\ a & 0 & b \end{bmatrix}$. Show that $c_A(x) = (x-b)(x-b-a)(x-b+a)$ and find an orthogonal matrix $P$ such that $P^{-1}AP$ is diagonal.

**Exercise 8.2.10** In each case find new variables $y_1$ and $y_2$ that diagonalize the quadratic form $q$.

a. $q = x_1^2 + 6x_1x_2 + x_2^2$ b. $q = x_1^2 + 4x_1x_2 - 2x_2^2$

**Exercise 8.2.11** Show that the following are equivalent for a symmetric matrix $A$.

a. $A$ is orthogonal. b. $A^2 = I$.

c. All eigenvalues of $A$ are $\pm 1$.

[Hint: For (b) if and only if (c), use Theorem 8.2.2.]

**Exercise 8.2.12** We call matrices $A$ and $B$ **orthogonally similar** (and write $A \overset{\circ}{\sim} B$) if $B = P^TAP$ for an orthogonal matrix $P$.

a. Show that $A \overset{\circ}{\sim} A$ for all $A$; $A \overset{\circ}{\sim} B \Rightarrow B \overset{\circ}{\sim} A$; and $A \overset{\circ}{\sim} B$ and $B \overset{\circ}{\sim} C \Rightarrow A \overset{\circ}{\sim} C$.

b. Show that the following are equivalent for two symmetric matrices $A$ and $B$.

   i. $A$ and $B$ are similar.

   ii. $A$ and $B$ are orthogonally similar.

   iii. $A$ and $B$ have the same eigenvalues.

**Exercise 8.2.13** Assume that $A$ and $B$ are orthogonally similar (Exercise 8.2.12).

a. If $A$ and $B$ are invertible, show that $A^{-1}$ and $B^{-1}$ are orthogonally similar.

b. Show that $A^2$ and $B^2$ are orthogonally similar.

c. Show that, if $A$ is symmetric, so is $B$.

**Exercise 8.2.14** If $A$ is symmetric, show that every eigenvalue of $A$ is nonnegative if and only if $A = B^2$ for some symmetric matrix $B$.

**Exercise 8.2.15** Prove the converse of Theorem 8.2.3:

If $(A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$ for all $n$-columns $\mathbf{x}$ and $\mathbf{y}$, then $A$ is symmetric.

**Exercise 8.2.16** Show that every eigenvalue of $A$ is zero if and only if $A$ is nilpotent ($A^k = 0$ for some $k \geq 1$).

**Exercise 8.2.17** If $A$ has real eigenvalues, show that $A = B + C$ where $B$ is symmetric and $C$ is nilpotent.
[*Hint*: Theorem 8.2.5.]

**Exercise 8.2.18** Let $P$ be an orthogonal matrix.

a. Show that $\det P = 1$ or $\det P = -1$.

b. Give $2 \times 2$ examples of $P$ such that $\det P = 1$ and $\det P = -1$.

c. If $\det P = -1$, show that $I + P$ has no inverse. [*Hint*: $P^T(I + P) = (I + P)^T$.]

d. If $P$ is $n \times n$ and $\det P \neq (-1)^n$, show that $I - P$ has no inverse.

[*Hint*: $P^T(I - P) = -(I - P)^T$.]

**Exercise 8.2.19** We call a square matrix $E$ a **projection matrix** if $E^2 = E = E^T$. (See Exercise 8.1.17.)

a. If $E$ is a projection matrix, show that $P = I - 2E$ is orthogonal and symmetric.

b. If $P$ is orthogonal and symmetric, show that $E = \frac{1}{2}(I - P)$ is a projection matrix.

c. If $U$ is $m \times n$ and $U^T U = I$ (for example, a unit column in $\mathbb{R}^n$), show that $E = UU^T$ is a projection matrix.

**Exercise 8.2.20** A matrix that we obtain from the identity matrix by writing its rows in a different order is called a **permutation matrix**. Show that every permutation matrix is orthogonal.

**Exercise 8.2.21** If the rows $\mathbf{r}_1, \ldots, \mathbf{r}_n$ of the $n \times n$ matrix $A = [a_{ij}]$ are orthogonal, show that the $(i, j)$-entry of $A^{-1}$ is $\frac{a_{ji}}{\|\mathbf{r}_j\|^2}$.

**Exercise 8.2.22**

a. Let $A$ be an $m \times n$ matrix. Show that the following are equivalent.

i. $A$ has orthogonal rows.

ii. $A$ can be factored as $A = DP$, where $D$ is invertible and diagonal and $P$ has orthonormal rows.

iii. $AA^T$ is an invertible, diagonal matrix.

b. Show that an $n \times n$ matrix $A$ has orthogonal rows if and only if $A$ can be factored as $A = DP$, where $P$ is orthogonal and $D$ is diagonal and invertible.

**Exercise 8.2.23** Let $A$ be a skew-symmetric matrix; that is, $A^T = -A$. Assume that $A$ is an $n \times n$ matrix.

a. Show that $I + A$ is invertible. [*Hint*: By Theorem 2.4.5, it suffices to show that $(I + A)\mathbf{x} = \mathbf{0}$, $\mathbf{x}$ in $\mathbb{R}^n$, implies $\mathbf{x} = \mathbf{0}$. Compute $\mathbf{x} \cdot \mathbf{x} = \mathbf{x}^T\mathbf{x}$, and use the fact that $A\mathbf{x} = -\mathbf{x}$ and $A^2\mathbf{x} = \mathbf{x}$.]

b. Show that $P = (I - A)(I + A)^{-1}$ is orthogonal.

c. Show that every orthogonal matrix $P$ such that $I + P$ is invertible arises as in part (b) from some skew-symmetric matrix $A$.
[*Hint*: Solve $P = (I - A)(I + A)^{-1}$ for $A$.]

**Exercise 8.2.24** Show that the following are equivalent for an $n \times n$ matrix $P$.

a. $P$ is orthogonal.

b. $\|P\mathbf{x}\| = \|\mathbf{x}\|$ for all columns $\mathbf{x}$ in $\mathbb{R}^n$.

c. $\|P\mathbf{x} - P\mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|$ for all columns $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.

d. $(P\mathbf{x}) \cdot (P\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ for all columns $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.

[*Hints*: For (c) $\Rightarrow$ (d), see Exercise 5.3.14(a). For (d) $\Rightarrow$ (a), show that column $i$ of $P$ equals $P\mathbf{e}_i$, where $\mathbf{e}_i$ is column $i$ of the identity matrix.]

**Exercise 8.2.25** Show that every $2 \times 2$ orthogonal matrix has the form $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ or $\begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$ for some angle $\theta$.

[*Hint*: If $a^2 + b^2 = 1$, then $a = \cos\theta$ and $b = \sin\theta$ for some angle $\theta$.]

**Exercise 8.2.26** Use Theorem 8.2.5 to show that every symmetric matrix is orthogonally diagonalizable.

## 8.3 Positive Definite Matrices

All the eigenvalues of any symmetric matrix are real; this section is about the case in which the eigenvalues are positive. These matrices, which arise whenever optimization (maximum and minimum) problems are encountered, have countless applications throughout science and engineering. They also arise in statistics (for example, in factor analysis used in the social sciences) and in geometry (see Section 8.9). We will encounter them again in Chapter 10 when describing all inner products in $\mathbb{R}^n$.

**Definition 8.5 Positive Definite Matrices**

*A square matrix is called* ***positive definite*** *if it is symmetric and all its eigenvalues $\lambda$ are positive, that is $\lambda > 0$.*

Because these matrices are symmetric, the principal axes theorem plays a central role in the theory.

**Theorem 8.3.1**

*If $A$ is positive definite, then it is invertible and $\det A > 0$.*

**Proof.** If $A$ is $n \times n$ and the eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_n$, then $\det A = \lambda_1\lambda_2\cdots\lambda_n > 0$ by the principal axes theorem (or the corollary to Theorem 8.2.5). □

If $\mathbf{x}$ is a column in $\mathbb{R}^n$ and $A$ is any real $n \times n$ matrix, we view the $1 \times 1$ matrix $\mathbf{x}^T A\mathbf{x}$ as a real number. With this convention, we have the following characterization of positive definite matrices.

**Theorem 8.3.2**

*A symmetric matrix $A$ is positive definite if and only if $\mathbf{x}^T A\mathbf{x} > 0$ for every column $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{R}^n$.*

**Proof.** $A$ is symmetric so, by the principal axes theorem, let $P^T AP = D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ where $P^{-1} = P^T$ and the $\lambda_i$ are the eigenvalues of $A$. Given a column $\mathbf{x}$ in $\mathbb{R}^n$, write $\mathbf{y} = P^T\mathbf{x} = \begin{bmatrix} y_1 & y_2 & \ldots & y_n \end{bmatrix}^T$. Then

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T(PDP^T)\mathbf{x} = \mathbf{y}^T D\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \tag{8.3}$$

If $A$ is positive definite and $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{x}^T A\mathbf{x} > 0$ by (8.3) because some $y_j \neq 0$ and every $\lambda_i > 0$. Conversely, if $\mathbf{x}^T A\mathbf{x} > 0$ whenever $\mathbf{x} \neq \mathbf{0}$, let $\mathbf{x} = P\mathbf{e}_j \neq \mathbf{0}$ where $\mathbf{e}_j$ is column $j$ of $I_n$. Then $\mathbf{y} = \mathbf{e}_j$, so (8.3) reads $\lambda_j = \mathbf{x}^T A\mathbf{x} > 0$. □

Note that Theorem 8.3.2 shows that the positive definite matrices are exactly the symmetric matrices $A$ for which the quadratic form $q = \mathbf{x}^T A\mathbf{x}$ takes only positive values.

**Example 8.3.1**

If $U$ is any invertible $n \times n$ matrix, show that $A = U^T U$ is positive definite.

**Solution.** If $\mathbf{x}$ is in $\mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{0}$, then

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T(U^T U)\mathbf{x} = (U\mathbf{x})^T(U\mathbf{x}) = \|U\mathbf{x}\|^2 > 0$$

because $U\mathbf{x} \neq \mathbf{0}$ ($U$ is invertible). Hence Theorem 8.3.2 applies.

It is remarkable that the converse to Example 8.3.1 is also true. In fact every positive definite matrix $A$ can be factored as $A = U^T U$ where $U$ is an upper triangular matrix with positive elements on the main diagonal. However, before verifying this, we introduce another concept that is central to any discussion of positive definite matrices.

If $A$ is any $n \times n$ matrix, let ${}^{(r)}A$ denote the $r \times r$ submatrix in the upper left corner of $A$; that is, ${}^{(r)}A$ is the matrix obtained from $A$ by deleting the last $n - r$ rows and columns. The matrices ${}^{(1)}A$, ${}^{(2)}A$, ${}^{(3)}A$, $\ldots$, ${}^{(n)}A = A$ are called the **principal submatrices** of $A$.

**Example 8.3.2**

If $A = \begin{bmatrix} 10 & 5 & 2 \\ 5 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$ then ${}^{(1)}A = [10]$, ${}^{(2)}A = \begin{bmatrix} 10 & 5 \\ 5 & 3 \end{bmatrix}$ and ${}^{(3)}A = A$.

**Lemma 8.3.1**

*If $A$ is positive definite, so is each principal submatrix ${}^{(r)}A$ for $r = 1, 2, \ldots, n$.*

**Proof.** Write $A = \begin{bmatrix} {}^{(r)}A & P \\ Q & R \end{bmatrix}$ in block form. If $\mathbf{y} \neq \mathbf{0}$ in $\mathbb{R}^r$, write $\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$ in $\mathbb{R}^n$.

Then $\mathbf{x} \neq \mathbf{0}$, so the fact that $A$ is positive definite gives

$$0 < \mathbf{x}^T A\mathbf{x} = \begin{bmatrix} \mathbf{y}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} {}^{(r)}A & P \\ Q & R \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \mathbf{y}^T({}^{(r)}A)\mathbf{y}$$

This shows that ${}^{(r)}A$ is positive definite by Theorem 8.3.2.[5] □

If $A$ is positive definite, Lemma 8.3.1 and Theorem 8.3.1 show that $\det({}^{(r)}A) > 0$ for every $r$. This proves part of the following theorem which contains the converse to Example 8.3.1, and characterizes the positive definite matrices among the symmetric ones.

[5] A similar argument shows that, if $B$ is any matrix obtained from a positive definite matrix $A$ by deleting certain rows and deleting the same columns, then $B$ is also positive definite.

**Theorem 8.3.3**

*The following conditions are equivalent for a symmetric $n \times n$ matrix $A$:*

1. *$A$ is positive definite.*
2. *$\det({}^{(r)}A) > 0$ for each $r = 1, 2, \ldots, n$.*
3. *$A = U^T U$ where $U$ is an upper triangular matrix with positive entries on the main diagonal.*

*Furthermore, the factorization in (3) is unique (called the **Cholesky factorization**[6] of $A$).*

**Proof.** First, (3) $\Rightarrow$ (1) by Example 8.3.1, and (1) $\Rightarrow$ (2) by Lemma 8.3.1 and Theorem 8.3.1.

(2) $\Rightarrow$ (3). Assume (2) and proceed by induction on $n$. If $n = 1$, then $A = [a]$ where $a > 0$ by (2), so take $U = [\sqrt{a}]$. If $n > 1$, write $B = {}^{(n-1)}A$. Then $B$ is symmetric and satisfies (2) so, by induction, we have $B = U^T U$ as in (3) where $U$ is of size $(n-1) \times (n-1)$. Then, as $A$ is symmetric, it has block form $A = \begin{bmatrix} B & \mathbf{p} \\ \mathbf{p}^T & b \end{bmatrix}$ where $\mathbf{p}$ is a column in $\mathbb{R}^{n-1}$ and $b$ is in $\mathbb{R}$. If we write $\mathbf{x} = (U^T)^{-1}\mathbf{p}$ and $c = b - \mathbf{x}^T\mathbf{x}$, block multiplication gives

$$A = \begin{bmatrix} U^T U & \mathbf{p} \\ \mathbf{p}^T & b \end{bmatrix} = \begin{bmatrix} U^T & 0 \\ \mathbf{x}^T & 1 \end{bmatrix} \begin{bmatrix} U & \mathbf{x} \\ 0 & c \end{bmatrix}$$

as the reader can verify. Taking determinants and applying Theorem 3.1.5 gives $\det A = \det(U^T) \det U \cdot c = c(\det U)^2$. Hence $c > 0$ because $\det A > 0$ by (2), so the above factorization can be written

$$A = \begin{bmatrix} U^T & 0 \\ \mathbf{x}^T & \sqrt{c} \end{bmatrix} \begin{bmatrix} U & \mathbf{x} \\ 0 & \sqrt{c} \end{bmatrix}$$

Since $U$ has positive diagonal entries, this proves (3).

As to the uniqueness, suppose that $A = U^T U = U_1^T U_1$ are two Cholesky factorizations. Now write $D = UU_1^{-1} = (U^T)^{-1}U_1^T$. Then $D$ is upper triangular, because $D = UU_1^{-1}$, and lower triangular, because $D = (U^T)^{-1}U_1^T$, and so it is a diagonal matrix. Thus $U = DU_1$ and $U_1 = DU$, so it suffices to show that $D = I$. But eliminating $U_1$ gives $U = D^2U$, so $D^2 = I$ because $U$ is invertible. Since the diagonal entries of $D$ are positive (this is true of $U$ and $U_1$), it follows that $D = I$. □

The remarkable thing is that the matrix $U$ in the Cholesky factorization is easy to obtain from $A$ using row operations. The key is that Step 1 of the following algorithm is *possible* for any positive definite matrix $A$. A proof of the algorithm is given following Example 8.3.3.

**Algorithm for the Cholesky Factorization**

*If $A$ is a positive definite matrix, the Cholesky factorization $A = U^T U$ can be obtained as follows:*

*Step 1.* *Carry $A$ to an upper triangular matrix $U_1$ with positive diagonal entries using row operations each of which adds a multiple of a row to a lower row.*

*Step 2.* *Obtain $U$ from $U_1$ by dividing each row of $U_1$ by the square root of the diagonal entry in that row.*

[6] Andre-Louis Cholesky (1875–1918), was a French mathematician who died in World War I. His factorization was published in 1924 by a fellow officer.

**Example 8.3.3**

Find the Cholesky factorization of $A = \begin{bmatrix} 10 & 5 & 2 \\ 5 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$.

**Solution.** The matrix $A$ is positive definite by Theorem 8.3.3 because $\det {}^{(1)}A = 10 > 0$, $\det {}^{(2)}A = 5 > 0$, and $\det {}^{(3)}A = \det A = 3 > 0$. Hence Step 1 of the algorithm is carried out as follows:

$$A = \begin{bmatrix} 10 & 5 & 2 \\ 5 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 10 & 5 & 2 \\ 0 & \frac{1}{2} & 1 \\ 0 & 1 & \frac{13}{5} \end{bmatrix} \rightarrow \begin{bmatrix} 10 & 5 & 2 \\ 0 & \frac{1}{2} & 1 \\ 0 & 0 & \frac{3}{5} \end{bmatrix} = U_1$$

Now carry out Step 2 on $U_1$ to obtain $U = \begin{bmatrix} \sqrt{10} & \frac{5}{\sqrt{10}} & \frac{2}{\sqrt{10}} \\ 0 & \frac{1}{\sqrt{2}} & \sqrt{2} \\ 0 & 0 & \frac{\sqrt{3}}{\sqrt{5}} \end{bmatrix}$.

The reader can verify that $U^TU = A$.

**Proof of the Cholesky Algorithm.** If $A$ is positive definite, let $A = U^TU$ be the Cholesky factorization, and let $D = \text{diag}\,(d_1, \ldots, d_n)$ be the common diagonal of $U$ and $U^T$. Then $U^TD^{-1}$ is lower triangular with ones on the diagonal (call such matrices LT-1). Hence $L = (U^TD^{-1})^{-1}$ is also LT-1, and so $I_n \rightarrow L$ by a sequence of row operations each of which adds a multiple of a row to a lower row (verify; modify columns right to left). But then $A \rightarrow LA$ by the same sequence of row operations (see the discussion preceding Theorem 2.5.1). Since $LA = [D(U^T)^{-1}][U^TU] = DU$ is upper triangular with positive entries on the diagonal, this shows that Step 1 of the algorithm is possible.

Turning to Step 2, let $A \rightarrow U_1$ as in Step 1 so that $U_1 = L_1A$ where $L_1$ is LT-1. Since A is symmetric, we get

$$L_1U_1^T = L_1(L_1A)^T = L_1A^TL_1^T = L_1AL_1^T = U_1L_1^T \tag{8.4}$$

Let $D_1 = \text{diag}\,(e_1, \ldots, e_n)$ denote the diagonal of $U_1$. Then (8.4) gives $L_1(U_1^TD_1^{-1}) = U_1L_1^TD_1^{-1}$. This is both upper triangular (right side) and LT-1 (left side), and so must equal $I_n$. In particular, $U_1^TD_1^{-1} = L_1^{-1}$. Now let $D_2 = \text{diag}\,(\sqrt{e_1}, \ldots, \sqrt{e_n})$, so that $D_2^2 = D_1$. If we write $U = D_2^{-1}U_1$ we have

$$U^TU = (U_1^TD_2^{-1})(D_2^{-1}U_1) = U_1^T(D_2^2)^{-1}U_1 = (U_1^TD_1^{-1})U_1 = (L_1^{-1})U_1 = A$$

This proves Step 2 because $U = D_2^{-1}U_1$ is formed by dividing each row of $U_1$ by the square root of its diagonal entry (verify). □

## Exercises for 8.3

**Exercise 8.3.1** Find the Cholesky decomposition of each of the following matrices.

a. $\begin{bmatrix} 4 & 3 \\ 3 & 5 \end{bmatrix}$ b. $\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$

c. $\begin{bmatrix} 12 & 4 & 3 \\ 4 & 2 & -1 \\ 3 & -1 & 7 \end{bmatrix}$ d. $\begin{bmatrix} 20 & 4 & 5 \\ 4 & 2 & 3 \\ 5 & 3 & 5 \end{bmatrix}$

**Exercise 8.3.2**

a. If $A$ is positive definite, show that $A^k$ is positive definite for all $k \geq 1$.

b. Prove the converse to (a) when $k$ is odd.

c. Find a symmetric matrix $A$ such that $A^2$ is positive definite but $A$ is not.

**Exercise 8.3.3** Let $A = \begin{bmatrix} 1 & a \\ a & b \end{bmatrix}$. If $a^2 < b$, show that $A$ is positive definite and find the Cholesky factorization.

**Exercise 8.3.4** If $A$ and $B$ are positive definite and $r > 0$, show that $A + B$ and $rA$ are both positive definite.

**Exercise 8.3.5** If $A$ and $B$ are positive definite, show that $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ is positive definite.

**Exercise 8.3.6** If $A$ is an $n \times n$ positive definite matrix and $U$ is an $n \times m$ matrix of rank $m$, show that $U^TAU$ is positive definite.

**Exercise 8.3.7** If $A$ is positive definite, show that each diagonal entry is positive.

**Exercise 8.3.8** Let $A_0$ be formed from $A$ by deleting rows 2 and 4 and deleting columns 2 and 4. If $A$ is positive definite, show that $A_0$ is positive definite.

**Exercise 8.3.9** If $A$ is positive definite, show that $A = CC^T$ where $C$ has orthogonal columns.

**Exercise 8.3.10** If $A$ is positive definite, show that $A = C^2$ where $C$ is positive definite.

**Exercise 8.3.11** Let $A$ be a positive definite matrix. If $a$ is a real number, show that $aA$ is positive definite if and only if $a > 0$.

**Exercise 8.3.12**

a. Suppose an invertible matrix $A$ can be factored in $\mathbf{M}_{nn}$ as $A = LDU$ where $L$ is lower triangular with 1s on the diagonal, $U$ is upper triangular with 1s on the diagonal, and $D$ is diagonal with positive diagonal entries. Show that the factorization is unique: If $A = L_1D_1U_1$ is another such factorization, show that $L_1 = L$, $D_1 = D$, and $U_1 = U$.

b. Show that a matrix $A$ is positive definite if and only if $A$ is symmetric and admits a factorization $A = LDU$ as in (a).

**Exercise 8.3.13** Let $A$ be positive definite and write $d_r = \det {}^{(r)}A$ for each $r = 1, 2, \ldots, n$. If $U$ is the upper triangular matrix obtained in step 1 of the algorithm, show that the diagonal elements $u_{11}, u_{22}, \ldots, u_{nn}$ of $U$ are given by $u_{11} = d_1$, $u_{jj} = d_j/d_{j-1}$ if $j > 1$. [*Hint*: If $LA = U$ where $L$ is lower triangular with 1s on the diagonal, use block multiplication to show that $\det {}^{(r)}A = \det {}^{(r)}U$ for each $r$.]

## 8.4 QR-Factorization[7]

One of the main virtues of orthogonal matrices is that they can be easily inverted—the transpose is the inverse. This fact, combined with the factorization theorem in this section, provides a useful way to simplify many matrix calculations (for example, in least squares approximation).

**Definition 8.6 QR-factorization**

*Let $A$ be an $m \times n$ matrix with independent columns. A* ***QR-factorization*** *of $A$ expresses it as $A = QR$ where $Q$ is $m \times n$ with orthonormal columns and $R$ is an invertible and upper triangular*

[7]This section is not used elsewhere in the book

*matrix with positive diagonal entries.*

The importance of the factorization lies in the fact that there are computer algorithms that accomplish it with good control over round-off error, making it particularly useful in matrix calculations. The factorization is a matrix version of the Gram-Schmidt process.

Suppose $A = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{bmatrix}$ is an $m \times n$ matrix with linearly independent columns $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$. The Gram-Schmidt algorithm can be applied to these columns to provide orthogonal columns $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ where $\mathbf{f}_1 = \mathbf{c}_1$ and

$$\mathbf{f}_k = \mathbf{c}_k - \frac{\mathbf{c}_k \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{c}_k \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\mathbf{c}_k \cdot \mathbf{f}_{k-1}}{\|\mathbf{f}_{k-1}\|^2}\mathbf{f}_{k-1}$$

for each $k = 2, 3, \ldots, n$. Now write $\mathbf{q}_k = \frac{1}{\|\mathbf{f}_k\|}\mathbf{f}_k$ for each $k$. Then $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ are orthonormal columns, and the above equation becomes

$$\|\mathbf{f}_k\|\mathbf{q}_k = \mathbf{c}_k - (\mathbf{c}_k \cdot \mathbf{q}_1)\mathbf{q}_1 - (\mathbf{c}_k \cdot \mathbf{q}_2)\mathbf{q}_2 - \cdots - (\mathbf{c}_k \cdot \mathbf{q}_{k-1})\mathbf{q}_{k-1}$$

Using these equations, express each $\mathbf{c}_k$ as a linear combination of the $\mathbf{q}_i$:

$$\begin{array}{rcl} \mathbf{c}_1 & = & \|\mathbf{f}_1\|\mathbf{q}_1 \\ \mathbf{c}_2 & = & (\mathbf{c}_2 \cdot \mathbf{q}_1)\mathbf{q}_1 + \|\mathbf{f}_2\|\mathbf{q}_2 \\ \mathbf{c}_3 & = & (\mathbf{c}_3 \cdot \mathbf{q}_1)\mathbf{q}_1 + (\mathbf{c}_3 \cdot \mathbf{q}_2)\mathbf{q}_2 + \|\mathbf{f}_3\|\mathbf{q}_3 \\ \vdots & & \vdots \\ \mathbf{c}_n & = & (\mathbf{c}_n \cdot \mathbf{q}_1)\mathbf{q}_1 + (\mathbf{c}_n \cdot \mathbf{q}_2)\mathbf{q}_2 + (\mathbf{c}_n \cdot \mathbf{q}_3)\mathbf{q}_3 + \cdots + \|\mathbf{f}_n\|\mathbf{q}_n \end{array}$$

These equations have a matrix form that gives the required factorization:

$$\begin{aligned} A &= \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \cdots & \mathbf{c}_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 & \cdots & \mathbf{q}_n \end{bmatrix} \begin{bmatrix} \|\mathbf{f}_1\| & \mathbf{c}_2 \cdot \mathbf{q}_1 & \mathbf{c}_3 \cdot \mathbf{q}_1 & \cdots & \mathbf{c}_n \cdot \mathbf{q}_1 \\ 0 & \|\mathbf{f}_2\| & \mathbf{c}_3 \cdot \mathbf{q}_2 & \cdots & \mathbf{c}_n \cdot \mathbf{q}_2 \\ 0 & 0 & \|\mathbf{f}_3\| & \cdots & \mathbf{c}_n \cdot \mathbf{q}_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \|\mathbf{f}_n\| \end{bmatrix} \end{aligned} \tag{8.5}$$

Here the first factor $Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 & \cdots & \mathbf{q}_n \end{bmatrix}$ has orthonormal columns, and the second factor is an $n \times n$ upper triangular matrix $R$ with positive diagonal entries (and so is invertible). We record this in the following theorem.

**Theorem 8.4.1: QR-Factorization**

*Every $m \times n$ matrix $A$ with linearly independent columns has a QR-factorization $A = QR$ where $Q$ has orthonormal columns and $R$ is upper triangular with positive diagonal entries.*

The matrices $Q$ and $R$ in Theorem 8.4.1 are uniquely determined by $A$; we return to this below.

**Example 8.4.1**

Find the QR-factorization of $A = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.

**Solution.** Denote the columns of $A$ as $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}_3$, and observe that $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ is independent. If we apply the Gram-Schmidt algorithm to these columns, the result is:

$$\mathbf{f}_1 = \mathbf{c}_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{f}_2 = \mathbf{c}_2 - \tfrac{1}{2}\mathbf{f}_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \text{ and } \quad \mathbf{f}_3 = \mathbf{c}_3 + \tfrac{1}{2}\mathbf{f}_1 - \mathbf{f}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Write $\mathbf{q}_j = \frac{1}{\|\mathbf{f}_j\|^2}\mathbf{f}_j$ for each $j$, so $\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\}$ is orthonormal. Then equation (8.5) preceding Theorem 8.4.1 gives $A = QR$ where

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & 0 \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{2}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{3} & 1 & 0 \\ -\sqrt{3} & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \sqrt{6} \end{bmatrix}$$

$$R = \begin{bmatrix} \|\mathbf{f}_1\| & \mathbf{c}_2 \cdot \mathbf{q}_1 & \mathbf{c}_3 \cdot \mathbf{q}_1 \\ 0 & \|\mathbf{f}_2\| & \mathbf{c}_3 \cdot \mathbf{q}_2 \\ 0 & 0 & \|\mathbf{f}_3\| \end{bmatrix} = \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{2}} \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 1 & -1 \\ 0 & \sqrt{3} & \sqrt{3} \\ 0 & 0 & \sqrt{2} \end{bmatrix}$$

The reader can verify that indeed $A = QR$.

