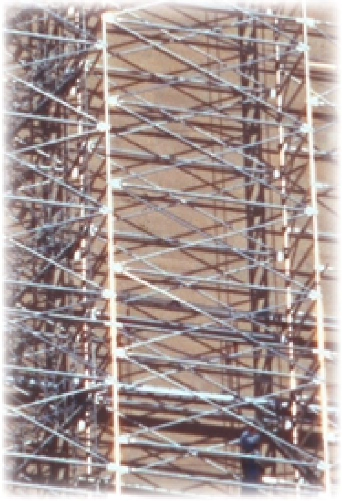


CHAPTER 5

Norms, Inner Products, and Orthogonality

5.1 VECTOR NORMS

A significant portion of linear algebra is in fact geometric in nature because much of the subject grew out of the need to generalize the basic geometry of $\Re^2$ and $\Re^3$ to nonvisual higher-dimensional spaces. The usual approach is to coordinatize geometric concepts in $\Re^2$ and $\Re^3$, and then extend statements concerning ordered pairs and triples to ordered n-tuples in $\Re^n$ and $\mathcal{C}^n$.

For example, the length of a vector $\mathbf{u} \in \Re^2$ or $\mathbf{v} \in \Re^3$ is obtained from the Pythagorean theorem by computing the length of the hypotenuse of a right triangle as shown in Figure 5.1.1.

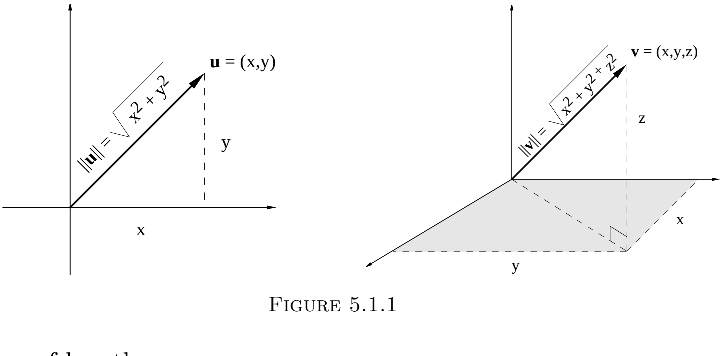


Figure 5.1.1

This measure of length,

$$\|\mathbf{u}\| = \sqrt{x^2 + y^2} \quad \text{and} \quad \|\mathbf{v}\| = \sqrt{x^2 + y^2 + z^2},$$

is called the *euclidean norm* in $\Re^2$ and $\Re^3$, and there is an obvious extension to higher dimensions.

Euclidean Vector Norm

For a vector $\mathbf{x}_{n\times 1}$, the ***euclidean norm*** of $\mathbf{x}$ is defined to be

- $\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} = \sqrt{\mathbf{x}^T\mathbf{x}}$ whenever $\mathbf{x}\in\Re^{n\times 1}$,
- $\|\mathbf{x}\| = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} = \sqrt{\mathbf{x}^*\mathbf{x}}$ whenever $\mathbf{x}\in\mathcal{C}^{n\times 1}$.

For example, if $\mathbf{u} = \begin{pmatrix} 0\\ -1\\ 2\\ -2\\ 4 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} \mathrm{i}\\ 2\\ 1-\mathrm{i}\\ 0\\ 1+\mathrm{i} \end{pmatrix}$, then

$$\|\mathbf{u}\| = \sqrt{\sum u_i^2} = \sqrt{\mathbf{u}^T\mathbf{u}} = \sqrt{0+1+4+4+16} = 5,$$

$$\|\mathbf{v}\| = \sqrt{\sum |v_i|^2} = \sqrt{\mathbf{v}^*\mathbf{v}} = \sqrt{1+4+2+0+2} = 3.$$

There are several points to note.[33]

- The complex version of $\|\mathbf{x}\|$ includes the real version as a special case because $|z|^2 = z^2$ whenever z is a real number. Recall that if $z = a + \mathrm{i}b$, then $\bar{z} = a - \mathrm{i}b$, and the magnitude of z is $|z| = \sqrt{\bar{z}z} = \sqrt{a^2+b^2}$. The fact that $|z|^2 = \bar{z}z = a^2 + b^2$ is a real number insures that $\|\mathbf{x}\|$ is real even if $\mathbf{x}$ has some complex components.
- The definition of euclidean norm guarantees that for all scalars α,

$$\|\mathbf{x}\| \ge 0, \quad \|\mathbf{x}\| = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}, \quad \text{and} \quad \|\alpha\mathbf{x}\| = |\alpha|\ \|\mathbf{x}\|. \tag{5.1.1}$$

- Given a vector $\mathbf{x} \ne \mathbf{0}$, it's frequently convenient to have another vector that points in the same direction as $\mathbf{x}$ (i.e., is a positive multiple of $\mathbf{x}$) but has unit length. To construct such a vector, we ***normalize*** $\mathbf{x}$ by setting $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$. From (5.1.1), it's easy to see that

$$\|\mathbf{u}\| = \left\|\frac{\mathbf{x}}{\|\mathbf{x}\|}\right\| = \frac{1}{\|\mathbf{x}\|}\|\mathbf{x}\| = 1. \tag{5.1.2}$$

[33] By convention, column vectors are used throughout this chapter. But there is nothing special about columns because, with the appropriate interpretation, all statements concerning columns will also hold for rows.

- The distance between vectors in $\Re^3$ can be visualized with the aid of the parallelogram law as shown in Figure 5.1.2, so for vectors in $\Re^n$ and $\mathcal{C}^n$, the ***distance*** between $\mathbf{u}$ and $\mathbf{v}$ is naturally defined to be $\|\mathbf{u}-\mathbf{v}\|$.

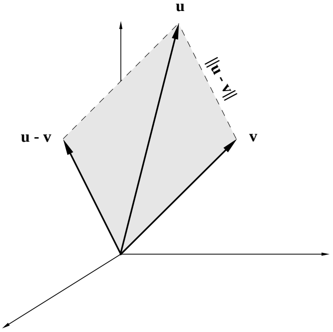


FIGURE 5.1.2

Standard Inner Product

The scalar terms defined by

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i \in \Re \quad \text{and} \quad \mathbf{x}^*\mathbf{y} = \sum_{i=1}^{n} \bar{x}_i y_i \in \mathcal{C}$$

are called the ***standard inner products*** for $\Re^n$ and $\mathcal{C}^n$, respectively.

The Cauchy–Bunyakovskii–Schwarz (CBS) inequality[34] is one of the most important inequalities in mathematics. It relates inner product to norm.

[34] The Cauchy–Bunyakovskii–Schwarz inequality is named in honor of the three men who played a role in its development. The basic inequality for real numbers is attributed to Cauchy in 1821, whereas Schwarz and Bunyakovskii contributed by later formulating useful generalizations of the inequality involving integrals of functions.

Augustin-Louis Cauchy (1789–1857) was a French mathematician who is generally regarded as being the founder of mathematical analysis—including the theory of complex functions. Although deeply embroiled in political turmoil for much of his life (he was a partisan of the Bourbons), Cauchy emerged as one of the most prolific mathematicians of all time. He authored at least 789 mathematical papers, and his collected works fill 27 volumes—this is on a par with Cayley and second only to Euler. It is said that more theorems, concepts, and methods bear Cauchy's name than any other mathematician.

Victor Bunyakovskii (1804–1889) was a Russian professor of mathematics at St. Petersburg, and in 1859 he extended Cauchy's inequality for discrete sums to integrals of continuous functions. His contribution was overlooked by western mathematicians for many years, and his name is often omitted in classical texts that simply refer to the *Cauchy–Schwarz inequality.*

Hermann Amandus Schwarz (1843–1921) was a student and successor of the famous German mathematician Karl Weierstrass at the University of Berlin. Schwarz independently generalized Cauchy's inequality just as Bunyakovskii had done earlier.

Cauchy–Bunyakovskii–Schwarz (CBS) Inequality

$$|\mathbf{x}^*\mathbf{y}| \le \|\mathbf{x}\|\ \|\mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathcal{C}^{n\times 1}. \tag{5.1.3}$$

Equality holds if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha = \mathbf{x}^*\mathbf{y}/\mathbf{x}^*\mathbf{x}$.

Proof. Set $\alpha = \mathbf{x}^*\mathbf{y}/\mathbf{x}^*\mathbf{x} = \mathbf{x}^*\mathbf{y}/\|\mathbf{x}\|^2$ (assume $\mathbf{x} \ne \mathbf{0}$ because there is nothing to prove if $\mathbf{x} = \mathbf{0}$) and observe that $\mathbf{x}^*(\alpha\mathbf{x} - \mathbf{y}) = 0$, so

$$\begin{aligned} 0 \le \|\alpha\mathbf{x} - \mathbf{y}\|^2 &= (\alpha\mathbf{x} - \mathbf{y})^*(\alpha\mathbf{x} - \mathbf{y}) = \bar{\alpha}\mathbf{x}^*(\alpha\mathbf{x} - \mathbf{y}) - \mathbf{y}^*(\alpha\mathbf{x} - \mathbf{y}) \\ &= -\mathbf{y}^*(\alpha\mathbf{x} - \mathbf{y}) = \mathbf{y}^*\mathbf{y} - \alpha\mathbf{y}^*\mathbf{x} = \frac{\|\mathbf{y}\|^2\|\mathbf{x}\|^2 - (\mathbf{x}^*\mathbf{y})(\mathbf{y}^*\mathbf{x})}{\|\mathbf{x}\|^2}. \end{aligned} \tag{5.1.4}$$

Since $\mathbf{y}^*\mathbf{x} = \overline{\mathbf{x}^*\mathbf{y}}$, it follows that $(\mathbf{x}^*\mathbf{y})(\mathbf{y}^*\mathbf{x}) = |\mathbf{x}^*\mathbf{y}|^2$, so

$$0 \le \frac{\|\mathbf{y}\|^2\|\mathbf{x}\|^2 - |\mathbf{x}^*\mathbf{y}|^2}{\|\mathbf{x}\|^2}.$$

Now, $0 < \|\mathbf{x}\|^2$ implies $0 \le \|\mathbf{y}\|^2\|\mathbf{x}\|^2 - |\mathbf{x}^*\mathbf{y}|^2$, and thus the CBS inequality is obtained. Establishing the conditions for equality is Exercise 5.1.9. ■

One reason that the CBS inequality is important is because it helps to establish that the geometry in higher-dimensional spaces is consistent with the geometry in the visual spaces $\Re^2$ and $\Re^3$. In particular, consider the situation depicted in Figure 5.1.3.

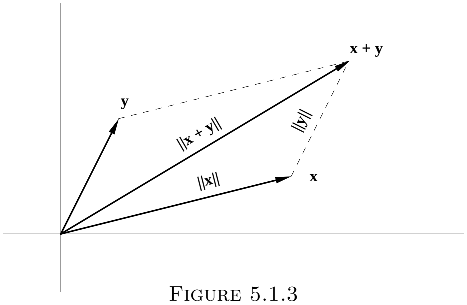


FIGURE 5.1.3

Imagine traveling from the origin to the point $\mathbf{x}$ and then moving from $\mathbf{x}$ to the point $\mathbf{x}+\mathbf{y}$. Clearly, you have traveled a distance that is at least as great as the direct distance from the origin to $\mathbf{x}+\mathbf{y}$ along the diagonal of the parallelogram. In other words, it's visually evident that $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$. This observation is known as the ***triangle inequality.*** In higher-dimensional spaces we do not have the luxury of visualizing the geometry with our eyes, and the question of whether or not the triangle inequality remains valid has no obvious answer. The CBS inequality is precisely what is required to prove that, in this respect, the geometry of higher dimensions is no different than that of the visual spaces.

Triangle Inequality

$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \text{for every } \mathbf{x},\, \mathbf{y} \in \mathcal{C}^n.$$

Proof. Consider $\mathbf{x}$ and $\mathbf{y}$ to be column vectors, and write

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= (\mathbf{x}+\mathbf{y})^*(\mathbf{x}+\mathbf{y}) = \mathbf{x}^*\mathbf{x} + \mathbf{x}^*\mathbf{y} + \mathbf{y}^*\mathbf{x} + \mathbf{y}^*\mathbf{y} \\ &= \|\mathbf{x}\|^2 + \mathbf{x}^*\mathbf{y} + \mathbf{y}^*\mathbf{x} + \|\mathbf{y}\|^2.\end{aligned} \tag{5.1.5}$$

Recall that if $z = a + \mathrm{i}b$, then $z + \bar{z} = 2a = 2\,\mathrm{Re}\,(z)$ and $|z|^2 = a^2 + b^2 \geq a^2$, so that $|z| \geq \mathrm{Re}\,(z)$. Using the fact that $\mathbf{y}^*\mathbf{x} = \overline{\mathbf{x}^*\mathbf{y}}$ together with the CBS inequality yields

$$\mathbf{x}^*\mathbf{y} + \mathbf{y}^*\mathbf{x} = 2\,\mathrm{Re}\,(\mathbf{x}^*\mathbf{y}) \leq 2\,|\mathbf{x}^*\mathbf{y}| \leq 2\,\|\mathbf{x}\|\;\|\mathbf{y}\|.$$

Consequently, we may infer from (5.1.5) that

$$\|\mathbf{x}+\mathbf{y}\|^2 \leq \|\mathbf{x}\|^2 + 2\,\|\mathbf{x}\|\;\|\mathbf{y}\| + \|\mathbf{y}\|^2 = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \quad \blacksquare$$

It's not difficult to see that the triangle inequality can be extended to any number of vectors in the sense that $\left\|\sum_i \mathbf{x}_i\right\| \leq \sum_i \|\mathbf{x}_i\|$. Furthermore, it follows as a corollary that for real or complex numbers, $\left|\sum_i \alpha_i\right| \leq \sum_i |\alpha_i|$ (the triangle inequality for scalars).

Example 5.1.1

Backward Triangle Inequality. The triangle inequality produces an upper bound for a sum, but it also yields the following lower bound for a difference:

$$\big|\,\|\mathbf{x}\| - \|\mathbf{y}\|\,\big| \leq \|\mathbf{x}-\mathbf{y}\|. \tag{5.1.6}$$

This is a consequence of the triangle inequality because

$$\|\mathbf{x}\| = \|\mathbf{x}-\mathbf{y}+\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{y}\| + \|\mathbf{y}\| \implies \|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{y}\|$$

and

$$\|\mathbf{y}\| = \|\mathbf{x}-\mathbf{y}\| + \|\mathbf{x}\| \implies -(\|\mathbf{x}\| - \|\mathbf{y}\|) \leq \|\mathbf{x}-\mathbf{y}\|.$$

There are notions of length other than the euclidean measure. For example, urban dwellers navigate on a grid of city blocks with one-way streets, so they are prone to measure distances in the city not as the crow flies but rather in terms of lengths on a directed grid. For example, instead of than saying that "it's a one-half mile straight-line (euclidean) trip from here to there," they are more apt to describe the length of the trip by saying, "it's two blocks north on Dan Allen Drive, four blocks west on Hillsborough Street, and five blocks south on Gorman Street." In other words, the length of the trip is $2+|-4|+|-5|=11$ blocks—absolute value is used to insure that southerly and westerly movement does not cancel the effect of northerly and easterly movement, respectively. This "grid norm" is better known as the 1-norm because it is a special case of a more general class of norms defined below.

p-Norms

For $p \geq 1$, the ***p-norm*** of $\mathbf{x} \in \mathcal{C}^n$ is defined as $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$.

It can be proven that the following properties of the euclidean norm are in fact valid for all p-norms:

$$\begin{aligned}
&\|\mathbf{x}\|_p \geq 0 \quad \text{and} \quad \|\mathbf{x}\|_p = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}, \\
&\|\alpha\mathbf{x}\|_p = |\alpha| \; \|\mathbf{x}\|_p \quad \text{for all scalars } \alpha, \\
&\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p \qquad \text{(see Exercise 5.1.13).}
\end{aligned} \tag{5.1.7}$$

The generalized version of the CBS inequality (5.1.3) for p-norms is ***Hölder's inequality*** (developed in Exercise 5.1.12), which states that if $p > 1$ and $q > 1$ are integers such that $1/p + 1/q = 1$, then

$$|\mathbf{x}^*\mathbf{y}| \leq \|\mathbf{x}\|_p \; \|\mathbf{y}\|_q \,. \tag{5.1.8}$$

In practice, only three of the p-norms are used, and they are

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \quad \text{(the grid norm)}, \quad \|\mathbf{x}\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2} \text{(the euclidean norm)},$$

and

$$\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p = \lim_{p\to\infty} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} = \max_i |x_i| \qquad \text{(the max norm).}$$

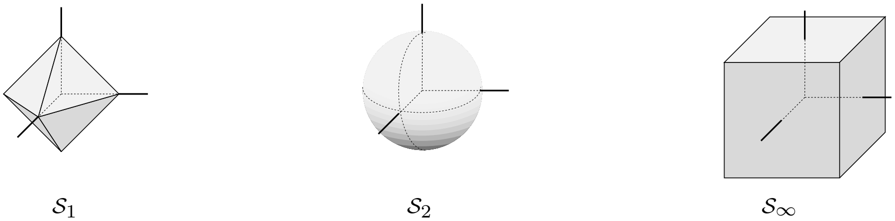
For example, if $\mathbf{x} = (3,\, 4-3\mathrm{i},\, 1)$, then $\|\mathbf{x}\|_1 = 9$, $\|\mathbf{x}\|_2 = \sqrt{35}$, and $\|\mathbf{x}\|_\infty = 5$.

To see that $\lim_{p\to\infty} \|\mathbf{x}\|_p = \max_i |x_i|$, proceed as follows. Relabel the entries of $\mathbf{x}$ by setting $\tilde{x}_1 = \max_i |x_i|$, and if there are other entries with this same maximal magnitude, label them $\tilde{x}_2, \ldots, \tilde{x}_k$. Label any remaining coordinates as $\tilde{x}_{k+1} \cdots \tilde{x}_n$. Consequently, $|\tilde{x}_i/\tilde{x}_1| < 1$ for $i = k+1, \ldots, n$, so, as $p \to \infty$,

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\tilde{x}_i|^p\right)^{1/p} = |\tilde{x}_1| \left(k + \left|\frac{\tilde{x}_{k+1}}{\tilde{x}_1}\right|^p + \cdots + \left|\frac{\tilde{x}_n}{\tilde{x}_1}\right|^p\right)^{1/p} \to |\tilde{x}_1|.$$

Example 5.1.2

To get a feel for the 1-, 2-, and ∞-norms, it helps to know the shapes and relative sizes of the ***unit p-spheres*** $\mathcal{S}_p = \{\mathbf{x} \mid \|\mathbf{x}\|_p = 1\}$ for $p = 1,\ 2,\ \infty$. As illustrated in Figure 5.1.4, the unit 1-, 2-, and ∞-spheres in $\Re^3$ are an octahedron, a ball, and a cube, respectively, and it's visually evident that $\mathcal{S}_1$ fits inside $\mathcal{S}_2$, which in turn fits inside $\mathcal{S}_\infty$. This means that $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2 \geq \|\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \Re^3$. In general, this is true in $\Re^n$ (Exercise 5.1.8).



$\mathcal{S}_1$ $\mathcal{S}_2$ $\mathcal{S}_\infty$

FIGURE 5.1.4

Because the p-norms are defined in terms of coordinates, their use is limited to coordinate spaces. But it's desirable to have a general notion of norm that works for *all* vector spaces. In other words, we need a coordinate-free definition of norm that includes the standard p-norms as a special case. Since all of the p-norms satisfy the properties (5.1.7), it's natural to use these properties to extend the concept of norm to general vector spaces.

General Vector Norms

A ***norm*** for a real or complex vector space $\mathcal{V}$ is a function $\|\star\|$ mapping $\mathcal{V}$ into $\Re$ that satisfies the following conditions.

$$\begin{aligned} &\|\mathbf{x}\| \geq 0 \quad \text{and} \quad \|\mathbf{x}\| = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}, \\ &\|\alpha\mathbf{x}\| = |\alpha|\ \|\mathbf{x}\| \quad \text{for all scalars } \alpha, \\ &\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|. \end{aligned} \tag{5.1.9}$$

Example 5.1.3

Equivalent Norms. Vector norms are basic tools for defining and analyzing limiting behavior in vector spaces $\mathcal{V}$. A sequence $\{\mathbf{x}_k\} \subset \mathcal{V}$ is said to converge to $\mathbf{x}$ (write $\mathbf{x}_k \rightarrow \mathbf{x}$) if $\|\mathbf{x}_k - \mathbf{x}\| \rightarrow 0$. This depends on the choice of the norm, so, ostensibly, we might have $\mathbf{x}_k \rightarrow \mathbf{x}$ with one norm but not with another. Fortunately, this is impossible in finite-dimensional spaces because all norms are *equivalent* in the following sense.

Problem: For each pair of norms, $\|\star\|_a$, $\|\star\|_b$, on an n-dimensional space $\mathcal{V}$, exhibit positive constants α and β (depending only on the norms) such that

$$\alpha \leq \frac{\|\mathbf{x}\|_a}{\|\mathbf{x}\|_b} \leq \beta \quad \text{for all nonzero vectors in } \mathcal{V}. \tag{5.1.10}$$

Solution: For $\mathcal{S}_b = \{\mathbf{y} \mid \|\mathbf{y}\|_b = 1\}$, let $\mu = \min_{\mathbf{y} \in \mathcal{S}_b} \|\mathbf{y}\|_a > 0$,[35] and write

$$\frac{\mathbf{x}}{\|\mathbf{x}\|_b} \in \mathcal{S}_b \quad \Longrightarrow \quad \|\mathbf{x}\|_a = \|\mathbf{x}\|_b \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_b} \right\|_a \geq \|\mathbf{x}\|_b \min_{\mathbf{y} \in \mathcal{S}_b} \|\mathbf{y}\|_a = \|\mathbf{x}\|_b \, \mu.$$

The same argument shows there is a $\nu > 0$ such that $\|\mathbf{x}\|_b \geq \nu \|\mathbf{x}\|_a$, so (5.1.10) is produced with $\alpha = \mu$ and $\beta = 1/\nu$. Note that (5.1.10) insures that $\|\mathbf{x}_k - \mathbf{x}\|_a \rightarrow 0$ if and only if $\|\mathbf{x}_k - \mathbf{x}\|_b \rightarrow 0$. Specific values for α and β are given in Exercises 5.1.8 and 5.12.3.

Exercises for section 5.1

5.1.1. Find the 1-, 2-, and ∞-norms of $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ -4 \\ -2 \end{pmatrix}$ and $\mathbf{x} = \begin{pmatrix} 1+\mathrm{i} \\ 1-\mathrm{i} \\ 1 \\ 4\mathrm{i} \end{pmatrix}$.

5.1.2. Consider the euclidean norm with $\mathbf{u} = \begin{pmatrix} 2 \\ 1 \\ -4 \\ -2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$.

(a) Determine the distance between $\mathbf{u}$ and $\mathbf{v}$.
(b) Verify that the triangle inequality holds for $\mathbf{u}$ and $\mathbf{v}$.
(c) Verify that the CBS inequality holds for $\mathbf{u}$ and $\mathbf{v}$.

5.1.3. Show that $(\alpha_1 + \alpha_2 + \cdots + \alpha_n)^2 \leq n\left(\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2\right)$ for $\alpha_i \in \Re$.

[35] An important theorem from analysis states that a continuous function mapping a closed and bounded subset $\mathcal{K} \subset \mathcal{V}$ into $\Re$ attains a minimum and maximum value at points in $\mathcal{K}$. Unit spheres in finite-dimensional spaces are closed and bounded, and every norm on $\mathcal{V}$ is continuous (Exercise 5.1.7), so this minimum is guaranteed to exist.

5.1.4. (a) Using the euclidean norm, describe the solid ball in $\Re^n$ centered at the origin with unit radius. (b) Describe a solid ball centered at the point $\mathbf{c} = (\,\xi_1 \quad \xi_2 \quad \cdots \quad \xi_n\,)$ with radius ρ.

5.1.5. If $\mathbf{x}, \mathbf{y} \in \Re^n$ such that $\|\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{x} + \mathbf{y}\|_2$, what is $\mathbf{x}^T\mathbf{y}$?

5.1.6. Explain why $\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{y} - \mathbf{x}\|$ is true for all norms.

5.1.7. For every vector norm on $\mathcal{C}^n$, prove that $\|\mathbf{v}\|$ depends continuously on the components of $\mathbf{v}$ in the sense that for each $\epsilon > 0$, there corresponds a $\delta > 0$ such that $\big|\, \|\mathbf{x}\| - \|\mathbf{y}\| \,\big| < \epsilon$ whenever $|x_i - y_i| < \delta$ for each i.

5.1.8. (a) For $\mathbf{x} \in \mathcal{C}^{n \times 1}$, explain why $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2 \geq \|\mathbf{x}\|_\infty$.

(b) For $\mathbf{x} \in \mathcal{C}^{n \times 1}$, show that $\|\mathbf{x}\|_i \leq \alpha \, \|\mathbf{x}\|_j$, where α is the (i, j)-entry in the following matrix. (See Exercise 5.12.3 for a similar statement regarding matrix norms.)

$$\begin{array}{cc} & \begin{array}{ccc} 1 & 2 & \infty \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \infty \end{array} & \left(\begin{array}{ccc} * & \sqrt{n} & n \\ 1 & * & \sqrt{n} \\ 1 & 1 & * \end{array}\right) \end{array}.$$

5.1.9. For $\mathbf{x}, \mathbf{y} \in \mathcal{C}^n$, $\mathbf{x} \neq \mathbf{0}$, explain why equality holds in the CBS inequality if and only if $\mathbf{y} = \alpha\mathbf{x}$, where $\alpha = \mathbf{x}^*\mathbf{y}/\mathbf{x}^*\mathbf{x}$. **Hint:** Use (5.1.4).

5.1.10. For nonzero vectors $\mathbf{x}, \mathbf{y} \in \mathcal{C}^n$ with the euclidean norm, prove that equality holds in the triangle inequality if and only if $\mathbf{y} = \alpha\mathbf{x}$, where α is real and positive. **Hint:** Make use of Exercise 5.1.9.

5.1.11. Use Hölder's inequality (5.1.8) to prove that if the components of $\mathbf{x} \in \Re^{n \times 1}$ sum to zero (i.e., $\mathbf{x}^T\mathbf{e} = 0$ for $\mathbf{e}^T = (1, 1, \ldots, 1)$), then

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_1 \left(\frac{y_{\max} - y_{\min}}{2} \right) \quad \text{for all } \mathbf{y} \in \Re^{n \times 1}.$$

Note: For "zero sum" vectors $\mathbf{x}$, this is at least as sharp and usually it's sharper than (5.1.8) because $(y_{\max} - y_{\min})/2 \leq \max_i |y_i| = \|\mathbf{y}\|_\infty$.

5.1.12. The classical form of ***Hölder's inequality***[36] states that if $p > 1$ and $q > 1$ are real numbers such that $1/p + 1/q = 1$, then

$$\sum_{i=1}^{n} |x_i y_i| \le \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q}.$$

Derive this inequality by executing the following steps:

(a) By considering the function $f(t) = (1-\lambda) + \lambda t - t^{\lambda}$ for $0 < \lambda < 1$, establish the inequality

$$\alpha^{\lambda}\beta^{1-\lambda} \le \lambda\alpha + (1-\lambda)\beta$$

for nonnegative real numbers α and β.

(b) Let $\hat{\mathbf{x}} = \mathbf{x}/\left\|\mathbf{x}\right\|_p$ and $\hat{\mathbf{y}} = \mathbf{y}/\left\|\mathbf{x}\right\|_q$, and apply the inequality of part (a) to obtain

$$\sum_{i=1}^{n} |\hat{x}_i \hat{y}_i| \le \frac{1}{p}\sum_{i=1}^{n} |\hat{x}_i|^p + \frac{1}{q}\sum_{i=1}^{n} |\hat{y}_i|^q = 1.$$

(c) Deduce the classical form of Hölder's inequality, and then explain why this means that

$$|\mathbf{x}^*\mathbf{y}| \le \left\|\mathbf{x}\right\|_p \left\|\mathbf{y}\right\|_q.$$

5.1.13. The triangle inequality $\left\|\mathbf{x}+\mathbf{y}\right\|_p \le \left\|\mathbf{x}\right\|_p + \left\|\mathbf{y}\right\|_p$ for a general p-norm is really the classical ***Minkowski inequality,***[37] which states that for $p \ge 1$,

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1/p} \le \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}.$$

Derive Minkowski's inequality. **Hint:** For $p > 1$, let q be the number such that $1/q = 1 - 1/p$. Verify that for scalars α and β,

$$|\alpha+\beta|^p = |\alpha+\beta|\,|\alpha+\beta|^{p/q} \le |\alpha|\,|\alpha+\beta|^{p/q} + |\beta|\,|\alpha+\beta|^{p/q},$$

and make use of Hölder's inequality in Exercise 5.1.12.

[36] Ludwig Otto Hölder (1859–1937) was a German mathematician who studied at Göttingen and lived in Leipzig. Although he made several contributions to analysis as well as algebra, he is primarily known for the development of the inequality that now bears his name.

[37] Hermann Minkowski (1864–1909) was born in Russia, but spent most of his life in Germany as a mathematician and professor at Königsberg and Göttingen. In addition to the inequality that now bears his name, he is known for providing a mathematical basis for the special theory of relativity. He died suddenly from a ruptured appendix at the age of 44.

5.2 MATRIX NORMS

Because $\mathcal{C}^{m\times n}$ is a vector space of dimension mn, magnitudes of matrices $\mathbf{A} \in \mathcal{C}^{m\times n}$ can be "measured" by employing any vector norm on $\mathcal{C}^{mn}$. For example, by stringing out the entries of $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ -4 & -2 \end{pmatrix}$ into a four-component vector, the euclidean norm on $\Re^4$ can be applied to write

$$\|\mathbf{A}\| = \left[2^2 + (-1)^2 + (-4)^2 + (-2)^2\right]^{1/2} = 5.$$

This is one of the simplest notions of a matrix norm, and it is called the *Frobenius* (p. 662) *norm* (older texts refer to it as the *Hilbert–Schmidt norm* or the *Schur norm*). There are several useful ways to describe the Frobenius matrix norm.

Frobenius Matrix Norm

The ***Frobenius norm*** of $\mathbf{A} \in \mathcal{C}^{m\times n}$ is defined by the equations

$$\|\mathbf{A}\|_F^2 = \sum_{i,j} |a_{ij}|^2 = \sum_i \|\mathbf{A}_{i*}\|_2^2 = \sum_j \|\mathbf{A}_{*j}\|_2^2 = trace\,(\mathbf{A}^*\mathbf{A}). \qquad (5.2.1)$$

The Frobenius matrix norm is fine for some problems, but it is not well suited for all applications. So, similar to the situation for vector norms, alternatives need to be explored. But before trying to develop different recipes for matrix norms, it makes sense to first formulate a general definition of a matrix norm. The goal is to start with the defining properties for a vector norm given in (5.1.9) on p. 275 and ask what, if anything, needs to be added to that list.

Matrix multiplication distinguishes matrix spaces from more general vector spaces, but the three vector-norm properties (5.1.9) say nothing about products. So, an extra property that relates $\|\mathbf{AB}\|$ to $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ is needed. The Frobenius norm suggests the nature of this extra property. The CBS inequality insures that $\|\mathbf{Ax}\|_2^2 = \sum_i |\mathbf{A}_{i*}\mathbf{x}|^2 \le \sum_i \|\mathbf{A}_{i*}\|_2^2 \|\mathbf{x}\|_2^2 = \|\mathbf{A}\|_F^2 \|\mathbf{x}\|_2^2$. That is,

$$\|\mathbf{Ax}\|_2 \le \|\mathbf{A}\|_F \|\mathbf{x}\|_2, \qquad (5.2.2)$$

and we express this by saying that the Frobenius matrix norm $\|\star\|_F$ and the euclidean vector norm $\|\star\|_2$ are ***compatible***. The compatibility condition (5.2.2) implies that for all conformable matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\begin{aligned}\|\mathbf{AB}\|_F^2 &= \sum_j \|[\mathbf{AB}]_{*j}\|_2^2 = \sum_j \|\mathbf{AB}_{*j}\|_2^2 \le \sum_j \|\mathbf{A}\|_F^2 \|\mathbf{B}_{*j}\|_2^2 \\ &= \|\mathbf{A}\|_F^2 \sum_j \|\mathbf{B}_{*j}\|_2^2 = \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 \quad \Longrightarrow \quad \|\mathbf{AB}\|_F \le \|\mathbf{A}\|_F \|\mathbf{B}\|_F.\end{aligned}$$

This suggests that the submultiplicative property $\|\mathbf{AB}\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|$ should be added to (5.1.9) to define a general matrix norm.

General Matrix Norms

A ***matrix norm*** is a function $\|\star\|$ from the set of all complex matrices (of all finite orders) into $\Re$ that satisfies the following properties.

$$\begin{aligned}
&\|\mathbf{A}\| \geq 0 \quad \text{and} \quad \|\mathbf{A}\| = 0 \Longleftrightarrow \mathbf{A} = \mathbf{0}.\\
&\|\alpha\mathbf{A}\| = |\alpha|\ \|\mathbf{A}\| \quad \text{for all scalars } \alpha.\\
&\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| \quad \text{for matrices of the same size.}\\
&\|\mathbf{AB}\| \leq \|\mathbf{A}\|\ \|\mathbf{B}\| \quad \text{for all conformable matrices.}
\end{aligned} \tag{5.2.3}$$

The Frobenius norm satisfies the above definition (it was built that way), but where do other useful matrix norms come from? In fact, every legitimate vector norm generates (or induces) a matrix norm as described below.

Induced Matrix Norms

A vector norm that is defined on $\mathcal{C}^p$ for $p = m, n$ ***induces*** a matrix norm on $\mathcal{C}^{m\times n}$ by setting

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\| \quad \text{for } \mathbf{A} \in \mathcal{C}^{m\times n},\ \ \mathbf{x} \in \mathcal{C}^{n\times 1}. \tag{5.2.4}$$

The footnote on p. 276 explains why this maximum value must exist.

- It's apparent that an induced matrix norm is compatible with its underlying vector norm in the sense that

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\|\,. \tag{5.2.5}$$

- When $\mathbf{A}$ is nonsingular, $\displaystyle\min_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\| = \frac{1}{\|\mathbf{A}^{-1}\|}.$ $\quad (5.2.6)$

Proof. Verifying that $\max_{\|\mathbf{x}\|=1}\|\mathbf{Ax}\|$ satisfies the first three conditions in (5.2.3) is straightforward, and (5.2.5) implies $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$ (see Exercise 5.2.5). Property (5.2.6) is developed in Exercise 5.2.7. ■

In words, an induced norm $\|\mathbf{A}\|$ represents the maximum extent to which a vector on the unit sphere can be stretched by $\mathbf{A}$, and $1/\left\|\mathbf{A}^{-1}\right\|$ measures the extent to which a nonsingular matrix $\mathbf{A}$ can shrink vectors on the unit sphere. Figure 5.2.1 depicts this in $\Re^3$ for the induced matrix 2-norm.

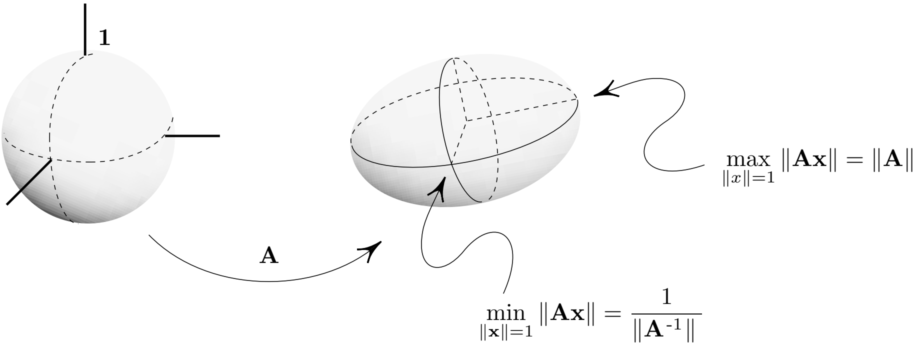

1
$\mathbf{A}$
$\max_{\|x\|=1} \|\mathbf{Ax}\| = \|\mathbf{A}\|$
$\min_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\| = \dfrac{1}{\|\mathbf{A}^{-1}\|}$

FIGURE 5.2.1. THE INDUCED MATRIX 2-NORM IN $\Re^3$.

Intuition might suggest that the euclidean vector norm should induce the Frobenius matrix norm (5.2.1), but something surprising happens instead.

Matrix 2-Norm

- The matrix norm induced by the euclidean vector norm is

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \sqrt{\lambda_{\max}}, \tag{5.2.7}$$

where $\lambda_{\max}$ is the largest number λ such that $\mathbf{A}^*\mathbf{A} - \lambda\mathbf{I}$ is singular.

- When $\mathbf{A}$ is nonsingular,

$$\|\mathbf{A}^{-1}\|_2 = \frac{1}{\min_{\|x\|_2=1} \|\mathbf{Ax}\|_2} = \frac{1}{\sqrt{\lambda_{\min}}}, \tag{5.2.8}$$

where $\lambda_{\max}$ is the smallest number λ such that $\mathbf{A}^*\mathbf{A} - \lambda\mathbf{I}$ is singular.

Note: If you are already familiar with eigenvalues, these say that $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $\mathbf{A}^*\mathbf{A}$ (Example 7.5.1, p. 549), while $(\lambda_{\max})^{1/2} = \sigma_1$ and $(\lambda_{\min})^{1/2} = \sigma_n$ are the largest and smallest singular values of $\mathbf{A}$ (p. 414).

Proof. To prove (5.2.7), assume that $\mathbf{A}_{m\times n}$ is real (a proof for complex matrices is given in Example 7.5.1 on p. 549). The strategy is to evaluate $\|\mathbf{A}\|_2^2$ by solving the problem

$$\text{maximize } f(\mathbf{x}) = \|\mathbf{Ax}\|_2^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} \quad \text{subject to } g(\mathbf{x}) = \mathbf{x}^T\mathbf{x} = 1$$

using the method of Lagrange multipliers. Introduce a new variable λ (the Lagrange multiplier), and consider the function $h(\mathbf{x},\lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$. The points at which f is maximized are contained in the set of solutions to the equations $\partial h/\partial x_i = 0$ $(i = 1, 2, \ldots, n)$ along with $g(\mathbf{x}) = 1$. Differentiating h with respect to the x_i's is essentially the same as described on p. 227, and the system generated by $\partial h/\partial x_i = 0$ $(i = 1, 2, \ldots, n)$ is $(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$. In other words, f is maximized at a vector $\mathbf{x}$ for which $(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ and $\|\mathbf{x}\|_2 = 1$. Consequently, λ must be a number such that $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ is singular (because $\mathbf{x} \neq \mathbf{0}$). Since

$$\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x} = \lambda,$$

it follows that

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = \max_{\|\mathbf{x}\|^2=1} \|\mathbf{A}\mathbf{x}\| = \left(\max_{\mathbf{x}^T\mathbf{x}=1} \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}\right)^{1/2} = \sqrt{\lambda_{\max}},$$

where $\lambda_{\max}$ is the largest number λ for which $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ is singular. A similar argument applied to (5.2.6) proves (5.2.8). Also, an independent development of (5.2.7) and (5.2.8) is contained in the discussion of singular values on p. 412. ■

Example 5.2.1

Problem: Determine the induced norm $\|\mathbf{A}\|_2$ as well as $\|\mathbf{A}^{-1}\|_2$ for the nonsingular matrix

$$\mathbf{A} = \frac{1}{\sqrt{3}}\begin{pmatrix} 3 & -1 \\ 0 & \sqrt{8} \end{pmatrix}.$$

Solution: Find the values of λ that make $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ singular by applying Gaussian elimination to produce

$$\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{pmatrix} \longrightarrow \begin{pmatrix} -1 & 3-\lambda \\ 3-\lambda & -1 \end{pmatrix} \longrightarrow \begin{pmatrix} -1 & 3-\lambda \\ 0 & -1+(3-\lambda)^2 \end{pmatrix}.$$

This shows that $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ is singular when $-1 + (3-\lambda)^2 = 0$ or, equivalently, when $\lambda = 2$ or $\lambda = 4$, so $\lambda_{\min} = 2$ and $\lambda_{\max} = 4$. Consequently, (5.2.7) and (5.2.8) say that

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}} = 2 \quad \text{and} \quad \|\mathbf{A}^{-1}\|_2 = \frac{1}{\sqrt{\lambda_{\min}}} = \frac{1}{\sqrt{2}}.$$

Note: As mentioned earlier, the values of λ that make $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ singular are called the *eigenvalues* of $\mathbf{A}^T\mathbf{A}$, and they are the focus of Chapter 7 where their determination is discussed in more detail. Using Gaussian elimination to determine the eigenvalues is not practical for larger matrices.

Some useful properties of the matrix 2-norm are stated below.

Properties of the 2-Norm

In addition to the properties shared by all induced norms, the 2-norm enjoys the following special properties.

- $\|\mathbf{A}\|_2 = \max\limits_{\|\mathbf{x}\|_2=1} \max\limits_{\|\mathbf{y}\|_2=1} |\mathbf{y}^*\mathbf{A}\mathbf{x}|.$ (5.2.9)
- $\|\mathbf{A}\|_2 = \|\mathbf{A}^*\|_2.$ (5.2.10)
- $\|\mathbf{A}^*\mathbf{A}\|_2 = \|\mathbf{A}\|_2^2.$ (5.2.11)
- $\left\|\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}\right\|_2 = \max\left\{ \|\mathbf{A}\|_2, \|\mathbf{B}\|_2 \right\}.$ (5.2.12)
- $\|\mathbf{U}^*\mathbf{A}\mathbf{V}\|_2 = \|\mathbf{A}\|_2$ when $\mathbf{U}\mathbf{U}^* = \mathbf{I}$ and $\mathbf{V}^*\mathbf{V} = \mathbf{I}.$ (5.2.13)

You are asked to verify the validity of these properties in Exercise 5.2.6 on p. 285. Furthermore, some additional properties of the matrix 2-norm are developed in Exercise 5.6.9 and on pp. 414 and 417.

Now that we understand how the euclidean vector norm induces the matrix 2-norm, let's investigate the nature of the matrix norms that are induced by the vector 1-norm and the vector ∞-norm.

Matrix 1-Norm and Matrix ∞-Norm

The matrix norms induced by the vector 1-norm and ∞-norm are as follows.

- $$\begin{aligned} \|\mathbf{A}\|_1 &= \max_{\|\mathbf{x}\|_1=1} \|\mathbf{A}\mathbf{x}\|_1 = \max_j \sum_i |a_{ij}| \\ &= \text{the largest absolute column sum.} \end{aligned} \qquad (5.2.14)$$
- $$\begin{aligned} \|\mathbf{A}\|_\infty &= \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_i \sum_j |a_{ij}| \\ &= \text{the largest absolute row sum.} \end{aligned} \qquad (5.2.15)$$

Proof of (5.2.14). For all $\mathbf{x}$ with $\|\mathbf{x}\|_1 = 1$, the scalar triangle inequality yields

$$\begin{aligned} \|\mathbf{A}\mathbf{x}\|_1 &= \sum_i |\mathbf{A}_{i*}\mathbf{x}| = \sum_i \Big|\sum_j a_{ij}x_j\Big| \le \sum_i \sum_j |a_{ij}|\ |x_j| = \sum_j \Big(|x_j| \sum_i |a_{ij}|\Big) \\ &\le \Big(\sum_j |x_j|\Big)\Big(\max_j \sum_i |a_{ij}|\Big) = \max_j \sum_i |a_{ij}|\,. \end{aligned}$$

Equality can be attained because if $\mathbf{A}_{*k}$ is the column with largest absolute sum, set $\mathbf{x} = \mathbf{e}_k$, and note that $\|\mathbf{e}_k\|_1 = 1$ and $\|\mathbf{A}\mathbf{e}_k\|_1 = \|\mathbf{A}_{*k}\|_1 = \max_j \sum_i |a_{ij}|$.

Proof of (5.2.15). For all $\mathbf{x}$ with $\|\mathbf{x}\|_\infty = 1$,

$$\|\mathbf{A}\mathbf{x}\|_\infty = \max_i \Big|\sum_j a_{ij}x_j\Big| \le \max_i \sum_j |a_{ij}|\ |x_j| \le \max_i \sum_j |a_{ij}|.$$

Equality can be attained because if $\mathbf{A}_{k*}$ is the row with largest absolute sum, and if $\mathbf{x}$ is the vector such that

$$x_j = \begin{cases} 1 & \text{if } a_{kj} \ge 0, \\ -1 & \text{if } a_{kj} < 0, \end{cases} \quad \text{then} \quad \begin{cases} |\mathbf{A}_{i*}\mathbf{x}| = |\sum_j a_{ij}x_j| \le \sum_j |a_{ij}| \text{ for all } i, \\ |\mathbf{A}_{k*}\mathbf{x}| = \sum_j |a_{kj}| = \max_i \sum_j |a_{ij}|, \end{cases}$$

so $\|\mathbf{x}\|_\infty = 1$, and $\|\mathbf{A}\mathbf{x}\|_\infty = \max_i |\mathbf{A}_{i*}\mathbf{x}| = \max_i \sum_j |a_{ij}|$. ■

Example 5.2.2

Problem: Determine the induced matrix norms $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_\infty$ for

$$\mathbf{A} = \frac{1}{\sqrt{3}}\begin{pmatrix} 3 & -1 \\ 0 & \sqrt{8} \end{pmatrix},$$

and compare the results with $\|\mathbf{A}\|_2$ (from Example 5.2.1) and $\|\mathbf{A}\|_F$.

Solution: Equation (5.2.14) says that $\|\mathbf{A}\|_1$ is the largest absolute column sum in $\mathbf{A}$, and (5.2.15) says that $\|\mathbf{A}\|_\infty$ is the largest absolute row sum, so

$$\|\mathbf{A}\|_1 = 1/\sqrt{3} + \sqrt{8}/\sqrt{3} \approx 2.21 \quad \text{and} \quad \|\mathbf{A}\|_\infty = 4/\sqrt{3} \approx 2.31.$$

Since $\|\mathbf{A}\|_2 = 2$ (Example 5.2.1) and $\|\mathbf{A}\|_F = \sqrt{\mathit{trace}\,(\mathbf{A}^T\mathbf{A})} = \sqrt{6} \approx 2.45$, we see that while $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_\infty$, and $\|\mathbf{A}\|_F$ are not equal, they are all in the same ballpark. This is true for all $n \times n$ matrices because it can be shown that $\|\mathbf{A}\|_i \le \alpha \|\mathbf{A}\|_j$, where α is the (i, j)-entry in the following matrix

$$\begin{array}{c} \\ 1 \\ 2 \\ \infty \\ F \end{array} \begin{array}{c} \begin{array}{cccc} 1 & \ \ 2 & \ \ \infty & \ \ F \end{array} \\ \begin{pmatrix} * & \sqrt{n} & n & \sqrt{n} \\ \sqrt{n} & * & \sqrt{n} & 1 \\ n & \sqrt{n} & * & \sqrt{n} \\ \sqrt{n} & \sqrt{n} & \sqrt{n} & * \end{pmatrix} \end{array}$$

(see Exercise 5.1.8 and Exercise 5.12.3 on p. 425). Since it's often the case that only the order of magnitude of $\|\mathbf{A}\|$ is needed and not the exact value (e.g., recall the rule of thumb in Example 3.8.2 on p. 129), and since $\|\mathbf{A}\|_2$ is difficult to compute in comparison with $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, and $\|\mathbf{A}\|_F$, you can see why any of these three might be preferred over $\|\mathbf{A}\|_2$ in spite of the fact that $\|\mathbf{A}\|_2$ is more "natural" by virtue of being induced by the euclidean vector norm.

Exercises for section 5.2

5.2.1. Evaluate the Frobenius matrix norm for each matrix below.

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

5.2.2. Evaluate the induced 1-, 2-, and ∞-matrix norm for each of the three matrices given in Exercise 5.2.1.

5.2.3. (a) Explain why $\|\mathbf{I}\| = 1$ for every induced matrix norm (5.2.4).
(b) What is $\|\mathbf{I}_{n\times n}\|_F$?

5.2.4. Explain why $\|\mathbf{A}\|_F = \|\mathbf{A}^*\|_F$ for Frobenius matrix norm (5.2.1).

5.2.5. For matrices $\mathbf{A}$ and $\mathbf{B}$ and for vectors $\mathbf{x}$, establish the following compatibility properties between a vector norm defined on every $\mathcal{C}^p$ and the associated induced matrix norm.

(a) Show that $\|\mathbf{Ax}\| \le \|\mathbf{A}\|\,\|\mathbf{x}\|$.
(b) Show that $\|\mathbf{AB}\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|$.
(c) Explain why $\|\mathbf{A}\| = \max_{\|\mathbf{x}\|\le 1} \|\mathbf{Ax}\|$.

5.2.6. Establish the following properties of the matrix 2-norm.

(a) $\|\mathbf{A}\|_2 = \max\limits_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} |\mathbf{y}^*\mathbf{Ax}|,$
(b) $\|\mathbf{A}\|_2 = \|\mathbf{A}^*\|_2,$
(c) $\|\mathbf{A}^*\mathbf{A}\|_2 = \|\mathbf{A}\|_2^2\,,$
(d) $\left\|\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}\right\|_2 = \max\left\{ \|\mathbf{A}\|_2\,, \|\mathbf{B}\|_2 \right\}$ (take $\mathbf{A}, \mathbf{B}$ to be real),
(e) $\|\mathbf{U}^*\mathbf{AV}\|_2 = \|\mathbf{A}\|_2$ when $\mathbf{UU}^* = \mathbf{I}$ and $\mathbf{V}^*\mathbf{V} = \mathbf{I}$.

5.2.7. Using the induced matrix norm (5.2.4), prove that if $\mathbf{A}$ is nonsingular, then

$$\|\mathbf{A}\| = \frac{1}{\min\limits_{\|\mathbf{x}\|=1} \|\mathbf{A}^{-1}\mathbf{x}\|} \quad \text{or, equivalently,} \quad \|\mathbf{A}^{-1}\| = \frac{1}{\min\limits_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|}.$$

5.2.8. For $\mathbf{A} \in \mathcal{C}^{n\times n}$ and a parameter $z \in \mathcal{C}$, the matrix $\mathbf{R}(z) = (z\mathbf{I} - \mathbf{A})^{-1}$ is called the ***resolvent of*** $\mathbf{A}$. Prove that if $|z| > \|\mathbf{A}\|$ for any induced matrix norm, then

$$\|\mathbf{R}(z)\| \le \frac{1}{|z| - \|\mathbf{A}\|}.$$

5.3 INNER-PRODUCT SPACES

The euclidean norm, which naturally came first, is a coordinate-dependent concept. But by isolating its important properties we quickly moved to the more general coordinate-free definition of a vector norm given in (5.1.9) on p. 275. The goal is to now do the same for inner products. That is, start with the standard inner product, which is a coordinate-dependent definition, and identify properties that characterize the basic essence of the concept. The ones listed below are those that have been distilled from the standard inner product to formulate a more general coordinate-free definition.

General Inner Product

An ***inner product*** on a real (or complex) vector space $\mathcal{V}$ is a function that maps each ordered pair of vectors $\mathbf{x}, \mathbf{y}$ to a real (or complex) scalar $\langle \mathbf{x}|\mathbf{y}\rangle$ such that the following four properties hold.

$$\begin{aligned}
&\langle \mathbf{x}|\mathbf{x}\rangle \text{ is real with } \langle \mathbf{x}|\mathbf{x}\rangle \geq 0, \text{ and } \langle \mathbf{x}|\mathbf{x}\rangle = 0 \text{ if and only if } \mathbf{x} = \mathbf{0},\\
&\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \alpha \langle \mathbf{x}|\mathbf{y}\rangle \text{ for all scalars } \alpha,\\
&\langle \mathbf{x}|\mathbf{y}+\mathbf{z}\rangle = \langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{x}|\mathbf{z}\rangle,\\
&\langle \mathbf{x}|\mathbf{y}\rangle = \overline{\langle \mathbf{y}|\mathbf{x}\rangle} \quad \text{(for real spaces, this becomes } \langle \mathbf{x}|\mathbf{y}\rangle = \langle \mathbf{y}|\mathbf{x}\rangle\text{)}.
\end{aligned} \tag{5.3.1}$$

Notice that for each fixed value of $\mathbf{x}$, the second and third properties say that $\langle \mathbf{x}|\mathbf{y}\rangle$ is a linear function of $\mathbf{y}$.

Any real or complex vector space that is equipped with an inner product is called an ***inner-product space***.

Example 5.3.1

- The standard inner products, $\langle \mathbf{x}|\mathbf{y}\rangle = \mathbf{x}^T\mathbf{y}$ for $\Re^{n\times 1}$ and $\langle \mathbf{x}|\mathbf{y}\rangle = \mathbf{x}^*\mathbf{y}$ for $\mathcal{C}^{n\times 1}$, each satisfy the four defining conditions (5.3.1) for a general inner product—this shouldn't be a surprise.

- If $\mathbf{A}_{n\times n}$ is a nonsingular matrix, then $\langle \mathbf{x}|\mathbf{y}\rangle = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{y}$ is an inner product for $\mathcal{C}^{n\times 1}$. This inner product is sometimes called an $\mathbf{A}$***-inner product*** or an ***elliptical inner product.***

- Consider the vector space of $m \times n$ matrices. The functions defined by

$$\langle \mathbf{A}|\mathbf{B}\rangle = trace\left(\mathbf{A}^T\mathbf{B}\right) \quad \text{and} \quad \langle \mathbf{A}|\mathbf{B}\rangle = trace\left(\mathbf{A}^*\mathbf{B}\right) \tag{5.3.2}$$

  are inner products for $\Re^{m\times n}$ and $\mathcal{C}^{m\times n}$, respectively. These are referred to as the ***standard inner products for matrices***. Notice that these reduce to the standard inner products for vectors when $n = 1$.

- If $\mathcal{V}$ is the vector space of real-valued continuous functions defined on the interval (a, b), then
$$\langle f|g\rangle = \int_a^b f(t)g(t)dt$$
is an inner product on $\mathcal{V}$.