If a matrix $A$ has independent rows and we apply QR-factorization to $A^T$, the result is:

**Corollary 8.4.1**

*If $A$ has independent rows, then $A$ factors uniquely as $A = LP$ where $P$ has orthonormal rows and $L$ is an invertible lower triangular matrix with positive main diagonal entries.*

Since a square matrix with orthonormal columns is orthogonal, we have

**Theorem 8.4.2**

*Every square, invertible matrix $A$ has factorizations $A = QR$ and $A = LP$ where $Q$ and $P$ are orthogonal, $R$ is upper triangular with positive diagonal entries, and $L$ is lower triangular with positive diagonal entries.*

***Remark***

In Section 5.6 we found how to find a best approximation $\mathbf{z}$ to a solution of a (possibly inconsistent) system $A\mathbf{x} = \mathbf{b}$ of linear equations: take $\mathbf{z}$ to be any solution of the "normal" equations $(A^TA)\mathbf{z} = A^T\mathbf{b}$. If $A$ has independent columns this $\mathbf{z}$ is unique ($A^TA$ is invertible by Theorem 5.4.3), so it is often desirable to compute $(A^TA)^{-1}$. This is particularly useful in least squares approximation (Section 5.6). This is simplified if we have a QR-factorization of $A$ (and is one of the main reasons for the importance of Theorem 8.4.1). For if $A = QR$ is such a factorization, then $Q^TQ = I_n$ because $Q$ has orthonormal columns (verify), so we obtain

$$A^TA = R^TQ^TQR = R^TR$$

Hence computing $(A^TA)^{-1}$ amounts to finding $R^{-1}$, and this is a routine matter because $R$ is upper triangular. Thus the difficulty in computing $(A^TA)^{-1}$ lies in obtaining the QR-factorization of $A$.

We conclude by proving the uniqueness of the QR-factorization.

**Theorem 8.4.3**

*Let $A$ be an $m \times n$ matrix with independent columns. If $A = QR$ and $A = Q_1R_1$ are QR-factorizations of $A$, then $Q_1 = Q$ and $R_1 = R$.*

**Proof.** Write $Q = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{bmatrix}$ and $Q_1 = \begin{bmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_n \end{bmatrix}$ in terms of their columns, and observe first that $Q^TQ = I_n = Q_1^TQ_1$ because $Q$ and $Q_1$ have orthonormal columns. Hence it suffices to show that $Q_1 = Q$ (then $R_1 = Q_1^TA = Q^TA = R$). Since $Q_1^TQ_1 = I_n$, the equation $QR = Q_1R_1$ gives $Q_1^TQ = R_1R^{-1}$; for convenience we write this matrix as

$$Q_1^TQ = R_1R^{-1} = \begin{bmatrix} t_{ij} \end{bmatrix}$$

This matrix is upper triangular with positive diagonal elements (since this is true for $R$ and $R_1$), so $t_{ii} > 0$ for each $i$ and $t_{ij} = 0$ if $i > j$. On the other hand, the $(i,\ j)$-entry of $Q_1^TQ$ is $\mathbf{d}_i^T\mathbf{c}_j = \mathbf{d}_i \cdot \mathbf{c}_j$, so we have $\mathbf{d}_i \cdot \mathbf{c}_j = t_{ij}$ for all $i$ and $j$. But each $\mathbf{c}_j$ is in $\text{span}\{\mathbf{d}_1,\ \mathbf{d}_2,\ \ldots,\ \mathbf{d}_n\}$ because $Q = Q_1(R_1R^{-1})$. Hence the expansion theorem gives

$$\mathbf{c}_j = (\mathbf{d}_1 \cdot \mathbf{c}_j)\mathbf{d}_1 + (\mathbf{d}_2 \cdot \mathbf{c}_j)\mathbf{d}_2 + \cdots + (\mathbf{d}_n \cdot \mathbf{c}_j)\mathbf{d}_n = t_{1j}\mathbf{d}_1 + t_{2j}\mathbf{d}_2 + \cdots + t_{jj}\mathbf{d}_i$$

because $\mathbf{d}_i \cdot \mathbf{c}_j = t_{ij} = 0$ if $i > j$. The first few equations here are

$$\begin{aligned} \mathbf{c}_1 &= t_{11}\mathbf{d}_1 \\ \mathbf{c}_2 &= t_{12}\mathbf{d}_1 + t_{22}\mathbf{d}_2 \\ \mathbf{c}_3 &= t_{13}\mathbf{d}_1 + t_{23}\mathbf{d}_2 + t_{33}\mathbf{d}_3 \\ \mathbf{c}_4 &= t_{14}\mathbf{d}_1 + t_{24}\mathbf{d}_2 + t_{34}\mathbf{d}_3 + t_{44}\mathbf{d}_4 \\ \vdots & \quad \vdots \end{aligned}$$

The first of these equations gives $1 = \|\mathbf{c}_1\| = \|t_{11}\mathbf{d}_1\| = |t_{11}|\|\mathbf{d}_1\| = t_{11}$, whence $\mathbf{c}_1 = \mathbf{d}_1$. But then we have $t_{12} = \mathbf{d}_1 \cdot \mathbf{c}_2 = \mathbf{c}_1 \cdot \mathbf{c}_2 = 0$, so the second equation becomes $\mathbf{c}_2 = t_{22}\mathbf{d}_2$. Now a similar argument gives $\mathbf{c}_2 = \mathbf{d}_2$, and then $t_{13} = 0$ and $t_{23} = 0$ follows in the same way. Hence $\mathbf{c}_3 = t_{33}\mathbf{d}_3$ and $\mathbf{c}_3 = \mathbf{d}_3$. Continue in this way to get $\mathbf{c}_i = \mathbf{d}_i$ for all $i$. This means that $Q_1 = Q$, which is what we wanted. □

## Exercises for 8.4

**Exercise 8.4.1** In each case find the QR-factorization of $A$.

a. $A = \begin{bmatrix} 1 & -1 \\ -1 & 0 \end{bmatrix}$ b. $A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$

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

**Exercise 8.4.2** Let $A$ and $B$ denote matrices.

a. If $A$ and $B$ have independent columns, show that $AB$ has independent columns. [*Hint*: Theorem 5.4.3.]

b. Show that $A$ has a QR-factorization if and only if $A$ has independent columns.

c. If $AB$ has a QR-factorization, show that the same is true of $B$ but not necessarily $A$.

[*Hint*: Consider $AA^T$ where $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$.]

**Exercise 8.4.3** If $R$ is upper triangular and invertible, show that there exists a diagonal matrix $D$ with diagonal entries $\pm 1$ such that $R_1 = DR$ is invertible, upper triangular, and has positive diagonal entries.

**Exercise 8.4.4** If $A$ has independent columns, let $A = QR$ where $Q$ has orthonormal columns and $R$ is invertible and upper triangular. [Some authors call *this* a QR-factorization of $A$.] Show that there is a diagonal matrix $D$ with diagonal entries $\pm 1$ such that $A = (QD)(DR)$ is the QR-factorization of $A$. [*Hint*: Preceding exercise.]

## 8.5 Computing Eigenvalues

In practice, the problem of finding eigenvalues of a matrix is virtually never solved by finding the roots of the characteristic polynomial. This is difficult for large matrices and iterative methods are much better. Two such methods are described briefly in this section.

### The Power Method

In Chapter 3 our initial rationale for diagonalizing matrices was to be able to compute the powers of a square matrix, and the eigenvalues were needed to do this. In this section, we are interested in efficiently computing eigenvalues, and it may come as no surprise that the first method we discuss uses the powers of a matrix.

Recall that an eigenvalue $\lambda$ of an $n \times n$ matrix $A$ is called a **dominant eigenvalue** if $\lambda$ has multiplicity 1, and

$$|\lambda| > |\mu| \quad \text{for all eigenvalues } \mu \neq \lambda$$

Any corresponding eigenvector is called a **dominant eigenvector** of $A$. When such an eigenvalue exists, one technique for finding it is as follows: Let $\mathbf{x}_0$ in $\mathbb{R}^n$ be a first approximation to a dominant eigenvector $\lambda$, and compute successive approximations $\mathbf{x}_1, \mathbf{x}_2, \ldots$ as follows:

$$\mathbf{x}_1 = A\mathbf{x}_0 \quad \mathbf{x}_2 = A\mathbf{x}_1 \quad \mathbf{x}_3 = A\mathbf{x}_2 \quad \cdots$$

In general, we define

$$\mathbf{x}_{k+1} = A\mathbf{x}_k \quad \text{for each } k \geq 0$$

If the first estimate $\mathbf{x}_0$ is good enough, these vectors $\mathbf{x}_n$ will approximate the dominant eigenvector $\lambda$ (see below). This technique is called the **power method** (because $\mathbf{x}_k = A^k\mathbf{x}_0$ for each $k \geq 1$). Observe that if $\mathbf{z}$ is any eigenvector corresponding to $\lambda$, then

$$\frac{\mathbf{z} \cdot (A\mathbf{z})}{\|\mathbf{z}\|^2} = \frac{\mathbf{z} \cdot (\lambda\mathbf{z})}{\|\mathbf{z}\|^2} = \lambda$$

Because the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots$ approximate dominant eigenvectors, this suggests that we define the **Rayleigh quotients** as follows:

$$r_k = \frac{\mathbf{x}_k \cdot \mathbf{x}_{k+1}}{\|\mathbf{x}_k\|^2} \quad \text{for } k \geq 1$$

Then the numbers $r_k$ approximate the dominant eigenvalue $\lambda$.

**Example 8.5.1**

Use the power method to approximate a dominant eigenvector and eigenvalue of $A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$.

**Solution.** The eigenvalues of $A$ are 2 and $-1$, with eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -2 \end{bmatrix}$. Take $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ as the first approximation and compute $\mathbf{x}_1, \mathbf{x}_2, \ldots,$ successively, from $\mathbf{x}_1 = A\mathbf{x}_0, \mathbf{x}_2 = A\mathbf{x}_1, \ldots$ . The result is

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \ \mathbf{x}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \ \mathbf{x}_3 = \begin{bmatrix} 5 \\ 6 \end{bmatrix}, \ \mathbf{x}_4 = \begin{bmatrix} 11 \\ 10 \end{bmatrix}, \ \mathbf{x}_3 = \begin{bmatrix} 21 \\ 22 \end{bmatrix}, \ \ldots$$

These vectors are approaching scalar multiples of the dominant eigenvector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Moreover, the Rayleigh quotients are

$$r_1 = \tfrac{7}{5}, \ r_2 = \tfrac{27}{13}, \ r_3 = \tfrac{115}{61}, \ r_4 = \tfrac{451}{221}, \ \ldots$$

and these are approaching the dominant eigenvalue 2.

To see why the power method works, let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be eigenvalues of $A$ with $\lambda_1$ dominant and let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m$ be corresponding eigenvectors. What is required is that the first approximation $\mathbf{x}_0$ be a linear combination of these eigenvectors:

$$\mathbf{x}_0 = a_1\mathbf{y}_1 + a_2\mathbf{y}_2 + \cdots + a_m\mathbf{y}_m \quad \text{with } a_1 \neq 0$$

If $k \geq 1$, the fact that $\mathbf{x}_k = A^k\mathbf{x}_0$ and $A^k\mathbf{y}_i = \lambda_i^k\mathbf{y}_i$ for each $i$ gives

$$\mathbf{x}_k = a_1\lambda_1^k\mathbf{y}_1 + a_2\lambda_2^k\mathbf{y}_2 + \cdots + a_m\lambda_m^k\mathbf{y}_m \quad \text{for } k \geq 1$$

Hence

$$\tfrac{1}{\lambda_1^k}\mathbf{x}_k = a_1\mathbf{y}_1 + a_2\left(\tfrac{\lambda_2}{\lambda_1}\right)^k\mathbf{y}_2 + \cdots + a_m\left(\tfrac{\lambda_m}{\lambda_1}\right)^k\mathbf{y}_m$$

The right side approaches $a_1\mathbf{y}_1$ as $k$ increases because $\lambda_1$ is dominant $\left(\left|\frac{\lambda_i}{\lambda_1}\right| < 1 \text{ for each } i > 1\right)$. Because $a_1 \neq 0$, this means that $\mathbf{x}_k$ approximates the dominant eigenvector $a_1\lambda_1^k\mathbf{y}_1$.

The power method requires that the first approximation $\mathbf{x}_0$ be a linear combination of eigenvectors. (In Example 8.5.1 the eigenvectors form a basis of $\mathbb{R}^2$.) But even in this case the method fails if $a_1 = 0$, where $a_1$ is the coefficient of the dominant eigenvector (try $\mathbf{x}_0 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ in Example 8.5.1). In general, the rate of convergence is quite slow if any of the ratios $\left|\frac{\lambda_i}{\lambda_1}\right|$ is near 1. Also, because the method requires repeated multiplications by $A$, it is not recommended unless these multiplications are easy to carry out (for example, if most of the entries of $A$ are zero).

## QR-Algorithm

A much better method for approximating the eigenvalues of an invertible matrix $A$ depends on the factorization (using the Gram-Schmidt algorithm) of $A$ in the form

$$A = QR$$

where $Q$ is orthogonal and $R$ is invertible and upper triangular (see Theorem 8.4.2). The **QR-algorithm** uses this repeatedly to create a sequence of matrices $A_1 = A$, $A_2$, $A_3$, $\ldots$, as follows:

1. Define $A_1 = A$ and factor it as $A_1 = Q_1R_1$.
2. Define $A_2 = R_1Q_1$ and factor it as $A_2 = Q_2R_2$.
3. Define $A_3 = R_2Q_2$ and factor it as $A_3 = Q_3R_3$.

$\vdots$

In general, $A_k$ is factored as $A_k = Q_kR_k$ and we define $A_{k+1} = R_kQ_k$. Then $A_{k+1}$ is similar to $A_k$ [in fact, $A_{k+1} = R_kQ_k = (Q_k^{-1}A_k)Q_k$], and hence each $A_k$ has the same eigenvalues as $A$. If the eigenvalues of $A$ are real and have distinct absolute values, the remarkable thing is that the sequence of matrices $A_1$, $A_2$, $A_3$, $\ldots$ converges to an upper triangular matrix with these eigenvalues on the main diagonal. [See below for the case of complex eigenvalues.]

**Example 8.5.2**

If $A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$ as in Example 8.5.1, use the QR-algorithm to approximate the eigenvalues.

**Solution.** The matrices $A_1$, $A_2$, and $A_3$ are as follows:

$$A_1 = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} = Q_1R_1 \quad \text{where } Q_1 = \tfrac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \text{ and } R_1 = \tfrac{1}{\sqrt{5}}\begin{bmatrix} 5 & 1 \\ 0 & 2 \end{bmatrix}$$

$$A_2 = \tfrac{1}{5}\begin{bmatrix} 7 & 9 \\ 4 & -2 \end{bmatrix} = \begin{bmatrix} 1.4 & -1.8 \\ -0.8 & -0.4 \end{bmatrix} = Q_2R_2$$

$$\text{where } Q_2 = \tfrac{1}{\sqrt{65}}\begin{bmatrix} 7 & 4 \\ 4 & -7 \end{bmatrix} \text{ and } R_2 = \tfrac{1}{\sqrt{65}}\begin{bmatrix} 13 & 11 \\ 0 & 10 \end{bmatrix}$$

$$A_3 = \tfrac{1}{13}\begin{bmatrix} 27 & -5 \\ 8 & -14 \end{bmatrix} = \begin{bmatrix} 2.08 & -0.38 \\ 0.62 & -1.08 \end{bmatrix}$$

This is converging to $\begin{bmatrix} 2 & * \\ 0 & -1 \end{bmatrix}$ and so is approximating the eigenvalues 2 and $-1$ on the main diagonal.

It is beyond the scope of this book to pursue a detailed discussion of these methods. The reader is referred to J. M. Wilkinson, *The Algebraic Eigenvalue Problem* (Oxford, England: Oxford University Press, 1965) or G. W. Stewart, *Introduction to Matrix Computations* (New York: Academic Press, 1973). We conclude with some remarks on the QR-algorithm.

**Shifting.** Convergence is accelerated if, at stage $k$ of the algorithm, a number $s_k$ is chosen and $A_k - s_k I$ is factored in the form $Q_k R_k$ rather than $A_k$ itself. Then

$$Q_k^{-1} A_k Q_k = Q_k^{-1}(Q_k R_k + s_k I)Q_k = R_k Q_k + s_k I$$

so we take $A_{k+1} = R_k Q_k + s_k I$. If the shifts $s_k$ are carefully chosen, convergence can be greatly improved.

**Preliminary Preparation.** A matrix such as

$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}$$

is said to be in **upper Hessenberg** form, and the QR-factorizations of such matrices are greatly simplified. Given an $n \times n$ matrix $A$, a series of orthogonal matrices $H_1$, $H_2$, ..., $H_m$ (called **Householder matrices**) can be easily constructed such that

$$B = H_m^T \cdots H_1^T A H_1 \cdots H_m$$

is in upper Hessenberg form. Then the QR-algorithm can be efficiently applied to $B$ and, because $B$ is similar to $A$, it produces the eigenvalues of $A$.

**Complex Eigenvalues.** If some of the eigenvalues of a real matrix $A$ are not real, the QR-algorithm converges to a block upper triangular matrix where the diagonal blocks are either $1 \times 1$ (the real eigenvalues) or $2 \times 2$ (each providing a pair of conjugate complex eigenvalues of $A$).

## Exercises for 8.5

**Exercise 8.5.1** In each case, find the exact eigenvalues and determine corresponding eigenvectors. Then start with $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and compute $\mathbf{x}_4$ and $r_3$ using the power method.

a. $A = \begin{bmatrix} 2 & -4 \\ -3 & 3 \end{bmatrix}$ b. $A = \begin{bmatrix} 5 & 2 \\ -3 & -2 \end{bmatrix}$

c. $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ d. $A = \begin{bmatrix} 3 & 1 \\ 1 & 0 \end{bmatrix}$

**Exercise 8.5.2** In each case, find the exact eigenvalues and then approximate them using the QR-algorithm.

a. $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ b. $A = \begin{bmatrix} 3 & 1 \\ 1 & 0 \end{bmatrix}$

**Exercise 8.5.3** Apply the power method to $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, starting at $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Does it converge? Explain.

**Exercise 8.5.4** If $A$ is symmetric, show that each matrix $A_k$ in the QR-algorithm is also symmetric. Deduce that they converge to a diagonal matrix.

**Exercise 8.5.5** Apply the QR-algorithm to $A = \begin{bmatrix} 2 & -3 \\ 1 & -2 \end{bmatrix}$. Explain.

**Exercise 8.5.6** Given a matrix $A$, let $A_k$, $Q_k$, and $R_k$, $k \geq 1$, be the matrices constructed in the QR-algorithm. Show that $A_k = (Q_1 Q_2 \cdots Q_k)(R_k \cdots R_2 R_1)$ for each $k \geq 1$ and hence that this is a QR-factorization of $A_k$.
[*Hint*: Show that $Q_k R_k = R_{k-1} Q_{k-1}$ for each $k \geq 2$, and use this equality to compute $(Q_1 Q_2 \cdots Q_k)(R_k \cdots R_2 R_1)$ "from the centre out." Use the fact that $(AB)^{n+1} = A(BA)^n B$ for any square matrices $A$ and $B$.]

## 8.6 The Singular Value Decomposition

When working with a square matrix $A$ it is clearly useful to be able to "diagonalize" $A$, that is to find a factorization $A = Q^{-1}DQ$ where $Q$ is invertible and $D$ is diagonal. Unfortunately such a factorization may not exist for $A$. However, even if $A$ is not square gaussian elimination provides a factorization of the form $A = PDQ$ where $P$ and $Q$ are invertible and $D$ is diagonal—the Smith Normal form (Theorem 2.5.3). However, if $A$ is real we can choose $P$ and $Q$ to be *orthogonal* real matrices and $D$ to be real. Such a factorization is called a **singular value decomposition (SVD)** for $A$, one of the most useful tools in applied linear algebra. In this Section we show how to explicitly compute an SVD for any real matrix $A$, and illustrate some of its many applications.

We need a fact about two subspaces associated with an $m \times n$ matrix $A$:

$$\text{im } A = \{A\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n\} \qquad \text{and} \qquad \text{col } A = \text{span}\{\mathbf{a} \mid \mathbf{a} \text{ is a column of } A\}$$

Then $\text{im } A$ is called the **image** of $A$ (so named because of the linear transformation $\mathbb{R}^n \to \mathbb{R}^m$ with $\mathbf{x} \mapsto A\mathbf{x}$); and $\text{col } A$ is called the **column space** of $A$ (Definition 5.10). Surprisingly, these spaces are equal:

**Lemma 8.6.1**

*For any $m \times n$ matrix $A$, $\text{im } A = \text{col } A$.*

**Proof.** Let $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}$ in terms of its columns. Let $\mathbf{x} \in \text{im } A$, say $\mathbf{x} = A\mathbf{y}$, $\mathbf{y}$ in $\mathbb{R}^n$. If $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T$, then $A\mathbf{y} = y_1\mathbf{a}_1 + y_2\mathbf{a}_2 + \cdots + y_n\mathbf{a}_n \in \text{col } A$ by Definition 2.5. This shows that $\text{im } A \subseteq \text{col } A$. For the other inclusion, each $\mathbf{a}_k = A\mathbf{e}_k$ where $\mathbf{e}_k$ is column $k$ of $I_n$. □

### 8.6.1 Singular Value Decompositions

We know a lot about any real symmetric matrix: Its eigenvalues are real (Theorem 5.5.7), and it is orthogonally diagonalizable by the Principal Axes Theorem (Theorem 8.2.2). So for any real matrix $A$ (square or not), the fact that both $A^TA$ and $AA^T$ are real and symmetric suggests that we can learn a lot about $A$ by studying them. This section shows just how true this is.

The following Lemma reveals some similarities between $A^TA$ and $AA^T$ which simplify the statement and the proof of the SVD we are constructing.

**Lemma 8.6.2**

*Let $A$ be a real $m \times n$ matrix. Then:*

1. *The eigenvalues of $A^TA$ and $AA^T$ are real and non-negative.*
2. *$A^TA$ and $AA^T$ have the same set of positive eigenvalues.*

**Proof.**

1. Let $\lambda$ be an eigenvalue of $A^TA$, with eigenvector $\mathbf{0} \neq \mathbf{q} \in \mathbb{R}^n$. Then:

$$\|A\mathbf{q}\|^2 = (A\mathbf{q})^T(A\mathbf{q}) = \mathbf{q}^T(A^TA\mathbf{q}) = \mathbf{q}^T(\lambda\mathbf{q}) = \lambda(\mathbf{q}^T\mathbf{q}) = \lambda\|\mathbf{q}\|^2$$

Then (1.) follows for $A^TA$, and the case $AA^T$ follows by replacing $A$ by $A^T$.

2. Write $N(B)$ for the set of positive eigenvalues of a matrix $B$. We must show that $N(A^TA) = N(AA^T)$. If $\lambda \in N(A^TA)$ with eigenvector $\mathbf{0} \neq \mathbf{q} \in \mathbb{R}^n$, then $A\mathbf{q} \in \mathbb{R}^m$ and
$$AA^T(A\mathbf{q}) = A[(A^TA)\mathbf{q}] = A(\lambda\mathbf{q}) = \lambda(A\mathbf{q})$$
Moreover, $A\mathbf{q} \neq \mathbf{0}$ since $A^TA\mathbf{q} = \lambda\mathbf{q} \neq \mathbf{0}$ and both $\lambda \neq 0$ and $\mathbf{q} \neq \mathbf{0}$. Hence $\lambda$ is an eigenvalue of $AA^T$, proving $N(A^TA) \subseteq N(AA^T)$. For the other inclusion replace $A$ by $A^T$. □

To analyze an $m \times n$ matrix $A$ we have two symmetric matrices to work with: $A^TA$ and $AA^T$. In view of Lemma 8.6.2, we choose $A^TA$ (sometimes called the **Gram** matrix of $A$), and derive a series of facts which we will need. This narrative is a bit long, but trust that it will be worth the effort. We parse it out in several steps:

**1.** The $n \times n$ matrix $A^TA$ is real and symmetric so, by the Principal Axes Theorem 8.2.2, let $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\} \subseteq \mathbb{R}^n$ be an orthonormal basis of eigenvectors of $A^TA$, with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. By Lemma 8.6.2(1), $\lambda_i$ is real for each $i$ and $\lambda_i \geq 0$. By re-ordering the $\mathbf{q}_i$ we may (and do) assume that
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 \quad \text{and}^{8} \quad \lambda_i = 0 \text{ if } i > r \tag{i}$$
By Theorems 8.2.1 and 3.3.4, the matrix
$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix} \text{ is orthogonal} \quad \text{and} \quad \text{orthogonally diagonalizes } A^TA \tag{ii}$$

**2.** Even though the $\lambda_i$ are the eigenvalues of $A^TA$, the number $r$ in (**i**) turns out to be rank $A$. To understand why, consider the vectors $A\mathbf{q}_i \in \text{im } A$. For all $i, j$:
$$A\mathbf{q}_i \cdot A\mathbf{q}_j = (A\mathbf{q}_i)^TA\mathbf{q}_j = \mathbf{q}_i^T(A^TA)\mathbf{q}_j = \mathbf{q}_i^T(\lambda_j\mathbf{q}_j) = \lambda_j(\mathbf{q}_i^T\mathbf{q}_j) = \lambda_j(\mathbf{q}_i \cdot \mathbf{q}_j)$$
Because $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ is an orthonormal set, this gives
$$A\mathbf{q}_i \cdot A\mathbf{q}_j = 0 \text{ if } i \neq j \qquad \text{and} \qquad \|A\mathbf{q}_i\|^2 = \lambda_i\|\mathbf{q}_i\|^2 = \lambda_i \text{ for each } i \tag{iii}$$
We can extract two conclusions from (**iii**) and (**i**):
$$\{A\mathbf{q}_1, A\mathbf{q}_2, \ldots, A\mathbf{q}_r\} \subseteq \text{im } A \text{ is an orthogonal set} \quad \text{and} \quad A\mathbf{q}_i = \mathbf{0} \text{ if } i > r \tag{iv}$$
With this write $U = \text{span}\{A\mathbf{q}_1, A\mathbf{q}_2, \ldots, A\mathbf{q}_r\} \subseteq \text{im } A$; we claim that $U = \text{im } A$, that is $\text{im } A \subseteq U$. For this we must show that $A\mathbf{x} \in U$ for each $\mathbf{x} \in \mathbb{R}^n$. Since $\{\mathbf{q}_1, \ldots, \mathbf{q}_r, \ldots, \mathbf{q}_n\}$ is a basis of $\mathbb{R}^n$ (it is orthonormal), we can write $\mathbf{x}_k = t_1\mathbf{q}_1 + \cdots + t_r\mathbf{q}_r + \cdots + t_n\mathbf{q}_n$ where each $t_j \in \mathbb{R}$. Then, using (**iv**) we obtain
$$A\mathbf{x} = t_1A\mathbf{q}_1 + \cdots + t_rA\mathbf{q}_r + \cdots + t_nA\mathbf{q}_n = t_1A\mathbf{q}_1 + \cdots + t_rA\mathbf{q}_r \in U$$
This shows that $U = \text{im } A$, and so
$$\{A\mathbf{q}_1, A\mathbf{q}_2, \ldots, A\mathbf{q}_r\} \text{ is an } \textit{orthogonal} \text{ basis of } \text{im}(A) \tag{v}$$
But $\text{col } A = \text{im } A$ by Lemma 8.6.1, and $\text{rank } A = \dim(\text{col } A)$ by Theorem 5.4.1, so
$$\text{rank } A = \dim(\text{col } A) = \dim(\text{im } A) \overset{(\mathbf{v})}{=} r \tag{vi}$$

**3.** Before proceeding, some definitions are in order:

[8] Of course they could all be positive ($r = n$) or all zero (so $A^TA = 0$, and hence $A = 0$ by Exercise 5.3.9).

**Definition 8.7**

*The real numbers* $\sigma_i = \sqrt{\lambda_i} \overset{\text{(iii)}}{=} \|A\bar{\mathbf{q}}_i\|$ *for* $i = 1, 2, \ldots, n$, *are called the* ***singular values*** *of the matrix* $A$.

Clearly $\sigma_1, \sigma_2, \ldots, \sigma_r$ are the *positive* singular values of $A$. By (**i**) we have

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \qquad \text{and} \qquad \sigma_i = 0 \text{ if } i > r \tag{vii}$$

With (**vi**) this makes the following definitions depend only upon $A$.

**Definition 8.8**

*Let* $A$ *be a real,* $m \times n$ *matrix of rank* $r$, *with positive singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ *and* $\sigma_i = 0$ *if* $i > r$. *Define:*

$$D_A = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \qquad \textit{and} \qquad \Sigma_A = \begin{bmatrix} D_A & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}$$

*Here* $\Sigma_A$ *is in block form and is called the* ***singular matrix*** *of* $A$.

The singular values $\sigma_i$ and the matrices $D_A$ and $\Sigma_A$ will be referred to frequently below.