Just as the standard inner product for $\mathcal{C}^{n\times 1}$ defines the euclidean norm on $\mathcal{C}^{n\times 1}$ by $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^*\mathbf{x}}$, every general inner product in an inner-product space $\mathcal{V}$ defines a norm on $\mathcal{V}$ by setting
$$\|\star\| = \sqrt{\langle\star|\star\rangle}. \tag{5.3.3}$$
It's straightforward to verify that this satisfies the first two conditions in (5.2.3) on p. 280 that define a general vector norm, but, just as in the case of euclidean norms, verifying that (5.3.3) satisfies the triangle inequality requires a generalized version of CBS inequality.

General CBS Inequality

If $\mathcal{V}$ is an inner-product space, and if we set $\|\star\| = \sqrt{\langle\star|\star\rangle}$, then
$$|\langle\mathbf{x}|\mathbf{y}\rangle| \leq \|\mathbf{x}\|\ \|\mathbf{y}\| \quad \text{for all } x, y \in \mathcal{V}. \tag{5.3.4}$$
Equality holds if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha = \langle\mathbf{x}|\mathbf{y}\rangle / \|\mathbf{x}\|^2$.

Proof. Set $\alpha = \langle\mathbf{x}|\mathbf{y}\rangle / \|\mathbf{x}\|^2$ (assume $\mathbf{x} \neq \mathbf{0}$, for otherwise there is nothing to prove), and observe that $\langle\mathbf{x}|\alpha\mathbf{x} - \mathbf{y}\rangle = 0$, so
$$\begin{aligned} 0 \leq \|\alpha\mathbf{x} - \mathbf{y}\|^2 &= \langle\alpha\mathbf{x} - \mathbf{y}|\alpha\mathbf{x} - \mathbf{y}\rangle \\ &= \bar{\alpha}\langle\mathbf{x}|\alpha\mathbf{x} - \mathbf{y}\rangle - \langle\mathbf{y}|\alpha\mathbf{x} - \mathbf{y}\rangle \quad \text{(see Exercise 5.3.2)} \\ &= -\langle\mathbf{y}|\alpha\mathbf{x} - \mathbf{y}\rangle = \langle\mathbf{y}|\mathbf{y}\rangle - \alpha\langle\mathbf{y}|\mathbf{x}\rangle = \frac{\|\mathbf{y}\|^2\|\mathbf{x}\|^2 - \langle\mathbf{x}|\mathbf{y}\rangle\langle\mathbf{y}|\mathbf{x}\rangle}{\|\mathbf{x}\|^2}. \end{aligned}$$
Since $\langle\mathbf{y}|\mathbf{x}\rangle = \overline{\langle\mathbf{x}|\mathbf{y}\rangle}$, it follows that $\langle\mathbf{x}|\mathbf{y}\rangle\langle\mathbf{y}|\mathbf{x}\rangle = |\langle\mathbf{x}|\mathbf{y}\rangle|^2$, so
$$0 \leq \frac{\|\mathbf{y}\|^2\|\mathbf{x}\|^2 - |\langle\mathbf{x}|\mathbf{y}\rangle|^2}{\|\mathbf{x}\|^2} \implies |\langle\mathbf{x}|\mathbf{y}\rangle| \leq \|\mathbf{x}\|\ \|\mathbf{y}\|.$$
Establishing the conditions for equality is the same as in Exercise 5.1.9. ■

Let's now complete the job of showing that $\|\star\| = \sqrt{\langle\star|\star\rangle}$ is indeed a vector norm as defined in (5.2.3) on p. 280.

Norms in Inner-Product Spaces

If $\mathcal{V}$ is an inner-product space with an inner product $\langle \mathbf{x}|\mathbf{y}\rangle$, then

$$\|\star\| = \sqrt{\langle \star|\star\rangle} \quad \text{defines a norm on } \mathcal{V}.$$

Proof. The fact that $\|\star\| = \sqrt{\langle \star|\star\rangle}$ satisfies the first two norm properties in (5.2.3) on p. 280 follows directly from the defining properties (5.3.1) for an inner product. You are asked to provide the details in Exercise 5.3.3. To establish the triangle inequality, use $\langle \mathbf{x}|\mathbf{y}\rangle \le |\langle \mathbf{x}|\mathbf{y}\rangle|$ and $\langle \mathbf{y}|\mathbf{x}\rangle = \overline{\langle \mathbf{x}|\mathbf{y}\rangle} \le |\langle \mathbf{x}|\mathbf{y}\rangle|$ together with the CBS inequality to write

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y}|\mathbf{x}+\mathbf{y}\rangle = \langle \mathbf{x}|\mathbf{x}\rangle + \langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{y}|\mathbf{x}\rangle + \langle \mathbf{y}|\mathbf{y}\rangle \\ &\le \|\mathbf{x}\|^2 + 2|\langle \mathbf{x}|\mathbf{y}\rangle| + \|\mathbf{y}\|^2 \le (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \quad \blacksquare\end{aligned}$$

Example 5.3.2

Problem: Describe the norms that are generated by the inner products presented in Example 5.3.1.

- Given a nonsingular matrix $\mathbf{A} \in \mathcal{C}^{n\times n}$, the $\mathbf{A}$***-norm*** (or ***elliptical norm***) generated by the $\mathbf{A}$-inner product on $\mathcal{C}^{n\times 1}$ is

$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\langle \mathbf{x}|\mathbf{x}\rangle} = \sqrt{\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}} = \|\mathbf{A}\mathbf{x}\|_2\,. \tag{5.3.5}$$

- The standard inner product for matrices generates the Frobenius matrix norm because

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}|\mathbf{A}\rangle} = \sqrt{\mathit{trace}\,(\mathbf{A}^*\mathbf{A})} = \|\mathbf{A}\|_F\,. \tag{5.3.6}$$

- For the space of real-valued continuous functions defined on (a,b), the norm of a function f generated by the inner product $\langle f|g\rangle = \int_a^b f(t)g(t)dt$ is

$$\|f\| = \sqrt{\langle f|f\rangle} = \left(\int_a^b f(t)^2 dt\right)^{1/2}.$$

Example 5.3.3

To illustrate the utility of the ideas presented above, consider the proposition

$$\mathit{trace}\left(\mathbf{A}^T\mathbf{B}\right)^2 \le \mathit{trace}\left(\mathbf{A}^T\mathbf{A}\right)\mathit{trace}\left(\mathbf{B}^T\mathbf{B}\right) \quad \text{for all } \mathbf{A},\mathbf{B}\in\Re^{m\times n}.$$

Problem: How would you know to formulate such a proposition and, second, how do you prove it?

Solution: The answer to both questions is the same. This is the CBS inequality in $\Re^{m\times n}$ equipped with the standard inner product $\langle\mathbf{A}|\mathbf{B}\rangle = \mathit{trace}\left(\mathbf{A}^T\mathbf{B}\right)$ and associated norm $\|\mathbf{A}\|_F = \sqrt{\langle\mathbf{A}|\mathbf{A}\rangle} = \sqrt{\mathit{trace}\left(\mathbf{A}^T\mathbf{A}\right)}$ because CBS says

$$\langle\mathbf{A}|\mathbf{B}\rangle^2 \le \|\mathbf{A}\|_F^2\,\|\mathbf{B}\|_F^2 \quad\Longrightarrow\quad \mathit{trace}\left(\mathbf{A}^T\mathbf{B}\right)^2 \le \mathit{trace}\left(\mathbf{A}^T\mathbf{A}\right)\mathit{trace}\left(\mathbf{B}^T\mathbf{B}\right).$$

The point here is that if your knowledge is limited to elementary matrix manipulations (which is all that is needed to understand the statement of the proposition), formulating the correct inequality might be quite a challenge to your intuition. And then proving the proposition using only elementary matrix manipulations would be a significant task—essentially, you would have to derive a version of CBS. But knowing the basic facts of inner-product spaces makes the proposition nearly trivial to conjecture and prove.

Since each inner product generates a norm by the rule $\|\star\| = \sqrt{\langle\star|\star\rangle}$, it's natural to ask if the reverse is also true. That is, for each vector norm $\|\star\|$ on a space $\mathcal{V}$, does there exist a corresponding inner product on $\mathcal{V}$ such that $\sqrt{\langle\star|\star\rangle} = \|\star\|^2$? If not, under what conditions will a given norm be generated by an inner product? These are tricky questions, and it took the combined efforts of Maurice R. Fréchet[38] (1878–1973) and John von Neumann (1903–1957) to provide the answer.

[38] Maurice René Fréchet began his illustrious career by writing an outstanding Ph.D. dissertation in 1906 under the direction of the famous French mathematician Jacques Hadamard (p. 469) in which the concepts of a metric space and compactness were first formulated. Fréchet developed into a versatile mathematical scientist, and he served as professor of mechanics at the University of Poitiers (1910–1919), professor of higher calculus at the University of Strasbourg (1920–1927), and professor of differential and integral calculus and professor of the calculus of probabilities at the University of Paris (1928–1948).

Born in Budapest, Hungary, John von Neumann was a child prodigy who could divide eight-digit numbers in his head when he was only six years old. Due to the political unrest in Europe, he came to America, where, in 1933, he became one of the six original professors of mathematics at the Institute for Advanced Study at Princeton University, a position he retained for the rest of his life. During his career, von Neumann's genius touched mathematics (pure and applied), chemistry, physics, economics, and computer science, and he is generally considered to be among the best scientists and mathematicians of the twentieth century.

Parallelogram Identity

For a given norm $\|\star\|$ on a vector space $\mathcal{V}$, there exists an inner product on $\mathcal{V}$ such that $\langle\star|\star\rangle = \|\star\|^2$ if and only if the ***parallelogram identity***

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\big(\,\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\,\big) \tag{5.3.7}$$

holds for all $\mathbf{x},\mathbf{y}\in\mathcal{V}$.

Proof. Consider real spaces—complex spaces are discussed in Exercise 5.3.6. If there exists an inner product such that $\langle\star|\star\rangle = \|\star\|^2$, then the parallelogram identity is immediate because $\langle\mathbf{x}+\mathbf{y}|\mathbf{x}+\mathbf{y}\rangle + \langle\mathbf{x}-\mathbf{y}|\mathbf{x}-\mathbf{y}\rangle = 2\langle\mathbf{x}|\mathbf{x}\rangle + 2\langle\mathbf{y}|\mathbf{y}\rangle$. The difficult part is establishing the converse. Suppose $\|\star\|$ satisfies the parallelogram identity, and prove that the function

$$\langle\mathbf{x}|\mathbf{y}\rangle = \frac{1}{4}\big(\,\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2\,\big) \tag{5.3.8}$$

is an inner product for $\mathcal{V}$ such that $\langle\mathbf{x}|\mathbf{x}\rangle = \|\mathbf{x}\|^2$ for all $\mathbf{x}$ by showing the four defining conditions (5.3.1) hold. The first and fourth conditions are immediate. To establish the third, use the parallelogram identity to write

$$\begin{aligned}
\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 &= \frac{1}{2}\big(\,\|\mathbf{x}+\mathbf{y}+\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}-\mathbf{z}\|^2\,\big),\\
\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{z}\|^2 &= \frac{1}{2}\big(\,\|\mathbf{x}-\mathbf{y}+\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{z}-\mathbf{y}\|^2\,\big),
\end{aligned}$$

and then subtract to obtain

$$\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2 = \frac{\|2\mathbf{x}+(\mathbf{y}+\mathbf{z})\|^2 - \|2\mathbf{x}-(\mathbf{y}+\mathbf{z})\|^2}{2}.$$

Consequently,

$$\begin{aligned}
\langle\mathbf{x}|\mathbf{y}\rangle + \langle\mathbf{x}|\mathbf{z}\rangle &= \frac{1}{4}\big(\,\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2\,\big)\\
&= \frac{1}{8}\big(\,\|2\mathbf{x}+(\mathbf{y}+\mathbf{z})\|^2 - \|2\mathbf{x}-(\mathbf{y}+\mathbf{z})\|^2\,\big)\\
&= \frac{1}{2}\left(\left\|\mathbf{x}+\frac{\mathbf{y}+\mathbf{z}}{2}\right\|^2 - \left\|\mathbf{x}-\frac{\mathbf{y}+\mathbf{z}}{2}\right\|^2\right) = 2\left\langle\mathbf{x}\middle|\frac{\mathbf{y}+\mathbf{z}}{2}\right\rangle,
\end{aligned} \tag{5.3.9}$$

and setting $\mathbf{z}=\mathbf{0}$ produces the statement that $\langle\mathbf{x}|\mathbf{y}\rangle = 2\langle\mathbf{x}|\mathbf{y}/2\rangle$ for all $\mathbf{y}\in\mathcal{V}$. Replacing $\mathbf{y}$ by $\mathbf{y}+\mathbf{z}$ yields $\langle\mathbf{x}|\mathbf{y}+\mathbf{z}\rangle = 2\langle\mathbf{x}|(\mathbf{y}+\mathbf{z})/2\rangle$, and thus (5.3.9) guarantees that $\langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{x}|\mathbf{z}\rangle = \langle \mathbf{x}|\mathbf{y}+\mathbf{z}\rangle$. Now prove that $\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \alpha\langle \mathbf{x}|\mathbf{y}\rangle$ for all real α. This is valid for integer values of α by the result just established, and it holds when α is rational because if β and γ are integers, then

$$\gamma^2\left\langle \mathbf{x}\middle|\frac{\beta}{\gamma}\mathbf{y}\right\rangle = \langle \gamma\mathbf{x}|\beta\mathbf{y}\rangle = \beta\gamma\langle \mathbf{x}|\mathbf{y}\rangle \quad\Longrightarrow\quad \left\langle \mathbf{x}\middle|\frac{\beta}{\gamma}\mathbf{y}\right\rangle = \frac{\beta}{\gamma}\langle \mathbf{x}|\mathbf{y}\rangle.$$

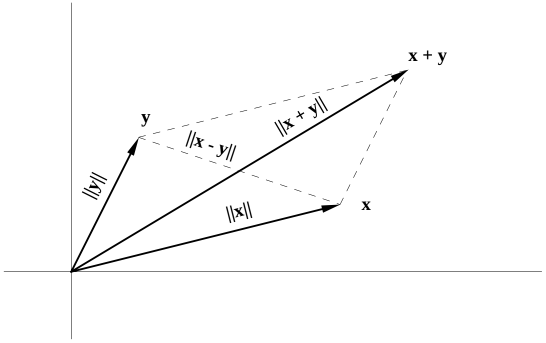
Because $\|\mathbf{x}+\alpha\mathbf{y}\|$ and $\|\mathbf{x}-\alpha\mathbf{y}\|$ are continuous functions of α (Exercise 5.1.7), equation (5.3.8) insures that $\langle \mathbf{x}|\alpha\mathbf{y}\rangle$ is a continuous function of α. Therefore, if α is irrational, and if $\{\alpha_n\}$ is a sequence of rational numbers such that $\alpha_n \to \alpha$, then $\langle \mathbf{x}|\alpha_n\mathbf{y}\rangle \to \langle \mathbf{x}|\alpha\mathbf{y}\rangle$ and $\langle \mathbf{x}|\alpha_n\mathbf{y}\rangle = \alpha_n\langle \mathbf{x}|\mathbf{y}\rangle \to \alpha\langle \mathbf{x}|\mathbf{y}\rangle$, so $\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \alpha\langle \mathbf{x}|\mathbf{y}\rangle$. ∎

Example 5.3.4

We already know that the euclidean vector norm on $\mathcal{C}^n$ is generated by the standard inner product, so the previous theorem guarantees that the parallelogram identity must hold for the 2-norm. This is easily corroborated by observing that

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_2^2 + \|\mathbf{x}-\mathbf{y}\|_2^2 &= (\mathbf{x}+\mathbf{y})^*(\mathbf{x}+\mathbf{y}) + (\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y})\\ &= 2\left(\mathbf{x}^*\mathbf{x}+\mathbf{y}^*\mathbf{y}\right) = \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2.\end{aligned}$$

The parallelogram identity is so named because it expresses the fact that the sum of the squares of the diagonals in a parallelogram is twice the sum of the squares of the sides. See the following diagram.



Example 5.3.5

Problem: Except for the euclidean norm, is any other vector p-norm generated by an inner product?

Solution: No, because the parallelogram identity (5.3.7) doesn't hold when $p \neq 2$. To see that $\|\mathbf{x}+\mathbf{y}\|_p^2 + \|\mathbf{x}-\mathbf{y}\|_p^2 = 2\big(\|\mathbf{x}\|_p^2 + \|\mathbf{y}\|_p^2\big)$ is not valid for all $\mathbf{x},\mathbf{y} \in \mathcal{C}^n$ when $p \neq 2$, consider $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{y} = \mathbf{e}_2$. It's apparent that $\|\mathbf{e_1}+\mathbf{e_2}\|_p^2 = 2^{2/p} = \|\mathbf{e_1}-\mathbf{e_2}\|_p^2$, so

$$\|\mathbf{e_1}+\mathbf{e_2}\|_p^2 + \|\mathbf{e_1}-\mathbf{e_2}\|_p^2 = 2^{(p+2)/p} \quad\text{and}\quad 2\big(\|\mathbf{e_1}\|_p^2 + \|\mathbf{e_2}\|_p^2\big) = 4.$$

Clearly, $2^{(p+2)/p} = 4$ only when $p = 2$. Details for the ∞-norm are asked for in Exercise 5.3.7.

Conclusion: For applications that are best analyzed in the context of an inner-product space (e.g., least squares problems), we are limited to the euclidean norm or else to one of its variation such as the elliptical norm in (5.3.5).

Virtually all important statements concerning $\Re^n$ or $\mathcal{C}^n$ with the standard inner product remain valid for general inner-product spaces—e.g., consider the statement and proof of the general CBS inequality. Advanced or more theoretical texts prefer a development in terms of general inner-product spaces. However, the focus of this text is matrices and the coordinate spaces $\Re^n$ and $\mathcal{C}^n$, so subsequent discussions will usually be phrased in terms of $\Re^n$ or $\mathcal{C}^n$ and their standard inner products. But remember that extensions to more general inner-product spaces are always lurking in the background, and we will not hesitate to use these generalities or general inner-product notation when they serve our purpose.

Exercises for section 5.3

5.3.1. For $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$, determine which of the following are inner products for $\Re^{3\times 1}$.

(a) $\langle \mathbf{x}|\mathbf{y}\rangle = x_1y_1 + x_3y_3$,
(b) $\langle \mathbf{x}|\mathbf{y}\rangle = x_1y_1 - x_2y_2 + x_3y_3$,
(c) $\langle \mathbf{x}|\mathbf{y}\rangle = 2x_1y_1 + x_2y_2 + 4x_3y_3$,
(d) $\langle \mathbf{x}|\mathbf{y}\rangle = x_1^2y_1^2 + x_2^2y_2^2 + x_3^2y_3^2$.

5.3.2. For a general inner-product space $\mathcal{V}$, explain why each of the following statements must be true.

(a) If $\langle \mathbf{x}|\mathbf{y}\rangle = 0$ for all $\mathbf{x} \in \mathcal{V}$, then $\mathbf{y} = \mathbf{0}$.
(b) $\langle \alpha\mathbf{x}|\mathbf{y}\rangle = \overline{\alpha}\,\langle \mathbf{x}|\mathbf{y}\rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ and for all scalars α.
(c) $\langle \mathbf{x}+\mathbf{y}|\mathbf{z}\rangle = \langle \mathbf{x}|\mathbf{z}\rangle + \langle \mathbf{y}|\mathbf{z}\rangle$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$.

5.3.3. Let $\mathcal{V}$ be an inner-product space with an inner product $\langle \mathbf{x}|\mathbf{y}\rangle$. Explain why the function defined by $\|\star\| = \sqrt{\langle\star|\star\rangle}$ satisfies the first two norm properties in (5.2.3) on p. 280.

5.3.4. For a real inner-product space with $\|\star\|^2 = \langle\star|\star\rangle$, derive the inequality

$$\langle \mathbf{x}|\mathbf{y}\rangle \le \frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}{2}. \qquad \textbf{Hint: } \text{Consider } \mathbf{x}-\mathbf{y}.$$

5.3.5. For $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, explain why each of the following inequalities is valid.

(a) $|trace\,(\mathbf{B})|^2 \leq n\,[trace\,(\mathbf{B}^*\mathbf{B})]\,.$

(b) $trace\left(\mathbf{B}^2\right) \leq trace\left(\mathbf{B}^T\mathbf{B}\right)$ for real matrices.

(c) $trace\left(\mathbf{A}^T\mathbf{B}\right) \leq \dfrac{trace\left(\mathbf{A}^T\mathbf{A}\right) + trace\left(\mathbf{B}^T\mathbf{B}\right)}{2}$ for real matrices.

5.3.6. Extend the proof given on p. 290 concerning the parallelogram identity (5.3.7) to include complex spaces. **Hint:** If $\mathcal{V}$ is a complex space with a norm $\|\star\|$ that satisfies the parallelogram identity, let

$$\langle \mathbf{x}|\mathbf{y}\rangle_r = \frac{\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2}{4},$$

and prove that

$$\langle \mathbf{x}|\mathbf{y}\rangle = \langle \mathbf{x}|\mathbf{y}\rangle_r + \mathrm{i}\,\langle \mathrm{i}\mathbf{x}|\mathbf{y}\rangle_r \quad \text{(the } \textbf{\textit{polarization identity}}\text{)} \qquad (5.3.10)$$

is an inner product on $\mathcal{V}$.

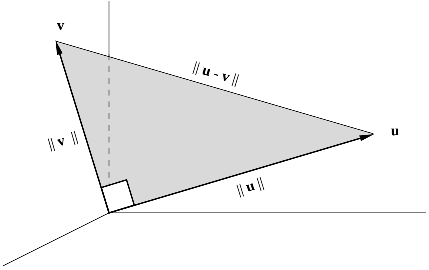
5.3.7. Explain why there does not exist an inner product on $\mathcal{C}^n$ $(n \geq 2)$ such that $\|\star\|_\infty = \sqrt{\langle \star|\star\rangle}$.

5.3.8. Explain why the Frobenius matrix norm on $\mathcal{C}^{n\times n}$ must satisfy the parallelogram identity.

5.3.9. For $n \geq 2$, is either the matrix 1-, 2-, or ∞-norm generated by an inner product on $\mathcal{C}^{n\times n}$?

5.4 ORTHOGONAL VECTORS

Two vectors in $\Re^3$ are ***orthogonal*** (perpendicular) if the angle between them is a right angle (90°). But the visual concept of a right angle is not at our disposal in higher dimensions, so we must dig a little deeper. The essence of perpendicularity in $\Re^2$ and $\Re^3$ is embodied in the classical Pythagorean theorem,



which says that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if $\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \|\mathbf{u}-\mathbf{v}\|^2$. But[39] $\|\mathbf{u}\|^2 = \mathbf{u}^T\mathbf{u}$ for all $\mathbf{u} \in \Re^3$, and $\mathbf{u}^T\mathbf{v} = \mathbf{v}^T\mathbf{u}$, so we can rewrite the Pythagorean statement as

$$\begin{aligned}\mathbf{0} &= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2 = \mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} - (\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v}) \\ &= \mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} - \left(\mathbf{u}^T\mathbf{u} - \mathbf{u}^T\mathbf{v} - \mathbf{v}^T\mathbf{u} + \mathbf{v}^T\mathbf{v}\right) = 2\mathbf{u}^T\mathbf{v}.\end{aligned}$$

Therefore, $\mathbf{u}$ and $\mathbf{v}$ are orthogonal vectors in $\Re^3$ if and only if $\mathbf{u}^T\mathbf{v} = 0$. The natural extension of this provides us with a definition in more general spaces.

Orthogonality

In an inner-product space $\mathcal{V}$, two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ are said to be ***orthogonal*** (to each other) whenever $\langle \mathbf{x}|\mathbf{y}\rangle = 0$, and this is denoted by writing $\mathbf{x} \perp \mathbf{y}$.

- For $\Re^n$ with the standard inner product, $\mathbf{x} \perp \mathbf{y} \iff \mathbf{x}^T\mathbf{y} = 0$.
- For $\mathcal{C}^n$ with the standard inner product, $\mathbf{x} \perp \mathbf{y} \iff \mathbf{x}^*\mathbf{y} = 0$.

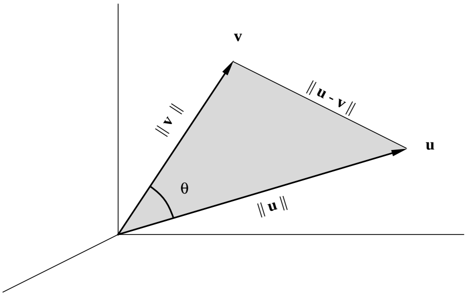
Example 5.4.1

$$\mathbf{x} = \begin{pmatrix} 1 \\ -2 \\ 3 \\ -1 \end{pmatrix} \text{ is orthogonal to } \mathbf{y} = \begin{pmatrix} 4 \\ 1 \\ -2 \\ -4 \end{pmatrix} \text{ because } \mathbf{x}^T\mathbf{y} = 0.$$

[39] Throughout this section, only norms generated by an underlying inner product $\|\star\|^2 = \langle\star|\star\rangle$ are used, so distinguishing subscripts on the norm notation can be omitted.

In spite of the fact that $\mathbf{u}^T\mathbf{v} = 0$, the vectors $\mathbf{u} = \begin{pmatrix} \mathrm{i} \\ 3 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} \mathrm{i} \\ 0 \\ 1 \end{pmatrix}$ are *not* orthogonal because $\mathbf{u}^*\mathbf{v} \neq 0$.

Now that "right angles" in higher dimensions make sense, how can more general angles be defined? Proceed just as before, but use the law of cosines rather than the Pythagorean theorem. Recall that



the ***law of cosines*** in $\Re^2$ or $\Re^3$ says $\|\mathbf{u}-\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\,\|\mathbf{u}\|\;\|\mathbf{v}\|\cos\theta$. If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, then this reduces to the Pythagorean theorem. But, in general,

$$\cos\theta = \frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2}{2\,\|\mathbf{u}\|\;\|\mathbf{v}\|} = \frac{\mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} - (\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})}{2\,\|\mathbf{u}\|\;\|\mathbf{v}\|}$$

$$= \frac{2\mathbf{u}^T\mathbf{v}}{2\,\|\mathbf{u}\|\;\|\mathbf{v}\|} = \frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|\;\|\mathbf{v}\|}.$$

This easily extends to higher dimensions because if $\mathbf{x}$, $\mathbf{y}$ are vectors from any real inner-product space, then the general CBS inequality (5.3.4) on p. 287 guarantees that $\langle\mathbf{x}|\mathbf{y}\rangle \,/\, \|\mathbf{x}\|\;\|\mathbf{y}\|$ is a number in the interval $[-1,\,1]$, and hence there is a unique value θ in $[0,\,\pi]$ such that $\cos\theta = \langle\mathbf{x}|\mathbf{y}\rangle \,/\, \|\mathbf{x}\|\;\|\mathbf{y}\|$.

Angles

In a real inner-product space $\mathcal{V}$, the radian measure of the ***angle*** between nonzero vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ is defined to be the number $\theta \in [0,\,\pi]$ such that

$$\cos\theta = \frac{\langle\mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|\;\|\mathbf{y}\|}. \qquad (5.4.1)$$

Example 5.4.2

In $\Re^n$, $\cos\theta = \mathbf{x}^T\mathbf{y}/\|\mathbf{x}\|\,\|\mathbf{y}\|$. For example, to determine the angle between $\mathbf{x} = \begin{pmatrix} -4 \\ 2 \\ 1 \\ 2 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \end{pmatrix}$, compute $\cos\theta = 2/(5)(3) = 2/15$, and use the inverse cosine function to conclude that $\theta = 1.437$ radians (rounded).

Example 5.4.3

Linear Correlation. Suppose that an experiment is conducted, and the resulting observations are recorded in two data vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \text{and let } \mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Problem: Determine to what extent the y_i 's are linearly related to the x_i 's. That is, measure how close $\mathbf{y}$ is to being a linear combination $\beta_0\mathbf{e} + \beta_1\mathbf{x}$.

Solution: The cosine as defined in (5.4.1) does the job. To understand how, let $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ be the ***mean*** and ***standard deviation*** of the data in $\mathbf{x}$. That is,

$$\mu_{\mathbf{x}} = \frac{\sum_i x_i}{n} = \frac{\mathbf{e}^T\mathbf{x}}{n} \quad \text{and} \quad \sigma_{\mathbf{x}} = \sqrt{\frac{\sum_i (x_i - \mu_{\mathbf{x}})^2}{n}} = \frac{\|\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e}\|_2}{\sqrt{n}}.$$

The mean is a measure of central tendency, and the standard deviation measures the extent to which the data is spread. Frequently, raw data from different sources is difficult to compare because the units of measure are different—e.g., one researcher may use the metric system while another uses American units. To compensate, data is almost always first "standardized" into unitless quantities. The ***standardization*** of a vector $\mathbf{x}$ for which $\sigma_{\mathbf{x}} \neq 0$ is defined to be

$$\mathbf{z}_{\mathbf{x}} = \frac{\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e}}{\sigma_{\mathbf{x}}}.$$

Entries in $\mathbf{z}_{\mathbf{x}}$ are often referred to as ***standard scores*** or ***z-scores.*** All standardized vectors have the properties that $\|\mathbf{z}\| = \sqrt{n}$, $\mu_{\mathbf{z}} = 0$, and $\sigma_{\mathbf{z}} = 1$. Furthermore, it's not difficult to verify that for vectors $\mathbf{x}$ and $\mathbf{y}$ such that $\sigma_{\mathbf{x}} \neq 0$ and $\sigma_{\mathbf{y}} \neq 0$, it's the case that

$$\begin{aligned} \mathbf{z}_{\mathbf{x}} = \mathbf{z}_{\mathbf{y}} &\Longleftrightarrow \exists \text{ constants } \beta_0, \beta_1 \text{ such that } \mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}, \quad \text{where} \quad \beta_1 > 0, \\ \mathbf{z}_{\mathbf{x}} = -\mathbf{z}_{\mathbf{y}} &\Longleftrightarrow \exists \text{ constants } \beta_0, \beta_1 \text{ such that } \mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}, \quad \text{where} \quad \beta_1 < 0. \end{aligned}$$

- In other words, $\mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}$ for some β_0 and β_1 if and only if $\mathbf{z}_{\mathbf{x}} = \pm\mathbf{z}_{\mathbf{y}}$, in which case we say $\mathbf{y}$ is ***perfectly linearly correlated*** with $\mathbf{x}$.

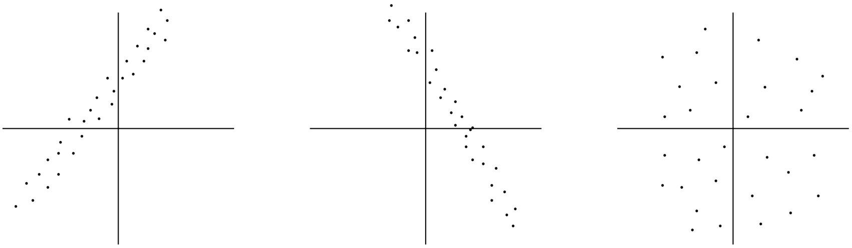
Since $\mathbf{z_x}$ varies continuously with $\mathbf{x}$, the existence of a “near” linear relationship between $\mathbf{x}$ and $\mathbf{y}$ is equivalent to $\mathbf{z_x}$ being “close” to $\pm\mathbf{z_y}$ in some sense. The fact that $\|\mathbf{z_x}\| = \|\pm\mathbf{z_y}\| = \sqrt{n}$ means $\mathbf{z_x}$ and $\pm\mathbf{z_y}$ differ only in orientation, so a natural measure of how close $\mathbf{z_x}$ is to $\pm\mathbf{z_y}$ is $\cos\theta$, where θ is the angle between $\mathbf{z_x}$ and $\mathbf{z_y}$. The number

$$\rho_{\mathbf{xy}} = \cos\theta = \frac{\mathbf{z_x}^T\mathbf{z_y}}{\|\mathbf{z_x}\|\;\|\mathbf{z_y}\|} = \frac{\mathbf{z_x}^T\mathbf{z_y}}{n} = \frac{(\mathbf{x}-\mu_{\mathbf{x}}\mathbf{e})^T(\mathbf{y}-\mu_{\mathbf{y}}\mathbf{e})}{\|\mathbf{x}-\mu_{\mathbf{x}}\mathbf{e}\|\;\|\mathbf{y}-\mu_{\mathbf{y}}\mathbf{e}\|}$$

is called the ***coefficient of linear correlation,*** and the following facts are now immediate.

- $\rho_{\mathbf{xy}} = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, in which case we say that $\mathbf{x}$ and $\mathbf{y}$ are ***completely uncorrelated.***
- $|\rho_{\mathbf{xy}}| = 1$ if and only if $\mathbf{y}$ is *perfectly* correlated with $\mathbf{x}$. That is, $|\rho_{\mathbf{xy}}| = 1$ if and only if there exists a linear relationship $\mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}$.
  - ▷ When $\beta_1 > 0$, we say that $\mathbf{y}$ is ***positively correlated*** with $\mathbf{x}$.
  - ▷ When $\beta_1 < 0$, we say that $\mathbf{y}$ is ***negatively correlated*** with $\mathbf{x}$.
- $|\rho_{\mathbf{xy}}|$ measures the degree to which $\mathbf{y}$ is linearly related to $\mathbf{x}$. In other words, $|\rho_{\mathbf{xy}}| \approx 1$ if and only if $\mathbf{y} \approx \beta_0\mathbf{e} + \beta_1\mathbf{x}$ for some β_0 and β_1.
  - ▷ Positive correlation is measured by the degree to which $\rho_{\mathbf{xy}} \approx 1$.
  - ▷ Negative correlation is measured by the degree to which $\rho_{\mathbf{xy}} \approx -1$.

If the data in $\mathbf{x}$ and $\mathbf{y}$ are plotted in $\Re^2$ as points (x_i, y_i), then, as depicted in Figure 5.4.1, $\rho_{\mathbf{xy}} \approx 1$ means that the points lie near a straight line with positive slope, while $\rho_{\mathbf{xy}} \approx -1$ means that the points lie near a line with negative slope, and $\rho_{\mathbf{xy}} \approx 0$ means that the points do not lie near a straight line.



$\rho_{\mathbf{xy}} \approx 1$ Positive Correlation

$\rho_{\mathbf{xy}} \approx -1$ Negative Correlation

$\rho_{\mathbf{xy}} \approx 0$ No Correlation

Figure 5.4.1

If $|\rho_{\mathbf{xy}}| \approx 1$, then the theory of least squares as presented in §4.6 can be used to determine a “best-fitting” straight line.

Orthonormal Sets

$\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is called an ***orthonormal set*** whenever $\|\mathbf{u}_i\| = 1$ for each i, and $\mathbf{u}_i \perp \mathbf{u}_j$ for all $i \neq j$. In other words,

$$\langle \mathbf{u}_i | \mathbf{u}_j \rangle = \begin{cases} 1 & \text{when } i = j, \\ 0 & \text{when } i \neq j. \end{cases}$$

- Every orthonormal set is linearly independent. (5.4.2)
- Every orthonormal set of n vectors from an n-dimensional space $\mathcal{V}$ is an orthonormal basis for $\mathcal{V}$.

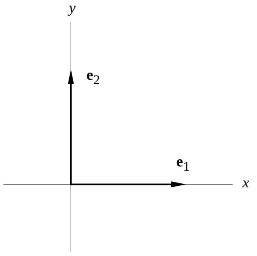
Proof. The second point follows from the first. To prove the first statement, suppose $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is orthonormal. If $\mathbf{0} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n$, use the properties of an inner product to write

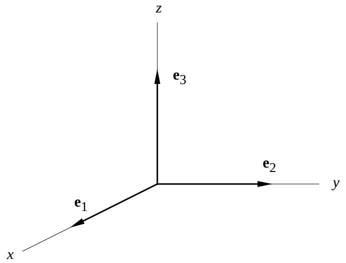
$$\begin{aligned} 0 &= \langle \mathbf{u}_i | \mathbf{0} \rangle = \langle \mathbf{u}_i | \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n \rangle \\ &= \alpha_1 \langle \mathbf{u}_i | \mathbf{u}_1 \rangle + \cdots + \alpha_i \langle \mathbf{u}_i | \mathbf{u}_i \rangle + \cdots + \alpha_n \langle \mathbf{u}_i | \mathbf{u}_n \rangle = \alpha_i \|\mathbf{u}_i\|^2 \\ &= \alpha_i \quad \text{for each } i. \ \blacksquare \end{aligned}$$

Example 5.4.4

The set $\mathcal{B}' = \left\{ \mathbf{u}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \right\}$ is a set of mutually orthogonal vectors because $\mathbf{u}_i^T\mathbf{u}_j = 0$ for $i \neq j$, but $\mathcal{B}'$ is *not* an orthonormal set—each vector does not have unit length. However, it's easy to convert an orthogonal set (not containing a zero vector) into an orthonormal set by simply normalizing each vector. Since $\|\mathbf{u}_1\| = \sqrt{2}$, $\|\mathbf{u}_2\| = \sqrt{3}$, and $\|\mathbf{u}_3\| = \sqrt{6}$, it follows that $\mathcal{B} = \left\{\mathbf{u}_1/\sqrt{2},\ \mathbf{u}_2/\sqrt{3},\ \mathbf{u}_3/\sqrt{6}\right\}$ is orthonormal.

The most common orthonormal basis is $\mathcal{S} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, the standard basis for $\Re^n$ and $\mathcal{C}^n$, and, as illustrated below for $\Re^2$ and $\Re^3$, these orthonormal vectors are directed along the standard coordinate axes.





Another orthonormal basis $\mathcal{B}$ need not be directed in the same way as $\mathcal{S}$, but that's the only significant difference because it's geometrically evident that $\mathcal{B}$ must amount to some rotation of $\mathcal{S}$. Consequently, we should expect general orthonormal bases to provide essentially the same advantages as the standard basis. For example, an important function of the standard basis $\mathcal{S}$ for $\Re^n$ is to provide coordinate representations by writing

$$\mathbf{x} = [\mathbf{x}]_{\mathcal{S}} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{to mean} \quad \mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n.$$

With respect to a general basis $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$, the coordinates of $\mathbf{x}$ are the scalars ξ_i in the representation $\mathbf{x} = \xi_1\mathbf{u}_1 + \xi_2\mathbf{u}_2 + \cdots + \xi_n\mathbf{u}_n$, and, as illustrated in Example 4.7.2, finding the ξ_i 's requires solving an $n \times n$ system, a nuisance we would like to avoid. But if $\mathcal{B}$ is an *orthonormal* basis, then the ξ_i 's are readily available because $\langle \mathbf{u}_i | \mathbf{x} \rangle = \langle \mathbf{u}_i | \xi_1\mathbf{u}_1 + \xi_2\mathbf{u}_2 + \cdots + \xi_n\mathbf{u}_n \rangle = \sum_{j=1}^{n} \xi_j \langle \mathbf{u}_i | \mathbf{u}_j \rangle = \xi_i \left\| \mathbf{u}_i \right\|^2 = \xi_i$. This yields the ***Fourier***[40] ***expansion*** of $\mathbf{x}$.

Fourier Expansions

If $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ is an orthonormal basis for an inner-product space $\mathcal{V}$, then each $\mathbf{x} \in \mathcal{V}$ can be expressed as

$$\mathbf{x} = \langle \mathbf{u}_1 | \mathbf{x} \rangle \mathbf{u}_1 + \langle \mathbf{u}_2 | \mathbf{x} \rangle \mathbf{u}_2 + \cdots + \langle \mathbf{u}_n | \mathbf{x} \rangle \mathbf{u}_n. \tag{5.4.3}$$

This is called the ***Fourier expansion*** of $\mathbf{x}$. The scalars $\xi_i = \langle \mathbf{u}_i | \mathbf{x} \rangle$ are the coordinates of $\mathbf{x}$ with respect to $\mathcal{B}$, and they are called the ***Fourier coefficients.*** Geometrically, the Fourier expansion resolves $\mathbf{x}$ into n mutually orthogonal vectors $\langle \mathbf{u}_i | \mathbf{x} \rangle \mathbf{u}_i$, each of which represents the orthogonal projection of $\mathbf{x}$ onto the space (line) spanned by $\mathbf{u}_i$. (More is said in Example 5.13.1 on p. 431 and Exercise 5.13.11.)

[40] Jean Baptiste Joseph Fourier (1768–1830) was a French mathematician and physicist who, while studying heat flow, developed expansions similar to (5.4.3). Fourier's work dealt with special infinite-dimensional inner-product spaces involving trigonometric functions as discussed in Example 5.4.6. Although they were apparently used earlier by Daniel Bernoulli (1700–1782) to solve problems concerned with vibrating strings, these orthogonal expansions became known as ***Fourier series,*** and they are now a fundamental tool in applied mathematics. Born the son of a tailor, Fourier was orphaned at the age of eight. Although he showed a great aptitude for mathematics at an early age, he was denied his dream of entering the French artillery because of his "low birth." Instead, he trained for the priesthood, but he never took his vows. However, his talents did not go unrecognized, and he later became a favorite of Napoleon. Fourier's work is now considered as marking an epoch in the history of both pure and applied mathematics. The next time you are in Paris, check out Fourier's plaque on the first level of the Eiffel Tower.

Example 5.4.5

Problem: Determine the Fourier expansion of $\mathbf{x} = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$ with respect to the standard inner product and the orthonormal basis given in Example 5.4.4

$$\mathcal{B} = \left\{ \mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \mathbf{u}_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \right\}.$$

Solution: The Fourier coefficients are

$$\xi_1 = \langle \mathbf{u}_1 | \mathbf{x} \rangle = \frac{-3}{\sqrt{2}}, \quad \xi_2 = \langle \mathbf{u}_2 | \mathbf{x} \rangle = \frac{2}{\sqrt{3}}, \quad \xi_3 = \langle \mathbf{u}_3 | \mathbf{x} \rangle = \frac{1}{\sqrt{6}},$$

so

$$\mathbf{x} = \xi_1 \mathbf{u}_1 + \xi_2 \mathbf{u}_2 + \xi_3 \mathbf{u}_3 = \frac{1}{2} \begin{pmatrix} -3 \\ 3 \\ 0 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} + \frac{1}{6} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

You may find it instructive to sketch a picture of these vectors in $\Re^3$.

Example 5.4.6

Fourier Series. Let $\mathcal{V}$ be the inner-product space of real-valued functions that are integrable on the interval $(-\pi, \pi)$ and where the inner product and norm are given by

$$\langle f | g \rangle = \int_{-\pi}^{\pi} f(t) g(t) dt \quad \text{and} \quad \|f\| = \left(\int_{-\pi}^{\pi} f^2(t) dt \right)^{1/2}.$$

It's straightforward to verify that the set of trigonometric functions

$$\mathcal{B}' = \{1, \cos t, \cos 2t, \ldots, \sin t, \sin 2t, \sin 3t, \ldots\}$$

is a set of mutually orthogonal vectors, so normalizing each vector produces the orthonormal set

$$\mathcal{B} = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{\cos t}{\sqrt{\pi}}, \frac{\cos 2t}{\sqrt{\pi}}, \ldots, \frac{\sin t}{\sqrt{\pi}}, \frac{\sin 2t}{\sqrt{\pi}}, \frac{\sin 3t}{\sqrt{\pi}}, \ldots \right\}.$$

Given an arbitrary $f \in \mathcal{V}$, we construct its Fourier expansion

$$F(t) = \alpha_0 \frac{1}{\sqrt{2\pi}} + \sum_{k=1}^{\infty} \alpha_k \frac{\cos kt}{\sqrt{\pi}} + \sum_{k=1}^{\infty} \beta_k \frac{\sin kt}{\sqrt{\pi}}, \tag{5.4.4}$$

where the Fourier coefficients are given by

$$\begin{aligned}
\alpha_0 &= \left\langle \frac{1}{\sqrt{2\pi}} \middle| f \right\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(t)dt\,, \\
\alpha_k &= \left\langle \frac{\cos kt}{\sqrt{\pi}} \middle| f \right\rangle = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} f(t)\cos kt\,dt \quad \text{for } k = 1, 2, 3, \ldots, \\
\beta_k &= \left\langle \frac{\sin kt}{\sqrt{\pi}} \middle| f \right\rangle = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} f(t)\sin kt\,dt \quad \text{for } k = 1, 2, 3, \ldots.
\end{aligned}$$

Substituting these coefficients in (5.4.4) produces the infinite series

$$F(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos nt + b_n \sin nt \right), \tag{5.4.5}$$

where

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)\cos nt\,dt \quad \text{and} \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)\sin nt\,dt. \tag{5.4.6}$$

The series $F(t)$ in (5.4.5) is called the ***Fourier series*** expansion for $f(t)$, but, unlike the situation in finite-dimensional spaces, $F(t)$ need not agree with the original function $f(t)$. After all, F is periodic, so there is no hope of agreement when f is not periodic. However, the following statement is true.

- If $f(t)$ is a periodic function with period 2π that is sectionally continuous[41] on the interval $(-\pi, \pi)$, then the Fourier series $F(t)$ converges to $f(t)$ at each $t \in (-\pi, \pi)$, where f is continuous. If f is discontinuous at t_0 but possesses left-hand and right-hand derivatives at t_0, then $F(t_0)$ converges to the average value
$$F(t_0) = \frac{f(t_0^-) + f(t_0^+)}{2},$$
where $f(t_0^-)$ and $f(t_0^+)$ denote the one-sided limits $f(t_0^-) = \lim_{t \to t_0^-} f(t)$ and $f(t_0^+) = \lim_{t \to t_0^+} f(t)$.

For example, the ***square wave function*** defined by

$$f(t) = \begin{cases} -1 & \text{when } -\pi < t < 0, \\ 1 & \text{when } \phantom{-\pi <{}} 0 < t < \pi, \end{cases}$$

[41] A function f is sectionally continuous on (a, b) when f has only a finite number of discontinuities in (a, b) and the one-sided limits exist at each point of discontinuity as well as at the end points a and b.

and illustrated in Figure 5.4.2, satisfies these conditions. The value of f at $t = 0$ is irrelevant—it's not even necessary that $f(0)$ be defined.

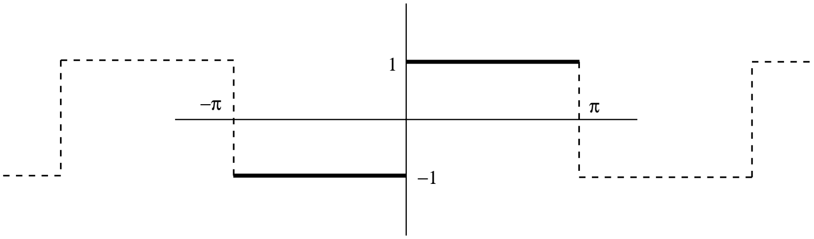


FIGURE 5.4.2

To find the Fourier series expansion for f, compute the coefficients in (5.4.6) as

$$\begin{aligned} a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos nt\,dt = \frac{1}{\pi}\int_{-\pi}^{0} -\cos nt\,dt + \frac{1}{\pi}\int_{0}^{\pi} \cos nt\,dt \\ &= 0, \\ b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin nt\,dt = \frac{1}{\pi}\int_{-\pi}^{0} -\sin nt\,dt + \frac{1}{\pi}\int_{0}^{\pi} \sin nt\,dt \\ &= \frac{2}{n\pi}(1-\cos n\pi) = \begin{cases} 0 & \text{when } n \text{ is even,} \\ 4/n\pi & \text{when } n \text{ is odd,} \end{cases} \end{aligned}$$

so that

$$F(t) = \frac{4}{\pi}\sin t + \frac{4}{3\pi}\sin 3t + \frac{4}{5\pi}\sin 5t + \cdots = \sum_{n=1}^{\infty} \frac{4}{(2n-1)\pi}\sin(2n-1)t.$$

For each $t \in (-\pi, \pi)$, except $t = 0$, it must be the case that $F(t) = f(t)$, and

$$F(0) = \frac{f(0^-) + f(0^+)}{2} = 0.$$

Not only does $F(t)$ agree with $f(t)$ everywhere f is defined, but F also provides a *periodic extension* of f in the sense that the graph of $F(t)$ is the entire square wave depicted in Figure 5.4.2—the values at the points of discontinuity (the jumps) are $F(\pm n\pi) = 0$.

Exercises for section 5.4

5.4.1. Using the standard inner product, determine which of the following pairs are orthogonal vectors in the indicated space.

$$\text{(a)}\quad \mathbf{x} = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} -2 \\ 2 \\ 2 \end{pmatrix} \quad \text{in } \Re^3,$$

$$\text{(b)}\quad \mathbf{x} = \begin{pmatrix} \mathrm{i} \\ 1+\mathrm{i} \\ 2 \\ 1-\mathrm{i} \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 0 \\ 1+\mathrm{i} \\ -2 \\ 1-\mathrm{i} \end{pmatrix} \quad \text{in } \mathcal{C}^4,$$

$$\text{(c)}\quad \mathbf{x} = \begin{pmatrix} 1 \\ -2 \\ 3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 4 \\ 2 \\ -1 \\ 1 \end{pmatrix} \quad \text{in } \Re^4,$$

$$\text{(d)}\quad \mathbf{x} = \begin{pmatrix} 1+\mathrm{i} \\ 1 \\ \mathrm{i} \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 1-\mathrm{i} \\ -3 \\ -\mathrm{i} \end{pmatrix} \quad \text{in } \mathcal{C}^3,$$

$$\text{(e)}\quad \mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{in } \Re^n.$$

5.4.2. Find two vectors of unit norm that are orthogonal to $\mathbf{u} = \left(\begin{smallmatrix} 3 \\ -2 \end{smallmatrix}\right)$.

5.4.3. Consider the following set of three vectors.

$$\left\{ \mathbf{x}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix},\ \mathbf{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix},\ \mathbf{x}_3 = \begin{pmatrix} -1 \\ -1 \\ 2 \\ 0 \end{pmatrix} \right\}.$$

(a) Using the standard inner product in $\Re^4$, verify that these vectors are mutually orthogonal.

(b) Find a nonzero vector $\mathbf{x}_4$ such that $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ is a set of mutually orthogonal vectors.

(c) Convert the resulting set into an orthonormal basis for $\Re^4$.

5.4.4. Using the standard inner product, determine the Fourier expansion of $\mathbf{x}$ with respect to $\mathcal{B}$, where

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \quad \text{and} \quad \mathcal{B} = \left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix},\ \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},\ \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \right\}.$$

5.4.5. With respect to the inner product for matrices given by (5.3.2), verify that the set

$$\mathcal{B} = \left\{ \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},\ \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},\ \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \right\}$$

is an orthonormal basis for $\Re^{2\times 2}$, and then compute the Fourier expansion of $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ with respect to $\mathcal{B}$.

5.4.6. Determine the angle between $\mathbf{x} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$.

5.4.7. Given an orthonormal basis $\mathcal{B}$ for a space $\mathcal{V}$, explain why the Fourier expansion for $\mathbf{x} \in \mathcal{V}$ is uniquely determined by $\mathcal{B}$.

5.4.8. Explain why the columns of $\mathbf{U}_{n\times n}$ are an orthonormal basis for $\mathcal{C}^n$ if and only if $\mathbf{U}^* = \mathbf{U}^{-1}$. Such matrices are said to be ***unitary***—their properties are studied in a later section.

5.4.9. Matrices with the property $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$ are said to be ***normal.*** Notice that hermitian matrices as well as real symmetric matrices are included in the class of normal matrices. Prove that if $\mathbf{A}$ is normal, then $R(\mathbf{A}) \perp N(\mathbf{A})$—i.e., every vector in $R(\mathbf{A})$ is orthogonal to every vector in $N(\mathbf{A})$. **Hint:** Recall equations (4.5.5) and (4.5.6).

5.4.10. Using the trace inner product described in Example 5.3.1, determine the angle between the following pairs of matrices.

(a) $\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 2 & -2 \\ 2 & 0 \end{pmatrix}$.

5.4.11. Why is the definition for $\cos\theta$ given in (5.4.1) not good for $\mathcal{C}^n$? Explain how to define $\cos\theta$ so that it makes sense in $\mathcal{C}^n$.

5.4.12. If $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is an orthonormal basis for an inner-product space $\mathcal{V}$, explain why

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum_i \langle \mathbf{x}|\mathbf{u}_i\rangle \langle \mathbf{u}_i|\mathbf{y}\rangle$$

holds for every $\mathbf{x}, \mathbf{y} \in \mathcal{V}$.

5.4.13. Consider a real inner-product space, where $\|\star\|^2 = \langle\star|\star\rangle$.

(a) Prove that if $\|\mathbf{x}\| = \|\mathbf{y}\|$, then $(\mathbf{x}+\mathbf{y}) \perp (\mathbf{x}-\mathbf{y})$.

(b) For the standard inner product in $\Re^2$, draw a picture of this. That is, sketch the location of $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ for two vectors with equal norms.

5.4.14. Pythagorean Theorem. Let $\mathcal{V}$ be a general inner-product space in which $\|\star\|^2 = \langle\star|\star\rangle$.

(a) When $\mathcal{V}$ is a *real* space, prove that $\mathbf{x} \perp \mathbf{y}$ if and only if $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$. (Something would be wrong if this were not true because this is where the definition of orthogonality originated.)

(b) Construct an example to show that one of the implications in part (a) does not hold when $\mathcal{V}$ is a *complex* space.

(c) When $\mathcal{V}$ is a complex space, prove that $\mathbf{x} \perp \mathbf{y}$ if and only if $\|\alpha\mathbf{x}+\beta\mathbf{y}\|^2 = \|\alpha\mathbf{x}\|^2 + \|\beta\mathbf{y}\|^2$ for all scalars α and β.

5.4.15. Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be an orthonormal basis for an inner-product space $\mathcal{V}$, and let $\mathbf{x} = \sum_i \xi_i \mathbf{u}_i$ be the Fourier expansion of $\mathbf{x} \in \mathcal{V}$.

(a) If $\mathcal{V}$ is a real space, and if θ_i is the angle between $\mathbf{u}_i$ and $\mathbf{x}$, explain why

$$\xi_i = \|\mathbf{x}\| \cos\theta_i.$$

Sketch a picture of this in $\Re^2$ or $\Re^3$ to show why the component $\xi_i\mathbf{u}_i$ represents the orthogonal projection of $\mathbf{x}$ onto the line determined by $\mathbf{u}_i$, and thus illustrate the fact that a Fourier expansion is nothing more than simply resolving $\mathbf{x}$ into mutually orthogonal components.

(b) Derive ***Parseval's identity***,[42] which says $\sum_{i=1}^n |\xi_i|^2 = \|\mathbf{x}\|^2$.