**4.** Returning to our narrative, normalize the vectors $A\mathbf{q}_1, A\mathbf{q}_2, \ldots, A\mathbf{q}_r$, by defining

$$\mathbf{p}_i = \tfrac{1}{\|A\mathbf{q}_i\|} A\mathbf{q}_i \in \mathbb{R}^m \quad \text{for each } i = 1, 2, \ldots, r \tag{viii}$$

By (**v**) and Lemma 8.6.1, we conclude that

$$\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_r\} \text{ is an } \textit{orthonormal} \text{ basis of } \operatorname{col} A \subseteq \mathbb{R}^m \tag{ix}$$

Employing the Gram-Schmidt algorithm (or otherwise), construct $\mathbf{p}_{r+1}, \ldots, \mathbf{p}_m$ so that

$$\{\mathbf{p}_1, \ldots, \mathbf{p}_r, \ldots, \mathbf{p}_m\} \text{ is an orthonormal basis of } \mathbb{R}^m \tag{x}$$

**5.** By (**x**) and (**ii**) we have *two* orthogonal matrices

$$P = \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \cdots & \mathbf{p}_m \end{bmatrix} \text{ of size } m \times m \quad \text{and} \quad Q = \begin{bmatrix} \mathbf{q}_1 & \cdots & \mathbf{q}_r & \cdots & \mathbf{q}_n \end{bmatrix} \text{ of size } n \times n$$

These matrices are related. In fact we have:

$$\sigma_i \mathbf{p}_i = \sqrt{\lambda_i}\mathbf{p}_i \overset{\text{(iii)}}{=} \|A\mathbf{q}_i\|\mathbf{p}_i \overset{\text{(viii)}}{=} A\mathbf{q}_i \quad \text{for each } i = 1, 2, \ldots, r \tag{xi}$$

This yields the following expression for $AQ$ in terms of its columns:

$$AQ = \begin{bmatrix} A\mathbf{q}_1 & \cdots & A\mathbf{q}_r & A\mathbf{q}_{r+1} & \cdots & A\mathbf{q}_n \end{bmatrix} \overset{\text{(iv)}}{=} \begin{bmatrix} \sigma_1\mathbf{p}_1 & \cdots & \sigma_r\mathbf{p}_r & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \tag{xii}$$

Then we compute:

$$P\Sigma_A = \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \mathbf{p}_{r+1} & \cdots & \mathbf{p}_m \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}$$
$$= \begin{bmatrix} \sigma_1\mathbf{p}_1 & \cdots & \sigma_r\mathbf{p}_r & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$
$$\overset{\text{(xii)}}{=} AQ$$

Finally, as $Q^{-1} = Q^T$ it follows that $A = P\Sigma_A Q^T$.

With this we can state the main theorem of this Section.

**Theorem 8.6.1**

*Let $A$ be a real $m \times n$ matrix, and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ be the positive singular values of $A$. Then $r$ is the rank of $A$ and we have the factorization*

$$A = P\Sigma_A Q^T \qquad \text{where } P \text{ and } Q \text{ are orthogonal matrices}$$

The factorization $A = P\Sigma_A Q^T$ in Theorem 8.6.1, where $P$ and $Q$ are orthogonal matrices, is called a *Singular Value Decomposition* (*SVD*) of $A$. This decomposition is not unique. For example if $r < m$ then the vectors $\mathbf{p}_{r+1}, \ldots, \mathbf{p}_m$ can be *any* extension of $\{\mathbf{p}_1, \ldots, \mathbf{p}_r\}$ to an orthonormal basis of $\mathbb{R}^m$, and each will lead to a different matrix $P$ in the decomposition. For a more dramatic example, if $A = I_n$ then $\Sigma_A = I_n$, and $A = P\Sigma_A P^T$ is a SVD of $A$ for *any* orthogonal $n \times n$ matrix $P$.

**Example 8.6.1**

Find a singular value decomposition for $A = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}$.

**Solution.** We have $A^TA = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$, so the characteristic polynomial is

$$c_{A^TA}(x) = \det \begin{bmatrix} x-2 & 1 & -1 \\ 1 & x-1 & 0 \\ -1 & 0 & x-1 \end{bmatrix} = (x-3)(x-1)x$$

Hence the eigenvalues of $A^TA$ (in descending order) are $\lambda_1 = 3$, $\lambda_2 = 1$ and $\lambda_3 = 0$ with, respectively, unit eigenvectors

$$\mathbf{q}_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \quad \mathbf{q}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{q}_3 = \frac{1}{\sqrt{3}} \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

It follows that the orthogonal matrix $Q$ in Theorem 8.6.1 is

$$Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix} = \frac{1}{\sqrt{6}} \begin{bmatrix} 2 & 0 & -\sqrt{2} \\ -1 & \sqrt{3} & -\sqrt{2} \\ 1 & \sqrt{3} & \sqrt{2} \end{bmatrix}$$

The singular values here are $\sigma_1 = \sqrt{3}$, $\sigma_2 = 1$ and $\sigma_3 = 0$, so $\operatorname{rank}(A) = 2$—clear in this case—and the singular matrix is

$$\Sigma_A = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

So it remains to find the $2 \times 2$ orthogonal matrix $P$ in Theorem 8.6.1. This involves the vectors

$$A\mathbf{q}_1 = \frac{\sqrt{6}}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad A\mathbf{q}_2 = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad A\mathbf{q}_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Normalize $A\mathbf{q}_1$ and $A\mathbf{q}_2$ to get

$$\mathbf{p}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \text{and} \quad \mathbf{p}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

In this case, $\{\mathbf{p}_1, \mathbf{p}_2\}$ is *already* a basis of $\mathbb{R}^2$ (so the Gram-Schmidt algorithm is not needed), and we have the $2 \times 2$ orthogonal matrix

$$P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

Finally (by Theorem 8.6.1) the singular value decomposition for $A$ is

$$A = P\Sigma_A Q^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \frac{1}{\sqrt{6}} \begin{bmatrix} 2 & -1 & 1 \\ 0 & \sqrt{3} & \sqrt{3} \\ -\sqrt{2} & -\sqrt{2} & \sqrt{2} \end{bmatrix}$$

Of course this can be confirmed by direct matrix multiplication.

Thus, computing an SVD for a real matrix $A$ is a routine matter, and we now describe a systematic procedure for doing so.

**SVD Algorithm**

*Given a real $m \times n$ matrix $A$, find an SVD $A = P\Sigma_A Q^T$ as follows:*

1. *Use the Diagonalization Algorithm (see page 179) to find the (real and non-negative) eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $A^T A$ with corresponding (orthonormal) eigenvectors $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$. Reorder the $\mathbf{q}_i$ (if necessary) to ensure that the nonzero eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and $\lambda_i = 0$ if $i > r$.*

2. *The integer $r$ is the rank of the matrix $A$.*

3. *The $n \times n$ orthogonal matrix $Q$ in the SVD is* $Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix}$.
4. *Define* $\mathbf{p}_i = \frac{1}{\|A\mathbf{q}_i\|} A\mathbf{q}_i$ *for* $i = 1, 2, \ldots, r$ *(where $r$ is as in step 1). Then* $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_r\}$ *is orthonormal in* $\mathbb{R}^m$ *so (using Gram-Schmidt or otherwise) extend it to an orthonormal basis* $\{\mathbf{p}_1, \ldots, \mathbf{p}_r, \ldots, \mathbf{p}_m\}$ *in* $\mathbb{R}^m$.
5. *The $m \times m$ orthogonal matrix $P$ in the SVD is* $P = \begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_r & \cdots & \mathbf{p}_m \end{bmatrix}$.
6. *The* singular values *for $A$ are* $\sigma_1, \sigma_2, \ldots, \sigma_n$ *where* $\sigma_i = \sqrt{\lambda_i}$ *for each $i$. Hence the nonzero singular values are* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, *and so the singular matrix of $A$ in the SVD is* $\Sigma_A = \begin{bmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}$.
7. *Thus* $A = P\Sigma Q^T$ *is a SVD for $A$.*

In practise the singular values $\sigma_i$, the matrices $P$ and $Q$, and even the rank of an $m \times n$ matrix are not calculated this way. There are sophisticated numerical algorithms for calculating them to a high degree of accuracy. The reader is referred to books on numerical linear algebra.

So the main virtue of Theorem 8.6.1 is that it provides a way of *constructing* an SVD for every real matrix $A$. In particular it shows that every real matrix $A$ *has* a singular value decomposition[9] in the following, more general, sense:

**Definition 8.9**

*A* ***Singular Value Decomposition (SVD)*** *of an $m \times n$ matrix $A$ of rank $r$ is a factorization* $A = U\Sigma V^T$ *where $U$ and $V$ are orthogonal and* $\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}_{m \times n}$ *in block form where* $D = \operatorname{diag}(d_1, d_2, \ldots, d_r)$ *where each* $d_i > 0$, *and* $r \leq m$ *and* $r \leq n$.

Note that for *any* SVD $A = U\Sigma V^T$ we immediately obtain some information about $A$:

**Lemma 8.6.3**

*If* $A = U\Sigma V^T$ *is any SVD for $A$ as in Definition 8.9, then:*

1. $r = \operatorname{rank} A$.
2. *The numbers* $d_1, d_2, \ldots, d_r$ *are the singular values of* $A^T A$ *in some order.*

**Proof.** Use the notation of Definition 8.9. We have

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

so $\Sigma^T\Sigma$ and $A^T A$ are similar $n \times n$ matrices (Definition 5.11). Hence $r = \operatorname{rank} A$ by Corollary 5.4.3, proving (1.). Furthermore, $\Sigma^T\Sigma$ and $A^T A$ have the same eigenvalues by Theorem 5.5.1; that is (using (1.)):

$$\{d_1^2, d_2^2, \ldots, d_r^2\} = \{\lambda_1, \lambda_2, \ldots, \lambda_r\} \quad \text{are equal as sets}$$

[9] In fact every complex matrix has an SVD [J.T. Scheick, Linear Algebra with Applications, McGraw-Hill, 1997]

where $\lambda_1, \lambda_2, \ldots, \lambda_r$ are the positive eigenvalues of $A^TA$. Hence there is a permutation $\tau$ of $\{1, 2, \cdots, r\}$ such that $d_i^2 = \lambda_{i\tau}$ for each $i = 1, 2, \ldots, r$. Hence $d_i = \sqrt{\lambda_{i\tau}} = \sigma_{i\tau}$ for each $i$ by Definition 8.7. This proves (2.). □

We note in passing that more is true. Let $A$ be $m \times n$ of rank $r$, and let $A = U\Sigma V^T$ be any SVD for $A$. Using the proof of Lemma 8.6.3 we have $d_i = \sigma_{i\tau}$ for some permutation $\tau$ of $\{1, 2, \ldots, r\}$. In fact, it can be shown that there exist orthogonal matrices $U_1$ and $V_1$ obtained from $U$ and $V$ by $\tau$-permuting columns and rows respectively, such that $A = U_1\Sigma_A V_1^T$ is an SVD of $A$.

### 8.6.2 Fundamental Subspaces

It turns out that any singular value decomposition contains a great deal of information about an $m \times n$ matrix $A$ and the subspaces associated with $A$. For example, in addition to Lemma 8.6.3, the set $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_r\}$ of vectors constructed in the proof of Theorem 8.6.1 is an orthonormal basis of $\operatorname{col} A$ (by (**v**) and (**viii**) in the proof). There are more such examples, which is the thrust of this subsection. In particular, there are four subspaces associated to a real $m \times n$ matrix $A$ that have come to be called fundamental:

**Definition 8.10**

*The **fundamental subspaces** of an $m \times n$ matrix $A$ are:*

$$\operatorname{row} A = \operatorname{span}\{\mathbf{x} \mid \mathbf{x} \text{ is a row of } A\}$$

$$\operatorname{col} A = \operatorname{span}\{\mathbf{x} \mid \mathbf{x} \text{ is a column of } A\}$$

$$\operatorname{null} A = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\}$$

$$\operatorname{null} A^T = \{\mathbf{x} \in \mathbb{R}^n \mid A^T\mathbf{x} = \mathbf{0}\}$$

If $A = U\Sigma V^T$ is *any* SVD for the real $m \times n$ matrix $A$, any orthonormal bases of $U$ and $V$ provide orthonormal bases for each of these fundamental subspaces. We are going to prove this, but first we need three properties related to the *orthogonal complement* $U^\perp$ of a subspace $U$ of $\mathbb{R}^n$, where (Definition 8.1):

$$U^\perp = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{u} \cdot \mathbf{x} = 0 \text{ for all } \mathbf{u} \in U\}$$

The orthogonal complement plays an important role in the Projection Theorem (Theorem 8.1.3), and we return to it in Section 10.2. For now we need:

**Lemma 8.6.4**

*If $A$ is any matrix then:*

1. $(\operatorname{row} A)^\perp = \operatorname{null} A$ *and* $(\operatorname{col} A)^\perp = \operatorname{null} A^T$.
2. *If $U$ is any subspace of $\mathbb{R}^n$ then $U^{\perp\perp} = U$.*
3. *Let $\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ be an orthonormal basis of $\mathbb{R}^m$. If $U = \operatorname{span}\{\mathbf{f}_1, \ldots, \mathbf{f}_k\}$, then*

$$U^\perp = \operatorname{span}\{\mathbf{f}_{k+1}, \ldots, \mathbf{f}_m\}$$

**Proof.**

1. Assume $A$ is $m \times n$, and let $\mathbf{b}_1, \ldots, \mathbf{b}_m$ be the rows of $A$. If $\mathbf{x}$ is a column in $\mathbb{R}^n$, then entry $i$ of $A\mathbf{x}$ is $\mathbf{b}_i \cdot \mathbf{x}$, so $A\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{b}_i \cdot \mathbf{x} = 0$ for each $i$. Thus:

$$\mathbf{x} \in \text{null } A \quad \Leftrightarrow \quad \mathbf{b}_i \cdot \mathbf{x} = 0 \text{ for each } i \quad \Leftrightarrow \quad \mathbf{x} \in (\text{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_m\})^\perp = (\text{row } A)^\perp$$

   Hence $\text{null } A = (\text{row } A)^\perp$. Now replace $A$ by $A^T$ to get $\text{null } A^T = (\text{row } A^T)^\perp = (\text{col } A)^\perp$, which is the other identity in (1).

2. If $\mathbf{x} \in U$ then $\mathbf{y} \cdot \mathbf{x} = 0$ for all $\mathbf{y} \in U^\perp$, that is $\mathbf{x} \in U^{\perp\perp}$. This proves that $U \subseteq U^{\perp\perp}$, so it is enough to show that $\dim U = \dim U^{\perp\perp}$. By Theorem 8.1.4 we see that $\dim V^\perp = n - \dim V$ for any subspace $V \subseteq \mathbb{R}^n$. Hence

$$\dim U^{\perp\perp} = n - \dim U^\perp = n - (n - \dim U) = \dim U, \text{ as required}$$

3. We have $\text{span}\{\mathbf{f}_{k+1}, \ldots, \mathbf{f}_m\} \subseteq U^\perp$ because $\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ is orthogonal. For the other inclusion, let $\mathbf{x} \in U^\perp$ so $\mathbf{f}_i \cdot \mathbf{x} = 0$ for $i = 1, 2, \ldots, k$. By the Expansion Theorem 5.3.6:

$$\begin{array}{rcccccccccc} \mathbf{x} & = & (\mathbf{f}_1 \cdot \mathbf{x})\mathbf{f}_1 & + & \cdots & + & (\mathbf{f}_k \cdot \mathbf{x})\mathbf{f}_k & + & (\mathbf{f}_{k+1} \cdot \mathbf{x})\mathbf{f}_{k+1} & + \cdots + & (\mathbf{f}_m \cdot \mathbf{x})\mathbf{f}_m \\ & = & \mathbf{0} & + & \cdots & + & \mathbf{0} & + & (\mathbf{f}_{k+1} \cdot \mathbf{x})\mathbf{f}_{k+1} & + \cdots + & (\mathbf{f}_m \cdot \mathbf{x})\mathbf{f}_m \end{array}$$

   Hence $U^\perp \subseteq \text{span}\{\mathbf{f}_{k+1}, \ldots, \mathbf{f}_m\}$.

□

With this we can see how *any* SVD for a matrix $A$ provides orthonormal bases for each of the four fundamental subspaces of $A$.

**Theorem 8.6.2**

*Let $A$ be an $m \times n$ real matrix, let $A = U\Sigma V^T$ be* any *SVD for $A$ where $U$ and $V$ are orthogonal of size $m \times m$ and $n \times n$ respectively, and let*

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}_{m \times n} \qquad \text{where} \qquad D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r), \text{ with each } \lambda_i > 0$$

*Write $U = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r & \cdots & \mathbf{u}_m \end{bmatrix}$ and $V = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_r & \cdots & \mathbf{v}_n \end{bmatrix}$, so $\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \ldots, \mathbf{u}_m\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_r, \ldots, \mathbf{v}_n\}$ are orthonormal bases of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Then*

1. *$r = \text{rank } A$, and the singular values of $A$ are $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r}$.*
2. *The fundamental spaces are described as follows:*
   a. *$\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ is an orthonormal basis of $\text{col } A$.*
   b. *$\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$ is an orthonormal basis of $\text{null } A^T$.*
   c. *$\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of $\text{null } A$.*
   d. *$\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ is an orthonormal basis of $\text{row } A$.*

**Proof.**

1. This is Lemma 8.6.3.

2. a. As $\operatorname{col} A = \operatorname{col}(AV)$ by Lemma 5.4.3 and $AV = U\Sigma$, (a.) follows from

$$U\Sigma = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r & \cdots & \mathbf{u}_m \end{bmatrix} \begin{bmatrix} \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r) & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \lambda_1\mathbf{u}_1 & \cdots & \lambda_r\mathbf{u}_r & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$

   b. We have $(\operatorname{col} A)^\perp \stackrel{(a.)}{=} (\operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\})^\perp = \operatorname{span}\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$ by Lemma 8.6.4(3). This proves (b.) because $(\operatorname{col} A)^\perp = \operatorname{null} A^T$ by Lemma 8.6.4(1).

   c. We have $\dim(\operatorname{null} A) + \dim(\operatorname{im} A) = n$ by the Dimension Theorem 7.2.4, applied to $T : \mathbb{R}^n \to \mathbb{R}^m$ where $T(\mathbf{x}) = A\mathbf{x}$. Since also $\operatorname{im} A = \operatorname{col} A$ by Lemma 8.6.1, we obtain

$$\dim(\operatorname{null} A) = n - \dim(\operatorname{col} A) = n - r = \dim(\operatorname{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\})$$

   So to prove (c.) it is enough to show that $\mathbf{v}_j \in \operatorname{null} A$ whenever $j > r$. To this end write

$$\lambda_{r+1} = \cdots = \lambda_n = 0, \quad \text{so} \quad E^TE = \operatorname{diag}(\lambda_1^2, \ldots, \lambda_r^2, \lambda_{r+1}^2, \ldots, \lambda_n^2)$$

   Observe that each $\lambda_j$ is an eigenvalue of $\Sigma^T\Sigma$ with eigenvector $\mathbf{e}_j =$ column $j$ of $I_n$. Thus $\mathbf{v}_j = V\mathbf{e}_j$ for each $j$. As $A^TA = V\Sigma^T\Sigma V^T$ (proof of Lemma 8.6.3), we obtain

$$(A^TA)\mathbf{v}_j = (V\Sigma^T\Sigma V^T)(V\mathbf{e}_j) = V(\Sigma^T\Sigma\mathbf{e}_j) = V\left(\lambda_j^2\mathbf{e}_j\right) = \lambda_j^2 V\mathbf{e}_j = \lambda_j^2\mathbf{v}_j$$

   for $1 \le j \le n$. Thus each $\mathbf{v}_j$ is an eigenvector of $A^TA$ corresponding to $\lambda_j^2$. But then

$$\|A\mathbf{v}_j\|^2 = (A\mathbf{v}_j)^TA\mathbf{v}_j = \mathbf{v}_j^T(A^TA\mathbf{v}_j) = \mathbf{v}_j^T(\lambda_j^2\mathbf{v}_j) = \lambda_j^2\|\mathbf{v}_j\|^2 = \lambda_j^2 \quad \text{for } i = 1, \ldots, n$$

   In particular, $A\mathbf{v}_j = \mathbf{0}$ whenever $j > r$, so $\mathbf{v}_j \in \operatorname{null} A$ if $j > r$, as desired. This proves (c).

   d. Observe that $\operatorname{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\} \stackrel{(c.)}{=} \operatorname{null} A = (\operatorname{row} A)^\perp$ by Lemma 8.6.4(1). But then parts (2) and (3) of Lemma 8.6.4 show

$$\operatorname{row} A = \left((\operatorname{row} A)^\perp\right)^\perp = (\operatorname{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\})^\perp = \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$$

   This proves (d.), and hence Theorem 8.6.2.

□

**Example 8.6.2**

Consider the homogeneous linear system

$$A\mathbf{x} = \mathbf{0} \text{ of } m \text{ equations in } n \text{ variables}$$

Then the set of all solutions is $\operatorname{null} A$. Hence if $A = U\Sigma V^T$ is any SVD for $A$ then (in the notation of Theorem 8.6.2) $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of the set of solutions for the system. As such they are a set of **basic solutions** for the system, the most basic notion in Chapter 1.

### 8.6.3 The Polar Decomposition of a Real Square Matrix

If $A$ is real and $n \times n$ the factorization in the title is related to the polar decomposition $A$. Unlike the SVD, in this case the decomposition is *uniquely* determined by $A$.

Recall (Section 8.3) that a symmetric matrix $A$ is called positive definite if and only if $\mathbf{x}^T A\mathbf{x} > 0$ for every column $\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n$. Before proceeding, we must explore the following weaker notion:

**Definition 8.11**

*A real $n \times n$ matrix $G$ is called* ***positive***[10] *if it is symmetric and*

$$\mathbf{x}^T G\mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

Clearly every positive definite matrix is positive, but the converse fails. Indeed, $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ is positive because, if $\mathbf{x} = \begin{bmatrix} a & b \end{bmatrix}^T$ in $\mathbb{R}^2$, then $\mathbf{x}^T A\mathbf{x} = (a+b)^2 \geq 0$. But $\mathbf{y}^T A\mathbf{y} = 0$ if $\mathbf{y} = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$, so $A$ is not positive definite.

**Lemma 8.6.5**

*Let $G$ denote an $n \times n$ positive matrix.*

1. *If $A$ is any $m \times n$ matrix and $G$ is positive, then $A^T GA$ is positive (and $m \times m$).*
2. *If $G = \operatorname{diag}(d_1, d_2, \cdots, d_n)$ and each $d_i \geq 0$ then $G$ is positive.*

**Proof.**

1. $\mathbf{x}^T(A^T GA)\mathbf{x} = (A\mathbf{x})^T G(A\mathbf{x}) \geq 0$ because $G$ is positive.
2. If $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$, then
$$\mathbf{x}^T G\mathbf{x} = d_1x_1^2 + d_2x_2^2 + \cdots + d_nx_n^2 \geq 0$$
because $d_i \geq 0$ for each $i$.

□

**Definition 8.12**

*If $A$ is a real $n \times n$ matrix, a factorization*

$$A = GQ \text{ where } G \text{ is positive and } Q \text{ is orthogonal}$$

*is called a* ***polar decomposition*** *for $A$.*

Any SVD for a real square matrix $A$ yields a polar form for $A$.

[10] Also called **positive semi-definite.**

**Theorem 8.6.3**

*Every square real matrix has a polar form.*

**Proof.** Let $A = U\Sigma V^T$ be a SVD for $A$ with $\Sigma$ as in Definition 8.9 and $m = n$. Since $U^TU = I_n$ here we have

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

So if we write $G = U\Sigma U^T$ and $Q = UV^T$, then $Q$ is orthogonal, and it remains to show that $G$ is positive. But this follows from Lemma 8.6.5. □

The SVD for a square matrix $A$ is not unique ($I_n = PI_nP^T$ for any orthogonal matrix $P$). But given the proof of Theorem 8.6.3 it is surprising that the polar decomposition *is* unique.[11] We omit the proof.

The name "polar form" is reminiscent of the same form for complex numbers (see Appendix A). This is no coincidence. To see why, we represent the complex numbers as real $2\times 2$ matrices. Write $\mathbf{M}_2(\mathbb{R})$ for the set of all real $2\times 2$ matrices, and define

$$\sigma : \mathbb{C} \to \mathbf{M}_2(\mathbb{R}) \quad \text{by} \quad \sigma(a+bi) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \text{ for all } a+bi \text{ in } \mathbb{C}$$

One verifies that $\sigma$ preserves addition and multiplication in the sense that

$$\sigma(zw) = \sigma(z)\sigma(w) \qquad \text{and} \qquad \sigma(z+w) = \sigma(z)+\sigma(w)$$

for all complex numbers $z$ and $w$. Since $\theta$ is one-to-one we may *identify* each complex number $a+bi$ with the matrix $\theta(a+bi)$, that is we write

$$a+bi = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \quad \text{for all } a+bi \text{ in } \mathbb{C}$$

Thus $0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, $1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$, $i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, and $r = \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}$ if $r$ is real.

If $z = a+bi$ is nonzero then the *absolute value* $r = |z| = \sqrt{a^2+b^2} \neq 0$. If $\theta$ is the *angle* of $z$ in standard position, then $\cos\theta = a/r$ and $\sin\theta = b/r$. Observe:

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix} \begin{bmatrix} a/r & -b/r \\ b/r & a/r \end{bmatrix} = \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = GQ \qquad \textbf{(xiii)}$$

where $G = \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}$ is positive and $Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ is orthogonal. But in $\mathbb{C}$ we have $G = r$ and $Q = \cos\theta + i\sin\theta$ so (**xiii**) reads $z = r(\cos\theta + i\sin\theta) = re^{i\theta}$ which is the *classical polar form* for the complex number $a+bi$. This is why (**xiii**) is called the polar form of the matrix $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$; Definition 8.12 simply adopts the terminology for $n\times n$ matrices.

[11] See J.T. Scheick, Linear Algebra with Applications, McGraw-Hill, 1997, page 379.

### 8.6.4 The Pseudoinverse of a Matrix

It is impossible for a non-square matrix $A$ to have an inverse (see the footnote to Definition 2.11). Nonetheless, one candidate for an "inverse" of $A$ is an $m \times n$ matrix $B$ such that

$$ABA = A \qquad \text{and} \qquad BAB = B$$

Such a matrix $B$ is called a *middle inverse* for $A$. If $A$ is invertible then $A^{-1}$ is the unique middle inverse for $A$, but a middle inverse is not unique in general, even for square matrices. For example, if $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ then $B = \begin{bmatrix} 1 & 0 & 0 \\ b & 0 & 0 \end{bmatrix}$ is a middle inverse for $A$ for any $b$.

If $ABA = A$ and $BAB = B$ it is easy to see that $AB$ and $BA$ are both idempotent matrices. In 1955 Roger Penrose observed that the middle inverse is unique if both $AB$ and $BA$ are symmetric. We omit the proof.

**Theorem 8.6.4: Penrose' Theorem**[12]

*Given any real $m \times n$ matrix $A$, there is exactly one $n \times m$ matrix $B$ such that $A$ and $B$ satisfy the following conditions:*

***P1*** *$ABA = A$ and $BAB = B$.*

***P2*** *Both $AB$ and $BA$ are symmetric.*

**Definition 8.13**

*Let $A$ be a real $m \times n$ matrix. The* ***pseudoinverse*** *of $A$ is the unique $n \times m$ matrix $A^+$ such that $A$ and $A^+$ satisfy* ***P1*** *and* ***P2****, that is:*

$$AA^+A = A, \qquad A^+AA^+ = A^+, \qquad \text{and both } AA^+ \text{ and } A^+A \text{ are symmetric}^{13}$$

If $A$ is invertible then $A^+ = A^{-1}$ as expected. In general, the symmetry in conditions P1 and P2 shows that $A$ is the pseudoinverse of $A^+$, that is $A^{++} = A$.

[12] R. Penrose, A generalized inverse for matrices, Proceedings of the Cambridge Philosophical Society **51** (1955), 406-413. In fact Penrose proved this for any complex matrix, where $AB$ and $BA$ are both required to be hermitian (see Definition 8.18 in the following section).

[13] Penrose called the matrix $A^+$ the generalized inverse of $A$, but the term pseudoinverse is now commonly used. The matrix $A^+$ is also called the **Moore-Penrose** inverse after E.H. Moore who had the idea in 1935 as part of a larger work on "General Analysis". Penrose independently re-discovered it 20 years later.

**Theorem 8.6.5**

*Let $A$ be an $m \times n$ matrix.*

1. *If* rank $A = m$ *then $AA^T$ is invertible and $A^+ = A^T(AA^T)^{-1}$.*
2. *If* rank $A = n$ *then $A^TA$ is invertible and $A^+ = (A^TA)^{-1}A^T$.*

**Proof.** Here $AA^T$ (respectively $A^TA$) is invertible by Theorem 5.4.4 (respectively Theorem 5.4.3). The rest is a routine verification. □

In general, given an $m \times n$ matrix $A$, the pseudoinverse $A^+$ can be computed from any SVD for $A$. To see how, we need some notation. Let $A = U\Sigma V^T$ be an SVD for $A$ (as in Definition 8.9) where $U$ and $V$ are orthogonal and $\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}_{m\times n}$ in block form where $D = \text{diag}\,(d_1, d_2, \ldots, d_r)$ where each $d_i > 0$. Hence $D$ is invertible, so we make:

**Definition 8.14**

$$\Sigma' = \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix}_{n\times m}.$$

A routine calculation gives:

**Lemma 8.6.6**

- $\Sigma\Sigma'\Sigma = \Sigma$
- $\Sigma'\Sigma\Sigma' = \Sigma'$
- $\Sigma\Sigma' = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}_{m\times m}$
- $\Sigma'\Sigma = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}_{n\times n}$

That is, $\Sigma'$ is the pseudoinverse of $\Sigma$.