5.4.16. Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ be an orthonormal set in an n-dimensional inner-product space $\mathcal{V}$. Derive ***Bessel's inequality***,[43] which says that if $\mathbf{x} \in \mathcal{V}$ and $\xi_i = \langle\mathbf{u}_i|\mathbf{x}\rangle$, then

$$\sum_{i=1}^k |\xi_i|^2 \le \|\mathbf{x}\|^2.$$

Explain why equality holds if and only if $\mathbf{x} \in \mathit{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$. **Hint:** Consider $\|\mathbf{x} - \sum_{i=1}^k \xi_i\mathbf{u}_i\|^2$.

[42] This result appeared in the second of the five mathematical publications by Marc-Antoine Parseval des Chênes (1755–1836). Parseval was a royalist who had to flee from France when Napoleon ordered his arrest for publishing poetry against the regime.

[43] This inequality is named in honor of the German astronomer and mathematician Friedrich Wilhelm Bessel (1784–1846), who devoted his life to understanding the motions of the stars. In the process he introduced several useful mathematical ideas.

5.4.17. Construct an example using the standard inner product in $\Re^n$ to show that two vectors $\mathbf{x}$ and $\mathbf{y}$ can have an angle between them that is close to $\pi/2$ without $\mathbf{x}^T\mathbf{y}$ being close to 0. **Hint:** Consider n to be large, and use the vector $\mathbf{e}$ of all 1's for one of the vectors.

5.4.18. It was demonstrated in Example 5.4.3 that $\mathbf{y}$ is linearly correlated with $\mathbf{x}$ in the sense that $\mathbf{y} \approx \beta_0\mathbf{e} + \beta_1\mathbf{x}$ if and only if the standardization vectors $\mathbf{z_x}$ and $\mathbf{z_y}$ are "close" in the sense that they are almost on the same line in $\Re^n$. Explain why simply measuring $\|\mathbf{z_x} - \mathbf{z_y}\|_2$ does not always gauge the degree of linear correlation.

5.4.19. Let θ be the angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ from a real inner-product space.

(a) Prove that $\cos\theta = 1$ if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha > 0$.

(b) Prove that $\cos\theta = -1$ if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha < 0$.

Hint: Use the generalization of Exercise 5.1.9.

5.4.20. With respect to the orthonormal set

$$\mathcal{B} = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{\cos t}{\sqrt{\pi}}, \frac{\cos 2t}{\sqrt{\pi}}, \ldots, \frac{\sin t}{\sqrt{\pi}}, \frac{\sin 2t}{\sqrt{\pi}}, \frac{\sin 3t}{\sqrt{\pi}}, \ldots \right\},$$

determine the Fourier series expansion of the ***saw-toothed function*** defined by $f(t) = t$ for $-\pi < t < \pi$. The periodic extension of this function is depicted in Figure 5.4.3.

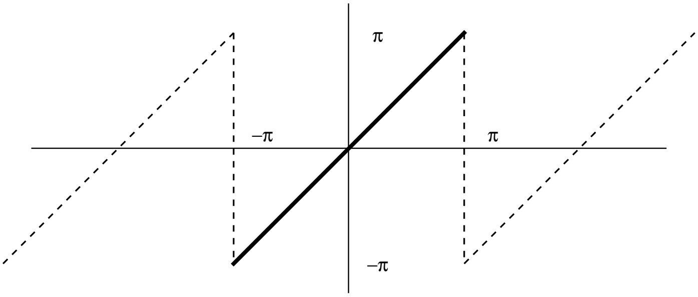


Figure 5.4.3

5.5 GRAM–SCHMIDT PROCEDURE

As discussed in §5.4, orthonormal bases possess significant advantages over bases that are not orthonormal. The spaces $\Re^n$ and $\mathcal{C}^n$ clearly possess orthonormal bases (e.g., the standard basis), but what about other spaces? Does every finite-dimensional space possess an orthonormal basis, and, if so, how can one be produced? The ***Gram–Schmidt***[44] orthogonalization procedure developed below answers these questions.

Let $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be an arbitrary basis (not necessarily orthonormal) for an n-dimensional inner-product space $\mathcal{S}$, and remember that $\|\star\| = \langle\star|\star\rangle^{1/2}$.

Objective: Use $\mathcal{B}$ to construct an orthonormal basis $\mathcal{O} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ for $\mathcal{S}$.

Strategy: Construct $\mathcal{O}$ sequentially so that $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is an orthonormal basis for $\mathcal{S}_k = span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for $k = 1, \ldots, n$.

For $k = 1$, simply take $\mathbf{u}_1 = \mathbf{x}_1/\|\mathbf{x}_1\|$. It's clear that $\mathcal{O}_1 = \{\mathbf{u}_1\}$ is an orthonormal set whose span agrees with that of $\mathcal{S}_1 = \{\mathbf{x}_1\}$. Now reason inductively. Suppose that $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is an orthonormal basis for $\mathcal{S}_k = span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, and consider the problem of finding one additional vector $\mathbf{u}_{k+1}$ such that $\mathcal{O}_{k+1} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k, \mathbf{u}_{k+1}\}$ is an orthonormal basis for $\mathcal{S}_{k+1} = span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k, \mathbf{x}_{k+1}\}$. For this to hold, the Fourier expansion (p. 299) of $\mathbf{x}_{k+1}$ with respect to $\mathcal{O}_{k+1}$ must be

$$\mathbf{x}_{k+1} = \sum_{i=1}^{k+1} \langle\mathbf{u}_i|\mathbf{x}_{k+1}\rangle\,\mathbf{u}_i,$$

which in turn implies that

$$\mathbf{u}_{k+1} = \frac{\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle\mathbf{u}_i|\mathbf{x}_{k+1}\rangle\,\mathbf{u}_i}{\langle\mathbf{u}_{k+1}|\mathbf{x}_{k+1}\rangle}. \tag{5.5.1}$$

Since $\|\mathbf{u}_{k+1}\| = 1$, it follows from (5.5.1) that

$$|\langle\mathbf{u}_{k+1}|\mathbf{x}_{k+1}\rangle| = \left\|\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle\mathbf{u}_i|\mathbf{x}_{k+1}\rangle\,\mathbf{u}_i\right\|,$$

[44] Jorgen P. Gram (1850–1916) was a Danish actuary who implicitly presented the essence of orthogonalization procedure in 1883. Gram was apparently unaware that Pierre-Simon Laplace (1749–1827) had earlier used the method. Today, Gram is remembered primarily for his development of this process, but in earlier times his name was also associated with the matrix product $\mathbf{A}^*\mathbf{A}$ that historically was referred to as the *Gram matrix* of $\mathbf{A}$.

Erhard Schmidt (1876–1959) was a student of Hermann Schwarz (of CBS inequality fame) and the great German mathematician David Hilbert. Schmidt explicitly employed the orthogonalization process in 1907 in his study of integral equations, which in turn led to the development of what are now called *Hilbert spaces*. Schmidt made significant use of the orthogonalization process to develop the geometry of Hilbert Spaces, and thus it came to bear Schmidt's name.

so $\langle \mathbf{u}_{k+1} | \mathbf{x}_{k+1} \rangle = e^{i\theta} \left\| \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right\|$ for some $0 \le \theta < 2\pi$, and

$$\mathbf{u}_{k+1} = \frac{\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i}{e^{i\theta} \left\| \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right\|}.$$

Since the value of θ in the scalar $e^{i\theta}$ neither affects $span\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k+1}\}$ nor the facts that $\|\mathbf{u}_{k+1}\| = 1$ and $\langle \mathbf{u}_{k+1} | \mathbf{u}_i \rangle = 0$ for all $i \le k$, we can arbitrarily define $\mathbf{u}_{k+1}$ to be the vector corresponding to the $\theta = 0$ or, equivalently, $e^{i\theta} = 1$. For the sake of convenience, let

$$\nu_{k+1} = \left\| \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right\|$$

so that we can write

$$\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \quad \text{and} \quad \mathbf{u}_{k+1} = \frac{\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i}{\nu_{k+1}} \text{ for } k > 0. \tag{5.5.2}$$

This sequence of vectors is called the ***Gram–Schmidt sequence.*** A straightforward induction argument proves that $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is indeed an orthonormal basis for $span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for each $k = 1, 2, \ldots$. Details are called for in Exercise 5.5.7.

The orthogonalization procedure defined by (5.5.2) is valid for any inner-product space, but if we concentrate on subspaces of $\Re^m$ or $\mathcal{C}^m$ with the standard inner product and euclidean norm, then we can formulate (5.5.2) in terms of matrices. Suppose that $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a basis for an n-dimensional subspace $\mathcal{S}$ of $\mathcal{C}^{m \times 1}$ so that the Gram–Schmidt sequence (5.5.2) becomes

$$\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \quad \text{and} \quad \mathbf{u}_k = \frac{\mathbf{x}_k - \sum_{i=1}^{k-1} (\mathbf{u}_i^* \mathbf{x}_k) \mathbf{u}_i}{\left\| \mathbf{x}_k - \sum_{i=1}^{k-1} (\mathbf{u}_i^* \mathbf{x}_k) \mathbf{u}_i \right\|} \quad \text{for } k = 2, 3, \ldots, n. \tag{5.5.3}$$

To express this in matrix notation, set

$$\mathbf{U}_1 = \mathbf{0}_{m \times 1} \quad \text{and} \quad \mathbf{U}_k = \big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\big)_{m \times k-1} \quad \text{for } k > 1,$$

and notice that

$$\mathbf{U}_k^* \mathbf{x}_k = \begin{pmatrix} \mathbf{u}_1^* \mathbf{x}_k \\ \mathbf{u}_2^* \mathbf{x}_k \\ \vdots \\ \mathbf{u}_{k-1}^* \mathbf{x}_k \end{pmatrix} \quad \text{and} \quad \mathbf{U}_k \mathbf{U}_k^* \mathbf{x}_k = \sum_{i=1}^{k-1} \mathbf{u}_i \left(\mathbf{u}_i^* \mathbf{x}_k\right) = \sum_{i=1}^{k-1} \left(\mathbf{u}_i^* \mathbf{x}_k\right) \mathbf{u}_i.$$

Since

$$\mathbf{x}_k - \sum_{i=1}^{k-1} \left(\mathbf{u}_i^* \mathbf{x}_k\right) \mathbf{u}_i = \mathbf{x}_k - \mathbf{U}_k \mathbf{U}_k^* \mathbf{x}_k = \left(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*\right) \mathbf{x}_k,$$

the vectors in (5.5.3) can be concisely written as

$$\mathbf{u}_k = \frac{\left(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*\right) \mathbf{x}_k}{\left\| \left(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*\right) \mathbf{x}_k \right\|} \quad \text{for } k = 1, 2, \ldots, n.$$

Below is a summary.

Gram–Schmidt Orthogonalization Procedure

If $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a basis for a general inner-product space $\mathcal{S}$, then the ***Gram–Schmidt sequence*** defined by

$$\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \quad \text{and} \quad \mathbf{u}_k = \frac{\mathbf{x}_k - \sum_{i=1}^{k-1} \langle \mathbf{u}_i | \mathbf{x}_k \rangle \mathbf{u}_i}{\left\| \mathbf{x}_k - \sum_{i=1}^{k-1} \langle \mathbf{u}_i | \mathbf{x}_k \rangle \mathbf{u}_i \right\|} \quad \text{for } k = 2, \ldots, n$$

is an orthonormal basis for $\mathcal{S}$. When $\mathcal{S}$ is an n-dimensional subspace of $\mathcal{C}^{m\times 1}$, the Gram–Schmidt sequence can be expressed as

$$\mathbf{u}_k = \frac{(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*)\,\mathbf{x}_k}{\|(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*)\,\mathbf{x}_k\|} \quad \text{for} \quad k = 1, 2, \ldots, n \tag{5.5.4}$$

in which $\mathbf{U}_1 = \mathbf{0}_{m\times 1}$ and $\mathbf{U}_k = \big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\big)_{m\times k-1}$ for $k > 1$.

Example 5.5.1

Classical Gram–Schmidt Algorithm. The following formal algorithm is the straightforward or "classical" implementation of the Gram–Schmidt procedure. Interpret $\mathbf{a} \leftarrow \mathbf{b}$ to mean that "$\mathbf{a}$ is defined to be (or overwritten by) $\mathbf{b}$."

```
For $k = 1$:
    $\mathbf{u}_1 \leftarrow \dfrac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$
For $k > 1$:
    $\mathbf{u}_k \leftarrow \mathbf{x}_k - \sum_{i=1}^{k-1} (\mathbf{u}_i^*\mathbf{x}_k)\mathbf{u}_i$
    $\mathbf{u}_k \leftarrow \dfrac{\mathbf{u}_k}{\|\mathbf{u}_k\|}$
```

(See Exercise 5.5.10 for other formulations of the Gram–Schmidt algorithm.)

Problem: Use the classical formulation of the Gram–Schmidt procedure given above to find an orthonormal basis for the space spanned by the following three linearly independent vectors.

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 3 \\ 1 \\ 1 \\ -1 \end{pmatrix}.$$

Solution:

$$k=1: \quad \mathbf{u}_1 \leftarrow \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1\\ 0\\ 0\\ -1 \end{pmatrix}$$

$$k=2: \quad \mathbf{u}_2 \leftarrow \mathbf{x}_2 - (\mathbf{u}_1^T\mathbf{x}_2)\mathbf{u}_1 = \begin{pmatrix} 0\\ 2\\ 0\\ 0 \end{pmatrix}, \quad \mathbf{u}_2 \leftarrow \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \begin{pmatrix} 0\\ 1\\ 0\\ 0 \end{pmatrix}$$

$$k=3: \quad \mathbf{u}_3 \leftarrow \mathbf{x}_3 - (\mathbf{u}_1^T\mathbf{x}_3)\mathbf{u}_1 - (\mathbf{u}_2^T\mathbf{x}_3)\mathbf{u}_2 = \begin{pmatrix} 1\\ 0\\ 1\\ 1 \end{pmatrix}, \quad \mathbf{u}_3 \leftarrow \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} = \frac{1}{\sqrt{3}}\begin{pmatrix} 1\\ 0\\ 1\\ 1 \end{pmatrix}$$

Thus

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1\\ 0\\ 0\\ -1 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0\\ 1\\ 0\\ 0 \end{pmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1\\ 0\\ 1\\ 1 \end{pmatrix}$$

is the desired orthonormal basis.

The Gram–Schmidt process frequently appears in the disguised form of a matrix factorization. To see this, let $\mathbf{A}_{m\times n} = \big(\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n\big)$ be a matrix with linearly independent columns. When Gram–Schmidt is applied to the columns of $\mathbf{A}$, the result is an orthonormal basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ for $R(\mathbf{A})$, where

$$\mathbf{q}_1 = \frac{\mathbf{a}_1}{\nu_1} \quad \text{and} \quad \mathbf{q}_k = \frac{\mathbf{a}_k - \sum_{i=1}^{k-1}\langle \mathbf{q}_i|\mathbf{a}_k\rangle\,\mathbf{q}_i}{\nu_k} \quad \text{for } k = 2, 3, \ldots, n,$$

where $\nu_1 = \|\mathbf{a}_1\|$ and $\nu_k = \big\|\mathbf{a}_k - \sum_{i=1}^{k-1}\langle \mathbf{q}_i|\mathbf{a}_k\rangle\,\mathbf{q}_i\big\|$ for $k > 1$. The above relationships can be rewritten as

$$\mathbf{a}_1 = \nu_1\mathbf{q}_1 \quad \text{and} \quad \mathbf{a}_k = \langle \mathbf{q}_1|\mathbf{a}_k\rangle\,\mathbf{q}_1 + \cdots + \langle \mathbf{q}_{k-1}|\mathbf{a}_k\rangle\,\mathbf{q}_{k-1} + \nu_k\mathbf{q}_k \quad \text{for } k > 1,$$

which in turn can be expressed in matrix form by writing

$$\big(\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n\big) = \big(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n\big) \begin{pmatrix} \nu_1 & \langle \mathbf{q}_1|\mathbf{a}_2\rangle & \langle \mathbf{q}_1|\mathbf{a}_3\rangle & \cdots & \langle \mathbf{q}_1|\mathbf{a}_n\rangle \\ 0 & \nu_2 & \langle \mathbf{q}_2|\mathbf{a}_3\rangle & \cdots & \langle \mathbf{q}_2|\mathbf{a}_n\rangle \\ 0 & 0 & \nu_3 & \cdots & \langle \mathbf{q}_3|\mathbf{a}_n\rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \nu_n \end{pmatrix}.$$

This says that it's possible to factor a matrix with independent columns as $\mathbf{A}_{m\times n} = \mathbf{Q}_{m\times n}\mathbf{R}_{n\times n}$, where the columns of $\mathbf{Q}$ are an orthonormal basis for $R(\mathbf{A})$ and $\mathbf{R}$ is an upper-triangular matrix with positive diagonal elements.

The factorization $\mathbf{A} = \mathbf{QR}$ is called the ***QR factorization*** for $\mathbf{A}$, and it is uniquely determined by $\mathbf{A}$ (Exercise 5.5.8). When $\mathbf{A}$ and $\mathbf{Q}$ are not square, some authors emphasize the point by calling $\mathbf{A} = \mathbf{QR}$ the *rectangular* QR factorization—the case when $\mathbf{A}$ and $\mathbf{Q}$ are square is further discussed on p. 345. Below is a summary of the above observations.

QR Factorization

Every matrix $\mathbf{A}_{m\times n}$ with linearly independent columns can be uniquely factored as $\mathbf{A} = \mathbf{QR}$ in which the columns of $\mathbf{Q}_{m\times n}$ are an orthonormal basis for $R(\mathbf{A})$ and $\mathbf{R}_{n\times n}$ is an upper-triangular matrix with positive diagonal entries.

- The QR factorization is the complete "road map" of the Gram–Schmidt process because the columns of $\mathbf{Q} = \left(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n\right)$ are the result of applying the Gram–Schmidt procedure to the columns of $\mathbf{A} = \left(\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n\right)$ and $\mathbf{R}$ is given by

$$\mathbf{R} = \begin{pmatrix} \nu_1 & \mathbf{q}_1^*\mathbf{a}_2 & \mathbf{q}_1^*\mathbf{a}_3 & \cdots & \mathbf{q}_1^*\mathbf{a}_n \\ 0 & \nu_2 & \mathbf{q}_2^*\mathbf{a}_3 & \cdots & \mathbf{q}_2^*\mathbf{a}_n \\ 0 & 0 & \nu_3 & \cdots & \mathbf{q}_3^*\mathbf{a}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \nu_n \end{pmatrix},$$

  where $\nu_1 = \|\mathbf{a}_1\|$ and $\nu_k = \left\|\mathbf{a}_k - \sum_{i=1}^{k-1} \langle \mathbf{q}_i|\mathbf{a}_k\rangle\, \mathbf{q}_i\right\|$ for $k > 1$.

Example 5.5.2

Problem: Determine the QR factors of

$$\mathbf{A} = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}.$$

Solution: Using the standard inner product for $\Re^n$, apply the Gram–Schmidt procedure to the columns of $\mathbf{A}$ by setting

$$\mathbf{q}_1 = \frac{\mathbf{a}_1}{\nu_1} \quad \text{and} \quad \mathbf{q}_k = \frac{\mathbf{a}_k - \sum_{i=1}^{k-1}\left(\mathbf{q}_i^T\mathbf{a}_k\right)\mathbf{q}_i}{\nu_k} \quad \text{for } k = 2, 3,$$

where $\nu_1 = \|\mathbf{a}_1\|$ and $\nu_k = \left\|\mathbf{a}_k - \sum_{i=1}^{k-1}\left(\mathbf{q}_i^T\mathbf{a}_k\right)\mathbf{q}_i\right\|$. The computation of these quantities can be organized as follows.

$$k = 1: \quad r_{11} \leftarrow \|\mathbf{a}_1\| = 5 \quad \text{and} \quad \mathbf{q}_1 \leftarrow \frac{\mathbf{a}_1}{r_{11}} = \begin{pmatrix} 0 \\ 3/5 \\ 4/5 \end{pmatrix}$$

$$k = 2: \quad r_{12} \leftarrow \mathbf{q}_1^T\mathbf{a}_2 = 25$$

$$\mathbf{q}_2 \leftarrow \mathbf{a}_2 - r_{12}\mathbf{q}_1 = \begin{pmatrix} -20 \\ 12 \\ -9 \end{pmatrix}$$

$$r_{22} \leftarrow \|\mathbf{q}_2\| = 25 \text{ and } \mathbf{q}_2 \leftarrow \frac{\mathbf{q}_2}{r_{22}} = \frac{1}{25}\begin{pmatrix} -20 \\ 12 \\ -9 \end{pmatrix}$$

$$k = 3: \quad r_{13} \leftarrow \mathbf{q}_1^T\mathbf{a}_3 = -4 \text{ and } r_{23} \leftarrow \mathbf{q}_2^T\mathbf{a}_3 = 10$$

$$\mathbf{q}_3 \leftarrow \mathbf{a}_3 - r_{13}\mathbf{q}_1 - r_{23}\mathbf{q}_2 = \frac{2}{5}\begin{pmatrix} -15 \\ -16 \\ 12 \end{pmatrix}$$

$$r_{33} \leftarrow \|\mathbf{q}_3\| = 10 \text{ and } \mathbf{q}_3 \leftarrow \frac{\mathbf{q}_3}{r_{33}} = \frac{1}{25}\begin{pmatrix} -15 \\ -16 \\ 12 \end{pmatrix}$$

Therefore,

$$\mathbf{Q} = \frac{1}{25}\begin{pmatrix} 0 & -20 & -15 \\ 15 & 12 & -16 \\ 20 & -9 & 12 \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}.$$

We now have two important matrix factorizations, namely, the LU factorization, discussed in §3.10 on p. 141 and the QR factorization. They are not the same, but some striking analogies exist.

- Each factorization represents a reduction to upper-triangular form—LU by Gaussian elimination, and QR by Gram–Schmidt. In particular, the LU factorization is the complete "road map" of Gaussian elimination applied to a square nonsingular matrix, whereas QR is the complete road map of Gram–Schmidt applied to a matrix with linearly independent columns.
- When they exist, both factorizations $\mathbf{A} = \mathbf{LU}$ and $\mathbf{A} = \mathbf{QR}$ are uniquely determined by $\mathbf{A}$.
- Once the LU factors (assuming they exist) of a nonsingular matrix $\mathbf{A}$ are known, the solution of $\mathbf{Ax} = \mathbf{b}$ is easily computed—solve $\mathbf{Ly} = \mathbf{b}$ by forward substitution, and then solve $\mathbf{Ux} = \mathbf{y}$ by back substitution (see p. 146). The QR factors can be used in a similar manner. If $\mathbf{A} \in \Re^{n\times n}$ is nonsingular, then $\mathbf{Q}^T = \mathbf{Q}^{-1}$ (because $\mathbf{Q}$ has orthonormal columns), so $\mathbf{Ax} = \mathbf{b} \iff \mathbf{QRx} = \mathbf{b} \iff \mathbf{Rx} = \mathbf{Q}^T\mathbf{b}$, which is also a triangular system that is solved by back substitution.

While the LU and QR factors can be used in more or less the same way to solve nonsingular systems, things are different for singular and rectangular cases because $\mathbf{Ax} = \mathbf{b}$ might be inconsistent, in which case a least squares solution as described in §4.6, (p. 223) may be desired. Unfortunately, the LU factors of $\mathbf{A}$ don't exist when $\mathbf{A}$ is rectangular. And even if $\mathbf{A}$ is square and has an LU factorization, the LU factors of $\mathbf{A}$ are not much help in solving the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ that produces least squares solutions. But the QR factors of $\mathbf{A}_{m\times n}$ always exist as long as $\mathbf{A}$ has linearly independent columns, and, as demonstrated in the following example, the QR factors provide the least squares solution of an inconsistent system in exactly the same way as they provide the solution of a consistent system.

Example 5.5.3

Application to the Least Squares Problem. If $\mathbf{Ax} = \mathbf{b}$ is a possibly inconsistent (real) system, then, as discussed on p. 226, the set of all least squares solutions is the set of solutions to the system of normal equations

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}. \tag{5.5.5}$$

But computing $\mathbf{A}^T\mathbf{A}$ and then performing an LU factorization of $\mathbf{A}^T\mathbf{A}$ to solve (5.5.5) is generally not advisable. First, it's inefficient and, second, as pointed out in Example 4.5.1, computing $\mathbf{A}^T\mathbf{A}$ with floating-point arithmetic can result in a loss of significant information. The QR approach doesn't suffer from either of these objections. Suppose that $rank\,(\mathbf{A}_{m\times n}) = n$ (so that there is a unique least squares solution), and let $\mathbf{A} = \mathbf{QR}$ be the QR factorization. Because the columns of $\mathbf{Q}$ are an orthonormal set, it follows that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n$, so

$$\mathbf{A}^T\mathbf{A} = (\mathbf{QR})^T(\mathbf{QR}) = \mathbf{R}^T\mathbf{Q}^T\mathbf{QR} = \mathbf{R}^T\mathbf{R}. \tag{5.5.6}$$

Consequently, the normal equations (5.5.5) can be written as

$$\mathbf{R}^T\mathbf{Rx} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b}. \tag{5.5.7}$$

But $\mathbf{R}^T$ is nonsingular (it is triangular with positive diagonal entries), so (5.5.7) simplifies to become

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b}. \tag{5.5.8}$$

This is just an upper-triangular system that is efficiently solved by back substitution. In other words, most of the work involved in solving the least squares problem is in computing the QR factorization of $\mathbf{A}$. Finally, notice that

$$\mathbf{x} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{b} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$$

is the solution of $\mathbf{Ax} = \mathbf{b}$ when the system is consistent as well as the least squares solution when the system is inconsistent (see p. 214). That is, with the QR approach, it makes no difference whether or not $\mathbf{Ax} = \mathbf{b}$ is consistent because in both cases things boil down to solving the same equation—namely, (5.5.8). Below is a formal summary.

Linear Systems and the QR Factorization

If $rank\,(\mathbf{A}_{m\times n}) = n$, and if $\mathbf{A} = \mathbf{QR}$ is the QR factorization, then the solution of the nonsingular triangular system

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b} \tag{5.5.9}$$

is either the solution or the least squares solution of $\mathbf{Ax} = \mathbf{b}$ depending on whether or not $\mathbf{Ax} = \mathbf{b}$ is consistent.

It's worthwhile to reemphasize that the QR approach to the least squares problem obviates the need to explicitly compute the product $\mathbf{A}^T\mathbf{A}$. But if $\mathbf{A}^T\mathbf{A}$ is ever needed, it is retrievable from the factorization $\mathbf{A}^T\mathbf{A} = \mathbf{R}^T\mathbf{R}$. In fact, this is the ***Cholesky factorization*** of $\mathbf{A}^T\mathbf{A}$ as discussed in Example 3.10.7, p. 154.

The Gram–Schmidt procedure is a powerful theoretical tool, but it's not a good numerical algorithm when implemented in the straightforward or "classical" sense. When floating-point arithmetic is used, the classical Gram–Schmidt algorithm applied to a set of vectors that is not already close to being an orthogonal set can produce a set of vectors that is far from being an orthogonal set. To see this, consider the following example.

Example 5.5.4

Problem: Using 3-digit floating-point arithmetic, apply the classical Gram–Schmidt algorithm to the set

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 10^{-3} \\ 0 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}.$$

Solution:

$k = 1$: $fl\,\|\mathbf{x}_1\| = 1$, so $\mathbf{u}_1 \leftarrow \mathbf{x}_1$.

$k = 2$: $fl\left(\mathbf{u}_1^T\mathbf{x}_2\right) = 1$, so

$$\mathbf{u}_2 \leftarrow \mathbf{x}_2 - \left(\mathbf{u}_1^T\mathbf{x}_2\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ 0 \\ -10^{-3} \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 \leftarrow fl\left(\frac{\mathbf{u}_2}{\|\mathbf{u}_2\|}\right) = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

$k = 3$: $fl\left(\mathbf{u}_1^T\mathbf{x}_3\right) = 1$ and $fl\left(\mathbf{u}_2^T\mathbf{x}_3\right) = -10^{-3}$, so

$$\mathbf{u}_3 \leftarrow \mathbf{x}_3 - \left(\mathbf{u}_1^T\mathbf{x}_3\right)\mathbf{u}_1 - \left(\mathbf{u}_2^T\mathbf{x}_3\right)\mathbf{u}_2 = \begin{pmatrix} 0 \\ -10^{-3} \\ -10^{-3} \end{pmatrix} \text{ and } \mathbf{u}_3 \leftarrow fl\left(\frac{\mathbf{u}_3}{\|\mathbf{u}_3\|}\right) = \begin{pmatrix} 0 \\ -.709 \\ -.709 \end{pmatrix}.$$

Therefore, classical Gram–Schmidt with 3-digit arithmetic returns

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0 \\ -.709 \\ -.709 \end{pmatrix}, \tag{5.5.10}$$

which is unsatisfactory because $\mathbf{u_2}$ and $\mathbf{u}_3$ are far from being orthogonal.

It's possible to improve the numerical stability of the orthogonalization process by rearranging the order of the calculations. Recall from (5.5.4) that

$$\mathbf{u}_k = \frac{(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*)\,\mathbf{x}_k}{\|(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*)\,\mathbf{x}_k\|}, \quad \text{where} \quad \mathbf{U}_1 = \mathbf{0} \text{ and } \mathbf{U}_k = \big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\big).$$

If $\mathbf{E}_1 = \mathbf{I}$ and $\mathbf{E}_i = \mathbf{I} - \mathbf{u}_{i-1}\mathbf{u}_{i-1}^*$ for $i > 1$, then the orthogonality of the $\mathbf{u}_i$'s insures that

$$\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1 = \mathbf{I} - \mathbf{u}_1\mathbf{u}_1^* - \mathbf{u}_2\mathbf{u}_2^* - \cdots - \mathbf{u}_{k-1}\mathbf{u}_{k-1}^* = \mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*,$$

so the Gram–Schmidt sequence can also be expressed as

$$\mathbf{u}_k = \frac{\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k}{\|\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k\|} \quad \text{for } k = 1, 2, \ldots, n.$$

This means that the Gram–Schmidt sequence can be generated as follows:

$$\begin{aligned}
\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} &\xrightarrow{\text{Normalize 1-st}} \{\mathbf{u}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \\
&\xrightarrow{\text{Apply } \mathbf{E}_2} \{\mathbf{u}_1,\ \mathbf{E}_2\mathbf{x}_2,\ \mathbf{E}_2\mathbf{x}_3,\ \ldots,\ \mathbf{E}_2\mathbf{x}_n\} \\
&\xrightarrow{\text{Normalize 2-nd}} \{\mathbf{u}_1, \mathbf{u}_2,\ \mathbf{E}_2\mathbf{x}_3,\ \ldots,\ \mathbf{E}_2\mathbf{x}_n\} \\
&\xrightarrow{\text{Apply } \mathbf{E}_3} \{\mathbf{u}_1, \mathbf{u}_2,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_3,\ \ldots,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_n\} \\
&\xrightarrow{\text{Normalize 3-rd}} \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_4, \ldots,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_n\}\,, \\
&\quad\ \text{etc.}
\end{aligned}$$

While there is no theoretical difference, this "modified" algorithm is numerically more stable than the classical algorithm when floating-point arithmetic is used. The k^{th} step of the classical algorithm alters only the k^{th} vector, but the k^{th} step of the modified algorithm "updates" all vectors from the k^{th} through the last, and conditioning the unorthogonalized tail in this way makes a difference.

Modified Gram–Schmidt Algorithm

For a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathcal{C}^{m\times 1}$, the Gram–Schmidt sequence given on p. 309 can be alternately described as

$$\mathbf{u}_k = \frac{\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k}{\|\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k\|} \text{ with } \mathbf{E}_1 = \mathbf{I}, \;\; \mathbf{E}_i = \mathbf{I} - \mathbf{u}_{i-1}\mathbf{u}_{i-1}^* \text{ for } i > 1,$$

and this sequence is generated by the following algorithm.

```
For  $k = 1$:   $\mathbf{u}_1 \leftarrow \mathbf{x}_1 / \|\mathbf{x}_1\|$   and   $\mathbf{u}_j \leftarrow \mathbf{x}_j$  for  $j = 2, 3, \ldots, n$
For  $k > 1$:   $\mathbf{u}_j \leftarrow \mathbf{E}_k\mathbf{u}_j = \mathbf{u}_j - \left(\mathbf{u}_{k-1}^*\mathbf{u}_j\right)\mathbf{u}_{k-1}$  for  $j = k, k+1, \ldots, n$
               $\mathbf{u}_k \leftarrow \mathbf{u}_k / \|\mathbf{u}_k\|$
```

(An alternate implementation is given in Exercise 5.5.10.)

To see that the modified version of Gram–Schmidt can indeed make a difference when floating-point arithmetic is used, consider the following example.

Example 5.5.5

Problem: Use 3-digit floating-point arithmetic, and apply the modified Gram–Schmidt algorithm to the set given in Example 5.5.4 (p. 314), and then compare the results of the modified algorithm with those of the classical algorithm.

Solution: $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 10^{-3} \\ 0 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}.$

$k = 1$: $\quad fl\,\|\mathbf{x}_1\| = 1$, so $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} \leftarrow \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$.

$k = 2$: $\quad fl\left(\mathbf{u}_1^T\mathbf{u}_2\right) = 1$ and $fl\left(\mathbf{u}_1^T\mathbf{u}_3\right) = 1$, so

$$\mathbf{u}_2 \leftarrow \mathbf{u}_2 - \left(\mathbf{u}_1^T\mathbf{u}_2\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ 0 \\ -10^{-3} \end{pmatrix}, \quad \mathbf{u}_3 \leftarrow \mathbf{u}_3 - \left(\mathbf{u}_1^T\mathbf{u}_3\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ -10^{-3} \\ 0 \end{pmatrix},$$

and

$$\mathbf{u}_2 \leftarrow \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

$k = 3$: $\quad \mathbf{u}_2^T\mathbf{u}_3 = 0$, so

$$\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \left(\mathbf{u}_2^T\mathbf{u}_3\right)\mathbf{u}_2 = \begin{pmatrix} 0 \\ -10^{-3} \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_3 \leftarrow \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}.$$

Thus the modified Gram–Schmidt algorithm produces

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \tag{5.5.11}$$

which is as good as one can expect using 3-digit arithmetic. Comparing (5.5.11) with the result (5.5.10) obtained in Example 5.5.4 illuminates the advantage possessed by modified Gram–Schmidt algorithm over the classical algorithm.

Below is a summary of some facts concerning the modified Gram–Schmidt algorithm compared with the classical implementation.

Summary

- When the Gram–Schmidt procedures (classical or modified) are applied to the columns of $\mathbf{A}$ using exact arithmetic, each produces an orthonormal basis for $R(\mathbf{A})$.
- For computing a QR factorization in floating-point arithmetic, the modified algorithm produces results that are at least as good as and often better than the classical algorithm, but the modified algorithm is not unconditionally stable—there are situations in which it fails to produce a set of columns that are nearly orthogonal.
- For solving the least square problem with floating-point arithmetic, the modified procedure is a numerically stable algorithm in the sense that the method described in Example 5.5.3 returns a result that is the exact solution of a nearby least squares problem. However, the Householder method described on p. 346 is just as stable and needs slightly fewer arithmetic operations.

Exercises for section 5.5

5.5.1. Let $\mathcal{S} = span\left\{ \mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 2 \\ -1 \\ -1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} -1 \\ 2 \\ 2 \\ 1 \end{pmatrix} \right\}.$

(a) Use the classical Gram–Schmidt algorithm (with exact arithmetic) to determine an orthonormal basis for $\mathcal{S}$.

(b) Verify directly that the Gram–Schmidt sequence produced in part (a) is indeed an orthonormal basis for $\mathcal{S}$.

(c) Repeat part (a) using the modified Gram–Schmidt algorithm, and compare the results.

5.5.2. Use the Gram–Schmidt procedure to find an orthonormal basis for the four fundamental subspaces of $\mathbf{A} = \begin{pmatrix} 1 & -2 & 3 & -1 \\ 2 & -4 & 6 & -2 \\ 3 & -6 & 9 & -3 \end{pmatrix}$.

5.5.3. Apply the Gram–Schmidt procedure with the standard inner product for $\mathcal{C}^3$ to $\left\{ \begin{pmatrix} \mathrm{i} \\ \mathrm{i} \\ \mathrm{i} \end{pmatrix}, \begin{pmatrix} 0 \\ \mathrm{i} \\ \mathrm{i} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \mathrm{i} \end{pmatrix} \right\}$.

5.5.4. Explain what happens when the Gram–Schmidt process is applied to an orthonormal set of vectors.

5.5.5. Explain what happens when the Gram–Schmidt process is applied to a linearly dependent set of vectors.

5.5.6. Let $\mathbf{A} = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & -3 \\ 0 & 1 & 1 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$.

(a) Determine the rectangular QR factorization of $\mathbf{A}$.

(b) Use the QR factors from part (a) to determine the least squares solution to $\mathbf{Ax} = \mathbf{b}$.

5.5.7. Given a linearly independent set of vectors $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ in an inner-product space, let $\mathcal{S}_k = \mathit{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for $k = 1, 2, \ldots, n$. Give an induction argument to prove that if $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is the Gram–Schmidt sequence defined in (5.5.2), then $\mathcal{O}_k$ is indeed an orthonormal basis for $\mathcal{S}_k = \mathit{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for each $k = 1, 2, \ldots, n$.

5.5.8. Prove that if $\mathit{rank}(\mathbf{A}_{m\times n}) = n$, then the rectangular QR factorization of $\mathbf{A}$ is unique. That is, if $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}_{m\times n}$ has orthonormal columns and $\mathbf{R}_{n\times n}$ is upper triangular with positive diagonal entries, then $\mathbf{Q}$ and $\mathbf{R}$ are unique. **Hint:** Recall Example 3.10.7, p. 154.

5.5.9. (a) Apply classical Gram–Schmidt with 3-digit floating-point arithmetic to $\left\{ \mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 1 \\ 10^{-3} \\ 0 \end{pmatrix} \right\}$. You may assume that $fl\left(\sqrt{2}\right) = 1.41$.

(b) Again using 3-digit floating-point arithmetic, apply the modified Gram–Schmidt algorithm to $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, and compare the result with that of part (a).

5.5.10. Depending on how the inner products r_{ij} are defined, verify that the following code implements both the classical and modified Gram–Schmidt algorithms applied to a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$.

```
For $j = 1$ to $n$
    $\mathbf{u}_j \longleftarrow \mathbf{x}_j$
        For $i = 1$ to $j - 1$
            $r_{ij} \longleftarrow \begin{cases} \langle \mathbf{u}_i|\mathbf{x}_j \rangle & \text{(classical Gram–Schmidt)} \\ \langle \mathbf{u}_i|\mathbf{u}_j \rangle & \text{(modified Gram–Schmidt)} \end{cases}$
            $\mathbf{u}_j \longleftarrow \mathbf{u}_j - r_{ij}\mathbf{u}_i$
        End
    $r_{jj} \longleftarrow \|\mathbf{u}_j\|$
        If $r_{jj} = 0$
            quit  (because $\mathbf{x}_j \in span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{j-1}\}$)
        Else $\mathbf{u}_j \longleftarrow \mathbf{u}_j / r_{jj}$
End
```

If exact arithmetic is used, will the inner products r_{ij} be the same for both implementations?

5.5.11. Let $\mathcal{V}$ be the inner-product space of real-valued continuous functions defined on the interval $[-1, 1]$, where the inner product is defined by

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

and let $\mathcal{S}$ be the subspace of $\mathcal{V}$ that is spanned by the three linearly independent polynomials $q_0 = 1, \quad q_1 = x, \quad q_2 = x^2$.

(a) Use the Gram–Schmidt process to determine an orthonormal set of polynomials $\{p_0, p_1, p_2\}$ that spans $\mathcal{S}$. These polynomials are the first three normalized ***Legendre***[45] ***polynomials.***

(b) Verify that p_n satisfies ***Legendre's differential equation***

$$(1 - x^2)y'' - 2xy' + n(n+1)y = 0$$

for $n = 0, 1, 2$. This equation and its solutions are of considerable importance in applied mathematics.

[45] Adrien–Marie Legendre (1752–1833) was one of the most eminent French mathematicians of the eighteenth century. His primary work in higher mathematics concerned number theory and the study of elliptic functions. But he was also instrumental in the development of the theory of least squares, and some people believe that Legendre should receive the credit that is often afforded to Gauss for the introduction of the method of least squares. Like Gauss and many other successful mathematicians, Legendre spent substantial time engaged in diligent and painstaking computation. It is reported that in 1824 Legendre refused to vote for the government's candidate for Institut National, so his pension was stopped, and he died in poverty.

5.6 UNITARY AND ORTHOGONAL MATRICES

The purpose of this section is to examine square matrices whose columns (or rows) are orthonormal. The standard inner product and the euclidean 2-norm are the only ones used in this section, so distinguishing subscripts are omitted.

Unitary and Orthogonal Matrices

- A ***unitary matrix*** is defined to be a *complex* matrix $\mathbf{U}_{n\times n}$ whose columns (or rows) constitute an orthonormal basis for $\mathcal{C}^n$.
- An ***orthogonal matrix*** is defined to be a *real* matrix $\mathbf{P}_{n\times n}$ whose columns (or rows) constitute an orthonormal basis for $\Re^n$.

Unitary and orthogonal matrices have some nice features, one of which is the fact that they are easy to invert. To see why, notice that the columns of $\mathbf{U}_{n\times n}$ are an orthonormal set if and only if

$$[\mathbf{U}^*\mathbf{U}]_{ij} = (\mathbf{U}_{*i})^*\mathbf{U}_{*j} = \begin{cases} 1 & \text{when } i = j, \\ 0 & \text{when } i \neq j. \end{cases}$$

In other words, $\mathbf{U}$ has orthonormal columns if and only if $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, which in turn is equivalent to saying that $\mathbf{U}^{-1} = \mathbf{U}^*$. Notice that because

$$\mathbf{U}^*\mathbf{U} = \mathbf{I} \Longleftrightarrow \mathbf{U}\mathbf{U}^* = \mathbf{I},$$

the columns of $\mathbf{U}$ are orthonormal if and only if the rows of $\mathbf{U}$ are orthonormal, and this is why the definitions of unitary and orthogonal matrices can be stated either in terms of orthonormal columns or orthonormal rows.

Another nice feature is that multiplication by a unitary matrix does not change the length of a vector—only the direction is altered. This is easy to see by writing

$$\|\mathbf{U}\mathbf{x}\|^2 = (\mathbf{U}\mathbf{x})^*\mathbf{U}\mathbf{x} = \mathbf{x}^*\mathbf{U}^*\mathbf{U}\mathbf{x} = \mathbf{x}^*\mathbf{x} = \|\mathbf{x}\|^2 \quad \forall\ \mathbf{x} \in \mathcal{C}^n. \tag{5.6.1}$$

Conversely, if (5.6.1) holds, then $\mathbf{U}$ must be unitary because

$$\begin{aligned} \|\mathbf{U}\mathbf{x}\|^2 = \|\mathbf{x}\|^2 \quad \forall\ \mathbf{x} \in \mathcal{C}^n \quad &\Longrightarrow \quad \mathbf{x}^*\mathbf{U}^*\mathbf{U}\mathbf{x} = \mathbf{x}^*\mathbf{x} \quad \forall\ \mathbf{x} \in \mathcal{C}^n \\ &\Longrightarrow \quad \mathbf{e}_i^T\mathbf{U}^*\mathbf{U}\mathbf{e}_j = \mathbf{e}_i^T\mathbf{e}_j = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{when } i \neq j \end{cases} \\ &\Longrightarrow \quad (\mathbf{U}_{*i})^*\mathbf{U}_{*j} = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{when } i \neq j. \end{cases} \end{aligned}$$

In the case of orthogonal matrices, everything is real so that $(\star)^*$ can be replaced by $(\star)^T$. Below is a summary of these observations.

Characterizations

- The following statements are equivalent to saying that a complex matrix $\mathbf{U}_{n\times n}$ is unitary.
  - ▷ $\mathbf{U}$ has orthonormal columns.
  - ▷ $\mathbf{U}$ has orthonormal rows.
  - ▷ $\mathbf{U}^{-1} = \mathbf{U}^*$.
  - ▷ $\|\mathbf{Ux}\|_2 = \|\mathbf{x}\|_2$ for every $\mathbf{x} \in \mathcal{C}^{n\times 1}$.
- The following statements are equivalent to saying that a real matrix $\mathbf{P}_{n\times n}$ is orthogonal.
  - ▷ $\mathbf{P}$ has orthonormal columns.
  - ▷ $\mathbf{P}$ has orthonormal rows.
  - ▷ $\mathbf{P}^{-1} = \mathbf{P}^T$.
  - ▷ $\|\mathbf{Px}\|_2 = \|\mathbf{x}\|_2$ for every $\mathbf{x} \in \Re^{n\times 1}$.

Example 5.6.1

- The identity matrix $\mathbf{I}$ is an orthogonal matrix.
- All permutation matrices (products of elementary interchange matrices) are orthogonal—recall Exercise 3.9.4.
- The matrix
$$\mathbf{P} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{3} & -1/\sqrt{6} \\ -1/\sqrt{2} & 1/\sqrt{3} & -1/\sqrt{6} \\ 0 & 1/\sqrt{3} & 2/\sqrt{6} \end{pmatrix}$$
is an orthogonal matrix because $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$ or, equivalently, because the columns (and rows) constitute an orthonormal set.
- The matrix $\mathbf{U} = \frac{1}{2}\begin{pmatrix} 1+\mathrm{i} & -1+\mathrm{i} \\ 1+\mathrm{i} & 1-\mathrm{i} \end{pmatrix}$ is unitary because $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$ or, equivalently, because the columns (and rows) are an orthonormal set.
- An orthogonal matrix can be considered to be unitary, but a unitary matrix is generally not orthogonal.

In general, a linear operator $\mathbf{T}$ on a vector space $\mathcal{V}$ with the property that $\|\mathbf{Tx}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathcal{V}$ is called an ***isometry*** on $\mathcal{V}$. The isometries on $\Re^n$ are precisely the orthogonal matrices, and the isometries on $\mathcal{C}^n$ are the unitary matrices. The term "isometry" has an advantage in that it can be used to treat the real and complex cases simultaneously, but for clarity we will often revert back to the more cumbersome "orthogonal" and "unitary" terminology.

The geometrical concepts of projection, reflection, and rotation are among the most fundamental of all linear transformations in $\Re^2$ and $\Re^3$ (see Example 4.7.1 for three simple examples), so pursuing these ideas in higher dimensions is only natural. The reflector and rotator given in Example 4.7.1 are isometries (because they preserve length), but the projector is not. We are about to see that the same is true in more general settings.

Elementary Orthogonal Projectors

For a vector $\mathbf{u} \in \mathcal{C}^{n\times 1}$ such that $\|\mathbf{u}\| = 1$, a matrix of the form

$$\mathbf{Q} = \mathbf{I} - \mathbf{u}\mathbf{u}^* \tag{5.6.2}$$

is called an ***elementary orthogonal projector.*** More general projectors are discussed on pp. 386 and 429.

To understand the nature of elementary projectors consider the situation in $\Re^3$. Suppose that $\|\mathbf{u}_{3\times 1}\| = 1$, and let $\mathbf{u}^\perp$ denote the space (the plane through the origin) consisting of all vectors that are perpendicular to $\mathbf{u}$—we call $\mathbf{u}^\perp$ the ***orthogonal complement*** of $\mathbf{u}$ (a more general definition appears on p. 403). The matrix $\mathbf{Q} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$ is the orthogonal projector onto $\mathbf{u}^\perp$ in the sense that $\mathbf{Q}$ maps each $\mathbf{x} \in \Re^{3\times 1}$ to its orthogonal projection in $\mathbf{u}^\perp$ as shown in Figure 5.6.1.

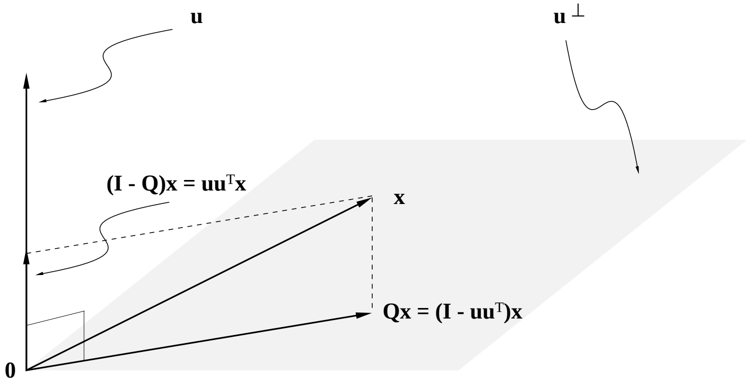


Figure 5.6.1

To see this, observe that each $\mathbf{x}$ can be resolved into two components

$$\mathbf{x} = (\mathbf{I} - \mathbf{Q})\mathbf{x} + \mathbf{Q}\mathbf{x}, \quad \text{where} \quad (\mathbf{I} - \mathbf{Q})\mathbf{x} \perp \mathbf{Q}\mathbf{x}.$$

The vector $(\mathbf{I} - \mathbf{Q})\mathbf{x} = \mathbf{u}(\mathbf{u}^T\mathbf{x})$ is on the line determined by $\mathbf{u}$, and $\mathbf{Q}\mathbf{x}$ is in the plane $\mathbf{u}^\perp$ because $\mathbf{u}^T\mathbf{Q}\mathbf{x} = \mathbf{0}$.

The situation is exactly as depicted in Figure 5.6.1. Notice that $(\mathbf{I}-\mathbf{Q})\mathbf{x} = \mathbf{u}\mathbf{u}^T\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the line determined by $\mathbf{u}$ and $\left\|\mathbf{u}\mathbf{u}^T\mathbf{x}\right\| = |\mathbf{u}^T\mathbf{x}|$. This provides a nice interpretation of the magnitude of the standard inner product. Below is a summary.

Geometry of Elementary Projectors

For vectors $\mathbf{u}, \mathbf{x} \in \mathcal{C}^{n\times 1}$ such that $\|\mathbf{u}\| = 1$,

- $(\mathbf{I} - \mathbf{u}\mathbf{u}^*)\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the orthogonal complement $\mathbf{u}^\perp$, the space of all vectors orthogonal to $\mathbf{u}$; (5.6.3)
- $\mathbf{u}\mathbf{u}^*\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the one-dimensional space $span\{\mathbf{u}\}$; (5.6.4)
- $|\mathbf{u}^*\mathbf{x}|$ represents the length of the orthogonal projection of $\mathbf{x}$ onto the one-dimensional space $span\{\mathbf{u}\}$. (5.6.5)

In passing, note that elementary projectors are never isometries—they can't be because they are not unitary matrices in the complex case and not orthogonal matrices in the real case. Furthermore, isometries are nonsingular but elementary projectors are singular.

Example 5.6.2

Problem: Determine the orthogonal projection of $\mathbf{x}$ onto $span\{\mathbf{u}\}$, and then find the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^\perp$ for $\mathbf{x} = \begin{pmatrix} 2\\ 0\\ 1 \end{pmatrix}$ and $\mathbf{u} = \begin{pmatrix} 2\\ -1\\ 3 \end{pmatrix}$.

Solution: We cannot apply (5.6.3) and (5.6.4) directly because $\|\mathbf{u}\| \neq 1$, but this is not a problem because

$$\left\|\frac{\mathbf{u}}{\|\mathbf{u}\|}\right\| = 1, \qquad span\{\mathbf{u}\} = span\left\{\frac{\mathbf{u}}{\|\mathbf{u}\|}\right\}, \quad \text{and} \quad \mathbf{u}^\perp = \left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)^\perp.$$

Consequently, the orthogonal projection of $\mathbf{x}$ onto $span\{\mathbf{u}\}$ is given by

$$\left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)\left(\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)^T \mathbf{x} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\mathbf{x} = \frac{1}{2}\begin{pmatrix} 2\\ -1\\ 3 \end{pmatrix},$$

and the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^\perp$ is

$$\left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{x} = \mathbf{x} - \frac{\mathbf{u}\mathbf{u}^T\mathbf{x}}{\mathbf{u}^T\mathbf{u}} = \frac{1}{2}\begin{pmatrix} 2\\ 1\\ -1 \end{pmatrix}.$$

There is nothing special about the numbers in this example. For every nonzero vector $\mathbf{u} \in \mathcal{C}^{n\times 1}$, the orthogonal projectors onto $span\{\mathbf{u}\}$ and $\mathbf{u}^{\perp}$ are

$$\mathbf{P_u} = \frac{\mathbf{uu}^*}{\mathbf{u}^*\mathbf{u}} \quad \text{and} \quad \mathbf{P}_{\mathbf{u}^{\perp}} = \mathbf{I} - \frac{\mathbf{uu}^*}{\mathbf{u}^*\mathbf{u}}. \tag{5.6.6}$$

Elementary Reflectors

For $\mathbf{u}_{n\times 1} \neq \mathbf{0}$, the ***elementary reflector*** about $\mathbf{u}^{\perp}$ is defined to be

$$\mathbf{R} = \mathbf{I} - 2\frac{\mathbf{uu}^*}{\mathbf{u}^*\mathbf{u}} \tag{5.6.7}$$

or, equivalently,

$$\mathbf{R} = \mathbf{I} - 2\mathbf{uu}^* \quad \text{when} \quad \|\mathbf{u}\| = 1. \tag{5.6.8}$$

Elementary reflectors are also called ***Householder transformations,***[46] and they are analogous to the simple reflector given in Example 4.7.1. To understand why, suppose $\mathbf{u} \in \Re^{3\times 1}$ and $\|\mathbf{u}\| = 1$ so that $\mathbf{Q} = \mathbf{I} - \mathbf{uu}^T$ is the orthogonal projector onto the plane $\mathbf{u}^{\perp}$. For each $\mathbf{x} \in \Re^{3\times 1}$, $\mathbf{Qx}$ is the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^{\perp}$ as shown in Figure 5.6.1. To locate $\mathbf{Rx} = (\mathbf{I} - 2\mathbf{uu}^T)\mathbf{x}$, notice that $\mathbf{Q}(\mathbf{Rx}) = \mathbf{Qx}$. In other words, $\mathbf{Qx}$ is simultaneously the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^{\perp}$ as well as the orthogonal projection of $\mathbf{Rx}$ onto $\mathbf{u}^{\perp}$. This together with $\|\mathbf{x} - \mathbf{Qx}\| = |\mathbf{u}^T\mathbf{x}| = \|\mathbf{Qx} - \mathbf{Rx}\|$ implies that $\mathbf{Rx}$ *is the reflection of* $\mathbf{x}$ *about the plane* $\mathbf{u}^{\perp}$, exactly as depicted in Figure 5.6.2. (Reflections about more general subspaces are examined in Exercise 5.13.21.)