Now given $A = U\Sigma V^T$, define $B = V\Sigma'U^T$. Then

$$ABA = (U\Sigma V^T)(V\Sigma'U^T)(U\Sigma V^T) = U(\Sigma\Sigma'\Sigma)V^T = U\Sigma V^T = A$$

by Lemma 8.6.6. Similarly $BAB = B$. Moreover $AB = U(\Sigma\Sigma')U^T$ and $BA = V(\Sigma'\Sigma)V^T$ are both symmetric again by Lemma 8.6.6. This proves

**Theorem 8.6.6**

*Let $A$ be real and $m \times n$, and let $A = U\Sigma V^T$ is any SVD for $A$ as in Definition 8.9. Then $A^+ = V\Sigma'U^T$.*

Of course we can always use the SVD constructed in Theorem 8.6.1 to find the pseudoinverse. If $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, we observed above that $B = \begin{bmatrix} 1 & 0 & 0 \\ b & 0 & 0 \end{bmatrix}$ is a middle inverse for $A$ for any $b$. Furthermore $AB$ is symmetric but $BA$ is not, so $B \neq A^+$.

**Example 8.6.3**

Find $A^+$ if $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$.

**Solution.** $A^TA = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ with eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 0$ and corresponding eigenvectors $\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{q}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Hence $Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix} = I_2$. Also $A$ has rank 1 with singular values $\sigma_1 = 1$ and $\sigma_2 = 0$, so $\Sigma_A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = A$ and $\Sigma'_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = A^T$ in this case.

Since $A\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $A\mathbf{q}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, we have $\mathbf{p}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ which extends to an orthonormal basis $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ of $\mathbb{R}^3$ where (say) $\mathbf{p}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ and $\mathbf{p}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Hence $P = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 \end{bmatrix} = I$, so the SVD for $A$ is $A = P\Sigma_A Q^T$. Finally, the pseudoinverse of $A$ is $A^+ = Q\Sigma'_A P^T = \Sigma'_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Note that $A^+ = A^T$ in this case.

The following Lemma collects some properties of the pseudoinverse that mimic those of the inverse. The verifications are left as exercises.

**Lemma 8.6.7**

*Let $A$ be an $m \times n$ matrix of rank $r$.*

1. $A^{++} = A$.
2. *If $A$ is invertible then $A^+ = A^{-1}$.*
3. $(A^T)^+ = (A^+)^T$.
4. $(kA)^+ = kA^+$ *for any real $k$.*
5. $(UAV)^+ = U^T(A^+)V^T$ *whenever $U$ and $V$ are orthogonal.*

## Exercises for 8.6

**Exercise 8.6.1** If $ACA = A$ show that $B = CAC$ is a middle inverse for $A$.

**Exercise 8.6.2** For any matrix $A$ show that

$$\Sigma_{A^T} = (\Sigma_A)^T$$

**Exercise 8.6.3** If $A$ is $m \times n$ with all singular values positive, what is rank $A$?

**Exercise 8.6.4** If $A$ has singular values $\sigma_1, \ldots, \sigma_r$, what are the singular values of:

a. $A^T$ b. $tA$ where $t > 0$ is real

c. $A^{-1}$ assuming $A$ is invertible.

**Exercise 8.6.5** If $A$ is square show that $|\det A|$ is the product of the singular values of $A$.

**Exercise 8.6.6** If $A$ is square and real, show that $A = 0$ if and only if every eigenvalue of $A^TA$ is 0.

**Exercise 8.6.7** Given a SVD for an invertible matrix $A$, find one for $A^{-1}$. How are $\Sigma_A$ and $\Sigma_{A^{-1}}$ related?

**Exercise 8.6.8** Let $A^{-1} = A = A^T$ where $A$ is $n \times n$. Given any orthogonal $n \times n$ matrix $U$, find an orthogonal matrix $V$ such that $A = U\Sigma_A V^T$ is an SVD for $A$.

If $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ do this for:

a. $U = \frac{1}{5}\begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix}$ b. $U = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

**Exercise 8.6.9** Find a SVD for the following matrices:

a. $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$ b. $\begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & -2 \\ 1 & 2 & 0 \end{bmatrix}$

**Exercise 8.6.10** Find an SVD for $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

**Exercise 8.6.11** If $A = U\Sigma V^T$ is an SVD for $A$, find an SVD for $A^T$.

**Exercise 8.6.12** Let $A$ be a real, $m \times n$ with positive singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$, and write

$$s(x) = (x - \sigma_1)(x - \sigma_2)\cdots(x - \sigma_r)$$

a. Show that $c_{A^TA}(x) = s(x)x^{n-r}$ and $c_{A^TA}(c) = s(x)x^{m-r}$.

b. If $m \leq n$ conclude that $c_{A^TA}(x) = s(x)x^{n-m}$.

**Exercise 8.6.13** If $G$ is positive show that:

a. $rG$ is positive if $r \geq 0$

b. $G + H$ is positive for any positive $H$.

**Exercise 8.6.14** If $G$ is positive and $\lambda$ is an eigenvalue, show that $\lambda \geq 0$.

**Exercise 8.6.15** If $G$ is positive show that $G = H^2$ for some positive matrix $H$. [Hint: Preceding exercise and Lemma 8.6.5]

**Exercise 8.6.16** If $A$ is $n \times n$ show that $AA^T$ and $A^TA$ are similar. [Hint: Start with an SVD for $A$.]

**Exercise 8.6.17** Find $A^+$ if:

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

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

**Exercise 8.6.18** Show that $(A^+)^T = (A^T)^+$.

## 8.7 Complex Matrices

If $A$ is an $n \times n$ matrix, the characteristic polynomial $c_A(x)$ is a polynomial of degree $n$ and the eigenvalues of $A$ are just the roots of $c_A(x)$. In most of our examples these roots have been *real* numbers (in fact, the examples have been carefully chosen so this will be the case!); but it need not happen, even when the characteristic polynomial has real coefficients. For example, if $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ then $c_A(x) = x^2 + 1$ has roots $i$ and $-i$, where $i$ is a complex number satisfying $i^2 = -1$. Therefore, we have to deal with the possibility that the eigenvalues of a (real) square matrix might be complex numbers.

In fact, nearly everything in this book would remain true if the phrase *real number* were replaced by *complex number* wherever it occurs. Then we would deal with matrices with complex entries, systems of linear equations with complex coefficients (and complex solutions), determinants of complex matrices, and vector spaces with scalar multiplication by any complex number allowed. Moreover, the proofs of most theorems about (the real version of) these concepts extend easily to the complex case. It is not our intention here to give a full treatment of complex linear algebra. However, we will carry the theory far enough to give another proof that the eigenvalues of a real symmetric matrix $A$ are real (Theorem 5.5.7) and to prove the spectral theorem, an extension of the principal axes theorem (Theorem 8.2.2).

The set of complex numbers is denoted $\mathbb{C}$ . We will use only the most basic properties of these numbers (mainly conjugation and absolute values), and the reader can find this material in Appendix A.

If $n \geq 1$, we denote the set of all $n$-tuples of complex numbers by $\mathbb{C}^n$. As with $\mathbb{R}^n$, these $n$-tuples will be written either as row or column matrices and will be referred to as **vectors**. We define vector operations on $\mathbb{C}^n$ as follows:

$$\begin{aligned}(v_1, v_2, \ldots, v_n) + (w_1, w_2, \ldots, w_n) &= (v_1 + w_1, v_2 + w_2, \ldots, v_n + w_n) \\ u(v_1, v_2, \ldots, v_n) &= (uv_1, uv_2, \ldots, uv_n) \quad \text{for } u \text{ in } \mathbb{C}\end{aligned}$$

With these definitions, $\mathbb{C}^n$ satisfies the axioms for a vector space (with complex scalars) given in Chapter 6. Thus we can speak of spanning sets for $\mathbb{C}^n$, of linearly independent subsets, and of bases. In all cases, the definitions are identical to the real case, except that the scalars are allowed to be complex numbers. In particular, the standard basis of $\mathbb{R}^n$ remains a basis of $\mathbb{C}^n$, called the **standard basis** of $\mathbb{C}^n$.

A matrix $A = \begin{bmatrix} a_{ij} \end{bmatrix}$ is called a **complex matrix** if every entry $a_{ij}$ is a complex number. The notion of conjugation for complex numbers extends to matrices as follows: Define the **conjugate** of $A = \begin{bmatrix} a_{ij} \end{bmatrix}$ to be the matrix

$$\overline{A} = \begin{bmatrix} \overline{a}_{ij} \end{bmatrix}$$

obtained from $A$ by conjugating every entry. Then (using Appendix A)

$$\overline{A + B} = \overline{A} + \overline{B} \quad \text{and} \quad \overline{AB} = \overline{A}\,\overline{B}$$

holds for all (complex) matrices of appropriate size.

## The Standard Inner Product

There is a natural generalization to $\mathbb{C}^n$ of the dot product in $\mathbb{R}^n$.

**Definition 8.15 Standard Inner Product in $\mathbb{R}^n$**

*Given $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ in $\mathbb{C}^n$, define their **standard inner product** $\langle \mathbf{z}, \mathbf{w} \rangle$ by*

$$\langle \mathbf{z}, \mathbf{w} \rangle = z_1\overline{w}_1 + z_2\overline{w}_2 + \cdots + z_n\overline{w}_n = \mathbf{z} \cdot \overline{\mathbf{w}}$$

*where $\overline{w}$ is the conjugate of the complex number $w$.*

Clearly, if $\mathbf{z}$ and $\mathbf{w}$ actually lie in $\mathbb{R}^n$, then $\langle \mathbf{z}, \mathbf{w} \rangle = \mathbf{z} \cdot \mathbf{w}$ is the usual dot product.

**Example 8.7.1**

If $\mathbf{z} = (2, 1-i, 2i, 3-i)$ and $\mathbf{w} = (1-i, -1, -i, 3+2i)$, then

$$\langle \mathbf{z}, \mathbf{w} \rangle = 2(1+i) + (1-i)(-1) + (2i)(i) + (3-i)(3-2i) = 6 - 6i$$
$$\langle \mathbf{z}, \mathbf{z} \rangle = 2 \cdot 2 + (1-i)(1+i) + (2i)(-2i) + (3-i)(3+i) = 20$$

Note that $\langle \mathbf{z}, \mathbf{w} \rangle$ is a complex number in general. However, if $\mathbf{w} = \mathbf{z} = (z_1, z_2, \ldots, z_n)$, the definition gives $\langle \mathbf{z}, \mathbf{z} \rangle = |z_1|^2 + \cdots + |z_n|^2$ which is a nonnegative real number, equal to 0 if and only if $\mathbf{z} = \mathbf{0}$. This explains the conjugation in the definition of $\langle \mathbf{z}, \mathbf{w} \rangle$, and it gives (4) of the following theorem.

**Theorem 8.7.1**

*Let $\mathbf{z}$, $\mathbf{z}_1$, $\mathbf{w}$, and $\mathbf{w}_1$ denote vectors in $\mathbb{C}^n$, and let $\lambda$ denote a complex number.*

1. $\langle \mathbf{z} + \mathbf{z}_1, \mathbf{w} \rangle = \langle \mathbf{z}, \mathbf{w} \rangle + \langle \mathbf{z}_1, \mathbf{w} \rangle$ *and* $\langle \mathbf{z}, \mathbf{w} + \mathbf{w}_1 \rangle = \langle \mathbf{z}, \mathbf{w} \rangle + \langle \mathbf{z}, \mathbf{w}_1 \rangle$.
2. $\langle \lambda \mathbf{z}, \mathbf{w} \rangle = \lambda \langle \mathbf{z}, \mathbf{w} \rangle$ *and* $\langle \mathbf{z}, \lambda \mathbf{w} \rangle = \overline{\lambda} \langle \mathbf{z}, \mathbf{w} \rangle$.
3. $\langle \mathbf{z}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{z} \rangle}$.
4. $\langle \mathbf{z}, \mathbf{z} \rangle \geq 0$, *and* $\langle \mathbf{z}, \mathbf{z} \rangle = 0$ *if and only if* $\mathbf{z} = \mathbf{0}$.

**Proof.** We leave (1) and (2) to the reader (Exercise 8.7.10), and (4) has already been proved. To prove (3), write $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)$. Then

$$\begin{aligned} \overline{\langle \mathbf{w}, \mathbf{z} \rangle} = \overline{(w_1\overline{z}_1 + \cdots + w_n\overline{z}_n)} &= \overline{w}_1\overline{\overline{z}}_1 + \cdots + \overline{w}_n\overline{\overline{z}}_n \\ &= z_1\overline{w}_1 + \cdots + z_n\overline{w}_n = \langle \mathbf{z}, \mathbf{w} \rangle \end{aligned}$$

□

**Definition 8.16 Norm and Length in $\mathbb{C}^n$**

*As for the dot product on $\mathbb{R}^n$, property (4) enables us to define the **norm** or **length** $\|\mathbf{z}\|$ of a vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ in $\mathbb{C}^n$:*

$$\|\mathbf{z}\| = \sqrt{\langle \mathbf{z}, \mathbf{z} \rangle} = \sqrt{|z_1|^2 + |z_2|^2 + \cdots + |z_n|^2}$$

The only properties of the norm function we will need are the following (the proofs are left to the reader):

**Theorem 8.7.2**

*If $\mathbf{z}$ is any vector in $\mathbb{C}^n$, then*

1. *$\|\mathbf{z}\| \geq 0$ and $\|\mathbf{z}\| = 0$ if and only if $\mathbf{z} = \mathbf{0}$.*
2. *$\|\lambda \mathbf{z}\| = |\lambda| \|\mathbf{z}\|$ for all complex numbers $\lambda$.*

A vector $\mathbf{u}$ in $\mathbb{C}^n$ is called a **unit vector** if $\|\mathbf{u}\| = 1$. Property (2) in Theorem 8.7.2 then shows that if $\mathbf{z} \neq \mathbf{0}$ is any nonzero vector in $\mathbb{C}^n$, then $\mathbf{u} = \frac{1}{\|\mathbf{z}\|}\mathbf{z}$ is a unit vector.

**Example 8.7.2**

In $\mathbb{C}^4$, find a unit vector $\mathbf{u}$ that is a positive real multiple of $\mathbf{z} = (1 - i, i, 2, 3 + 4i)$.

**Solution.** $\|\mathbf{z}\| = \sqrt{2 + 1 + 4 + 25} = \sqrt{32} = 4\sqrt{2}$, so take $\mathbf{u} = \frac{1}{4\sqrt{2}}\mathbf{z}$.

Transposition of complex matrices is defined just as in the real case, and the following notion is fundamental.

**Definition 8.17 Conjugate Transpose in $\mathbb{C}^n$**

*The **conjugate transpose** $A^H$ of a complex matrix $A$ is defined by*

$$A^H = (\overline{A})^T = \overline{(A^T)}$$

Observe that $A^H = A^T$ when $A$ is real.[14]

**Example 8.7.3**

$$\begin{bmatrix} 3 & 1-i & 2+i \\ 2i & 5+2i & -i \end{bmatrix}^H = \begin{bmatrix} 3 & -2i \\ 1+i & 5-2i \\ 2-i & i \end{bmatrix}$$

[14] Other notations for $A^H$ are $A^*$ and $A^\dagger$.

The following properties of $A^H$ follow easily from the rules for transposition of real matrices and extend these rules to complex matrices. Note the conjugate in property (3).

**Theorem 8.7.3**

*Let $A$ and $B$ denote complex matrices, and let $\lambda$ be a complex number.*

1. $(A^H)^H = A$.
2. $(A+B)^H = A^H + B^H$.
3. $(\lambda A)^H = \overline{\lambda} A^H$.
4. $(AB)^H = B^H A^H$.

## Hermitian and Unitary Matrices

If $A$ is a real symmetric matrix, it is clear that $A^H = A$. The complex matrices that satisfy this condition turn out to be the most natural generalization of the real symmetric matrices:

**Definition 8.18 Hermitian Matrices**

*A square complex matrix $A$ is called* ***hermitian***[15] *if $A^H = A$, equivalently if $\overline{A} = A^T$.*

Hermitian matrices are easy to recognize because the entries on the main diagonal must be real, and the "reflection" of each nondiagonal entry in the main diagonal must be the conjugate of that entry.

**Example 8.7.4**

$\begin{bmatrix} 3 & i & 2+i \\ -i & -2 & -7 \\ 2-i & -7 & 1 \end{bmatrix}$ is hermitian, whereas $\begin{bmatrix} 1 & i \\ i & -2 \end{bmatrix}$ and $\begin{bmatrix} 1 & i \\ -i & i \end{bmatrix}$ are not.

The following Theorem extends Theorem 8.2.3, and gives a very useful characterization of hermitian matrices in terms of the standard inner product in $\mathbb{C}^n$.

**Theorem 8.7.4**

*An $n \times n$ complex matrix $A$ is hermitian if and only if*

$$\langle A\mathbf{z},\ \mathbf{w}\rangle = \langle \mathbf{z},\ A\mathbf{w}\rangle$$

*for all $n$-tuples $\mathbf{z}$ and $\mathbf{w}$ in $\mathbb{C}^n$.*

[15] The name hermitian honours Charles Hermite (1822–1901), a French mathematician who worked primarily in analysis and is remembered as the first to show that the number $e$ from calculus is transcendental—that is, $e$ is not a root of any polynomial with integer coefficients.

**Proof.** If $A$ is hermitian, we have $A^T = \overline{A}$. If $\mathbf{z}$ and $\mathbf{w}$ are columns in $\mathbb{C}^n$, then $\langle \mathbf{z},\ \mathbf{w}\rangle = \mathbf{z}^T\overline{\mathbf{w}}$, so

$$\langle A\mathbf{z},\ \mathbf{w}\rangle = (A\mathbf{z})^T\overline{\mathbf{w}} = \mathbf{z}^TA^T\overline{\mathbf{w}} = \mathbf{z}^T\overline{A}\overline{\mathbf{w}} = \mathbf{z}^T(\overline{A\mathbf{w}}) = \langle \mathbf{z},\ A\mathbf{w}\rangle$$

To prove the converse, let $\mathbf{e}_j$ denote column $j$ of the identity matrix. If $A = \left[a_{ij}\right]$, the condition gives

$$\overline{a}_{ij} = \langle \mathbf{e}_i,\ A\mathbf{e}_j\rangle = \langle A\mathbf{e}_i,\ \mathbf{e}_j\rangle = a_{ij}$$

Hence $\overline{A} = A^T$, so $A$ is hermitian. □

Let $A$ be an $n \times n$ complex matrix. As in the real case, a complex number $\lambda$ is called an **eigenvalue** of $A$ if $A\mathbf{x} = \lambda\mathbf{x}$ holds for some column $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{C}^n$. In this case $\mathbf{x}$ is called an **eigenvector** of $A$ corresponding to $\lambda$. The **characteristic polynomial** $c_A(x)$ is defined by

$$c_A(x) = \det(xI - A)$$

This polynomial has complex coefficients (possibly nonreal). However, the proof of Theorem 3.3.2 goes through to show that the eigenvalues of $A$ are the roots (possibly complex) of $c_A(x)$.

It is at this point that the advantage of working with complex numbers becomes apparent. The real numbers are incomplete in the sense that the characteristic polynomial of a real matrix may fail to have all its roots real. However, this difficulty does not occur for the complex numbers. The so-called fundamental theorem of algebra ensures that *every* polynomial of positive degree with complex coefficients has a complex root. Hence every square complex matrix $A$ has a (complex) eigenvalue. Indeed (Appendix A), $c_A(x)$ factors completely as follows:

$$c_A(x) = (x - \lambda_1)(x - \lambda_2)\cdots(x - \lambda_n)$$

where $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_n$ are the eigenvalues of $A$ (with possible repetitions due to multiple roots).

The next result shows that, for hermitian matrices, the eigenvalues are actually real. Because symmetric real matrices are hermitian, this re-proves Theorem 5.5.7. It also extends Theorem 8.2.4, which asserts that eigenvectors of a symmetric real matrix corresponding to distinct eigenvalues are actually orthogonal. In the complex context, two $n$-tuples $\mathbf{z}$ and $\mathbf{w}$ in $\mathbb{C}^n$ are said to be **orthogonal** if $\langle \mathbf{z},\ \mathbf{w}\rangle = 0$.

**Theorem 8.7.5**

*Let $A$ denote a hermitian matrix.*

1. *The eigenvalues of $A$ are real.*
2. *Eigenvectors of $A$ corresponding to distinct eigenvalues are orthogonal.*

**Proof.** Let $\lambda$ and $\mu$ be eigenvalues of $A$ with (nonzero) eigenvectors $\mathbf{z}$ and $\mathbf{w}$. Then $A\mathbf{z} = \lambda\mathbf{z}$ and $A\mathbf{w} = \mu\mathbf{w}$, so Theorem 8.7.4 gives

$$\lambda\langle \mathbf{z},\ \mathbf{w}\rangle = \langle \lambda\mathbf{z},\ \mathbf{w}\rangle = \langle A\mathbf{z},\ \mathbf{w}\rangle = \langle \mathbf{z},\ A\mathbf{w}\rangle = \langle \mathbf{z},\ \mu\mathbf{w}\rangle = \overline{\mu}\langle \mathbf{z},\ \mathbf{w}\rangle \tag{8.6}$$

If $\mu = \lambda$ and $\mathbf{w} = \mathbf{z}$, this becomes $\lambda\langle \mathbf{z},\ \mathbf{z}\rangle = \overline{\lambda}\langle \mathbf{z},\ \mathbf{z}\rangle$. Because $\langle \mathbf{z},\ \mathbf{z}\rangle = \|\mathbf{z}\|^2 \neq 0$, this implies $\lambda = \overline{\lambda}$. Thus $\lambda$ is real, proving (1). Similarly, $\mu$ is real, so equation (8.6) gives $\lambda\langle \mathbf{z},\ \mathbf{w}\rangle = \mu\langle \mathbf{z},\ \mathbf{w}\rangle$. If $\lambda \neq \mu$, this implies $\langle \mathbf{z},\ \mathbf{w}\rangle = 0$, proving (2). □

The principal axes theorem (Theorem 8.2.2) asserts that every real symmetric matrix $A$ is orthogonally diagonalizable—that is $P^TAP$ is diagonal where $P$ is an orthogonal matrix ($P^{-1} = P^T$). The next theorem identifies the complex analogs of these orthogonal real matrices.

**Definition 8.19 Orthogonal and Orthonormal Vectors in $\mathbb{C}^n$**

*As in the real case, a set of nonzero vectors $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$ in $\mathbb{C}^n$ is called **orthogonal** if $\langle \mathbf{z}_i, \mathbf{z}_j \rangle = 0$ whenever $i \neq j$, and it is **orthonormal** if, in addition, $\|\mathbf{z}_i\| = 1$ for each $i$.*

**Theorem 8.7.6**

*The following are equivalent for an $n \times n$ complex matrix $A$.*

1. *$A$ is invertible and $A^{-1} = A^H$.*
2. *The rows of $A$ are an orthonormal set in $\mathbb{C}^n$.*
3. *The columns of $A$ are an orthonormal set in $\mathbb{C}^n$.*

**Proof.** If $A = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \end{bmatrix}$ is a complex matrix with $j$th column $\mathbf{c}_j$, then $A^T\overline{A} = \left[\langle \mathbf{c}_i, \mathbf{c}_j \rangle\right]$, as in Theorem 8.2.1. Now (1) $\Leftrightarrow$ (2) follows, and (1) $\Leftrightarrow$ (3) is proved in the same way. □

**Definition 8.20 Unitary Matrices**

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

Thus a real matrix is unitary if and only if it is orthogonal.

**Example 8.7.5**

The matrix $A = \begin{bmatrix} 1+i & 1 \\ 1-i & i \end{bmatrix}$ has orthogonal columns, but the rows are not orthogonal.

Normalizing the columns gives the unitary matrix $\frac{1}{2}\begin{bmatrix} 1+i & \sqrt{2} \\ 1-i & \sqrt{2}i \end{bmatrix}$.

Given a real symmetric matrix $A$, the diagonalization algorithm in Section 3.3 leads to a procedure for finding an orthogonal matrix $P$ such that $P^TAP$ is diagonal (see Example 8.2.4). The following example illustrates Theorem 8.7.5 and shows that the technique works for complex matrices.

**Example 8.7.6**

Consider the hermitian matrix $A = \begin{bmatrix} 3 & 2+i \\ 2-i & 7 \end{bmatrix}$. Find the eigenvalues of $A$, find two orthonormal eigenvectors, and so find a unitary matrix $U$ such that $U^HAU$ is diagonal.

**Solution.** The characteristic polynomial of $A$ is

$$c_A(x) = \det(xI - A) = \det \begin{bmatrix} x-3 & -2-i \\ -2+i & x-7 \end{bmatrix} = (x-2)(x-8)$$

Hence the eigenvalues are 2 and 8 (both real as expected), and corresponding eigenvectors are

$\begin{bmatrix} 2+i \\ -1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 2-i \end{bmatrix}$ (orthogonal as expected). Each has length $\sqrt{6}$ so, as in the (real) diagonalization algorithm, let $U = \frac{1}{\sqrt{6}} \begin{bmatrix} 2+i & 1 \\ -1 & 2-i \end{bmatrix}$ be the unitary matrix with the normalized eigenvectors as columns.
Then $U^H AU = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}$ is diagonal.

## Unitary Diagonalization

An $n \times n$ complex matrix $A$ is called **unitarily diagonalizable** if $U^H AU$ is diagonal for some unitary matrix $U$. As Example 8.7.6 suggests, we are going to prove that every hermitian matrix is unitarily diagonalizable. However, with only a little extra effort, we can get a very important theorem that has this result as an easy consequence.

A complex matrix is called **upper triangular** if every entry below the main diagonal is zero. We owe the following theorem to Issai Schur.[16]

**Theorem 8.7.7: Schur's Theorem**

*If $A$ is any $n \times n$ complex matrix, there exists a unitary matrix $U$ such that*

$$U^H AU = T$$

*is upper triangular. Moreover, the entries on the main diagonal of $T$ are the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $A$ (including multiplicities).*

**Proof.** We use induction on $n$. If $n = 1$, $A$ is already upper triangular. If $n > 1$, assume the theorem is valid for $(n-1) \times (n-1)$ complex matrices. Let $\lambda_1$ be an eigenvalue of $A$, and let $\mathbf{y}_1$ be an eigenvector with $\|\mathbf{y}_1\| = 1$. Then $\mathbf{y}_1$ is part of a basis of $\mathbb{C}^n$ (by the analog of Theorem 6.4.1), so the (complex analog of the) Gram-Schmidt process provides $\mathbf{y}_2, \ldots, \mathbf{y}_n$ such that $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ is an orthonormal basis of $\mathbb{C}^n$. If $U_1 = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n \end{bmatrix}$ is the matrix with these vectors as its columns, then (see Lemma 5.4.3)

$$U_1^H AU_1 = \begin{bmatrix} \lambda_1 & X_1 \\ 0 & A_1 \end{bmatrix}$$

in block form. Now apply induction to find a unitary $(n-1) \times (n-1)$ matrix $W_1$ such that $W_1^H A_1 W_1 = T_1$ is upper triangular. Then $U_2 = \begin{bmatrix} 1 & 0 \\ 0 & W_1 \end{bmatrix}$ is a unitary $n \times n$ matrix. Hence $U = U_1 U_2$ is unitary (using Theorem 8.7.6), and

$$\begin{aligned} U^H AU &= U_2^H (U_1^H AU_1) U_2 \\ &= \begin{bmatrix} 1 & 0 \\ 0 & W_1^H \end{bmatrix} \begin{bmatrix} \lambda_1 & X_1 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & W_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & X_1 W_1 \\ 0 & T_1 \end{bmatrix} \end{aligned}$$

[16] Issai Schur (1875–1941) was a German mathematician who did fundamental work in the theory of representations of groups as matrices.

is upper triangular. Finally, $A$ and $U^HAU = T$ have the same eigenvalues by (the complex version of) Theorem 5.5.1, and they are the diagonal entries of $T$ because $T$ is upper triangular. □

The fact that similar matrices have the same traces and determinants gives the following consequence of Schur's theorem.

**Corollary 8.7.1**

*Let $A$ be an $n \times n$ complex matrix, and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the eigenvalues of $A$, including multiplicities. Then*

$$\det A = \lambda_1\lambda_2\cdots\lambda_n \quad \text{and} \quad \operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

Schur's theorem asserts that every complex matrix can be "unitarily triangularized." However, we cannot substitute "unitarily diagonalized" here. In fact, if $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, there is no invertible complex matrix $U$ at all such that $U^{-1}AU$ is diagonal. However, the situation is much better for hermitian matrices.