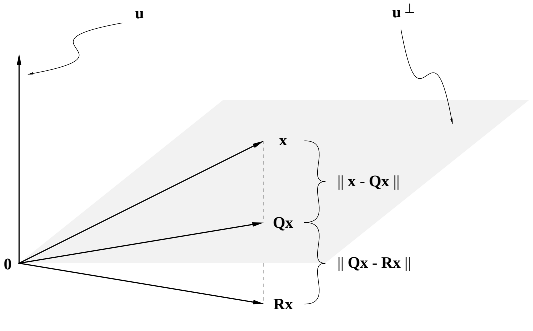


FIGURE 5.6.2

[46] Alston Scott Householder (1904–1993) was one of the first people to appreciate and promote the use of elementary reflectors for numerical applications. Although his 1937 Ph.D. dissertation at University of Chicago concerned the calculus of variations, Householder's passion was mathematical biology, and this was the thrust of his career until it was derailed by the war effort in 1944. Householder joined the Mathematics Division of Oak Ridge National Laboratory in 1946 and became its director in 1948. He stayed at Oak Ridge for the remainder of his career, and he became a leading figure in numerical analysis and matrix computations. Like his counterpart J. Wallace Givens (p. 333) at the Argonne National Laboratory, Householder was one of the early presidents of SIAM.

Properties of Elementary Reflectors

- All elementary reflectors $\mathbf{R}$ are unitary, hermitian, and involutory $(\mathbf{R}^2 = \mathbf{I})$. That is,
$$\mathbf{R} = \mathbf{R}^* = \mathbf{R}^{-1}. \tag{5.6.9}$$
- If $\mathbf{x}_{n\times 1}$ is a vector whose first entry is $x_1 \neq 0$, and if
$$\mathbf{u} = \mathbf{x} \pm \mu \, \|\mathbf{x}\| \, \mathbf{e}_1, \quad \text{where} \quad \mu = \begin{cases} 1 & \text{if } x_1 \text{ is real,} \\ x_1/|x_1| & \text{if } x_1 \text{ is not real,} \end{cases} \tag{5.6.10}$$
is used to build the elementary reflector $\mathbf{R}$ in (5.6.7), then
$$\mathbf{R}\mathbf{x} = \mp\mu \, \|\mathbf{x}\| \, \mathbf{e}_1. \tag{5.6.11}$$
In other words, this $\mathbf{R}$ "reflects" $\mathbf{x}$ onto the first coordinate axis.
Computational Note: To avoid cancellation when using floating-point arithmetic for real matrices, set $\mathbf{u} = \mathbf{x} + \operatorname{sign}(x_1) \, \|\mathbf{x}\| \, \mathbf{e}_1$.

Proof of (5.6.9). It is clear that $\mathbf{R} = \mathbf{R}^*$, and the fact that $\mathbf{R} = \mathbf{R}^{-1}$ is established simply by verifying that $\mathbf{R}^2 = \mathbf{I}$.

Proof of (5.6.10). Observe that $\mathbf{R} = \mathbf{I} - 2\hat{\mathbf{u}}\hat{\mathbf{u}}^*$, where $\hat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|$.

Proof of (5.6.11). Write $\mathbf{R}\mathbf{x} = \mathbf{x} - 2\mathbf{u}\mathbf{u}^*\mathbf{x}/\mathbf{u}^*\mathbf{u} = \mathbf{x} - (2\mathbf{u}^*\mathbf{x}/\mathbf{u}^*\mathbf{u})\mathbf{u}$ and verify that $2\mathbf{u}^*\mathbf{x} = \mathbf{u}^*\mathbf{u}$ to conclude $\mathbf{R}\mathbf{x} = \mathbf{x} - \mathbf{u} = \mp\mu \, \|\mathbf{x}\| \, \mathbf{e}_1$. ■

Example 5.6.3

Problem: Given $\mathbf{x} \in \mathcal{C}^{n\times 1}$ such that $\|\mathbf{x}\| = 1$, construct an orthonormal basis for $\mathcal{C}^n$ that contains $\mathbf{x}$.

Solution: An efficient solution is to build a unitary matrix that contains $\mathbf{x}$ as its first column. Set $\mathbf{u} = \mathbf{x} \pm \mu\mathbf{e}_1$ in $\mathbf{R} = \mathbf{I} - 2(\mathbf{u}\mathbf{u}^*/\mathbf{u}^*\mathbf{u})$ and notice that (5.6.11) guarantees $\mathbf{R}\mathbf{x} = \mp\mu\mathbf{e}_1$, so multiplication on the left by $\mathbf{R}$ (remembering that $\mathbf{R}^2 = \mathbf{I}$) produces $\mathbf{x} = \mp\mu\mathbf{R}\mathbf{e}_1 = [\mp\mu\mathbf{R}]_{*1}$. Since $|\mp\mu| = 1$, $\mathbf{U} = \mp\mu\mathbf{R}$ is a unitary matrix with $\mathbf{U}_{*1} = \mathbf{x}$, so the columns of $\mathbf{U}$ provide the desired orthonormal basis. For example, to construct an orthonormal basis for $\Re^4$ that includes $\mathbf{x} = (1/3)\,(-1 \quad 2 \quad 0 - 2)^T$, set
$$\mathbf{u} = \mathbf{x} - \mathbf{e}_1 = \frac{1}{3}\begin{pmatrix} -4 \\ 2 \\ 0 \\ -2 \end{pmatrix} \text{ and compute } \mathbf{R} = \mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \frac{1}{3}\begin{pmatrix} -1 & 2 & 0 & -2 \\ 2 & 2 & 0 & 1 \\ 0 & 0 & 3 & 0 \\ -2 & 1 & 0 & 2 \end{pmatrix}.$$
The columns of $\mathbf{R}$ do the job.

Now consider rotation, and begin with a basic problem in $\Re^2$. If a nonzero vector $\mathbf{u} = (u_1,\, u_2)$ is rotated counterclockwise through an angle θ to produce $\mathbf{v} = (v_1,\, v_2)$, how are the coordinates of $\mathbf{v}$ related to the coordinates of $\mathbf{u}$? To answer this question, refer to Figure 5.6.3, and use the fact that $\|\mathbf{u}\| = \nu = \|\mathbf{v}\|$ (rotation is an isometry) together with some elementary trigonometry to obtain

$$\begin{aligned} v_1 &= \nu\cos(\phi+\theta) = \nu(\cos\theta\cos\phi - \sin\theta\sin\phi),\\ v_2 &= \nu\sin(\phi+\theta) = \nu(\sin\theta\cos\phi + \cos\theta\sin\phi). \end{aligned} \tag{5.6.12}$$

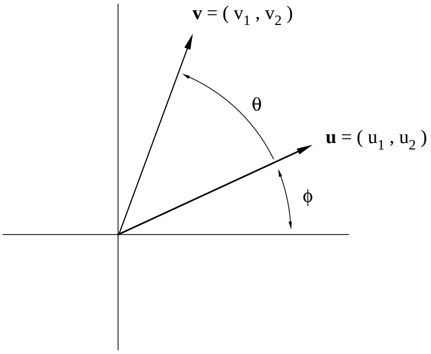


FIGURE 5.6.3

Substituting $\cos\phi = u_1/\nu$ and $\sin\phi = u_2/\nu$ into (5.6.12) yields

$$\begin{aligned} v_1 &= (\cos\theta)u_1 - (\sin\theta)u_2,\\ v_2 &= (\sin\theta)u_1 + (\cos\theta)u_2, \end{aligned} \quad \text{or} \quad \begin{pmatrix} v_1\\ v_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta\\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u_1\\ u_2 \end{pmatrix}. \tag{5.6.13}$$

In other words, $\mathbf{v} = \mathbf{P}\mathbf{u}$, where $\mathbf{P}$ is the ***rotator*** (rotation operator)

$$\mathbf{P} = \begin{pmatrix} \cos\theta & -\sin\theta\\ \sin\theta & \cos\theta \end{pmatrix}. \tag{5.6.14}$$

Notice that $\mathbf{P}$ is an orthogonal matrix because $\mathbf{P}^T\mathbf{P} = \mathbf{I}$. This means that if $\mathbf{v} = \mathbf{P}\mathbf{u}$, then $\mathbf{u} = \mathbf{P}^T\mathbf{v}$, and hence $\mathbf{P}^T$ is also a rotator, but in the opposite direction of that associated with $\mathbf{P}$. That is, $\mathbf{P}^T$ is the rotator associated with the angle $-\theta$. This is confirmed by the fact that if θ is replaced by $-\theta$ in (5.6.14), then $\mathbf{P}^T$ is produced.

Rotating vectors in $\Re^3$ around any one of the coordinate axes is similar. For example, consider rotation around the z-axis. Suppose that $\mathbf{v} = (v_1,\, v_2,\, v_3)$ is obtained by rotating $\mathbf{u} = (u_1,\, u_2,\, u_3)$ counterclockwise[47] through an angle θ around the z-axis. Just as before, the goal is to determine the relationship between the coordinates of $\mathbf{u}$ and $\mathbf{v}$. Since we are rotating around the z-axis,

[47] This is from the perspective of looking down the z-axis onto the xy-plane.

it is evident (see Figure 5.6.4) that the third coordinates are unaffected—i.e., $v_3 = u_3$. To see how the xy-coordinates of $\mathbf{u}$ and $\mathbf{v}$ are related, consider the orthogonal projections

$$\mathbf{u}_p = (u_1,\, u_2,\, 0) \quad \text{and} \quad \mathbf{v}_p = (v_1,\, v_2,\, 0)$$

of $\mathbf{u}$ and $\mathbf{v}$ onto the xy-plane.

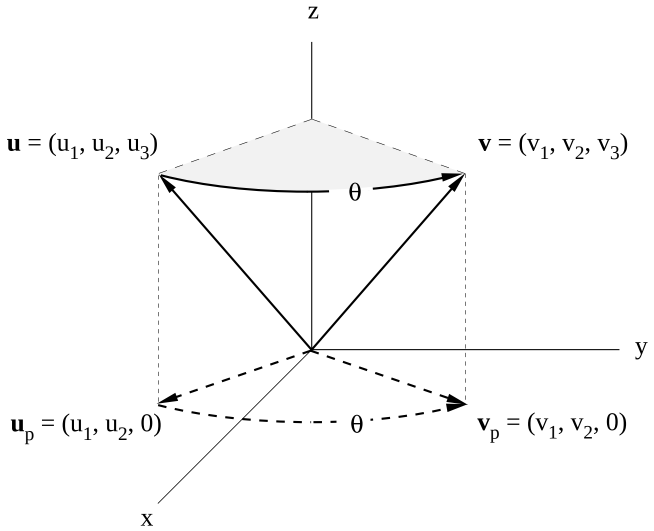


Figure 5.6.4

It's apparent from Figure 5.6.4 that the problem has been reduced to rotation in the xy-plane, and we already know how to do this. Combining (5.6.13) with the fact that $v_3 = u_3$ produces the equation

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix},$$

so

$$\mathbf{P}_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is the matrix that rotates vectors in $\Re^3$ counterclockwise around the z-axis through an angle θ. It is easy to verify that $\mathbf{P}_z$ is an orthogonal matrix and that $\mathbf{P}_z^{-1} = \mathbf{P}_z^T$ rotates vectors *clockwise* around the z-axis.

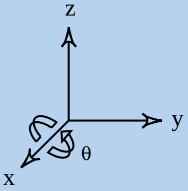
By using similar techniques, it is possible to derive orthogonal matrices that rotate vectors around the x-axis or around the y-axis. Below is a summary of these rotations in $\Re^3$.

Rotations in $\mathbf{R}^3$

A vector $\mathbf{u} \in \Re^3$ can be rotated counterclockwise through an angle θ around a coordinate axis by means of a multiplication $\mathbf{P}_\star \mathbf{u}$ in which $\mathbf{P}_\star$ is an appropriate orthogonal matrix as described below.

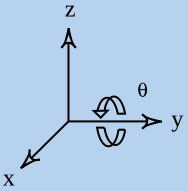
Rotation around the x-Axis

$$\mathbf{P}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}$$



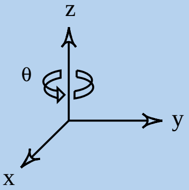
Rotation around the y-Axis

$$\mathbf{P}_y = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}$$



Rotation around the z-Axis

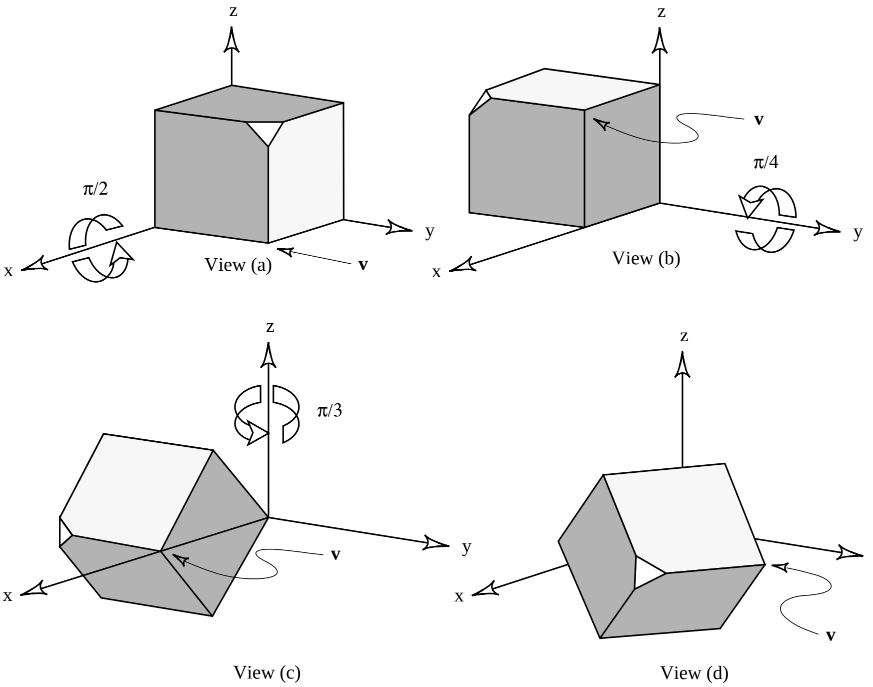
$$\mathbf{P}_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$



Note: The minus sign appears above the diagonal in $\mathbf{P}_x$ and $\mathbf{P}_,$ but below the diagonal in $\mathbf{P}_y$. This is not a mistake—it's due to the orientation of the positive x-axis with respect to the yz-plane.

Example 5.6.4

3-D Rotational Coordinates. Suppose that three counterclockwise rotations are performed on the three-dimensional solid shown in Figure 5.6.5. First rotate the solid in View (a) $90°$ around the x-axis to obtain the orientation shown in View (b). Then rotate View (b) $45°$ around the y-axis to produce View (c) and, finally, rotate View (c) $60°$ around the z-axis to end up with View (d). You can follow the process by watching how the notch, the vertex $\mathbf{v}$, and the lighter shaded face move.



View (a)

View (b)

View (c)

View (d)

FIGURE 5.6.5

Problem: If the coordinates of each vertex in View (a) are specified, what are the coordinates of each vertex in View (d)?

Solution: If $\mathbf{P}_x$ is the rotator that maps points in View (a) to corresponding points in View (b), and if $\mathbf{P}_y$ and $\mathbf{P}_z$ are the respective rotators carrying View (b) to View (c) and View (c) to View (d), then

$$\mathbf{P}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \ \mathbf{P}_y = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ -1 & 0 & 1 \end{pmatrix}, \ \mathbf{P}_z = \begin{pmatrix} 1/2 & -\sqrt{3}/2 & 0 \\ \sqrt{3}/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

so

$$\mathbf{P} = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x = \frac{1}{2\sqrt{2}} \begin{pmatrix} 1 & 1 & \sqrt{6} \\ \sqrt{3} & \sqrt{3} & -\sqrt{2} \\ -2 & 2 & 0 \end{pmatrix} \tag{5.6.15}$$

is the orthogonal matrix that maps points in View (a) to their corresponding images in View (d). For example, focus on the vertex labeled $\mathbf{v}$ in View (a), and let $\mathbf{v}_a$, $\mathbf{v}_b$, $\mathbf{v}_c$, and $\mathbf{v}_d$ denote its respective coordinates in Views (a), (b), (c), and (d). If $\mathbf{v}_a = (\,1 \quad 1 \quad 0\,)^T$, then $\mathbf{v}_b = \mathbf{P}_x\mathbf{v}_a = (\,1 \quad 0 \quad 1\,)^T$,

$$\mathbf{v}_c = \mathbf{P}_y\mathbf{v}_b = \mathbf{P}_y\mathbf{P}_x\mathbf{v}_a = \begin{pmatrix} \sqrt{2} \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{v}_d = \mathbf{P}_z\mathbf{v}_c = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x\mathbf{v}_a = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{6}/2 \\ 0 \end{pmatrix}.$$

More generally, if the coordinates of each of the ten vertices in View (a) are placed as columns in a ***vertex matrix,***

$$\mathbf{V}_a = \begin{array}{c} \begin{array}{cccc} \mathbf{v}_1 & \mathbf{v}_2 & & \mathbf{v}_{10} \\ \downarrow & \downarrow & & \downarrow \end{array} \\ \begin{pmatrix} x_1 & x_2 & \cdots & x_{10} \\ y_1 & y_2 & \cdots & y_{10} \\ z_1 & z_2 & \cdots & z_{10} \end{pmatrix} \end{array}, \quad \text{then} \quad \mathbf{V}_d = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x\mathbf{V}_a = \begin{array}{c} \begin{array}{cccc} \hat{\mathbf{v}}_1 & \hat{\mathbf{v}}_2 & & \hat{\mathbf{v}}_{10} \\ \downarrow & \downarrow & & \downarrow \end{array} \\ \begin{pmatrix} \hat{x}_1 & \hat{x}_2 & \cdots & \hat{x}_{10} \\ \hat{y}_1 & \hat{y}_2 & \cdots & \hat{y}_{10} \\ \hat{z}_1 & \hat{z}_2 & \cdots & \hat{z}_{10} \end{pmatrix} \end{array}$$

is the vertex matrix for the orientation shown in View (d). The polytope in View (d) is drawn by identifying pairs of vertices $(\mathbf{v}_i, \mathbf{v}_j)$ in $\mathbf{V}_a$ that have an edge between them, and by drawing an edge between the corresponding vertices $(\hat{\mathbf{v}}_i, \hat{\mathbf{v}}_j)$ in $\mathbf{V}_d$.

Example 5.6.5

3-D Computer Graphics. Consider the problem of displaying and manipulating views of a three-dimensional solid on a two-dimensional computer display monitor. One simple technique is to use a *wire-frame representation* of the solid consisting of a mesh of points (vertices) on the solid's surface connected by straight line segments (edges). Once these vertices and edges have been defined, the resulting polytope can be oriented in any desired manner as described in Example 5.6.4, so all that remains are the following problems.

Problem: How should the vertices and edges of a three-dimensional polytope be plotted on a two-dimensional computer monitor?

Solution: Assume that the screen represents the yz-plane, and suppose the x-axis is orthogonal to the screen so that it points toward the viewer's eye as shown in Figure 5.6.6.

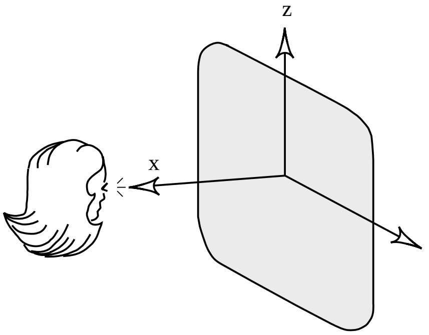


Figure 5.6.6

A solid in the xyz-coordinate system appears to the viewer as the orthogonal projection of the solid onto the yz-plane, and the projection of a polytope is easy to draw. Just set the x-coordinate of each vertex to 0 (i.e., ignore the x-coordinates), plot the (y, z)-coordinates on the yz-plane (the screen), and draw edges between appropriate vertices. For example, suppose that the vertices of the polytope in Figure 5.6.5 are numbered as indicated below in Figure 5.6.7,

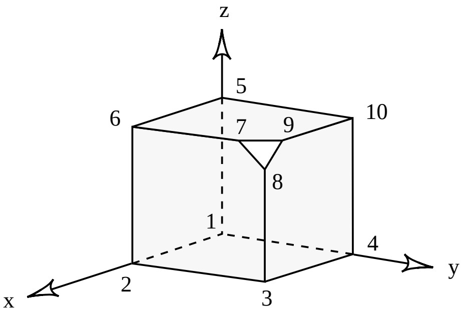


FIGURE 5.6.7

and suppose that the associated vertex matrix is

$$\mathbf{V} = \begin{array}{cc} & \begin{array}{cccccccccc} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 & \mathbf{v}_5 & \mathbf{v}_6 & \mathbf{v}_7 & \mathbf{v}_8 & \mathbf{v}_9 & \mathbf{v}_{10} \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & .8 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & .8 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & .8 & 1 & 1 \end{pmatrix} \end{array}.$$

There are 15 edges, and they can be recorded in an ***edge matrix***

$$\mathbf{E} = \begin{array}{c} \begin{array}{ccccccccccccccc} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 & \mathbf{e}_5 & \mathbf{e}_6 & \mathbf{e}_7 & \mathbf{e}_8 & \mathbf{e}_9 & \mathbf{e}_{10} & \mathbf{e}_{11} & \mathbf{e}_{12} & \mathbf{e}_{13} & \mathbf{e}_{14} & \mathbf{e}_{15} \end{array} \\ \begin{pmatrix} 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 7 & 8 & 9 & 10 \\ 2 & 3 & 4 & 1 & 5 & 6 & 8 & 10 & 6 & 7 & 8 & 9 & 9 & 10 & 5 \end{pmatrix} \end{array}$$

in which the k^{th} column represents an edge between the indicated pair of vertices. To display the image of the polytope in Figure 5.6.7 on a monitor, (i) drop the first row from $\mathbf{V}$, (ii) plot the remaining yz-coordinates on the screen, (iii) draw edges between appropriate vertices as dictated by the information in the edge matrix $\mathbf{E}$. To display the image of the polytope after it has been rotated counterclockwise around the x-, y-, and z-axes by 90°, 45°, and 60°, respectively, use the orthogonal matrix $\mathbf{P} = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x$ determined in (5.6.15) and compute the product

$$\mathbf{PV} = \begin{pmatrix} 0 & .354 & .707 & .354 & .866 & 1.22 & 1.5 & 1.4 & 1.5 & 1.22 \\ 0 & .612 & 1.22 & .612 & -.5 & .112 & .602 & .825 & .602 & .112 \\ 0 & -.707 & 0 & .707 & 0 & -.707 & -.141 & 0 & .141 & .707 \end{pmatrix}.$$

Now proceed as before—(i) ignore the first row of $\mathbf{PV}$, (ii) plot the points in the second and third row of $\mathbf{PV}$ as yz-coordinates on the monitor, (iii) draw edges between appropriate vertices as indicated by the edge matrix $\mathbf{E}$.

Problem: In addition to rotation, how can a polytope (or its image on a computer monitor) be translated?

Solution: Translation of a polytope to a different point in space is accomplished by adding a constant to each of its coordinates. For example, to translate the polytope shown in Figure 5.6.7 to the location where vertex 1 is at $\mathbf{p}^T = (x_0, y_0, z_0)$ instead of at the origin, just add $\mathbf{p}$ to every point. In particular, if $\mathbf{e}$ is the column of 1's, the translated vertex matrix is

$$\mathbf{V}_{trans} = \mathbf{V}_{orig} + \begin{pmatrix} x_0 & x_0 & \cdots & x_0 \\ y_0 & y_0 & \cdots & y_0 \\ z_0 & z_0 & \cdots & z_0 \end{pmatrix} = \mathbf{V}_{orig} + \mathbf{p}\mathbf{e}^T \quad \text{(a rank-1 update)}.$$

Of course, the edge matrix is not affected by translation.

Problem: How can a polytope (or its image on a computer monitor) be scaled?

Solution: Simply multiply every coordinate by the desired scaling factor. For example, to scale an image by a factor α, form the scaled vertex matrix

$$\mathbf{V}_{scaled} = \alpha \mathbf{V}_{orig},$$

and then connect the scaled vertices with appropriate edges as dictated by the edge matrix $\mathbf{E}$.

Problem: How can the faces of a polytope that are hidden from the viewer's perspective be detected so that they can be omitted from the drawing on the screen?

Solution: A complete discussion of this tricky problem would carry us too far astray, but one clever solution relying on the cross product of vectors in $\Re^3$ is presented in Exercise 5.6.21 for the case of *convex* polytopes.

Rotations in higher dimensions are straightforward generalizations of rotations in $\Re^3$. Recall from p. 328 that rotation around any particular axis in $\Re^3$ amounts to rotation in the complementary plane, and the associated 3×3 rotator is constructed by embedding a 2×2 rotator in the appropriate position in a 3×3 identity matrix. For example, rotation around the y-axis is rotation in the xz-plane, and the corresponding rotator is produced by embedding

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

in the "xz-position" of $\mathbf{I}_{3\times 3}$ to form

$$\mathbf{P}_y = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}.$$

These observations directly extend to higher dimensions.

Plane Rotations

Orthogonal matrices of the form

$$\mathbf{P}_{ij} = \begin{array}{cc}
\begin{array}{ccccccccc} & & \text{col } i & & & \text{col } j & & & \\ & & \downarrow & & & \downarrow & & & \end{array} & \\
\left(\begin{array}{ccccccccc} 1 &&&&&&&& \\ & \ddots &&&&&&& \\ && c &&& s &&& \\ &&& 1 &&&&& \\ &&&& \ddots &&&& \\ && -s &&& c &&& \\ &&&&&& 1 && \\ &&&&&&& \ddots & \\ &&&&&&&& 1 \end{array}\right) & \begin{array}{l} \\ \\ \longleftarrow \text{row } i \\ \\ \\ \longleftarrow \text{row } j \\ \\ \\ \\ \end{array}
\end{array}$$

in which $c^2 + s^2 = 1$ are called ***plane rotation matrices*** because they perform a rotation in the (i, j)-plane of $\Re^n$. The entries c and s are meant to suggest cosine and sine, respectively, but designating a rotation angle θ as is done in $\Re^2$ and $\Re^3$ is not useful in higher dimensions.