**Theorem 8.7.8: Spectral Theorem**

*If $A$ is hermitian, there is a unitary matrix $U$ such that $U^HAU$ is diagonal.*

**Proof.** By Schur's theorem, let $U^HAU = T$ be upper triangular where $U$ is unitary. Since $A$ is hermitian, this gives

$$T^H = (U^HAU)^H = U^HA^HU^{HH} = U^HAU = T$$

This means that $T$ is both upper and lower triangular. Hence $T$ is actually diagonal. □

The principal axes theorem asserts that a real matrix $A$ is symmetric if and only if it is orthogonally diagonalizable (that is, $P^TAP$ is diagonal for some real orthogonal matrix $P$). Theorem 8.7.8 is the complex analog of half of this result. However, the converse is false for complex matrices: There exist unitarily diagonalizable matrices that are not hermitian.

**Example 8.7.7**

Show that the non-hermitian matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ is unitarily diagonalizable.

**Solution.** The characteristic polynomial is $c_A(x) = x^2 + 1$. Hence the eigenvalues are $i$ and $-i$, and it is easy to verify that $\begin{bmatrix} i \\ -1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ i \end{bmatrix}$ are corresponding eigenvectors. Moreover, these eigenvectors are orthogonal and both have length $\sqrt{2}$, so $U = \frac{1}{\sqrt{2}}\begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix}$ is a unitary matrix such that $U^HAU = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$ is diagonal.

There is a very simple way to characterize those complex matrices that are unitarily diagonalizable. To this end, an $n \times n$ complex matrix $N$ is called **normal** if $NN^H = N^HN$. It is clear that every hermitian

or unitary matrix is normal, as is the matrix $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ in Example 8.7.7. In fact we have the following result.

**Theorem 8.7.9**

*An $n \times n$ complex matrix $A$ is unitarily diagonalizable if and only if $A$ is normal.*

**Proof.** Assume first that $U^H AU = D$, where $U$ is unitary and $D$ is diagonal. Then $DD^H = D^H D$ as is easily verified. Because $DD^H = U^H(AA^H)U$ and $D^H D = U^H(A^H A)U$, it follows by cancellation that $AA^H = A^H A$.

Conversely, assume $A$ is normal—that is, $AA^H = A^H A$. By Schur's theorem, let $U^H AU = T$, where $T$ is upper triangular and $U$ is unitary. Then $T$ is normal too:

$$TT^H = U^H(AA^H)U = U^H(A^H A)U = T^H T$$

Hence it suffices to show that a normal $n \times n$ upper triangular matrix $T$ must be diagonal. We induct on $n$; it is clear if $n = 1$. If $n > 1$ and $T = \begin{bmatrix} t_{ij} \end{bmatrix}$, then equating $(1, 1)$-entries in $TT^H$ and $T^H T$ gives

$$|t_{11}|^2 + |t_{12}|^2 + \cdots + |t_{1n}|^2 = |t_{11}|^2$$

This implies $t_{12} = t_{13} = \cdots = t_{1n} = 0$, so $T = \begin{bmatrix} t_{11} & 0 \\ 0 & T_1 \end{bmatrix}$ in block form. Hence $T = \begin{bmatrix} \overline{t}_{11} & 0 \\ 0 & T_1^H \end{bmatrix}$ so $TT^H = T^H T$ implies $T_1 T_1^H = T_1 T_1^H$. Thus $T_1$ is diagonal by induction, and the proof is complete. □

We conclude this section by using Schur's theorem (Theorem 8.7.7) to prove a famous theorem about matrices. Recall that the characteristic polynomial of a square matrix $A$ is defined by $c_A(x) = \det(xI - A)$, and that the eigenvalues of $A$ are just the roots of $c_A(x)$.

**Theorem 8.7.10: Cayley-Hamilton Theorem**[17]

*If $A$ is an $n \times n$ complex matrix, then $c_A(A) = 0$; that is, $A$ is a root of its characteristic polynomial.*

**Proof.** If $p(x)$ is any polynomial with complex coefficients, then $p(P^{-1}AP) = P^{-1}p(A)P$ for any invertible complex matrix $P$. Hence, by Schur's theorem, we may assume that $A$ is upper triangular. Then the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $A$ appear along the main diagonal, so

$$c_A(x) = (x - \lambda_1)(x - \lambda_2)(x - \lambda_3) \cdots (x - \lambda_n)$$

Thus

$$c_A(A) = (A - \lambda_1 I)(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)$$

Note that each matrix $A - \lambda_i I$ is upper triangular. Now observe:

1. $A - \lambda_1 I$ has zero first column because column 1 of $A$ is $(\lambda_1, 0, 0, \ldots, 0)^T$.

[17] Named after the English mathematician Arthur Cayley (1821–1895) and William Rowan Hamilton (1805–1865), an Irish mathematician famous for his work on physical dynamics.

2. Then $(A-\lambda_1 I)(A-\lambda_2 I)$ has the first two columns zero because the second column of $(A-\lambda_2 I)$ is $(b,\ 0,\ 0,\ \ldots,\ 0)^T$ for some constant $b$.

3. Next $(A-\lambda_1 I)(A-\lambda_2 I)(A-\lambda_3 I)$ has the first three columns zero because column 3 of $(A-\lambda_3 I)$ is $(c,\ d,\ 0,\ \ldots,\ 0)^T$ for some constants $c$ and $d$.

Continuing in this way we see that $(A-\lambda_1 I)(A-\lambda_2 I)(A-\lambda_3 I)\cdots(A-\lambda_n I)$ has all $n$ columns zero; that is, $c_A(A)=0$. □

## Exercises for 8.7

**Exercise 8.7.1** In each case, compute the norm of the complex vector.

a. $(1,\ 1-i,\ -2,\ i)$

b. $(1-i,\ 1+i,\ 1,\ -1)$

c. $(2+i,\ 1-i,\ 2,\ 0,\ -i)$

d. $(-2,\ -i,\ 1+i,\ 1-i,\ 2i)$

**Exercise 8.7.2** In each case, determine whether the two vectors are orthogonal.

a. $(4,\ -3i,\ 2+i)$, $(i,\ 2,\ 2-4i)$

b. $(i,\ -i,\ 2+i)$, $(i,\ i,\ 2-i)$

c. $(1,\ 1,\ i,\ i)$, $(1,\ i,\ -i,\ 1)$

d. $(4+4i,\ 2+i,\ 2i)$, $(-1+i,\ 2,\ 3-2i)$

**Exercise 8.7.3** A subset $U$ of $\mathbb{C}^n$ is called a **complex subspace** of $\mathbb{C}^n$ if it contains 0 and if, given $\mathbf{v}$ and $\mathbf{w}$ in $U$, both $\mathbf{v}+\mathbf{w}$ and $z\mathbf{v}$ lie in $U$ ($z$ any complex number). In each case, determine whether $U$ is a complex subspace of $\mathbb{C}^3$.

a. $U=\{(w,\ \overline{w},\ 0)\mid w \text{ in } \mathbb{C}\}$

b. $U=\{(w,\ 2w,\ a)\mid w \text{ in } \mathbb{C},\ a \text{ in } \mathbb{R}\}$

c. $U=\mathbb{R}^3$

d. $U=\{(v+w,\ v-2w,\ v)\mid v,\ w \text{ in } \mathbb{C}\}$

**Exercise 8.7.4** In each case, find a basis over $\mathbb{C}$, and determine the dimension of the complex subspace $U$ of $\mathbb{C}^3$ (see the previous exercise).

a. $U=\{(w,\ v+w,\ v-iw)\mid v,\ w \text{ in } \mathbb{C}\}$

b. $U=\{(iv+w,\ 0,\ 2v-w)\mid v,\ w \text{ in } \mathbb{C}\}$

c. $U=\{(u,\ v,\ w)\mid iu-3v+(1-i)w=0;\ u,\ v,\ w \text{ in } \mathbb{C}\}$

d. $U=\{(u,\ v,\ w)\mid 2u+(1+i)v-iw=0;\ u,\ v,\ w \text{ in } \mathbb{C}\}$

**Exercise 8.7.5** In each case, determine whether the given matrix is hermitian, unitary, or normal.

a. $\begin{bmatrix} 1 & -i \\ i & i \end{bmatrix}$ b. $\begin{bmatrix} 2 & 3 \\ -3 & 2 \end{bmatrix}$

c. $\begin{bmatrix} 1 & i \\ -i & 2 \end{bmatrix}$ d. $\begin{bmatrix} 1 & -i \\ i & -1 \end{bmatrix}$

e. $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ f. $\begin{bmatrix} 1 & 1+i \\ 1+i & i \end{bmatrix}$

g. $\begin{bmatrix} 1+i & 1 \\ -i & -1+i \end{bmatrix}$ h. $\frac{1}{\sqrt{2}|z|}\begin{bmatrix} z & z \\ \overline{z} & -\overline{z} \end{bmatrix}$, $z\neq 0$

**Exercise 8.7.6** Show that a matrix $N$ is normal if and only if $\overline{N}N^T=N^T\overline{N}$.

**Exercise 8.7.7** Let $A=\begin{bmatrix} z & \overline{v} \\ v & w \end{bmatrix}$ where $v$, $w$, and $z$ are complex numbers. Characterize in terms of $v$, $w$, and $z$ when $A$ is

a. hermitian b. unitary

c. normal.

**Exercise 8.7.8** In each case, find a unitary matrix $U$ such that $U^HAU$ is diagonal.

a. $A=\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$

b. $A = \begin{bmatrix} 4 & 3-i \\ 3+i & 1 \end{bmatrix}$

c. $A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$; $a$, $b$, real

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

e. $A = \begin{bmatrix} 1 & 0 & 1+i \\ 0 & 2 & 0 \\ 1-i & 0 & 0 \end{bmatrix}$

f. $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1+i \\ 0 & 1-i & 2 \end{bmatrix}$

**Exercise 8.7.9** Show that $\langle A\mathbf{x},\ \mathbf{y}\rangle = \langle \mathbf{x},\ A^H\mathbf{y}\rangle$ holds for all $n \times n$ matrices $A$ and for all $n$-tuples $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{C}^n$.

**Exercise 8.7.10**

a. Prove (1) and (2) of Theorem 8.7.1.

b. Prove Theorem 8.7.2.

c. Prove Theorem 8.7.3.

**Exercise 8.7.11**

a. Show that $A$ is hermitian if and only if $\overline{A} = A^T$.

b. Show that the diagonal entries of any hermitian matrix are real.

**Exercise 8.7.12**

a. Show that every complex matrix $Z$ can be written uniquely in the form $Z = A + iB$, where $A$ and $B$ are real matrices.

b. If $Z = A + iB$ as in (a), show that $Z$ is hermitian if and only if $A$ is symmetric, and $B$ is skew-symmetric (that is, $B^T = -B$).

**Exercise 8.7.13** If $Z$ is any complex $n \times n$ matrix, show that $ZZ^H$ and $Z + Z^H$ are hermitian.

**Exercise 8.7.14** A complex matrix $B$ is called **skew-hermitian** if $B^H = -B$.

a. Show that $Z - Z^H$ is skew-hermitian for any square complex matrix $Z$.

b. If $B$ is skew-hermitian, show that $B^2$ and $iB$ are hermitian.

c. If $B$ is skew-hermitian, show that the eigenvalues of $B$ are pure imaginary ($i\lambda$ for real $\lambda$).

d. Show that every $n \times n$ complex matrix $Z$ can be written uniquely as $Z = A + B$, where $A$ is hermitian and $B$ is skew-hermitian.

**Exercise 8.7.15** Let $U$ be a unitary matrix. Show that:

a. $\|U\mathbf{x}\| = \|\mathbf{x}\|$ for all columns $\mathbf{x}$ in $\mathbb{C}^n$.

b. $|\lambda| = 1$ for every eigenvalue $\lambda$ of $U$.

**Exercise 8.7.16**

a. If $Z$ is an invertible complex matrix, show that $Z^H$ is invertible and that $(Z^H)^{-1} = (Z^{-1})^H$.

b. Show that the inverse of a unitary matrix is again unitary.

c. If $U$ is unitary, show that $U^H$ is unitary.

**Exercise 8.7.17** Let $Z$ be an $m \times n$ matrix such that $Z^HZ = I_n$ (for example, $Z$ is a unit column in $\mathbb{C}^n$).

a. Show that $V = ZZ^H$ is hermitian and satisfies $V^2 = V$.

b. Show that $U = I - 2ZZ^H$ is both unitary and hermitian (so $U^{-1} = U^H = U$).

**Exercise 8.7.18**

a. If $N$ is normal, show that $zN$ is also normal for all complex numbers $z$.

b. Show that (a) fails if *normal* is replaced by *hermitian*.

**Exercise 8.7.19** Show that a real $2 \times 2$ normal matrix is either symmetric or has the form $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$.

**Exercise 8.7.20** If $A$ is hermitian, show that all the coefficients of $c_A(x)$ are real numbers.

**Exercise 8.7.21**

a. If $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, show that $U^{-1}AU$ is not diagonal for any invertible complex matrix $U$.

b. If $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, show that $U^{-1}AU$ is not upper triangular for any *real* invertible matrix $U$.

**Exercise 8.7.22** If $A$ is any $n \times n$ matrix, show that $U^H AU$ is lower triangular for some unitary matrix $U$.

**Exercise 8.7.23** If $A$ is a $3 \times 3$ matrix, show that $A^2 = 0$ if and only if there exists a unitary matrix $U$ such that $U^H AU$ has the form $\begin{bmatrix} 0 & 0 & u \\ 0 & 0 & v \\ 0 & 0 & 0 \end{bmatrix}$ or the form $\begin{bmatrix} 0 & u & v \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

**Exercise 8.7.24** If $A^2 = A$, show that $\operatorname{rank} A = \operatorname{tr} A$. [*Hint*: Use Schur's theorem.]

**Exercise 8.7.25** Let $A$ be any $n \times n$ complex matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Show that $A = P + N$ where $N^n = 0$ and $P = UDU^T$ where $U$ is unitary and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. [Hint: Schur's theorem]

## 8.8 An Application to Linear Codes over Finite Fields

For centuries mankind has been using codes to transmit messages. In many cases, for example transmitting financial, medical, or military information, the message is disguised in such a way that it cannot be understood by an intruder who intercepts it, but can be easily "decoded" by the intended receiver. This subject is called *cryptography* and, while intriguing, is not our focus here. Instead, we investigate methods for detecting and correcting errors in the transmission of the message.

The stunning photos of the planet Saturn sent by the space probe are a very good example of how successful these methods can be. These messages are subject to "noise" such as solar interference which causes errors in the message. The signal is received on Earth with errors that must be detected and corrected before the high-quality pictures can be printed. This is done using error-correcting codes. To see how, we first discuss a system of adding and multiplying integers while ignoring multiples of a fixed integer.

### Modular Arithmetic

We work in the set $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$ of **integers**, that is the set of whole numbers. Everyone is familiar with the process of "long division" from arithmetic. For example, we can divide an integer $a$ by 5 and leave a remainder "modulo 5" in the set $\{0, 1, 2, 3, 4\}$. As an illustration

$$19 = 3 \cdot 5 + 4$$

so the remainder of 19 modulo 5 is 4. Similarly, the remainder of 137 modulo 5 is 2 because we have $137 = 27 \cdot 5 + 2$. This works even for negative integers: For example,

$$-17 = (-4) \cdot 5 + 3$$

so the remainder of $-17$ modulo 5 is 3.

This process is called the **division algorithm**. More formally, let $n \geq 2$ denote an integer. Then every integer $a$ can be written uniquely in the form

$$a = qn + r \quad \text{where } q \text{ and } r \text{ are integers and } 0 \leq r \leq n - 1$$

Here $q$ is called the **quotient** of $a$ **modulo** $n$, and $r$ is called the **remainder** of $a$ **modulo** $n$. We refer to $n$ as the **modulus**. Thus, if $n = 6$, the fact that $134 = 22 \cdot 6 + 2$ means that 134 has quotient 22 and remainder 2 modulo 6.

Our interest here is in the set of *all* possible remainders modulo $n$. This set is denoted

$$\mathbb{Z}_n = \{0,\ 1,\ 2,\ 3,\ \ldots,\ n-1\}$$

and is called the set of **integers modulo** $n$. Thus every integer is uniquely represented in $\mathbb{Z}_n$ by its remainder modulo $n$.

We are going to show how to do arithmetic in $\mathbb{Z}_n$ by adding and multiplying modulo $n$. That is, we add or multiply two numbers in $\mathbb{Z}_n$ by calculating the usual sum or product in $\mathbb{Z}$ and taking the remainder modulo $n$. It is proved in books on abstract algebra that the usual laws of arithmetic hold in $\mathbb{Z}_n$ for any modulus $n \geq 2$. This seems remarkable until we remember that these laws are true for ordinary addition and multiplication and all we are doing is reducing modulo $n$.

To illustrate, consider the case $n = 6$, so that $\mathbb{Z}_6 = \{0,\ 1,\ 2,\ 3,\ 4,\ 5\}$. Then $2+5 = 1$ in $\mathbb{Z}_6$ because 7 leaves a remainder of 1 when divided by 6. Similarly, $2 \cdot 5 = 4$ in $\mathbb{Z}_6$, while $3+5 = 2$, and $3+3 = 0$. In this way we can fill in the addition and multiplication tables for $\mathbb{Z}_6$; the result is:

Tables for $\mathbb{Z}_6$

| $+$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 2 | 3 | 4 | 5 |
| 1 | 1 | 2 | 3 | 4 | 5 | 0 |
| 2 | 2 | 3 | 4 | 5 | 0 | 1 |
| 3 | 3 | 4 | 5 | 0 | 1 | 2 |
| 4 | 4 | 5 | 0 | 1 | 2 | 3 |
| 5 | 5 | 0 | 1 | 2 | 3 | 4 |

| $\times$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 2 | 3 | 4 | 5 |
| 2 | 0 | 2 | 4 | 0 | 2 | 4 |
| 3 | 0 | 3 | 0 | 3 | 0 | 3 |
| 4 | 0 | 4 | 2 | 0 | 4 | 2 |
| 5 | 0 | 5 | 4 | 3 | 2 | 1 |

Calculations in $\mathbb{Z}_6$ are carried out much as in $\mathbb{Z}$ . As an illustration, consider the familiar "distributive law" $a(b+c) = ab+ac$ from ordinary arithmetic. This holds for all $a$, $b$, and $c$ in $\mathbb{Z}_6$; we verify a particular case:

$$3(5+4) = 3 \cdot 5 + 3 \cdot 4 \quad \text{in } \mathbb{Z}_6$$

In fact, the left side is $3(5+4) = 3 \cdot 3 = 3$, and the right side is $(3 \cdot 5) + (3 \cdot 4) = 3+0 = 3$ too. Hence doing arithmetic in $\mathbb{Z}_6$ is familiar. However, there are differences. For example, $3 \cdot 4 = 0$ in $\mathbb{Z}_6$, in contrast to the fact that $a \cdot b = 0$ in $\mathbb{Z}$ can only happen when either $a = 0$ or $b = 0$. Similarly, $3^2 = 3$ in $\mathbb{Z}_6$, unlike $\mathbb{Z}$.

Note that we will make statements like $-30 = 19$ in $\mathbb{Z}_7$; it means that $-30$ and 19 leave the same remainder 5 when divided by 7, and so are equal in $\mathbb{Z}_7$ because they both equal 5. In general, if $n \geq 2$ is any modulus, the operative fact is that

$$a = b \text{ in } \mathbb{Z}_n \quad \text{if and only if} \quad a - b \text{ is a multiple of } n$$

In this case we say that $a$ and $b$ are **equal modulo** $n$, and write $a = b (\bmod n)$.

Arithmetic in $\mathbb{Z}_n$ is, in a sense, simpler than that for the integers. For example, consider negatives. Given the element 8 in $\mathbb{Z}_{17}$, what is $-8$? The answer lies in the observation that $8+9 = 0$ in $\mathbb{Z}_{17}$, so $-8 = 9$ (and $-9 = 8$). In the same way, finding negatives is not difficult in $\mathbb{Z}_n$ for any modulus $n$.

## Finite Fields

In our study of linear algebra so far the scalars have been real (possibly complex) numbers. The set $\mathbb{R}$ of real numbers has the property that it is closed under addition and multiplication, that the usual laws of arithmetic hold, and that every nonzero real number has an inverse in $\mathbb{R}$. Such a system is called a **field**. Hence the real numbers $\mathbb{R}$ form a field, as does the set $\mathbb{C}$ of complex numbers. Another example is the set $\mathbb{Q}$ of all rational numbers (fractions); however the set $\mathbb{Z}$ of integers is *not* a field—for example, 2 has no inverse *in* the set $\mathbb{Z}$ because $2 \cdot x = 1$ has no solution $x$ in $\mathbb{Z}$ .

Our motivation for isolating the concept of a field is that nearly everything we have done remains valid if the scalars are restricted to some field: The gaussian algorithm can be used to solve systems of linear equations with coefficients in the field; a square matrix with entries from the field is invertible if and only if its determinant is nonzero; the matrix inversion algorithm works in the same way; and so on. The reason is that the field has all the properties used in the proofs of these results for the field $\mathbb{R}$, so all the theorems remain valid.

It turns out that there are *finite* fields—that is, finite sets that satisfy the usual laws of arithmetic and in which every nonzero element $a$ has an **inverse**, that is an element $b$ in the field such that $ab = 1$. If $n \geq 2$ is an integer, the modular system $\mathbb{Z}_n$ certainly satisfies the basic laws of arithmetic, but it need not be a field. For example we have $2 \cdot 3 = 0$ in $\mathbb{Z}_6$ so 3 has no inverse in $\mathbb{Z}_6$ (if $3a = 1$ then $2 = 2 \cdot 1 = 2(3a) = 0a = 0$ in $\mathbb{Z}_6$, a contradiction). The problem is that $6 = 2 \cdot 3$ can be properly factored in $\mathbb{Z}$.

An integer $p \geq 2$ is called a **prime** if $p$ *cannot* be factored as $p = ab$ where $a$ and $b$ are positive integers and neither $a$ nor $b$ equals 1. Thus the first few primes are 2, 3, 5, 7, 11, 13, 17, .... If $n \geq 2$ is not a prime and $n = ab$ where $2 \leq a$, $b \leq n - 1$, then $ab = 0$ in $\mathbb{Z}_n$ and it follows (as above in the case $n = 6$) that $b$ cannot have an inverse in $\mathbb{Z}_n$, and hence that $\mathbb{Z}_n$ is not a field. In other words, if $\mathbb{Z}_n$ is a field, then $n$ must be a prime. Surprisingly, the converse is true:

**Theorem 8.8.1**

*If $p$ is a prime, then $\mathbb{Z}_p$ is a field using addition and multiplication modulo $p$.*

The proof can be found in books on abstract algebra.[18] If $p$ is a prime, the field $\mathbb{Z}_p$ is called the **field of integers modulo** $p$.

For example, consider the case $n = 5$. Then $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$ and the addition and multiplication tables are:

| $+$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 2 | 3 | 4 |
| 1 | 1 | 2 | 3 | 4 | 0 |
| 2 | 2 | 3 | 4 | 0 | 1 |
| 3 | 3 | 4 | 0 | 1 | 2 |
| 4 | 4 | 0 | 1 | 2 | 3 |

| $\times$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 2 | 3 | 4 |
| 2 | 0 | 2 | 4 | 1 | 3 |
| 3 | 0 | 3 | 1 | 4 | 2 |
| 4 | 0 | 4 | 3 | 2 | 1 |

Hence 1 and 4 are self-inverse in $\mathbb{Z}_5$, and 2 and 3 are inverses of each other, so $\mathbb{Z}_5$ is indeed a field. Here is another important example.

[18] See, for example, W. Keith Nicholson, Introduction to Abstract Algebra, 4th ed., (New York: Wiley, 2012).

**Example 8.8.1**

If $p = 2$, then $\mathbb{Z}_2 = \{0, 1\}$ is a field with addition and multiplication modulo 2 given by the tables

$$\begin{array}{c|cc} + & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 1 & 0 \end{array} \quad \text{and} \quad \begin{array}{c|cc} \times & 0 & 1 \\ \hline 0 & 0 & 0 \\ 1 & 0 & 1 \end{array}$$

This is binary arithmetic, the basic algebra of computers.

While it is routine to find negatives of elements of $\mathbb{Z}_p$, it is a bit more difficult to find inverses in $\mathbb{Z}_p$. For example, how does one find $14^{-1}$ in $\mathbb{Z}_{17}$? Since we want $14^{-1} \cdot 14 = 1$ in $\mathbb{Z}_{17}$, we are looking for an integer $a$ with the property that $a \cdot 14 = 1$ modulo 17. Of course we can try all possibilities in $\mathbb{Z}_{17}$ (there are only 17 of them!), and the result is $a = 11$ (verify). However this method is of little use for large primes $p$, and it is a comfort to know that there is a systematic procedure (called the **euclidean algorithm**) for finding inverses in $\mathbb{Z}_p$ for any prime $p$. Furthermore, this algorithm is easy to program for a computer. To illustrate the method, let us once again find the inverse of 14 in $\mathbb{Z}_{17}$.

**Example 8.8.2**

Find the inverse of 14 in $\mathbb{Z}_{17}$.

**Solution.** The idea is to first divide $p = 17$ by 14:

$$17 = 1 \cdot 14 + 3$$

Now divide (the previous divisor) 14 by the new remainder 3 to get

$$14 = 4 \cdot 3 + 2$$

and then divide (the previous divisor) 3 by the new remainder 2 to get

$$3 = 1 \cdot 2 + 1$$

It is a theorem of number theory that, because 17 is a prime, this procedure will *always* lead to a remainder of 1. At this point we eliminate remainders in these equations from the bottom up:

$$\begin{aligned} 1 &= 3 - 1 \cdot 2 && \text{since } 3 = 1 \cdot 2 + 1 \\ &= 3 - 1 \cdot (14 - 4 \cdot 3) = 5 \cdot 3 - 1 \cdot 14 && \text{since } 2 = 14 - 4 \cdot 3 \\ &= 5 \cdot (17 - 1 \cdot 14) - 1 \cdot 14 = 5 \cdot 17 - 6 \cdot 14 && \text{since } 3 = 17 - 1 \cdot 14 \end{aligned}$$

Hence $(-6) \cdot 14 = 1$ in $\mathbb{Z}_{17}$, that is, $11 \cdot 14 = 1$. So $14^{-1} = 11$ in $\mathbb{Z}_{17}$.

As mentioned above, nearly everything we have done with matrices over the field of real numbers can be done in the same way for matrices with entries from $\mathbb{Z}_p$. We illustrate this with one example. Again the reader is referred to books on abstract algebra.

**Example 8.8.3**

Determine if the matrix $A = \begin{bmatrix} 1 & 4 \\ 6 & 5 \end{bmatrix}$ from $\mathbb{Z}_7$ is invertible and, if so, find its inverse.

**Solution.** Working in $\mathbb{Z}_7$ we have $\det A = 1 \cdot 5 - 6 \cdot 4 = 5 - 3 = 2 \neq 0$ in $\mathbb{Z}_7$, so $A$ is invertible. Hence Example 2.4.4 gives $A^{-1} = 2^{-1} \begin{bmatrix} 5 & -4 \\ -6 & 1 \end{bmatrix}$. Note that $2^{-1} = 4$ in $\mathbb{Z}_7$ (because $2 \cdot 4 = 1$ in $\mathbb{Z}_7$). Note also that $-4 = 3$ and $-6 = 1$ in $\mathbb{Z}_7$, so finally $A^{-1} = 4 \begin{bmatrix} 5 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 4 & 4 \end{bmatrix}$. The reader can verify that indeed $\begin{bmatrix} 1 & 4 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} 6 & 5 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ in $\mathbb{Z}_7$.

While we shall not use them, there are finite fields other than $\mathbb{Z}_p$ for the various primes $p$. Surprisingly, for every prime $p$ and every integer $n \geq 1$, there *exists* a field with exactly $p^n$ elements, and this field is *unique*.[19] It is called the **Galois field** of order $p^n$, and is denoted $GF(p^n)$.

## Error Correcting Codes

Coding theory is concerned with the transmission of information over a *channel* that is affected by *noise*. The noise causes errors, so the aim of the theory is to find ways to detect such errors and correct at least some of them. General coding theory originated with the work of Claude Shannon (1916–2001) who showed that information can be transmitted at near optimal rates with arbitrarily small chance of error.

Let $F$ denote a finite field and, if $n \geq 1$, let

$$F^n \text{ denote the } F\text{-vector space of } 1 \times n \text{ row matrices over } F$$

with the usual componentwise addition and scalar multiplication. In this context, the rows in $F^n$ are called **words** (or $n$-**words**) and, as the name implies, will be written as $[a\ b\ c\ d] = abcd$. The individual components of a word are called its **digits**. A nonempty subset $C$ of $F^n$ is called a **code** (or an $n$-**code**), and the elements in $C$ are called **code words**. If $F = \mathbb{Z}_2$, these are called **binary** codes.

If a code word $\mathbf{w}$ is transmitted and an error occurs, the resulting word $\mathbf{v}$ is decoded as the code word "closest" to $\mathbf{v}$ in $F^n$. To make sense of what "closest" means, we need a distance function on $F^n$ analogous to that in $\mathbb{R}^n$ (see Theorem 5.3.3). The usual definition in $\mathbb{R}^n$ does not work in this situation. For example, if $\mathbf{w} = 1111$ in $(\mathbb{Z}_2)^4$ then the square of the distance of $\mathbf{w}$ from $\mathbf{0}$ is

$$(1-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 = 0$$

even though $\mathbf{w} \neq \mathbf{0}$.

However there is a satisfactory notion of distance in $F^n$ due to Richard Hamming (1915–1998). Given a word $\mathbf{w} = a_1 a_2 \cdots a_n$ in $F^n$, we first define the **Hamming weight** $wt(\mathbf{w})$ to be the number of nonzero digits in $\mathbf{w}$:

$$wt(\mathbf{w}) = wt(a_1 a_2 \cdots a_n) = |\{i \mid a_i \neq 0\}|$$

Clearly, $0 \leq wt(\mathbf{w}) \leq n$ for every word $\mathbf{w}$ in $F^n$. Given another word $\mathbf{v} = b_1 b_2 \cdots b_n$ in $F^n$, the **Hamming distance** $d(\mathbf{v}, \mathbf{w})$ between $\mathbf{v}$ and $\mathbf{w}$ is defined by

$$d(\mathbf{v}, \mathbf{w}) = wt(\mathbf{v} - \mathbf{w}) = |\{i \mid b_i \neq a_i\}|$$

[19] See, for example, W. K. Nicholson, Introduction to Abstract Algebra, 4th ed., (New York: Wiley, 2012).

In other words, $d(\mathbf{v}, \mathbf{w})$ is the number of places at which the digits of $\mathbf{v}$ and $\mathbf{w}$ differ. The next result justifies using the term *distance* for this function $d$.

**Theorem 8.8.2**

*Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ denote words in $F^n$. Then:*

1. $d(\mathbf{v}, \mathbf{w}) \geq 0$.
2. $d(\mathbf{v}, \mathbf{w}) = 0$ *if and only if* $\mathbf{v} = \mathbf{w}$.
3. $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v})$.
4. $d(\mathbf{v}, \mathbf{w}) \leq d(\mathbf{v}, \mathbf{u}) + d(\mathbf{u}, \mathbf{w})$

**Proof.** (1) and (3) are clear, and (2) follows because $wt(\mathbf{v}) = 0$ if and only if $\mathbf{v} = \mathbf{0}$. To prove (4), write $\mathbf{x} = \mathbf{v} - \mathbf{u}$ and $\mathbf{y} = \mathbf{u} - \mathbf{w}$. Then (4) reads $wt(\mathbf{x} + \mathbf{y}) \leq wt(\mathbf{x}) + wt(\mathbf{y})$. If $\mathbf{x} = a_1 a_2 \cdots a_n$ and $\mathbf{y} = b_1 b_2 \cdots b_n$, this follows because $a_i + b_i \neq 0$ implies that either $a_i \neq 0$ or $b_i \neq 0$. □

Given a word $\mathbf{w}$ in $F^n$ and a real number $r > 0$, define the **ball** $B_r(\mathbf{w})$ of radius $r$ (or simply the $r$-**ball**) about $\mathbf{w}$ as follows:

$$B_r(\mathbf{w}) = \{\mathbf{x} \in F^n \mid d(\mathbf{w}, \mathbf{x}) \leq r\}$$

Using this we can describe one of the most useful decoding methods.

**Nearest Neighbour Decoding**

*Let $C$ be an $n$-code, and suppose a word $\mathbf{v}$ is transmitted and $\mathbf{w}$ is received. Then $\mathbf{w}$ is decoded as the code word in $C$ closest to it. (If there is a tie, choose arbitrarily.)*

Using this method, we can describe how to construct a code $C$ that can detect (or correct) $t$ errors. Suppose a code word $\mathbf{c}$ is transmitted and a word $\mathbf{w}$ is received with $s$ errors where $1 \leq s \leq t$. Then $s$ is the number of places at which the $\mathbf{c}$- and $\mathbf{w}$-digits differ, that is, $s = d(\mathbf{c}, \mathbf{w})$. Hence $B_t(\mathbf{c})$ consists of all possible received words where at most $t$ errors have occurred.

Assume first that $C$ has the property that no code word lies in the $t$-ball of another code word. Because $\mathbf{w}$ is in $B_t(\mathbf{c})$ and $\mathbf{w} \neq \mathbf{c}$, this means that $\mathbf{w}$ is not a code word and the error has been detected. If we strengthen the assumption on $C$ to require that the $t$-balls about code words are pairwise disjoint, then $\mathbf{w}$ belongs to a unique ball (the one about $\mathbf{c}$), and so $\mathbf{w}$ will be correctly decoded as $\mathbf{c}$.

To describe when this happens, let $C$ be an $n$-code. The **minimum distance** $d$ of $C$ is defined to be the smallest distance between two distinct code words in $C$; that is,

$$d = \min\{d(\mathbf{v}, \mathbf{w}) \mid \mathbf{v} \text{ and } \mathbf{w} \text{ in } C; \mathbf{v} \neq \mathbf{w}\}$$

**Theorem 8.8.3**

*Let $C$ be an $n$-code with minimum distance $d$. Assume that nearest neighbour decoding is used. Then:*

1. *If $t < d$, then $C$ can detect $t$ errors.*[20]
2. *If $2t < d$, then $C$ can correct $t$ errors.*

**Proof.**

1. Let $\mathbf{c}$ be a code word in $C$. If $\mathbf{w} \in B_t(\mathbf{c})$, then $d(\mathbf{w}, \mathbf{c}) \leq t < d$ by hypothesis. Thus the $t$-ball $B_t(\mathbf{c})$ contains no other code word, so $C$ can detect $t$ errors by the preceding discussion.
2. If $2t < d$, it suffices (again by the preceding discussion) to show that the $t$-balls about distinct code words are pairwise disjoint. But if $\mathbf{c} \neq \mathbf{c}'$ are code words in $C$ and $\mathbf{w}$ is in $B_t(\mathbf{c}') \cap B_t(\mathbf{c})$, then Theorem 8.8.2 gives
$$d(\mathbf{c}, \mathbf{c}') \leq d(\mathbf{c}, \mathbf{w}) + d(\mathbf{w}, \mathbf{c}') \leq t + t = 2t < d$$
by hypothesis, contradicting the minimality of $d$. □

**Example 8.8.4**

If $F = \mathbb{Z}_3 = \{0, 1, 2\}$, the 6-code $\{111111, 111222, 222111\}$ has minimum distance 3 and so can detect 2 errors and correct 1 error.

Let $\mathbf{c}$ be any word in $F^n$. A word $\mathbf{w}$ satisfies $d(\mathbf{w}, \mathbf{c}) = r$ if and only if $\mathbf{w}$ and $\mathbf{c}$ differ in exactly $r$ digits. If $|F| = q$, there are exactly $\binom{n}{r}(q-1)^r$ such words where $\binom{n}{r}$ is the binomial coefficient. Indeed, choose the $r$ places where they differ in $\binom{n}{r}$ ways, and then fill those places in $\mathbf{w}$ in $(q-1)^r$ ways. It follows that the number of words in the $t$-ball about $\mathbf{c}$ is

$$|B_t(\mathbf{c})| = \binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \cdots + \binom{n}{t}(q-1)^t = \sum_{i=0}^{t} \binom{n}{i}(q-1)^i$$

This leads to a useful bound on the size of error-correcting codes.

**Theorem 8.8.4: Hamming Bound**

*Let $C$ be an $n$-code over a field $F$ that can correct $t$ errors using nearest neighbour decoding. If $|F| = q$, then*
$$|C| \leq \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i}(q-1)^i}$$

**Proof.** Write $k = \sum_{i=0}^{t} \binom{n}{i}(q-1)^i$. The $t$-balls centred at distinct code words each contain $k$ words, and there are $|C|$ of them. Moreover they are pairwise disjoint because the code corrects $t$ errors (see the discussion preceding Theorem 8.8.3). Hence they contain $k \cdot |C|$ distinct words, and so $k \cdot |C| \leq |F^n| = q^n$, proving the theorem. □

A code is called **perfect** if there is equality in the Hamming bound; equivalently, if every word in $F^n$ lies in exactly one $t$-ball about a code word. For example, if $F = \mathbb{Z}_2$, $n = 3$, and $t = 1$, then $q = 2$ and $\binom{3}{0} + \binom{3}{1} = 4$, so the Hamming bound is $\frac{2^3}{4} = 2$. The 3-code $C = \{000, 111\}$ has minimum distance 3 and so can correct 1 error by Theorem 8.8.3. Hence $C$ is perfect.

[20] We say that $C$ detects (corrects) $t$ errors if $C$ can detect (or correct) $t$ or fewer errors.

## Linear Codes

Up to this point we have been regarding *any* nonempty subset of the $F$-vector space $F^n$ as a code. However many important codes are actually subspaces. A subspace $C \subseteq F^n$ of dimension $k \geq 1$ over $F$ is called an $(n, k)$**-linear code**, or simply an $(n, k)$**-code**. We do not regard the zero subspace (that is, $k = 0$) as a code.

**Example 8.8.5**

If $F = \mathbb{Z}_2$ and $n \geq 2$, the $n$**-parity-check code** is constructed as follows: An extra digit is added to each word in $F^{n-1}$ to make the number of 1s in the resulting word even (we say such words have **even parity**). The resulting $(n, n-1)$-code is linear because the sum of two words of even parity again has even parity.

Many of the properties of general codes take a simpler form for linear codes. The following result gives a much easier way to find the minimal distance of a linear code, and sharpens the results in Theorem 8.8.3.

**Theorem 8.8.5**

*Let $C$ be an $(n, k)$-code with minimum distance $d$ over a finite field $F$, and use nearest neighbour decoding.*

1. $d = \min\{wt(\mathbf{w}) \mid \mathbf{0} \neq \mathbf{w} \in C\}$.
2. *$C$ can detect $t \geq 1$ errors if and only if $t < d$.*
3. *$C$ can correct $t \geq 1$ errors if and only if $2t < d$.*
4. *If $C$ can correct $t \geq 1$ errors and $|F| = q$, then*

$$\binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \cdots + \binom{n}{t}(q-1)^t \leq q^{n-k}$$

**Proof.**

1. Write $d' = \min\{wt(\mathbf{w}) \mid \mathbf{0} \neq \mathbf{w} \text{ in } C\}$. If $\mathbf{v} \neq \mathbf{w}$ are words in $C$, then $d(\mathbf{v}, \mathbf{w}) = wt(\mathbf{v} - \mathbf{w}) \geq d'$ because $\mathbf{v} - \mathbf{w}$ is in the subspace $C$. Hence $d \geq d'$. Conversely, given $\mathbf{w} \neq \mathbf{0}$ in $C$ then, since $\mathbf{0}$ is in $C$, we have $wt(\mathbf{w}) = d(\mathbf{w}, \mathbf{0}) \geq d$ by the definition of $d$. Hence $d' \geq d$ and (1) is proved.

2. Assume that $C$ can detect $t$ errors. Given $\mathbf{w} \neq \mathbf{0}$ in $C$, the $t$-ball $B_t(\mathbf{w})$ about $\mathbf{w}$ contains no other code word (see the discussion preceding Theorem 8.8.3). In particular, it does not contain the code word $\mathbf{0}$, so $t < d(\mathbf{w}, \mathbf{0}) = wt(\mathbf{w})$. Hence $t < d$ by (1). The converse is part of Theorem 8.8.3.

3. We require a result of interest in itself.

   *Claim.* Suppose $\mathbf{c}$ in $C$ has $wt(\mathbf{c}) \leq 2t$. Then $B_t(\mathbf{0}) \cap B_t(\mathbf{c})$ is nonempty.

   *Proof.* If $wt(\mathbf{c}) \leq t$, then $\mathbf{c}$ itself is in $B_t(\mathbf{0}) \cap B_t(\mathbf{c})$. So assume $t < wt(\mathbf{c}) \leq 2t$. Then $\mathbf{c}$ has more than $t$ nonzero digits, so we can form a new word $\mathbf{w}$ by changing exactly $t$ of these nonzero digits to zero.

Then $d(\mathbf{w}, \mathbf{c}) = t$, so $\mathbf{w}$ is in $B_t(\mathbf{c})$. But $wt(\mathbf{w}) = wt(\mathbf{c}) - t \leq t$, so $\mathbf{w}$ is also in $B_t(\mathbf{0})$. Hence $\mathbf{w}$ is in $B_t(\mathbf{0}) \cap B_t(\mathbf{c})$, proving the Claim.

If $C$ corrects $t$ errors, the $t$-balls about code words are pairwise disjoint (see the discussion preceding Theorem 8.8.3). Hence the claim shows that $wt(\mathbf{c}) > 2t$ for all $\mathbf{c} \neq \mathbf{0}$ in $C$, from which $d > 2t$ by (1). The other inequality comes from Theorem 8.8.3.

4. We have $|C| = q^k$ because $\dim_F C = k$, so this assertion restates Theorem 8.8.4. □

**Example 8.8.6**

If $F = \mathbb{Z}_2$, then

$$C = \{0000000,\ 0101010,\ 1010101,\ 1110000,\ 1011010,\ 0100101,\ 0001111,\ 1111111\}$$

is a $(7,\ 3)$-code; in fact $C = \text{span}\{0101010,\ 1010101,\ 1110000\}$. The minimum distance for $C$ is 3, the minimum weight of a nonzero word in $C$.

## Matrix Generators

Given a linear $n$-code $C$ over a finite field $F$, the way encoding works in practice is as follows. A message stream is blocked off into segments of length $k \leq n$ called **messages**. Each message $\mathbf{u}$ in $F^k$ is encoded as a code word, the code word is transmitted, the receiver decodes the received word as the nearest code word, and then re-creates the original message. A fast and convenient method is needed to encode the incoming messages, to decode the received word after transmission (with or without error), and finally to retrieve messages from code words. All this can be achieved for any linear code using matrix multiplication.

Let $G$ denote a $k \times n$ matrix over a finite field $F$, and encode each message $\mathbf{u}$ in $F^k$ as the word $\mathbf{u}G$ in $F^n$ using matrix multiplication (thinking of words as rows). This amounts to saying that the set of code words is the subspace $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$ of $F^n$. This subspace need not have dimension $k$ for every $k \times n$ matrix $G$. But, if $\{\mathbf{e}_1,\ \mathbf{e}_2,\ \ldots,\ \mathbf{e}_k\}$ is the standard basis of $F^k$, then $\mathbf{e}_iG$ is row $i$ of $G$ for each $I$ and $\{\mathbf{e}_1G,\ \mathbf{e}_2G,\ \ldots,\ \mathbf{e}_kG\}$ spans $C$. Hence $\dim C = k$ if and only if the rows of $G$ are independent in $F^n$, and these matrices turn out to be exactly the ones we need. For reference, we state their main properties in Lemma 8.8.1 below (see Theorem 5.4.4).

**Lemma 8.8.1**

*The following are equivalent for a $k \times n$ matrix $G$ over a finite field $F$:*

1. $\text{rank } G = k$.
2. *The columns of $G$ span $F^k$.*
3. *The rows of $G$ are independent in $F^n$.*
4. *The system $GX = B$ is consistent for every column $B$ in $\mathbb{R}^k$.*
5. *$GK = I_k$ for some $n \times k$ matrix $K$.*

**Proof.** (1) $\Rightarrow$ (2). This is because $\dim(\operatorname{col} G) = k$ by (1).
(2) $\Rightarrow$ (4). $G\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T = x_1\mathbf{c}_1 + \cdots + x_n\mathbf{c}_n$ where $\mathbf{c}_j$ is column $j$ of $G$.
(4) $\Rightarrow$ (5). $G\begin{bmatrix} \mathbf{k}_1 & \cdots & \mathbf{k}_k \end{bmatrix} = \begin{bmatrix} G\mathbf{k}_1 & \cdots & G\mathbf{k}_k \end{bmatrix}$ for columns $\mathbf{k}_j$.
(5) $\Rightarrow$ (3). If $a_1R_1 + \cdots + a_kR_k = 0$ where $R_i$ is row $i$ of $G$, then $\begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix} G = 0$, so by (5), $\begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix} = 0$. Hence each $a_i = 0$, proving (3).
(3) $\Rightarrow$ (1). $\operatorname{rank} G = \dim(\operatorname{row} G) = k$ by (3). □

Note that Theorem 5.4.4 asserts that, over the real field $\mathbb{R}$, the properties in Lemma 8.8.1 hold if and only if $GG^T$ is invertible. But this need not be true in general. For example, if $F = \mathbb{Z}_2$ and $G = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$, then $GG^T = 0$. The reason is that the dot product $\mathbf{w} \cdot \mathbf{w}$ can be zero for $\mathbf{w}$ in $F^n$ even if $\mathbf{w} \neq \mathbf{0}$. However, even though $GG^T$ is not invertible, we do have $GK = I_2$ for some $4 \times 2$ matrix $K$ over $F$ as Lemma 8.8.1 asserts (in fact, $K = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}^T$ is one such matrix).

Let $C \subseteq F^n$ be an $(n, k)$-code over a finite field $F$. If $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ is a basis of $C$, let $G = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_k \end{bmatrix}$ be the $k \times n$ matrix with the $\mathbf{w}_i$ as its rows. Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is the standard basis of $F^k$ regarded as rows. Then $\mathbf{w}_i = \mathbf{e}_iG$ for each $i$, so $C = \operatorname{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\} = \operatorname{span}\{\mathbf{e}_1G, \ldots, \mathbf{e}_kG\}$. It follows (verify) that

$$C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$$

Because of this, the $k \times n$ matrix $G$ is called a **generator** of the code $C$, and $G$ has rank $k$ by Lemma 8.8.1 because its rows $\mathbf{w}_i$ are independent.

In fact, every linear code $C$ in $F^n$ has a generator of a simple, convenient form. If $G$ is a generator matrix for $C$, let $R$ be the reduced row-echelon form of $G$. We claim that $C$ is also generated by $R$. Since $G \to R$ by row operations, Theorem 2.5.1 shows that these same row operations $\begin{bmatrix} G & I_k \end{bmatrix} \to \begin{bmatrix} R & W \end{bmatrix}$, performed on $\begin{bmatrix} G & I_k \end{bmatrix}$, produce an invertible $k \times k$ matrix $W$ such that $R = WG$. Then $C = \{\mathbf{u}R \mid \mathbf{u} \text{ in } F^k\}$. [In fact, if $\mathbf{u}$ is in $F^k$, then $\mathbf{u}G = \mathbf{u}_1R$ where $\mathbf{u}_1 = \mathbf{u}W^{-1}$ is in $F^k$, and $\mathbf{u}R = \mathbf{u}_2G$ where $\mathbf{u}_2 = \mathbf{u}W$ is in $F^k$]. Thus $R$ is a generator of $C$, so we may assume that $G$ is in reduced row-echelon form.

In that case, $G$ has no row of zeros (since $\operatorname{rank} G = k$) and so contains all the columns of $I_k$. Hence a series of *column* interchanges will carry $G$ to the block form $G'' = \begin{bmatrix} I_k & A \end{bmatrix}$ for some $k \times (n-k)$ matrix $A$. Hence the code $C'' = \{\mathbf{u}G'' \mid \mathbf{u} \text{ in } F^k\}$ is essentially the same as $C$; the code words in $C''$ are obtained from those in $C$ by a series of column interchanges. Hence if $C$ is a linear $(n, k)$-code, we may (and shall) assume that the generator matrix $G$ has the form

$$G = \begin{bmatrix} I_k & A \end{bmatrix} \quad \text{for some} \quad k \times (n-k) \text{ matrix } A$$

Such a matrix is called a **standard generator**, or a **systematic generator**, for the code $C$. In this case, if $\mathbf{u}$ is a message word in $F^k$, the first $k$ digits of the encoded word $\mathbf{u}G$ are just the first $k$ digits of $\mathbf{u}$, so retrieval of $\mathbf{u}$ from $\mathbf{u}G$ is very simple indeed. The last $n-k$ digits of $\mathbf{u}G$ are called **parity digits**.

## Parity-Check Matrices

We begin with an important theorem about matrices over a finite field.

**Theorem 8.8.6**

*Let $F$ be a finite field, let $G$ be a $k \times n$ matrix of rank $k$, let $H$ be an $(n-k) \times n$ matrix of rank $n-k$, and let $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$ and $D = \{\mathbf{v}H \mid \mathbf{V} \text{ in } F^{n-k}\}$ be the codes they generate. Then the following conditions are equivalent:*

1. $GH^T = 0$.
2. $HG^T = 0$.
3. $C = \{\mathbf{w} \text{ in } F^n \mid \mathbf{w}H^T = \mathbf{0}\}$.
4. $D = \{\mathbf{w} \text{ in } F^n \mid \mathbf{w}G^T = \mathbf{0}\}$.

**Proof.** First, (1) $\Leftrightarrow$ (2) holds because $HG^T$ and $GH^T$ are transposes of each other.

(1) $\Rightarrow$ (3) Consider the linear transformation $T : F^n \to F^{n-k}$ defined by $T(\mathbf{w}) = \mathbf{w}H^T$ for all $\mathbf{w}$ in $F^n$. To prove (3) we must show that $C = \ker T$. We have $C \subseteq \ker T$ by (1) because $T(\mathbf{u}G) = \mathbf{u}GH^T = \mathbf{0}$ for all $\mathbf{u}$ in $F^k$. Since $\dim C = \operatorname{rank} G = k$, it is enough (by Theorem 6.4.2) to show $\dim(\ker T) = k$. However the dimension theorem (Theorem 7.2.4) shows that $\dim(\ker T) = n - \dim(\operatorname{im} T)$, so it is enough to show that $\dim(\operatorname{im} T) = n-k$. But if $R_1, \ldots, R_n$ are the rows of $H^T$, then block multiplication gives

$$\operatorname{im} T = \{\mathbf{w}H^T \mid \mathbf{w} \text{ in } \mathbb{R}^n\} = \operatorname{span}\{R_1, \ldots, R_n\} = \operatorname{row}(H^T)$$

Hence $\dim(\operatorname{im} T) = \operatorname{rank}(H^T) = \operatorname{rank} H = n-k$, as required. This proves (3).

(3) $\Rightarrow$ (1) If $\mathbf{u}$ is in $F^k$, then $\mathbf{u}G$ is in $C$ so, by (3), $\mathbf{u}(GH^T) = (\mathbf{u}G)H^T = \mathbf{0}$. Since $\mathbf{u}$ is arbitrary in $F^k$, it follows that $GH^T = 0$.

(2) $\Leftrightarrow$ (4) The proof is analogous to (1) $\Leftrightarrow$ (3). □

The relationship between the codes $C$ and $D$ in Theorem 8.8.6 will be characterized in another way in the next subsection.

If $C$ is an $(n, k)$-code, an $(n-k) \times n$ matrix $H$ is called a **parity-check matrix** for $C$ if $C = \{\mathbf{w} \mid \mathbf{w}H^T = \mathbf{0}\}$ as in Theorem 8.8.6. Such matrices are easy to find for a given code $C$. If $G = \begin{bmatrix} I_k & A \end{bmatrix}$ is a standard generator for $C$ where $A$ is $k \times (n-k)$, the $(n-k) \times n$ matrix

$$H = \begin{bmatrix} -A^T & I_{n-k} \end{bmatrix}$$

is a parity-check matrix for $C$. Indeed, $\operatorname{rank} H = n-k$ because the rows of $H$ are independent (due to the presence of $I_{n-k}$), and

$$GH^T = \begin{bmatrix} I_k & A \end{bmatrix} \begin{bmatrix} -A \\ I_{n-k} \end{bmatrix} = -A + A = 0$$

by block multiplication. Hence $H$ is a parity-check matrix for $C$ and we have $C = \{\mathbf{w} \text{ in } F^n \mid \mathbf{w}H^T = \mathbf{0}\}$. Since $\mathbf{w}H^T$ and $H\mathbf{w}^T$ are transposes of each other, this shows that $C$ can be characterized as follows:

$$C = \{\mathbf{w} \text{ in } F^n \mid H\mathbf{w}^T = \mathbf{0}\}$$

by Theorem 8.8.6.

This is useful in decoding. The reason is that decoding is done as follows: If a code word $\mathbf{c}$ is transmitted and $\mathbf{v}$ is received, then $\mathbf{z} = \mathbf{v} - \mathbf{c}$ is called the **error**. Since $H\mathbf{c}^T = \mathbf{0}$, we have $H\mathbf{z}^T = H\mathbf{v}^T$ and this word

$$\mathbf{s} = H\mathbf{z}^T = H\mathbf{v}^T$$

is called the **syndrome**. The receiver knows $\mathbf{v}$ and $\mathbf{s} = H\mathbf{v}^T$, and wants to recover $\mathbf{c}$. Since $\mathbf{c} = \mathbf{v} - \mathbf{z}$, it is enough to find $\mathbf{z}$. But the possibilities for $\mathbf{z}$ are the solutions of the linear system

$$H\mathbf{z}^T = \mathbf{s}$$

where $\mathbf{s}$ is known. Now recall that Theorem 2.2.3 shows that these solutions have the form $\mathbf{z} = \mathbf{x} + \mathbf{s}$ where $\mathbf{x}$ is any solution of the homogeneous system $H\mathbf{x}^T = \mathbf{0}$, that is, $\mathbf{x}$ is any word in $C$ (by Lemma 8.8.1). In other words, the errors $\mathbf{z}$ are the elements of the set

$$C + \mathbf{s} = \{\mathbf{c} + \mathbf{s} \mid \mathbf{c} \text{ in } C\}$$

The set $C + \mathbf{s}$ is called a **coset** of $C$. Let $|F| = q$. Since $|C + \mathbf{s}| = |C| = q^{n-k}$ the search for $\mathbf{z}$ is reduced from $q^n$ possibilities in $F^n$ to $q^{n-k}$ possibilities in $C + \mathbf{s}$. This is called **syndrome decoding**, and various methods for improving efficiency and accuracy have been devised. The reader is referred to books on coding for more details.[21]

## Orthogonal Codes

Let $F$ be a finite field. Given two words $\mathbf{v} = a_1 a_2 \cdots a_n$ and $\mathbf{w} = b_1 b_2 \cdots b_n$ in $F^n$, the dot product $\mathbf{v} \cdot \mathbf{w}$ is defined (as in $\mathbb{R}^n$) by

$$\mathbf{v} \cdot \mathbf{w} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

Note that $\mathbf{v} \cdot \mathbf{w}$ is an element of $F$, and it can be computed as a matrix product: $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}\mathbf{w}^T$.

If $C \subseteq F^n$ is an $(n, k)$-code, the **orthogonal complement** $C^\perp$ is defined as in $\mathbb{R}^n$:

$$C^\perp = \{\mathbf{v} \text{ in } F^n \mid \mathbf{v} \cdot \mathbf{c} = 0 \text{ for all } \mathbf{c} \text{ in } C\}$$

This is easily seen to be a subspace of $F^n$, and it turns out to be an $(n, n-k)$-code. This follows when $F = \mathbb{R}$ because we showed (in the projection theorem) that $n = \dim U^\perp + \dim U$ for any subspace $U$ of $\mathbb{R}^n$. However the proofs break down for a finite field $F$ because the dot product in $F^n$ has the property that $\mathbf{w} \cdot \mathbf{w} = 0$ can happen even if $\mathbf{w} \neq \mathbf{0}$. Nonetheless, the result remains valid.

**Theorem 8.8.7**

*Let $C$ be an $(n, k)$-code over a finite field $F$, let $G = \begin{bmatrix} I_k & A \end{bmatrix}$ be a standard generator for $C$ where $A$ is $k \times (n-k)$, and write $H = \begin{bmatrix} -A^T & I_{n-k} \end{bmatrix}$ for the parity-check matrix. Then:*

1. *$H$ is a generator of $C^\perp$.*
2. *$\dim(C^\perp) = n - k = \operatorname{rank} H$.*
3. *$C^{\perp\perp} = C$ and $\dim(C^\perp) + \dim C = n$.*

[21] For an elementary introduction, see V. Pless, Introduction to the Theory of Error-Correcting Codes, 3rd ed., (New York: Wiley, 1998).

**Proof.** As in Theorem 8.8.6, let $D = \{\mathbf{v}H \mid \mathbf{v} \text{ in } F^{n-k}\}$ denote the code generated by $H$. Observe first that, for all $\mathbf{w}$ in $F^n$ and all $\mathbf{u}$ in $F^k$, we have

$$\mathbf{w} \cdot (\mathbf{u}G) = \mathbf{w}(\mathbf{u}G)^T = \mathbf{w}(G^T\mathbf{u}^T) = (\mathbf{w}G^T) \cdot \mathbf{u}$$

Since $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^k\}$, this shows that $\mathbf{w}$ is in $C^\perp$ if and only if $(\mathbf{w}G^T) \cdot \mathbf{u} = 0$ for all $\mathbf{u}$ in $F^k$; if and only if[22] $\mathbf{w}G^T = \mathbf{0}$; if and only if $\mathbf{w}$ is in $D$ (by Theorem 8.8.6). Thus $C^\perp = D$ and a similar argument shows that $D^\perp = C$.

1. $H$ generates $C^\perp$ because $C^\perp = D = \{\mathbf{v}H \mid \mathbf{v} \text{ in } F^{n-k}\}$.
2. This follows from (1) because, as we observed above, $\operatorname{rank} H = n - k$.
3. Since $C^\perp = D$ and $D^\perp = C$, we have $C^{\perp\perp} = (C^\perp)^\perp = D^\perp = C$. Finally the second equation in (3) restates (2) because $\dim C = k$. □

We note in passing that, if $C$ is a subspace of $\mathbb{R}^k$, we have $C + C^\perp = \mathbb{R}^k$ by the projection theorem (Theorem 8.1.3), and $C \cap C^\perp = \{\mathbf{0}\}$ because any vector $\mathbf{x}$ in $C \cap C^\perp$ satisfies $\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x} = 0$. However, this fails in general. For example, if $F = \mathbb{Z}_2$ and $C = \operatorname{span}\{1010,\ 0101\}$ in $F^4$ then $C^\perp = C$, so $C + C^\perp = C = C \cap C^\perp$.

We conclude with one more example. If $F = \mathbb{Z}_2$, consider the standard matrix $G$ below, and the corresponding parity-check matrix $H$:

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

The code $C = \{\mathbf{u}G \mid \mathbf{u} \text{ in } F^4\}$ generated by $G$ has dimension $k = 4$, and is called the **Hamming (7, 4)-code**. The vectors in $C$ are listed in the first table below. The dual code generated by $H$ has dimension $n - k = 3$ and is listed in the second table.

$C$:

| $\mathbf{u}$ | $\mathbf{u}G$ |
|---|---|
| 0000 | 0000000 |
| 0001 | 0001011 |
| 0010 | 0010101 |
| 0011 | 0011110 |
| 0100 | 0100110 |
| 0101 | 0101101 |
| 0110 | 0110011 |
| 0111 | 0111000 |
| 1000 | 1000111 |
| 1001 | 1001100 |
| 1010 | 1010010 |
| 1011 | 1011001 |
| 1100 | 1100001 |
| 1101 | 1101010 |
| 1110 | 1110100 |
| 1111 | 1111111 |

$C^\perp$:

| $\mathbf{v}$ | $\mathbf{v}H$ |
|---|---|
| 000 | 0000000 |
| 001 | 1011001 |
| 010 | 1101010 |
| 011 | 0110011 |
| 100 | 1110100 |
| 101 | 0101101 |
| 110 | 0011110 |
| 111 | 1000111 |

[22] If $\mathbf{v} \cdot \mathbf{u} = 0$ for every $\mathbf{u}$ in $F^k$, then $\mathbf{v} = \mathbf{0}$—let $\mathbf{u}$ range over the standard basis of $F^k$.

Clearly each nonzero code word in $C$ has weight at least 3, so $C$ has minimum distance $d = 3$. Hence $C$ can detect two errors and correct one error by Theorem 8.8.5. The dual code has minimum distance 4 and so can detect 3 errors and correct 1 error.

## Exercises for 8.8

**Exercise 8.8.1** Find all $a$ in $\mathbb{Z}_{10}$ such that:

a. $a^2 = a$.

b. $a$ has an inverse (and find the inverse).

c. $a^k = 0$ for some $k \geq 1$.

d. $a = 2^k$ for some $k \geq 1$.

e. $a = b^2$ for some $b$ in $\mathbb{Z}_{10}$.

**Exercise 8.8.2**

a. Show that if $3a = 0$ in $\mathbb{Z}_{10}$, then necessarily $a = 0$ in $\mathbb{Z}_{10}$.

b. Show that $2a = 0$ in $\mathbb{Z}_{10}$ holds in $\mathbb{Z}_{10}$ if and only if $a = 0$ or $a = 5$.

**Exercise 8.8.3** Find the inverse of:

a. 8 in $\mathbb{Z}_{13}$; b. 11 in $\mathbb{Z}_{19}$.

**Exercise 8.8.4** If $ab = 0$ in a field $F$, show that either $a = 0$ or $b = 0$.

**Exercise 8.8.5** Show that the entries of the last column of the multiplication table of $\mathbb{Z}_n$ are

$$0,\ n-1,\ n-2,\ \ldots,\ 2,\ 1$$

in that order.

**Exercise 8.8.6** In each case show that the matrix $A$ is invertible over the given field, and find $A^{-1}$.

a. $A = \begin{bmatrix} 1 & 4 \\ 2 & 1 \end{bmatrix}$ over $\mathbb{Z}_5$.

b. $A = \begin{bmatrix} 5 & 6 \\ 4 & 3 \end{bmatrix}$ over $\mathbb{Z}_7$.

**Exercise 8.8.7** Consider the linear system $\begin{matrix} 3x + & y + 4z = 3 \\ 4x + 3y + & z = 1 \end{matrix}$. In each case solve the system by reducing the augmented matrix to reduced row-echelon form over the given field:

a. $\mathbb{Z}_5$ b. $\mathbb{Z}_7$

**Exercise 8.8.8** Let $K$ be a vector space over $\mathbb{Z}_2$ with basis $\{1,\ t\}$, so $K = \{a + bt \mid a,\ b,\ \text{in } \mathbb{Z}_2\}$. It is known that $K$ becomes a field of four elements if we define $t^2 = 1 + t$. Write down the multiplication table of $K$.

**Exercise 8.8.9** Let $K$ be a vector space over $\mathbb{Z}_3$ with basis $\{1,\ t\}$, so $K = \{a + bt \mid a,\ b,\ \text{in } \mathbb{Z}_3\}$. It is known that $K$ becomes a field of nine elements if we define $t^2 = -1$ in $\mathbb{Z}_3$. In each case find the inverse of the element $x$ of $K$:

a. $x = 1 + 2t$ b. $x = 1 + t$

**Exercise 8.8.10** How many errors can be detected or corrected by each of the following binary linear codes?

a. $C = \{0000000,\ 0011110,\ 0100111,\ 0111001,\ 1001011,\ 1010101,\ 1101100,\ 1110010\}$

b. $C = \{0000000000,\ 0010011111,\ 0101100111,\ 0111111000,\ 1001110001,\ 1011101110,\ 1100010110,\ 1110001001\}$

**Exercise 8.8.11**

a. If a binary linear $(n,\ 2)$-code corrects one error, show that $n \geq 5$. [*Hint*: Hamming bound.]

b. Find a $(5,\ 2)$-code that corrects one error.

**Exercise 8.8.12**

a. If a binary linear $(n,\ 3)$-code corrects two errors, show that $n \geq 9$. [*Hint*: Hamming bound.]

b. If $G = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$, show that the binary $(10,\ 3)$-code generated by $G$ corrects two errors. [It can be shown that no binary $(9,\ 3)$-code corrects two errors.]

**Exercise 8.8.13**

a. Show that no binary linear $(4,\ 2)$-code can correct single errors.

b. Find a binary linear $(5,\ 2)$-code that can correct one error.

**Exercise 8.8.14** Find the standard generator matrix $G$ and the parity-check matrix $H$ for each of the following systematic codes:

a. $\{00000,\ 11111\}$ over $\mathbb{Z}_2$.

b. Any systematic $(n,\ 1)$-code where $n \geq 2$.

c. The code in Exercise 8.8.10(a).

d. The code in Exercise 8.8.10(b).

**Exercise 8.8.15** Let $\mathbf{c}$ be a word in $F^n$. Show that $B_t(\mathbf{c}) = \mathbf{c} + B_t(\mathbf{0})$, where we write

$$\mathbf{c} + B_t(\mathbf{0}) = \{\mathbf{c} + \mathbf{v} \mid \mathbf{v} \text{ in } B_t(\mathbf{0})\}$$

**Exercise 8.8.16** If a $(n,\ k)$-code has two standard generator matrices $G$ and $G_1$, show that $G = G_1$.

**Exercise 8.8.17** Let $C$ be a binary linear $n$-code (over $\mathbb{Z}_2$). Show that either each word in $C$ has even weight, or half the words in $C$ have even weight and half have odd weight. [*Hint*: The dimension theorem.]

## 8.9 An Application to Quadratic Forms

An expression like $x_1^2 + x_2^2 + x_3^2 - 2x_1x_3 + x_2x_3$ is called a quadratic form in the variables $x_1$, $x_2$, and $x_3$. In this section we show that new variables $y_1$, $y_2$, and $y_3$ can always be found so that the quadratic form, when expressed in terms of the new variables, has no cross terms $y_1y_2$, $y_1y_3$, or $y_2y_3$. Moreover, we do this for forms involving any finite number of variables using orthogonal diagonalization. This has far-reaching applications; quadratic forms arise in such diverse areas as statistics, physics, the theory of functions of several variables, number theory, and geometry.

**Definition 8.21 Quadratic Form**

*A **quadratic form** $q$ in the $n$ variables $x_1,\ x_2,\ \ldots,\ x_n$ is a linear combination of terms $x_1^2,\ x_2^2,\ \ldots,\ x_n^2$, and cross terms $x_1x_2,\ x_1x_3,\ x_2x_3,\ \ldots$ .*

If $n = 3$, $q$ has the form

$$q = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{13}x_1x_3 + a_{31}x_3x_1 + a_{23}x_2x_3 + a_{32}x_3x_2$$

In general

$$q = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + a_{12}x_1x_2 + a_{13}x_1x_3 + \cdots$$

This sum can be written compactly as a matrix product

$$q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$$

where $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_n)$ is thought of as a column, and $A = \left[a_{ij}\right]$ is a real $n \times n$ matrix. Note that if $i \neq j$, two separate terms $a_{ij}x_ix_j$ and $a_{ji}x_jx_i$ are listed, each of which involves $x_ix_j$, and they can (rather cleverly) be replaced by

$$\tfrac{1}{2}(a_{ij} + a_{ji})x_ix_j \quad \text{and} \quad \tfrac{1}{2}(a_{ij} + a_{ji})x_jx_i$$

respectively, *without altering the quadratic form*. Hence there is no loss of generality in assuming that $x_ix_j$ and $x_jx_i$ have the same coefficient in the sum for $q$. In other words, **we may assume that $A$ is symmetric**.

**Example 8.9.1**

Write $q = x_1^2 + 3x_3^2 + 2x_1x_2 - x_1x_3$ in the form $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, where $A$ is a symmetric $3 \times 3$ matrix.

**Solution.** The cross terms are $2x_1x_2 = x_1x_2 + x_2x_1$ and $-x_1x_3 = -\frac{1}{2}x_1x_3 - \frac{1}{2}x_3x_1$.
Of course, $x_2x_3$ and $x_3x_2$ both have coefficient zero, as does $x_2^2$. Hence

$$q(\mathbf{x}) = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & 1 & -\frac{1}{2} \\ 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

is the required form (verify).

We shall assume from now on that all quadratic forms are given by

$$q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$$

where $A$ is symmetric. Given such a form, the problem is to find new variables $y_1, y_2, \ldots, y_n$, related to $x_1, x_2, \ldots, x_n$, with the property that when $q$ is expressed in terms of $y_1, y_2, \ldots, y_n$, there are no cross terms. If we write

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$$

this amounts to asking that $q = \mathbf{y}^T D\mathbf{y}$ where $D$ is diagonal. It turns out that this can always be accomplished and, not surprisingly, that $D$ is the matrix obtained when the symmetric matrix $A$ is orthogonally diagonalized. In fact, as Theorem 8.2.2 shows, a matrix $P$ can be found that is orthogonal (that is, $P^{-1} = P^T$) and diagonalizes $A$:

$$P^T AP = D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

The diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the (not necessarily distinct) eigenvalues of $A$, repeated according to their multiplicities in $c_A(x)$, and the columns of $P$ are corresponding (orthonormal) eigenvectors of $A$. As $A$ is symmetric, the $\lambda_i$ are real by Theorem 5.5.7.

Now define new variables $\mathbf{y}$ by the equations

$$\mathbf{x} = P\mathbf{y} \quad \text{equivalently} \quad \mathbf{y} = P^T\mathbf{x}$$

Then substitution in $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ gives

$$q = (P\mathbf{y})^T A(P\mathbf{y}) = \mathbf{y}^T(P^T AP)\mathbf{y} = \mathbf{y}^T D\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$$

Hence this change of variables produces the desired simplification in $q$.

**Theorem 8.9.1: Diagonalization Theorem**

*Let $q = \mathbf{x}^T A\mathbf{x}$ be a quadratic form in the variables $x_1, x_2, \ldots, x_n$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ and $A$ is a symmetric $n \times n$ matrix. Let $P$ be an orthogonal matrix such that $P^T AP$ is diagonal, and*

*define new variables* $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ *by*

$$\mathbf{x} = P\mathbf{y} \quad \textit{equivalently} \quad \mathbf{y} = P^T\mathbf{x}$$

*If* $q$ *is expressed in terms of these new variables* $y_1, y_2, \ldots, y_n$*, the result is*

$$q = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$$

*where* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *are the eigenvalues of* $A$ *repeated according to their multiplicities.*

Let $q = \mathbf{x}^T A\mathbf{x}$ be a quadratic form where $A$ is a symmetric matrix and let $\lambda_1, \ldots, \lambda_n$ be the (real) eigenvalues of $A$ repeated according to their multiplicities. A corresponding set $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ of orthonormal eigenvectors for $A$ is called a set of **principal axes** for the quadratic form $q$. (The reason for the name will become clear later.) The orthogonal matrix $P$ in Theorem 8.9.1 is given as $P = \begin{bmatrix} \mathbf{f}_1 & \cdots & \mathbf{f}_n \end{bmatrix}$, so the variables $X$ and $Y$ are related by

$$\mathbf{x} = P\mathbf{y} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \cdots & \mathbf{f}_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y_1\mathbf{f}_1 + y_2\mathbf{f}_2 + \cdots + y_n\mathbf{f}_n$$

Thus the new variables $y_i$ are the coefficients when $\mathbf{x}$ is expanded in terms of the orthonormal basis $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ of $\mathbb{R}^n$. In particular, the coefficients $y_i$ are given by $y_i = \mathbf{x} \cdot \mathbf{f}_i$ by the expansion theorem (Theorem 5.3.6). Hence $q$ itself is easily computed from the eigenvalues $\lambda_i$ and the principal axes $\mathbf{f}_i$:

$$q = q(\mathbf{x}) = \lambda_1(\mathbf{x} \cdot \mathbf{f}_1)^2 + \cdots + \lambda_n(\mathbf{x} \cdot \mathbf{f}_n)^2$$

**Example 8.9.2**

Find new variables $y_1$, $y_2$, $y_3$, and $y_4$ such that

$$q = 3(x_1^2 + x_2^2 + x_3^2 + x_4^2) + 2x_1x_2 - 10x_1x_3 + 10x_1x_4 + 10x_2x_3 - 10x_2x_4 + 2x_3x_4$$

has diagonal form, and find the corresponding principal axes.

**Solution.** The form can be written as $q = \mathbf{x}^T A\mathbf{x}$, where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 3 & 1 & -5 & 5 \\ 1 & 3 & 5 & -5 \\ -5 & 5 & 3 & 1 \\ 5 & -5 & 1 & 3 \end{bmatrix}$$

A routine calculation yields

$$c_A(x) = \det(xI - A) = (x - 12)(x + 8)(x - 4)^2$$

so the eigenvalues are $\lambda_1 = 12$, $\lambda_2 = -8$, and $\lambda_3 = \lambda_4 = 4$. Corresponding orthonormal

eigenvectors are the principal axes:

$$\mathbf{f}_1 = \tfrac{1}{2}\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \quad \mathbf{f}_2 = \tfrac{1}{2}\begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \quad \mathbf{f}_3 = \tfrac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{f}_4 = \tfrac{1}{2}\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

The matrix

$$P = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \mathbf{f}_3 & \mathbf{f}_4 \end{bmatrix} = \tfrac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

is thus orthogonal, and $P^{-1}AP = P^TAP$ is diagonal. Hence the new variables $\mathbf{y}$ and the old variables $\mathbf{x}$ are related by $\mathbf{y} = P^T\mathbf{x}$ and $\mathbf{x} = P\mathbf{y}$. Explicitly,

$$\begin{aligned} y_1 &= \tfrac{1}{2}(x_1 - x_2 - x_3 + x_4) & x_1 &= \tfrac{1}{2}(y_1 + y_2 + y_3 + y_4) \\ y_2 &= \tfrac{1}{2}(x_1 - x_2 + x_3 - x_4) & x_2 &= \tfrac{1}{2}(-y_1 - y_2 + y_3 + y_4) \\ y_3 &= \tfrac{1}{2}(x_1 + x_2 + x_3 + x_4) & x_3 &= \tfrac{1}{2}(-y_1 + y_2 + y_3 - y_4) \\ y_4 &= \tfrac{1}{2}(x_1 + x_2 - x_3 - x_4) & x_4 &= \tfrac{1}{2}(y_1 - y_2 + y_3 - y_4) \end{aligned}$$

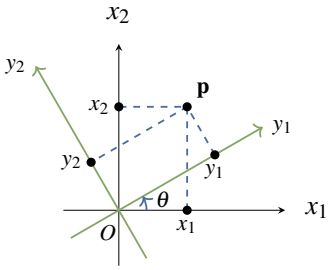
If these $x_i$ are substituted in the original expression for $q$, the result is

$$q = 12y_1^2 - 8y_2^2 + 4y_3^2 + 4y_4^2$$

This is the required diagonal form.

It is instructive to look at the case of quadratic forms in two variables $x_1$ and $x_2$. Then the principal axes can always be found by rotating the $x_1$ and $x_2$ axes counterclockwise about the origin through an angle $\theta$. This rotation is a linear transformation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$, and it is shown in Theorem 2.6.4 that $R_\theta$ has matrix $P = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. If $\{\mathbf{e}_1, \mathbf{e}_2\}$ denotes the standard basis of $\mathbb{R}^2$, the rotation produces a new basis $\{\mathbf{f}_1, \mathbf{f}_2\}$ given by

$$\mathbf{f}_1 = R_\theta(\mathbf{e}_1) = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \quad \text{and} \quad \mathbf{f}_2 = R_\theta(\mathbf{e}_2) = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \tag{8.7}$$




Given a point $\mathbf{p} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2$ in the original system, let $y_1$ and $y_2$ be the coordinates of $\mathbf{p}$ in the new system (see the diagram). That is,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{p} = y_1\mathbf{f}_1 + y_2\mathbf{f}_2 = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{8.8}$$

Writing $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, this reads $\mathbf{x} = P\mathbf{y}$ so, since $P$ is orthogonal, this is the change of variables formula for the rotation as in Theorem 8.9.1.

If $r \neq 0 \neq s$, the graph of the equation $rx_1^2 + sx_2^2 = 1$ is called an **ellipse** if $rs > 0$ and a **hyperbola** if $rs < 0$. More generally, given a quadratic form

$$q = ax_1^2 + bx_1x_2 + cx_2^2 \quad \text{where not all of } a,\ b,\ \text{and } c \text{ are zero}$$

the graph of the equation $q = 1$ is called a **conic**. We can now completely describe this graph. There are two special cases which we leave to the reader.

1. *If exactly one of $a$ and $c$ is zero, then the graph of $q = 1$ is a* **parabola**.

So we assume that $a \neq 0$ and $c \neq 0$. In this case, the description depends on the quantity $b^2 - 4ac$, called the **discriminant** of the quadratic form $q$.

2. *If $b^2 - 4ac = 0$, then either both $a \geq 0$ and $c \geq 0$, or both $a \leq 0$ and $c \leq 0$.*
   *Hence $q = (\sqrt{a}x_1 + \sqrt{c}x_2)^2$ or $q = (\sqrt{-a}x_1 + \sqrt{-c}x_2)^2$, so the graph of $q = 1$ is a* **pair of straight lines** *in either case.*

So we also assume that $b^2 - 4ac \neq 0$. But then the next theorem asserts that there exists a rotation of the plane about the origin which transforms the equation $ax_1^2 + bx_1x_2 + cx_2^2 = 1$ into either an ellipse or a hyperbola, and the theorem also provides a simple way to decide which conic it is.

**Theorem 8.9.2**

*Consider the quadratic form $q = ax_1^2 + bx_1x_2 + cx_2^2$ where $a$, $c$, and $b^2 - 4ac$ are all nonzero.*

1. *There is a counterclockwise rotation of the coordinate axes about the origin such that, in the new coordinate system, $q$ has no cross term.*
2. *The graph of the equation*
$$ax_1^2 + bx_1x_2 + cx_2^2 = 1$$
*is an ellipse if $b^2 - 4ac < 0$ and an hyperbola if $b^2 - 4ac > 0$.*

**Proof.** If $b = 0$, $q$ *already* has no cross term and (1) and (2) are clear. So assume $b \neq 0$. The matrix $A = \begin{bmatrix} a & \frac{1}{2}b \\ \frac{1}{2}b & c \end{bmatrix}$ of $q$ has characteristic polynomial $c_A(x) = x^2 - (a+c)x - \frac{1}{4}(b^2 - 4ac)$. If we write $d = \sqrt{b^2 + (a-c)^2}$ for convenience; then the quadratic formula gives the eigenvalues

$$\lambda_1 = \tfrac{1}{2}[a + c - d] \quad \text{and} \quad \lambda_2 = \tfrac{1}{2}[a + c + d]$$

with corresponding principal axes

$$\mathbf{f}_1 = \frac{1}{\sqrt{b^2 + (a-c-d)^2}} \begin{bmatrix} a - c - d \\ b \end{bmatrix} \quad \text{and}$$
$$\mathbf{f}_2 = \frac{1}{\sqrt{b^2 + (a-c-d)^2}} \begin{bmatrix} -b \\ a - c - d \end{bmatrix}$$

as the reader can verify. These agree with equation (8.7) above if $\theta$ is an angle such that

$$\cos\theta = \frac{a-c-d}{\sqrt{b^2+(a-c-d)^2}} \quad \text{and} \quad \sin\theta = \frac{b}{\sqrt{b^2+(a-c-d)^2}}$$

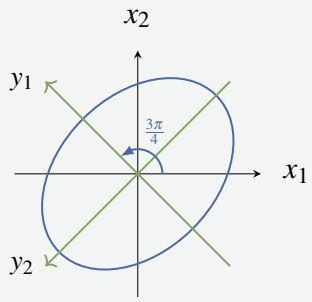
Then $P = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ diagonalizes $A$ and equation (8.8) becomes the formula $\mathbf{x} = P\mathbf{y}$ in Theorem 8.9.1. This proves (1).

Finally, $A$ is similar to $\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ so $\lambda_1\lambda_2 = \det A = \frac{1}{4}(4ac-b^2)$. Hence the graph of $\lambda_1 y_1^2 + \lambda_2 y_2^2 = 1$ is an ellipse if $b^2 < 4ac$ and an hyperbola if $b^2 > 4ac$. This proves (2). □

**Example 8.9.3**

Consider the equation $x^2 + xy + y^2 = 1$. Find a rotation so that the equation has no cross term.

**Solution.**




Here $a = b = c = 1$ in the notation of Theorem 8.9.2, so $\cos\theta = \frac{-1}{\sqrt{2}}$ and $\sin\theta = \frac{1}{\sqrt{2}}$. Hence $\theta = \frac{3\pi}{4}$ will do it. The new variables are $y_1 = \frac{1}{\sqrt{2}}(x_2 - x_1)$ and $y_2 = \frac{-1}{\sqrt{2}}(x_2 + x_1)$ by (8.8), and the equation becomes $y_1^2 + 3y_2^2 = 2$. The angle $\theta$ has been chosen such that the new $y_1$ and $y_2$ axes are the axes of symmetry of the ellipse (see the diagram). The eigenvectors $\mathbf{f}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $\mathbf{f}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ point along these axes of symmetry, and this is the reason for the name *principal axes*.

The determinant of any orthogonal matrix $P$ is either 1 or $-1$ (because $PP^T = I$). The orthogonal matrices $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ arising from rotations all have determinant 1. More generally, given any quadratic form $q = \mathbf{x}^T A\mathbf{x}$, the orthogonal matrix $P$ such that $P^T AP$ is diagonal can always be chosen so that $\det P = 1$ by interchanging two eigenvalues (and hence the corresponding columns of $P$). It is shown in Theorem 10.4.4 that orthogonal $2 \times 2$ matrices with determinant 1 correspond to rotations. Similarly, it can be shown that orthogonal $3 \times 3$ matrices with determinant 1 correspond to rotations about a line through the origin. This extends Theorem 8.9.2: Every quadratic form in two or three variables can be diagonalized by a rotation of the coordinate system.

## Congruence

We return to the study of quadratic forms in general.

**Theorem 8.9.3**

*If $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ is a quadratic form given by a symmetric matrix $A$, then $A$ is uniquely determined by $q$.*

**Proof.** Let $q(\mathbf{x}) = \mathbf{x}^T B\mathbf{x}$ for all $\mathbf{x}$ where $B^T = B$. If $C = A - B$, then $C^T = C$ and $\mathbf{x}^T C\mathbf{x} = 0$ for all $\mathbf{x}$. We must show that $C = 0$. Given $\mathbf{y}$ in $\mathbb{R}^n$,

$$\begin{aligned} 0 = (\mathbf{x}+\mathbf{y})^T C(\mathbf{x}+\mathbf{y}) &= \mathbf{x}^T C\mathbf{x} + \mathbf{x}^T C\mathbf{y} + \mathbf{y}^T C\mathbf{x} + \mathbf{y}^T C\mathbf{y} \\ &= \mathbf{x}^T C\mathbf{y} + \mathbf{y}^T C\mathbf{x} \end{aligned}$$

But $\mathbf{y}^T C\mathbf{x} = (\mathbf{x}^T C\mathbf{y})^T = \mathbf{x}^T C\mathbf{y}$ (it is $1 \times 1$). Hence $\mathbf{x}^T C\mathbf{y} = 0$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$. If $\mathbf{e}_j$ is column $j$ of $I_n$, then the $(i,\ j)$-entry of $C$ is $\mathbf{e}_i^T C\mathbf{e}_j = 0$. Thus $C = 0$. □

Hence we can speak of *the* symmetric matrix of a quadratic form.

On the other hand, a quadratic form $q$ in variables $x_i$ can be written in several ways as a linear combination of squares of new variables, even if the new variables are required to be linear combinations of the $x_i$. For example, if $q = 2x_1^2 - 4x_1x_2 + x_2^2$ then

$$q = 2(x_1 - x_2)^2 - x_2^2 \quad \text{and} \quad q = -2x_1^2 + (2x_1 - x_2)^2$$

The question arises: How are these changes of variables related, and what properties do they share? To investigate this, we need a new concept.

Let a quadratic form $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ be given in terms of variables $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_n)^T$. If the new variables $\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_n)^T$ are to be linear combinations of the $x_i$, then $\mathbf{y} = A\mathbf{x}$ for some $n \times n$ matrix $A$. Moreover, since we want to be able to solve for the $x_i$ in terms of the $y_i$, we ask that the matrix $A$ be invertible. Hence suppose $U$ is an invertible matrix and that the new variables $\mathbf{y}$ are given by

$$\mathbf{y} = U^{-1}\mathbf{x}, \qquad \text{equivalently } \mathbf{x} = U\mathbf{y}$$

In terms of these new variables, $q$ takes the form

$$q = q(\mathbf{x}) = (U\mathbf{y})^T A(U\mathbf{y}) = \mathbf{y}^T(U^T AU)\mathbf{y}$$

That is, $q$ has matrix $U^T AU$ with respect to the new variables $\mathbf{y}$. Hence, to study changes of variables in quadratic forms, we study the following relationship on matrices: Two $n \times n$ matrices $A$ and $B$ are called **congruent**, written $A \overset{c}{\sim} B$, if $B = U^T AU$ for some invertible matrix $U$. Here are some properties of congruence:

1. *$A \overset{c}{\sim} A$ for all $A$.*
2. *If $A \overset{c}{\sim} B$, then $B \overset{c}{\sim} A$.*
3. *If $A \overset{c}{\sim} B$ and $B \overset{c}{\sim} C$, then $A \overset{c}{\sim} C$.*

4. *If* $A \overset{c}{\sim} B$, *then* $A$ *is symmetric if and only if* $B$ *is symmetric*.
5. *If* $A \overset{c}{\sim} B$, *then* $\text{rank } A = \text{rank } B$.

The converse to (5) can fail even for symmetric matrices.

**Example 8.9.4**

The symmetric matrices $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ have the same rank but are not congruent. Indeed, if $A \overset{c}{\sim} B$, an invertible matrix $U$ exists such that $B = U^T AU = U^T U$. But then $-1 = \det B = (\det U)^2$, a contradiction.

The key distinction between $A$ and $B$ in Example 8.9.4 is that $A$ has two positive eigenvalues (counting multiplicities) whereas $B$ has only one.

**Theorem 8.9.4: Sylvester's Law of Inertia**

*If* $A \overset{c}{\sim} B$, *then* $A$ *and* $B$ *have the same number of positive eigenvalues, counting multiplicities.*

The proof is given at the end of this section.

The **index** of a symmetric matrix $A$ is the number of positive eigenvalues of $A$. If $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ is a quadratic form, the **index** and **rank** of $q$ are defined to be, respectively, the index and rank of the matrix $A$. As we saw before, if the variables expressing a quadratic form $q$ are changed, the new matrix is congruent to the old one. Hence the index and rank depend only on $q$ and not on the way it is expressed.

Now let $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ be any quadratic form in $n$ variables, of index $k$ and rank $r$, where $A$ is symmetric. We claim that new variables $\mathbf{z}$ can be found so that $q$ is **completely diagonalized**—that is,

$$q(\mathbf{z}) = z_1^2 + \cdots + z_k^2 - z_{k+1}^2 - \cdots - z_r^2$$

If $k \leq r \leq n$, let $D_n(k,\ r)$ denote the $n \times n$ diagonal matrix whose main diagonal consists of $k$ ones, followed by $r-k$ minus ones, followed by $n-r$ zeros. Then we seek new variables $\mathbf{z}$ such that

$$q(\mathbf{z}) = \mathbf{z}^T D_n(k,\ r)\mathbf{z}$$

To determine $\mathbf{z}$, first diagonalize $A$ as follows: Find an orthogonal matrix $P_0$ such that

$$P_0^T AP_0 = D = \text{diag}\,(\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_r,\ 0,\ \ldots,\ 0)$$

is diagonal with the nonzero eigenvalues $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_r$ of $A$ on the main diagonal (followed by $n-r$ zeros). By reordering the columns of $P_0$, if necessary, we may assume that $\lambda_1,\ \ldots,\ \lambda_k$ are positive and $\lambda_{k+1},\ \ldots,\ \lambda_r$ are negative. This being the case, let $D_0$ be the $n \times n$ diagonal matrix

$$D_0 = \text{diag}\left(\frac{1}{\sqrt{\lambda_1}},\ \ldots,\ \frac{1}{\sqrt{\lambda_k}},\ \frac{1}{\sqrt{-\lambda_{k+1}}},\ \ldots,\ \frac{1}{\sqrt{-\lambda_r}},\ 1,\ \ldots,\ 1\right)$$

Then $D_0^T DD_0 = D_n(k,\ r)$, so if new variables $\mathbf{z}$ are given by $\mathbf{x} = (P_0 D_0)\mathbf{z}$, we obtain

$$q(\mathbf{z}) = \mathbf{z}^T D_n(k,\ r)\mathbf{z} = z_1^2 + \cdots + z_k^2 - z_{k+1}^2 - \cdots - z_r^2$$

as required. Note that the change-of-variables matrix $P_0 D_0$ from $\mathbf{z}$ to $\mathbf{x}$ has orthogonal columns (in fact, scalar multiples of the columns of $P_0$).

**Example 8.9.5**

Completely diagonalize the quadratic form $q$ in Example 8.9.2 and find the index and rank.

**Solution.** In the notation of Example 8.9.2, the eigenvalues of the matrix $A$ of $q$ are $12, -8, 4, 4$; so the index is 3 and the rank is 4. Moreover, the corresponding orthogonal eigenvectors are $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$ (see Example 8.9.2), and $\mathbf{f}_4$. Hence $P_0 = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_3 & \mathbf{f}_4 & \mathbf{f}_2 \end{bmatrix}$ is orthogonal and

$$P_0^T A P_0 = \operatorname{diag}(12,\ 4,\ 4,\ -8)$$

As before, take $D_0 = \operatorname{diag}(\frac{1}{\sqrt{12}},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{\sqrt{8}})$ and define the new variables $\mathbf{z}$ by $\mathbf{x} = (P_0 D_0)\mathbf{z}$. Hence the new variables are given by $\mathbf{z} = D_0^{-1} P_0^T \mathbf{x}$. The result is

$$\begin{aligned} z_1 &= \sqrt{3}(x_1 - x_2 - x_3 + x_4) \\ z_2 &= x_1 + x_2 + x_3 + x_4 \\ z_3 &= x_1 + x_2 - x_3 - x_4 \\ z_4 &= \sqrt{2}(x_1 - x_2 + x_3 - x_4) \end{aligned}$$

This discussion gives the following information about symmetric matrices.

**Theorem 8.9.5**

*Let $A$ and $B$ be symmetric $n \times n$ matrices, and let $0 \le k \le r \le n$.*

1. *$A$ has index $k$ and* rank *$r$ if and only if $A \overset{c}{\sim} D_n(k,\ r)$.*
2. *$A \overset{c}{\sim} B$ if and only if they have the same* rank *and index.*

**Proof.**

1. If $A$ has index $k$ and rank $r$, take $U = P_0 D_0$ where $P_0$ and $D_0$ are as described prior to Example 8.9.5. Then $U^T A U = D_n(k,\ r)$. The converse is true because $D_n(k,\ r)$ has index $k$ and rank $r$ (using Theorem 8.9.4).
2. If $A$ and $B$ both have index $k$ and rank $r$, then $A \overset{c}{\sim} D_n(k,\ r) \overset{c}{\sim} B$ by (1). The converse was given earlier.

□

**Proof of Theorem 8.9.4.**

By Theorem 8.9.1, $A \overset{C}{\sim} D_1$ and $B \overset{C}{\sim} D_2$ where $D_1$ and $D_2$ are diagonal and have the same eigenvalues as $A$ and $B$, respectively. We have $D_1 \overset{C}{\sim} D_2$, (because $A \overset{C}{\sim} B$), so we may assume that $A$ and $B$ are both diagonal. Consider the quadratic form $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$. If $A$ has $k$ positive eigenvalues, $q$ has the form

$$q(\mathbf{x}) = a_1x_1^2 + \cdots + a_kx_k^2 - a_{k+1}x_{k+1}^2 - \cdots - a_rx_r^2, \quad a_i > 0$$

where $r = \text{rank } A = \text{rank } B$. The subspace $W_1 = \{\mathbf{x} \mid x_{k+1} = \cdots = x_r = 0\}$ of $\mathbb{R}^n$ has dimension $n - r + k$ and satisfies $q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in $W_1$.

On the other hand, if $B = U^T AU$, define new variables $\mathbf{y}$ by $\mathbf{x} = U\mathbf{y}$. If $B$ has $k'$ positive eigenvalues, $q$ has the form

$$q(\mathbf{x}) = b_1y_1^2 + \cdots + b_{k'}y_{k'}^2 - b_{k'+1}y_{k'+1}^2 - \cdots - b_ry_r^2, \quad b_i > 0$$

Let $\mathbf{f}_1, \ldots, \mathbf{f}_n$ denote the columns of $U$. They are a basis of $\mathbb{R}^n$ and

$$\mathbf{x} = U\mathbf{y} = \begin{bmatrix} \mathbf{f}_1 & \cdots & \mathbf{f}_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = y_1\mathbf{f}_1 + \cdots + y_n\mathbf{f}_n$$

Hence the subspace $W_2 = \text{span}\{\mathbf{f}_{k'+1}, \ldots, \mathbf{f}_r\}$ satisfies $q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$ in $W_2$. Note $\dim W_2 = r - k'$. It follows that $W_1$ and $W_2$ have only the zero vector in common. Hence, if $B_1$ and $B_2$ are bases of $W_1$ and $W_2$, respectively, then (Exercise 6.3.33) $B_1 \cup B_2$ is an independent set of $(n - r + k) + (r - k') = n + k - k'$ vectors in $\mathbb{R}^n$. This implies that $k \leq k'$, and a similar argument shows $k' \leq k$. □

## Exercises for 8.9

**Exercise 8.9.1** In each case, find a symmetric matrix $A$ such that $q = \mathbf{x}^T B\mathbf{x}$ takes the form $q = \mathbf{x}^T A\mathbf{x}$.

a. $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$

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

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

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

f. $q = 5x_1^2 + 8x_2^2 + 5x_3^2 - 4(x_1x_2 + 2x_1x_3 + x_2x_3)$

g. $q = x_1^2 - x_3^2 - 4x_1x_2 + 4x_2x_3$

h. $q = x_1^2 + x_3^2 - 2x_1x_2 + 2x_2x_3$

**Exercise 8.9.2** In each case, find a change of variables that will diagonalize the quadratic form $q$. Determine the index and rank of $q$.

a. $q = x_1^2 + 2x_1x_2 + x_2^2$

b. $q = x_1^2 + 4x_1x_2 + x_2^2$

c. $q = x_1^2 + x_2^2 + x_3^2 - 4(x_1x_2 + x_1x_3 + x_2x_3)$

d. $q = 7x_1^2 + x_2^2 + x_3^2 + 8x_1x_2 + 8x_1x_3 - 16x_2x_3$

e. $q = 2(x_1^2 + x_2^2 + x_3^2 - x_1x_2 + x_1x_3 - x_2x_3)$

**Exercise 8.9.3** For each of the following, write the equation in terms of new variables so that it is in standard position, and identify the curve.

a. $xy = 1$

b. $3x^2 - 4xy = 2$

c. $6x^2 + 6xy - 2y^2 = 5$

d. $2x^2 + 4xy + 5y^2 = 1$

**Exercise 8.9.4** Consider the equation $ax^2+bxy+cy^2=d$, where $b\neq 0$. Introduce new variables $x_1$ and $y_1$ by rotating the axes counterclockwise through an angle $\theta$. Show that the resulting equation has no $x_1y_1$-term if $\theta$ is given by

$$\cos 2\theta = \frac{a-c}{\sqrt{b^2+(a-c)^2}}$$
$$\sin 2\theta = \frac{b}{\sqrt{b^2+(a-c)^2}}$$

[*Hint*: Use equation (8.8) preceding Theorem 8.9.2 to get $x$ and $y$ in terms of $x_1$ and $y_1$, and substitute.]

**Exercise 8.9.5** Prove properties (1)–(5) preceding Example 8.9.4.

**Exercise 8.9.6** If $A \overset{c}{\sim} B$ show that $A$ is invertible if and only if $B$ is invertible.

**Exercise 8.9.7** If $\mathbf{x}=(x_1,\ \ldots,\ x_n)^T$ is a column of variables, $A=A^T$ is $n\times n$, $B$ is $1\times n$, and $c$ is a constant, $\mathbf{x}^TA\mathbf{x}+B\mathbf{x}=c$ is called a **quadratic equation** in the variables $x_i$.

a. Show that new variables $y_1,\ \ldots,\ y_n$ can be found such that the equation takes the form

$$\lambda_1y_1^2+\cdots+\lambda_ry_r^2+k_1y_1+\cdots+k_ny_n=c$$

b. Put $x_1^2+3x_2^2+3x_3^2+4x_1x_2-4x_1x_3+5x_1-6x_3=7$ in this form and find variables $y_1$, $y_2$, $y_3$ as in (a).

**Exercise 8.9.8** Given a symmetric matrix $A$, define $q_A(\mathbf{x})=\mathbf{x}^TA\mathbf{x}$. Show that $B \overset{c}{\sim} A$ if and only if $B$ is symmetric and there is an invertible matrix $U$ such that $q_B(\mathbf{x})=q_A(U\mathbf{x})$ for all $\mathbf{x}$. [*Hint*: Theorem 8.9.3.]

**Exercise 8.9.9** Let $q(\mathbf{x})=\mathbf{x}^TA\mathbf{x}$ be a quadratic form where $A=A^T$.

a. Show that $q(\mathbf{x})>0$ for all $\mathbf{x}\neq\mathbf{0}$, if and only if $A$ is positive definite (all eigenvalues are positive). In this case, $q$ is called **positive definite**.

b. Show that new variables $\mathbf{y}$ can be found such that $q=\|\mathbf{y}\|^2$ and $\mathbf{y}=U\mathbf{x}$ where $U$ is upper triangular with positive diagonal entries. [*Hint*: Theorem 8.3.3.]

**Exercise 8.9.10** A **bilinear form** $\beta$ on $\mathbb{R}^n$ is a function that assigns to every pair $\mathbf{x}$, $\mathbf{y}$ of columns in $\mathbb{R}^n$ a number $\beta(\mathbf{x},\ \mathbf{y})$ in such a way that

$$\beta(r\mathbf{x}+s\mathbf{y},\ \mathbf{z})=r\beta(\mathbf{x},\ \mathbf{z})+s\beta(\mathbf{y},\ \mathbf{z})$$
$$\beta(\mathbf{x},\ r\mathbf{y}+s\mathbf{z})=r\beta(\mathbf{x},\ \mathbf{z})+s\beta(\mathbf{x},\ \mathbf{z})$$

for all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ in $\mathbb{R}^n$ and $r$, $s$ in $\mathbb{R}$. If $\beta(\mathbf{x},\ \mathbf{y})=\beta(\mathbf{y},\ \mathbf{x})$ for all $\mathbf{x}$, $\mathbf{y}$, $\beta$ is called **symmetric**.

a. If $\beta$ is a bilinear form, show that an $n\times n$ matrix $A$ exists such that $\beta(\mathbf{x},\ \mathbf{y})=\mathbf{x}^TA\mathbf{y}$ for all $\mathbf{x}$, $\mathbf{y}$.

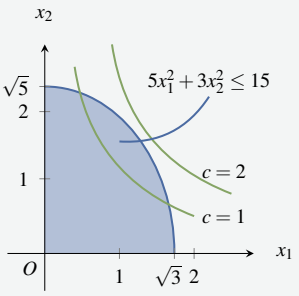
b. Show that $A$ is uniquely determined by $\beta$.

c. Show that $\beta$ is symmetric if and only if $A=A^T$.

## 8.10 An Application to Constrained Optimization

It is a frequent occurrence in applications that a function $q=q(x_1,\ x_2,\ \ldots,\ x_n)$ of $n$ variables, called an **objective function**, is to be made as large or as small as possible among all vectors $\mathbf{x}=(x_1,\ x_2,\ \ldots,\ x_n)$ lying in a certain region of $\mathbb{R}^n$ called the **feasible region**. A wide variety of objective functions $q$ arise in practice; our primary concern here is to examine one important situation where $q$ is a quadratic form. The next example gives some indication of how such problems arise.

**Example 8.10.1**




A politician proposes to spend $x_1$ dollars annually on health care and $x_2$ dollars annually on education. She is constrained in her spending by various budget pressures, and one model of this is that the expenditures $x_1$ and $x_2$ should satisfy a constraint like

$$5x_1^2 + 3x_2^2 \leq 15$$

Since $x_i \geq 0$ for each $i$, the feasible region is the shaded area shown in the diagram. Any choice of feasible point $(x_1,\ x_2)$ in this region will satisfy the budget constraints. However, these choices have different effects on voters, and the politician wants to choose $\mathbf{x} = (x_1,\ x_2)$ to maximize some measure $q = q(x_1,\ x_2)$ of voter satisfaction. Thus the assumption is that, for any value of $c$, all points on the graph of $q(x_1,\ x_2) = c$ have the same appeal to voters. Hence the goal is to find the largest value of $c$ for which the graph of $q(x_1,\ x_2) = c$ contains a feasible point.

The choice of the function $q$ depends upon many factors; we will show how to solve the problem for any quadratic form $q$ (even with more than two variables). In the diagram the function $q$ is given by

$$q(x_1,\ x_2) = x_1x_2$$

and the graphs of $q(x_1,\ x_2) = c$ are shown for $c = 1$ and $c = 2$. As $c$ increases the graph of $q(x_1,\ x_2) = c$ moves up and to the right. From this it is clear that there will be a solution for some value of $c$ between 1 and 2 (in fact the largest value is $c = \frac{1}{2}\sqrt{15} = 1.94$ to two decimal places).

The constraint $5x_1^2 + 3x_2^2 \leq 15$ in Example 8.10.1 can be put in a standard form. If we divide through by 15, it becomes $\left(\frac{x_1}{\sqrt{3}}\right)^2 + \left(\frac{x_2}{\sqrt{5}}\right)^2 \leq 1$. This suggests that we introduce new variables $\mathbf{y} = (y_1,\ y_2)$ where $y_1 = \frac{x_1}{\sqrt{3}}$ and $y_2 = \frac{x_2}{\sqrt{5}}$. Then the constraint becomes $\|\mathbf{y}\|^2 \leq 1$, equivalently $\|\mathbf{y}\| \leq 1$. In terms of these new variables, the objective function is $q = \sqrt{15}y_1y_2$, and we want to maximize this subject to $\|\mathbf{y}\| \leq 1$. When this is done, the maximizing values of $x_1$ and $x_2$ are obtained from $x_1 = \sqrt{3}y_1$ and $x_2 = \sqrt{5}y_2$.

Hence, for constraints like that in Example 8.10.1, there is no real loss in generality in assuming that the constraint takes the form $\|\mathbf{x}\| \leq 1$. In this case the principal axes theorem solves the problem. Recall that a vector in $\mathbb{R}^n$ of length 1 is called a *unit vector*.

**Theorem 8.10.1**

*Consider the quadratic form $q = q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ where $A$ is an $n \times n$ symmetric matrix, and let $\lambda_1$ and $\lambda_n$ denote the largest and smallest eigenvalues of $A$, respectively. Then:*

1. $\max\{q(\mathbf{x}) \mid \|\mathbf{x}\| \leq 1\} = \lambda_1$, *and* $q(\mathbf{f}_1) = \lambda_1$ *where $\mathbf{f}_1$ is any unit $\lambda_1$-eigenvector.*
2. $\min\{q(\mathbf{x}) \mid \|\mathbf{x}\| \leq 1\} = \lambda_n$, *and* $q(\mathbf{f}_n) = \lambda_n$ *where $\mathbf{f}_n$ is any unit $\lambda_n$-eigenvector.*

**Proof.** Since $A$ is symmetric, let the (real) eigenvalues $\lambda_i$ of $A$ be ordered as to size as follows:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

By the principal axes theorem, let $P$ be an orthogonal matrix such that $P^TAP = D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Define $\mathbf{y} = P^T\mathbf{x}$, equivalently $\mathbf{x} = P\mathbf{y}$, and note $\|\mathbf{y}\| = \|\mathbf{x}\|$ because $\|\mathbf{y}\|^2 = \mathbf{y}^T\mathbf{y} = \mathbf{x}^T(PP^T)\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2$. If we write $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, then

$$\begin{aligned} q(\mathbf{x}) &= q(P\mathbf{y}) = (P\mathbf{y})^T A(P\mathbf{y}) \\ &= \mathbf{y}^T(P^TAP)\mathbf{y} = \mathbf{y}^T D\mathbf{y} \\ &= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \end{aligned} \tag{8.9}$$

Now assume that $\|\mathbf{x}\| \leq 1$. Since $\lambda_i \leq \lambda_1$ for each $i$, (8.9) gives

$$q(\mathbf{x}) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \leq \lambda_1 y_1^2 + \lambda_1 y_2^2 + \cdots + \lambda_1 y_n^2 = \lambda_1 \|\mathbf{y}\|^2 \leq \lambda_1$$

because $\|\mathbf{y}\| = \|\mathbf{x}\| \leq 1$. This shows that $q(\mathbf{x})$ cannot exceed $\lambda_1$ when $\|\mathbf{x}\| \leq 1$. To see that this maximum is actually achieved, let $\mathbf{f}_1$ be a unit eigenvector corresponding to $\lambda_1$. Then

$$q(\mathbf{f}_1) = \mathbf{f}_1^T A\mathbf{f}_1 = \mathbf{f}_1^T(\lambda_1\mathbf{f}_1) = \lambda_1(\mathbf{f}_1^T\mathbf{f}_1) = \lambda_1\|\mathbf{f}_1\|^2 = \lambda_1$$

Hence $\lambda_1$ is the maximum value of $q(\mathbf{x})$ when $\|\mathbf{x}\| \leq 1$, proving (1). The proof of (2) is analogous. □

The set of all vectors $\mathbf{x}$ in $\mathbb{R}^n$ such that $\|\mathbf{x}\| \leq 1$ is called the **unit ball**. If $n = 2$, it is often called the unit disk and consists of the unit circle and its interior; if $n = 3$, it is the unit sphere and its interior. It is worth noting that the maximum value of a quadratic form $q(\mathbf{x})$ as $\mathbf{x}$ ranges *throughout* the unit ball is (by Theorem 8.10.1) actually attained for a unit vector $\mathbf{x}$ on the *boundary* of the unit ball.

Theorem 8.10.1 is important for applications involving vibrations in areas as diverse as aerodynamics and particle physics, and the maximum and minimum values in the theorem are often found using advanced calculus to minimize the quadratic form on the unit ball. The algebraic approach using the principal axes theorem gives a geometrical interpretation of the optimal values because they are eigenvalues.

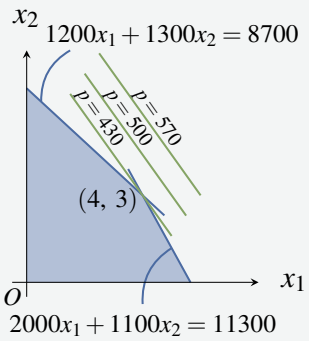
**Example 8.10.2**

Maximize and minimize the form $q(\mathbf{x}) = 3x_1^2 + 14x_1x_2 + 3x_2^2$ subject to $\|\mathbf{x}\| \leq 1$.

**Solution.** The matrix of $q$ is $A = \begin{bmatrix} 3 & 7 \\ 7 & 3 \end{bmatrix}$, with eigenvalues $\lambda_1 = 10$ and $\lambda_2 = -4$, and corresponding unit eigenvectors $\mathbf{f}_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $\mathbf{f}_2 = \frac{1}{\sqrt{2}}(1, -1)$. Hence, among all unit vectors $\mathbf{x}$ in $\mathbb{R}^2$, $q(\mathbf{x})$ takes its maximal value 10 at $\mathbf{x} = \mathbf{f}_1$, and the minimum value of $q(\mathbf{x})$ is $-4$ when $\mathbf{x} = \mathbf{f}_2$.

As noted above, the objective function in a constrained optimization problem need not be a quadratic form. We conclude with an example where the objective function is linear, and the feasible region is determined by linear constraints.

**Example 8.10.3**




A manufacturer makes $x_1$ units of product 1, and $x_2$ units of product 2, at a profit of \$70 and \$50 per unit respectively, and wants to choose $x_1$ and $x_2$ to maximize the total profit $p(x_1, x_2) = 70x_1 + 50x_2$. However $x_1$ and $x_2$ are not arbitrary; for example, $x_1 \geq 0$ and $x_2 \geq 0$. Other conditions also come into play. Each unit of product 1 costs \$1200 to produce and requires 2000 square feet of warehouse space; each unit of product 2 costs \$1300 to produce and requires 1100 square feet of space. If the total warehouse space is 11 300 square feet, and if the total production budget is \$8700, $x_1$ and $x_2$ must also satisfy the conditions

$$2000x_1 + 1100x_2 \leq 11300$$
$$1200x_1 + 1300x_2 \leq 8700$$

The feasible region in the plane satisfying these constraints (and $x_1 \geq 0$, $x_2 \geq 0$) is shaded in the diagram. If the profit equation $70x_1 + 50x_2 = p$ is plotted for various values of $p$, the resulting lines are parallel, with $p$ increasing with distance from the origin. Hence the best choice occurs for the line $70x_1 + 50x_2 = 430$ that touches the shaded region at the point $(4, 3)$. So the profit $p$ has a maximum of $p = 430$ for $x_1 = 4$ units and $x_2 = 3$ units.

Example 8.10.3 is a simple case of the general **linear programming** problem[23] which arises in economic, management, network, and scheduling applications. Here the objective function is a linear combination $q = a_1x_1 + a_2x_2 + \cdots + a_nx_n$ of the variables, and the feasible region consists of the vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ in $\mathbb{R}^n$ which satisfy a set of linear inequalities of the form $b_1x_1 + b_2x_2 + \cdots + b_nx_n \leq b$. There is a good method (an extension of the gaussian algorithm) called the **simplex algorithm** for finding the maximum and minimum values of $q$ when $\mathbf{x}$ ranges over such a feasible set. As Example 8.10.3 suggests, the optimal values turn out to be vertices of the feasible set. In particular, they are on the boundary of the feasible region, as is the case in Theorem 8.10.1.

## 8.11 An Application to Statistical Principal Component Analysis

Linear algebra is important in multivariate analysis in statistics, and we conclude with a very short look at one application of diagonalization in this area. A main feature of probability and statistics is the idea of a *random variable* $X$, that is a real-valued function which takes its values according to a probability law (called its *distribution*). Random variables occur in a wide variety of contexts; examples include the number of meteors falling per square kilometre in a given region, the price of a share of a stock, or the duration of a long distance telephone call from a certain city.

The values of a random variable $X$ are distributed about a central number $\mu$, called the *mean* of $X$. The mean can be calculated from the distribution as the *expectation* $E(X) = \mu$ of the random variable $X$. Functions of a random variable are again random variables. In particular, $(X - \mu)^2$ is a random variable,

[23] More information is available in "Linear Programming and Extensions" by N. Wu and R. Coppins, McGraw-Hill, 1981.

and the *variance* of the random variable $X$, denoted $\text{var}(X)$, is defined to be the number

$$\text{var}(X) = E\{(X-\mu)^2\} \quad \text{where } \mu = E(X)$$

It is not difficult to see that $\text{var}(X) \geq 0$ for every random variable $X$. The number $\sigma = \sqrt{\text{var}(X)}$ is called the *standard deviation* of $X$, and is a measure of how much the values of $X$ are spread about the mean $\mu$ of $X$. A main goal of statistical inference is finding reliable methods for estimating the mean and the standard deviation of a random variable $X$ by sampling the values of $X$.

If two random variables $X$ and $Y$ are given, and their joint distribution is known, then functions of $X$ and $Y$ are also random variables. In particular, $X+Y$ and $aX$ are random variables for any real number $a$, and we have

$$E(X+Y) = E(X)+E(Y) \quad \text{and} \quad E(aX) = aE(X).^{24}$$

An important question is how much the random variables $X$ and $Y$ depend on each other. One measure of this is the *covariance* of $X$ and $Y$, denoted $\text{cov}(X, Y)$, defined by

$$\text{cov}(X, Y) = E\{(X-\mu)(Y-\upsilon)\} \quad \text{where } \mu = E(X) \text{ and } \upsilon = E(Y)$$

Clearly, $\text{cov}(X, X) = \text{var}(X)$. If $\text{cov}(X, Y) = 0$ then $X$ and $Y$ have little relationship to each other and are said to be *uncorrelated*.[25]

Multivariate statistical analysis deals with a family $X_1, X_2, \ldots, X_n$ of random variables with means $\mu_i = E(X_i)$ and variances $\sigma_i^2 = \text{var}(X_i)$ for each $i$. Let $\sigma_{ij} = \text{cov}(X_i, X_j)$ denote the covariance of $X_i$ and $X_j$. Then the *covariance matrix* of the random variables $X_1, X_2, \ldots, X_n$ is defined to be the $n \times n$ matrix

$$\Sigma = [\sigma_{ij}]$$

whose $(i, j)$-entry is $\sigma_{ij}$. The matrix $\Sigma$ is clearly symmetric; in fact it can be shown that $\Sigma$ is **positive semidefinite** in the sense that $\lambda \geq 0$ for every eigenvalue $\lambda$ of $\Sigma$. (In reality, $\Sigma$ is positive definite in most cases of interest.) So suppose that the eigenvalues of $\Sigma$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. The principal axes theorem (Theorem 8.2.2) shows that an orthogonal matrix $P$ exists such that

$$P^T \Sigma P = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

If we write $\overline{X} = (X_1, X_2, \ldots, X_n)$, the procedure for diagonalizing a quadratic form gives new variables $\overline{Y} = (Y_1, Y_2, \ldots, Y_n)$ defined by

$$\overline{Y} = P^T \overline{X}$$

These new random variables $Y_1, Y_2, \ldots, Y_n$ are called the **principal components** of the original random variables $X_i$, and are linear combinations of the $X_i$. Furthermore, it can be shown that

$$\text{cov}(Y_i, Y_j) = 0 \text{ if } i \neq j \quad \text{and} \quad \text{var}(Y_i) = \lambda_i \quad \text{for each } i$$

Of course the principal components $Y_i$ point along the principal axes of the quadratic form $q = \overline{X}^T \Sigma \overline{X}$.

The sum of the variances of a set of random variables is called the **total variance** of the variables, and determining the source of this total variance is one of the benefits of principal component analysis. The fact that the matrices $\Sigma$ and $\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ are similar means that they have the same trace, that is,

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

[24] Hence $E(\ )$ is a linear transformation from the vector space of all random variables to the space of real numbers.

[25] If $X$ and $Y$ are independent in the sense of probability theory, then they are uncorrelated; however, the converse is not true in general.

This means that the principal components $Y_i$ have the same total variance as the original random variables $X_i$. Moreover, the fact that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ means that most of this variance resides in the first few $Y_i$. In practice, statisticians find that studying these first few $Y_i$ (and ignoring the rest) gives an accurate analysis of the total system variability. This results in substantial data reduction since often only a few $Y_i$ suffice for all practical purposes. Furthermore, these $Y_i$ are easily obtained as linear combinations of the $X_i$. Finally, the analysis of the principal components often reveals relationships among the $X_i$ that were not previously suspected, and so results in interpretations that would not otherwise have been made.