Plane rotations matrices $\mathbf{P}_{ij}$ are also called ***Givens***[48] ***rotations***. Applying $\mathbf{P}_{ij}$ to $\mathbf{0} \neq \mathbf{x} \in \Re^n$ rotates the (i, j)-coordinates of $\mathbf{x}$ in the sense that

$$\mathbf{P}_{ij}\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ cx_i + sx_j \\ \vdots \\ -sx_i + cx_j \\ \vdots \\ x_n \end{pmatrix} \begin{array}{l} \\ \\ \longleftarrow i \\ \\ \longleftarrow j \\ \\ \\ \end{array}.$$

If x_i and x_j are not both zero, and if we set

$$c = \frac{x_i}{\sqrt{x_i^2 + x_j^2}} \quad \text{and} \quad s = \frac{x_j}{\sqrt{x_i^2 + x_j^2}}, \tag{5.6.16}$$

[48] J. Wallace Givens, Jr. (1910–1993) pioneered the use of plane rotations in the early days of automatic matrix computations. Givens graduated from Lynchburg College in 1928, and he completed his Ph.D. at Princeton University in 1936. After spending three years at the Institute for Advanced Study in Princeton as an assistant of O. Veblen, Givens accepted an appointment at Cornell University but later moved to Northwestern University. In addition to his academic career, Givens was the Director of the Applied Mathematics Division at Argonne National Laboratory and, like his counterpart A. S. Householder (p. 324) at Oak Ridge National Laboratory, Givens served as an early president of SIAM.

then

$$\mathbf{P}_{ij}\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ \sqrt{x_i^2 + x_j^2} \\ \vdots \\ 0 \\ \vdots \\ x_n \end{pmatrix} \begin{matrix} \\ \\ \longleftarrow i \\ \\ \longleftarrow j \\ \\ \\ \end{matrix}.$$

This means that we can selectively annihilate any component—the j^{th} in this case—by a rotation in the (i, j)-plane without affecting any entry except x_i and x_j. Consequently, plane rotations can be applied to annihilate *all* components below any particular "pivot." For example, to annihilate all entries below the first position in $\mathbf{x}$, apply a sequence of plane rotations as follows:

$$\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} \sqrt{x_1^2+x_2^2} \\ 0 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{P}_{13}\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} \sqrt{x_1^2+x_2^2+x_3^2} \\ 0 \\ 0 \\ x_4 \\ \vdots \\ x_n \end{pmatrix}, \quad \ldots, \quad \mathbf{P}_{1n}\cdots\mathbf{P}_{13}\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} \|\mathbf{x}\| \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The product of plane rotations is generally not another plane rotation, but such a product is always an orthogonal matrix, and hence it is an isometry. If we are willing to interpret "rotation in $\Re^n$" as a sequence of plane rotations, then we can say that it is always possible to "rotate" each nonzero vector onto the first coordinate axis. Recall from (5.6.11) that we can also do this with a reflection. More generally, the following statement is true.

Rotations in $\Re^n$

Every nonzero vector $\mathbf{x} \in \Re^n$ can be rotated to the i^{th} coordinate axis by a sequence of $n-1$ plane rotations. In other words, there is an orthogonal matrix $\mathbf{P}$ such that

$$\mathbf{P}\mathbf{x} = \|\mathbf{x}\|\,\mathbf{e}_i, \tag{5.6.17}$$

where $\mathbf{P}$ has the form

$$\mathbf{P} = \mathbf{P}_{in} \cdots \mathbf{P}_{i,i+1}\mathbf{P}_{i,i-1} \cdots \mathbf{P}_{i1}.$$

Example 5.6.6

Problem: If $\mathbf{x} \in \Re^n$ is a vector such that $\|\mathbf{x}\| = 1$, explain how to use plane rotations to construct an orthonormal basis for $\Re^n$ that contains $\mathbf{x}$.

Solution: This is almost the same problem as that posed in Example 5.6.3, and, as explained there, the goal is to construct an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}_{*1} = \mathbf{x}$. But this time we need to use plane rotations rather than an elementary reflector. Equation (5.6.17) asserts that we can build an orthogonal matrix from a sequence of plane rotations $\mathbf{P} = \mathbf{P}_{1n} \cdots \mathbf{P}_{13}\mathbf{P}_{12}$ such that $\mathbf{P}\mathbf{x} = \mathbf{e}_1$. Thus $\mathbf{x} = \mathbf{P}^T\mathbf{e}_1 = \mathbf{P}^T_{*1}$, and hence the columns of $\mathbf{Q} = \mathbf{P}^T$ serve the purpose. For example, to extend

$$\mathbf{x} = \frac{1}{3}\begin{pmatrix} -1 \\ 2 \\ 0 \\ -2 \end{pmatrix}$$

to an orthonormal basis for $\Re^4$, sequentially annihilate the second and fourth components of $\mathbf{x}$ by using (5.6.16) to construct the following plane rotations:

$$\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} -1/\sqrt{5} & 2/\sqrt{5} & 0 & 0 \\ -2/\sqrt{5} & -1/\sqrt{5} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \frac{1}{3}\begin{pmatrix} -1 \\ 2 \\ 0 \\ -2 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} \sqrt{5} \\ 0 \\ 0 \\ -2 \end{pmatrix},$$

$$\mathbf{P}_{14}\Big(\mathbf{P}_{12}\mathbf{x}\Big) = \begin{pmatrix} \sqrt{5}/3 & 0 & 0 & -2/3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2/3 & 0 & 0 & \sqrt{5}/3 \end{pmatrix} \frac{1}{3}\begin{pmatrix} \sqrt{5} \\ 0 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Therefore, the columns of

$$\mathbf{Q} = \left(\mathbf{P}_{14}\mathbf{P}_{12}\right)^T = \mathbf{P}_{12}^T\mathbf{P}_{14}^T = \begin{pmatrix} -1/3 & -2/\sqrt{5} & 0 & -2/3\sqrt{5} \\ 2/3 & -1/\sqrt{5} & 0 & 4/3\sqrt{5} \\ 0 & 0 & 1 & 0 \\ -2/3 & 0 & 0 & \sqrt{5}/3 \end{pmatrix}$$

are an orthonormal set containing the specified vector $\mathbf{x}$.

Exercises for section 5.6

5.6.1. Determine which of the following matrices are isometries.

(a) $\begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix}$. (b) $\begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$.

(c) $\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$. (d) $\begin{pmatrix} e^{i\theta_1} & 0 & \cdots & 0 \\ 0 & e^{i\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{i\theta_n} \end{pmatrix}$.

5.6.2. Is $\begin{pmatrix} \dfrac{1+i}{\sqrt{3}} & \dfrac{1+i}{\sqrt{6}} \\ \dfrac{i}{\sqrt{3}} & \dfrac{-2\,i}{\sqrt{6}} \end{pmatrix}$ a unitary matrix?

5.6.3. (a) How many 3×3 matrices are both diagonal and orthogonal?

(b) How many $n \times n$ matrices are both diagonal and orthogonal?

(c) How many $n \times n$ matrices are both diagonal and unitary?

5.6.4. (a) Under what conditions on the real numbers α and β will

$$\mathbf{P} = \begin{pmatrix} \alpha+\beta & \beta-\alpha \\ \alpha-\beta & \beta+\alpha \end{pmatrix}$$

be an orthogonal matrix?

(b) Under what conditions on the real numbers α and β will

$$\mathbf{U} = \begin{pmatrix} 0 & \alpha & 0 & \mathrm{i}\beta \\ \alpha & 0 & \mathrm{i}\beta & 0 \\ 0 & \mathrm{i}\beta & 0 & \alpha \\ \mathrm{i}\beta & 0 & \alpha & 0 \end{pmatrix}$$

be a unitary matrix?

5.6.5. Let $\mathbf{U}$ and $\mathbf{V}$ be two $n \times n$ unitary (orthogonal) matrices.

(a) Explain why the product $\mathbf{UV}$ must be unitary (orthogonal).

(b) Explain why the sum $\mathbf{U}+\mathbf{V}$ need not be unitary (orthogonal).

(c) Explain why $\begin{pmatrix} \mathbf{U}_{n\times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{m\times m} \end{pmatrix}$ must be unitary (orthogonal).

5.6.6. Cayley Transformation. Prove, as Cayley did in 1846, that if $\mathbf{A}$ is skew hermitian (or real skew symmetric), then

$$\mathbf{U} = (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A})$$

is unitary (orthogonal) by first showing that $(\mathbf{I}+\mathbf{A})^{-1}$ exists for skew-hermitian matrices, and $(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A})$ (recall Exercise 3.7.6). **Note:** There is a more direct approach, but it requires the diagonalization theorem for normal matrices—see Exercise 7.5.5.

5.6.7. Suppose that $\mathbf{R}$ and $\mathbf{S}$ are elementary reflectors.

(a) Is $\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$ an elementary reflector?

(b) Is $\begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}$ an elementary reflector?

5.6.8. (a) Explain why the standard inner product is invariant under a unitary transformation. That is, if $\mathbf{U}$ is any unitary matrix, and if $\mathbf{u} = \mathbf{Ux}$ and $\mathbf{v} = \mathbf{Uy}$, then

$$\mathbf{u}^*\mathbf{v} = \mathbf{x}^*\mathbf{y}.$$

(b) Given any two vectors $\mathbf{x}, \mathbf{y} \in \Re^n$, explain why the angle between them is invariant under an orthogonal transformation. That is, if $\mathbf{u} = \mathbf{Px}$ and $\mathbf{v} = \mathbf{Py}$, where $\mathbf{P}$ is an orthogonal matrix, then

$$\cos\theta_{\mathbf{u},\mathbf{v}} = \cos\theta_{\mathbf{x},\mathbf{y}}.$$

5.6.9. Let $\mathbf{U}_{m\times r}$ be a matrix with orthonormal columns, and let $\mathbf{V}_{k\times n}$ be a matrix with orthonormal rows. For an arbitrary $\mathbf{A} \in \mathcal{C}^{r\times k}$, solve the following problems using the matrix 2-norm (p. 281) and the Frobenius matrix norm (p. 279).

(a) Determine the values of $\|\mathbf{U}\|_2$, $\|\mathbf{V}\|_2$, $\|\mathbf{U}\|_F$, and $\|\mathbf{V}\|_F$.

(b) Show that $\|\mathbf{UAV}\|_2 = \|\mathbf{A}\|_2$. (**Hint:** Start with $\|\mathbf{UA}\|_2$.)

(c) Show that $\|\mathbf{UAV}\|_F = \|\mathbf{A}\|_F$.

Note: In particular, these properties are valid when $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. Because of parts (b) and (c), the 2-norm and the F-norm are said to be ***unitarily invariant norms.***

5.6.10. Let $\mathbf{u} = \begin{pmatrix} -2 \\ 1 \\ 3 \\ -1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ 4 \\ 0 \\ -1 \end{pmatrix}$.

(a) Determine the orthogonal projection of $\mathbf{u}$ onto $span\{\mathbf{v}\}$.

(b) Determine the orthogonal projection of $\mathbf{v}$ onto $span\{\mathbf{u}\}$.

(c) Determine the orthogonal projection of $\mathbf{u}$ onto $\mathbf{v}^\perp$.

(d) Determine the orthogonal projection of $\mathbf{v}$ onto $\mathbf{u}^\perp$.

5.6.11. Consider elementary orthogonal projectors $\mathbf{Q} = \mathbf{I} - \mathbf{uu}^*$.

(a) Prove that $\mathbf{Q}$ is singular.

(b) Now prove that if $\mathbf{Q}$ is $n \times n$, then $rank(\mathbf{Q}) = n - 1$.
Hint: Recall Exercise 4.4.10.

5.6.12. For vectors $\mathbf{u}, \mathbf{x} \in \mathcal{C}^n$ such that $\|\mathbf{u}\| = 1$, let $\mathbf{p}$ be the orthogonal projection of $\mathbf{x}$ onto $span\{\mathbf{u}\}$. Explain why $\|\mathbf{p}\| \le \|\mathbf{x}\|$ with equality holding if and only if $\mathbf{x}$ is a scalar multiple of $\mathbf{u}$.

5.6.13. Let $\mathbf{x} = (1/3)\begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix}$.

(a) Determine an elementary reflector $\mathbf{R}$ such that $\mathbf{Rx}$ lies on the x-axis.

(b) Verify by direct computation that your reflector $\mathbf{R}$ is symmetric, orthogonal, and involutory.

(c) Extend $\mathbf{x}$ to an orthonormal basis for $\Re^3$ by using an elementary reflector.

5.6.14. Let $\mathbf{R} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^*$, where $\|\mathbf{u}_{n\times 1}\| = 1$. If $\mathbf{x}$ is a *fixed point* for $\mathbf{R}$ in the sense that $\mathbf{Rx} = \mathbf{x}$, and if $n > 1$, prove that $\mathbf{x}$ must be orthogonal to $\mathbf{u}$, and then sketch a picture of this situation in $\Re^3$.

5.6.15. Let $\mathbf{x}, \mathbf{y} \in \Re^{n\times 1}$ be vectors such that $\|\mathbf{x}\| = \|\mathbf{y}\|$ but $\mathbf{x} \neq \mathbf{y}$. Explain how to construct an elementary reflector $\mathbf{R}$ such that $\mathbf{Rx} = \mathbf{y}$.
Hint: The vector $\mathbf{u}$ that defines $\mathbf{R}$ can be determined visually in $\Re^3$ by considering Figure 5.6.2.

5.6.16. Let $\mathbf{x}_{n\times 1}$ be a vector such that $\|\mathbf{x}\| = 1$, and partition $\mathbf{x}$ as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \tilde{\mathbf{x}} \end{pmatrix}, \qquad \text{where } \tilde{\mathbf{x}} \text{ is } n-1\times 1.$$

(a) If the entries of $\mathbf{x}$ are real, and if $x_1 \neq 1$, show that

$$\mathbf{P} = \begin{pmatrix} x_1 & \tilde{\mathbf{x}}^T \\ \tilde{\mathbf{x}} & \mathbf{I} - \alpha\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T \end{pmatrix}, \qquad \text{where} \quad \alpha = \frac{1}{1-x_1}$$

is an orthogonal matrix.

(b) Suppose that the entries of $\mathbf{x}$ are complex. If $|x_1| \neq 1$, and if μ is the number defined in (5.6.10), show that the matrix

$$\mathbf{U} = \begin{pmatrix} x_1 & \mu^2\tilde{\mathbf{x}}^* \\ \tilde{\mathbf{x}} & \mu(\mathbf{I} - \alpha\tilde{\mathbf{x}}\tilde{\mathbf{x}}^*) \end{pmatrix}, \qquad \text{where} \quad \alpha = \frac{1}{1-|x_1|}$$

is unitary. **Note:** These results provide an easy way to extend a given vector to an orthonormal basis for the entire space $\Re^n$ or $\mathcal{C}^n$.

5.6.17. Perform the following sequence of rotations in $\Re^3$ beginning with

$$\mathbf{v}_0 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

1. Rotate $\mathbf{v}_0$ counterclockwise 45° around the x-axis to produce $\mathbf{v}_1$.
2. Rotate $\mathbf{v}_1$ clockwise 90° around the y-axis to produce $\mathbf{v}_2$.
3. Rotate $\mathbf{v}_2$ counterclockwise 30° around the z-axis to produce $\mathbf{v}_3$.

Determine the coordinates of $\mathbf{v}_3$ as well as an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{v}_0 = \mathbf{v}_3$.

5.6.18. Does it matter in what order rotations in $\Re^3$ are performed? For example, suppose that a vector $\mathbf{v} \in \Re^3$ is first rotated counterclockwise around the x-axis through an angle θ, and then that vector is rotated counterclockwise around the y-axis through an angle ϕ. Is the result the same as first rotating $\mathbf{v}$ counterclockwise around the y-axis through an angle ϕ followed by a rotation counterclockwise around the x-axis through an angle θ?

5.6.19. For each nonzero vector $\mathbf{u} \in \mathcal{C}^n$, prove that $\dim \mathbf{u}^{\perp} = n - 1$.

5.6.20. A matrix satisfying $\mathbf{A}^2 = \mathbf{I}$ is said to be an ***involution*** or an ***involutory matrix***, and a matrix $\mathbf{P}$ satisfying $\mathbf{P}^2 = \mathbf{P}$ is called a ***projector*** or is said to be an ***idempotent matrix***—properties of such matrices are developed on p. 386. Show that there is a one-to-one correspondence between the set of involutions and the set of projectors in $\mathcal{C}^{n\times n}$. **Hint:** Consider the relationship between the projectors in (5.6.6) and the reflectors (which are involutions) in (5.6.7) on p. 324.

5.6.21. When using a computer to generate and display a three-dimensional convex polytope such as the one in Example 5.6.4, it is desirable to not draw those faces that should be hidden from the perspective of a viewer positioned as shown in Figure 5.6.6. The operation of ***cross product*** in $\Re^3$ (usually introduced in elementary calculus courses) can be used to decide which faces are visible and which are not. Recall that if

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad \text{then} \quad \mathbf{u} \times \mathbf{v} = \begin{pmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{pmatrix},$$

and $\mathbf{u}\times\mathbf{v}$ is a vector orthogonal to both $\mathbf{u}$ and $\mathbf{v}$. The direction of $\mathbf{u}\times\mathbf{v}$ is determined from the so-called right-hand rule as illustrated in Figure 5.6.8.

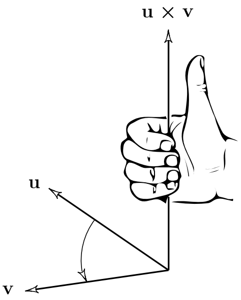


Figure 5.6.8

Assume the origin is interior to the polytope, and consider a particular face and three vertices $\mathbf{p}_0$, $\mathbf{p}_1$, and $\mathbf{p}_2$ on the face that are positioned as shown in Figure 5.6.9. The vector $\mathbf{n} = (\mathbf{p}_1 - \mathbf{p}_0) \times (\mathbf{p}_2 - \mathbf{p}_1)$ is orthogonal to the face, and it points in the outward direction.

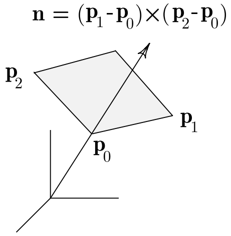


Figure 5.6.9

Explain why the outside of the face is visible from the perspective indicated in Figure 5.6.6 if and only if the first component of the outward normal vector $\mathbf{n}$ is positive. In other words, the face is drawn if and only if $n_1 > 0$.

5.7 ORTHOGONAL REDUCTION

We know that a matrix $\mathbf{A}$ can be reduced to row echelon form by elementary row operations. This is Gaussian elimination, and, as explained on p. 143, the basic "Gaussian transformation" is an elementary lower triangular matrix $\mathbf{T}_k$ whose action annihilates all entries below the k^{th} pivot at the k^{th} elimination step. But Gaussian elimination is not the only way to reduce a matrix. Elementary reflectors $\mathbf{R}_k$ can be used in place of elementary lower triangular matrices $\mathbf{T}_k$ to annihilate all entries below the k^{th} pivot at the k^{th} elimination step, or a sequence of plane rotation matrices can accomplish the same purpose.

When reflectors are used, the process is usually called ***Householder reduction***, and it proceeds as follows. For $\mathbf{A}_{m\times n} = [\mathbf{A}_{*1}\,|\,\mathbf{A}_{*2}\,|\,\cdots\,|\,\mathbf{A}_{*n}]$, use $\mathbf{x} = \mathbf{A}_{*1}$ in (5.6.10) to construct the elementary reflector

$$\mathbf{R}_1 = \mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}, \quad \text{where} \quad \mathbf{u} = \mathbf{A}_{*1} \pm \mu\,\|\mathbf{A}_{*1}\|\,\mathbf{e}_1, \tag{5.7.1}$$

so that

$$\mathbf{R}_1\mathbf{A}_{*1} = \mp\mu\,\|\mathbf{A}_{*1}\|\,\mathbf{e}_1 = \begin{pmatrix} t_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{5.7.2}$$

Applying $\mathbf{R}_1$ to $\mathbf{A}$ yields

$$\mathbf{R}_1\mathbf{A} = [\mathbf{R}_1\mathbf{A}_{*1}\,|\,\mathbf{R}_1\mathbf{A}_{*2}\,|\,\cdots\,|\,\mathbf{R}_1\mathbf{A}_{*n}] = \left(\begin{array}{c|ccc} t_{11} & t_{12} & \cdots & t_{1n} \\ \hline 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{array}\right) = \begin{pmatrix} t_{11} & \mathbf{t}_1^T \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix},$$

where $\mathbf{A}_2$ is $m-1 \times n-1$. Thus all entries below the $(1,1)$-position are annihilated. Now apply the same procedure to $\mathbf{A}_2$ to construct an elementary reflector $\hat{\mathbf{R}}_2$ that annihilates all entries below the $(1,1)$-position in $\mathbf{A}_2$. If we set $\mathbf{R}_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$, then $\mathbf{R}_2\mathbf{R}_1$ is an orthogonal matrix (Exercise 5.6.5) such that

$$\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} t_{11} & \mathbf{t}_1^T \\ \mathbf{0} & \hat{\mathbf{R}}_2\mathbf{A}_2 \end{pmatrix} = \left(\begin{array}{c|c|ccc} t_{11} & t_{12} & t_{13} & \cdots & t_{1n} \\ \hline 0 & t_{22} & t_{23} & \cdots & t_{2n} \\ \hline 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{array}\right).$$

The result after $k-1$ steps is $\mathbf{R}_{k-1}\cdots\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} \mathbf{T}_{k-1} & \tilde{\mathbf{T}}_{k-1} \\ \mathbf{0} & \mathbf{A}_k \end{pmatrix}$. At step k an elementary reflector $\hat{\mathbf{R}}_k$ is constructed in a manner similar to (5.7.1) to annihilate all entries below the $(1,1)$-position in $\mathbf{A}_k$, and $\mathbf{R}_k$ is defined as $\mathbf{R}_k = \begin{pmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_k \end{pmatrix}$, which is another elementary reflector (Exercise 5.6.7). Eventually, all of the rows or all of the columns will be exhausted, so the final result is one of the two following ***upper-trapezoidal forms***:

$$\mathbf{R}_n \cdots \mathbf{R}_2\mathbf{R}_1\mathbf{A}_{m\times n} = \left(\begin{array}{cccc} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & * \\ \hline 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{array}\right) \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} n\times n \qquad \text{when} \quad m > n,$$

$$\mathbf{R}_{m-1} \cdots \mathbf{R}_2\mathbf{R}_1\mathbf{A}_{m\times n} = \left(\begin{array}{cccc|ccc} * & * & \cdots & * & * & \cdots & * \\ 0 & * & \cdots & * & * & \cdots & * \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & * & * & \cdots & * \end{array}\right) \qquad \text{when} \quad m < n.$$
$$\underbrace{\qquad\qquad\qquad}_{m\times m}$$

If $m = n$, then the final form is an upper-triangular matrix.

A product of elementary reflectors is not necessarily another elementary reflector, but a product of unitary (orthogonal) matrices is again unitary (orthogonal) (Exercise 5.6.5). The elementary reflectors $\mathbf{R}_i$ described above are unitary (orthogonal in the real case) matrices, so every product $\mathbf{R}_k\mathbf{R}_{k-1}\cdots\mathbf{R}_2\mathbf{R}_1$ is a unitary matrix, and thus we arrive at the following important conclusion.

Orthogonal Reduction

- For every $\mathbf{A} \in \mathcal{C}^{m\times n}$, there exists a unitary matrix $\mathbf{P}$ such that

$$\mathbf{PA} = \mathbf{T} \tag{5.7.3}$$

  has an upper-trapezoidal form. When $\mathbf{P}$ is constructed as a product of elementary reflectors as described above, the process is called ***Householder reduction.***

- If $\mathbf{A}$ is square, then $\mathbf{T}$ is upper triangular, and if $\mathbf{A}$ is real, then the $\mathbf{P}$ can be taken to be an orthogonal matrix.

Example 5.7.1

Problem: Use Householder reduction to find an orthogonal matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{T}$ is upper triangular with positive diagonal entries, where

$$\mathbf{A} = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}.$$

Solution: To annihilate the entries below the $(1,1)$-position and to guarantee that t_{11} is positive, equations (5.7.1) and (5.7.2) dictate that we set

$$\mathbf{u}_1 = \mathbf{A}_{*1} - \|\mathbf{A}_{*1}\|\,\mathbf{e}_1 = \mathbf{A}_{*1} - 5\mathbf{e}_1 = \begin{pmatrix} -5 \\ 3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_1 = \mathbf{I} - 2\frac{\mathbf{u}_1\mathbf{u}_1^T}{\mathbf{u}_1^T\mathbf{u}_1}.$$

To compute a reflector-by-matrix product $\mathbf{RA} = [\mathbf{RA}_{*1} \,|\, \mathbf{RA}_{*2} \,|\, \cdots \,|\, \mathbf{RA}_{*n}]$, it's wasted effort to actually determine the entries in $\mathbf{R} = \mathbf{I} - 2\mathbf{uu}^T/\mathbf{u}^T\mathbf{u}$. Simply compute $\mathbf{u}^T\mathbf{A}_{*j}$ and then

$$\mathbf{RA}_{*j} = \mathbf{A}_{*j} - 2\left(\frac{\mathbf{u}^T\mathbf{A}_{*j}}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u} \quad \text{for each } j = 1, 2, \ldots, n. \tag{5.7.4}$$

By using this observation we obtain

$$\mathbf{R}_1\mathbf{A} = [\mathbf{R}_1\mathbf{A}_{*1} \,|\, \mathbf{R}_1\mathbf{A}_{*2} \,|\, \mathbf{R}_1\mathbf{A}_{*3}] = \left(\begin{array}{c|cc} 5 & 25 & -4 \\ \hline 0 & 0 & -10 \\ 0 & -25 & -10 \end{array}\right).$$

To annihilate the entry below the $(2,2)$-position, set

$$\mathbf{A}_2 = \begin{pmatrix} 0 & -10 \\ -25 & -10 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 = [\mathbf{A}_2]_{*1} - \left\| [\mathbf{A}_2]_{*1} \right\| \mathbf{e}_1 = 25\begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

If $\hat{\mathbf{R}}_2 = \mathbf{I} - 2\mathbf{u}_2\mathbf{u}_2^T/\mathbf{u}_2^T\mathbf{u}_2$ and $\mathbf{R}_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$ (neither is explicitly computed), then

$$\hat{\mathbf{R}}_2\mathbf{A}_2 = \begin{pmatrix} 25 & 10 \\ 0 & 10 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_2\mathbf{R}_1\mathbf{A} = \mathbf{T} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}.$$

If $\hat{\mathbf{R}}_k = \mathbf{I} - 2\hat{\mathbf{u}}\hat{\mathbf{u}}^T/\hat{\mathbf{u}}^T\hat{\mathbf{u}}$ is an elementary reflector, then so is

$$\mathbf{R}_k = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_k \end{pmatrix} = \mathbf{I} - 2\frac{\mathbf{uu}^T}{\mathbf{u}^T\mathbf{u}} \quad \text{with} \quad \mathbf{u} = \begin{pmatrix} \mathbf{0} \\ \hat{\mathbf{u}} \end{pmatrix},$$

and consequently the product of any sequence of these $\mathbf{R}_k$'s can be formed by using the observation (5.7.4). In this example,

$$\mathbf{P} = \mathbf{R}_2\mathbf{R}_1 = \frac{1}{25}\begin{pmatrix} 0 & 15 & 20 \\ -20 & 12 & -9 \\ -15 & -16 & 12 \end{pmatrix}.$$

You may wish to check that $\mathbf{P}$ really is an orthogonal matrix and $\mathbf{PA} = \mathbf{T}$.

Elementary reflectors are not the only type of orthogonal matrices that can be used to reduce a matrix to an upper-trapezoidal form. Plane rotation matrices are also orthogonal, and, as explained on p. 334, plane rotation matrices can be used to selectively annihilate any component in a given column, so a sequence of plane rotations can be used to annihilate all elements below a particular pivot. This means that a matrix $\mathbf{A} \in \Re^{m\times n}$ can be reduced to an upper-trapezoidal form strictly by using plane rotations—such a process is usually called a ***Givens reduction.***

Example 5.7.2

Problem: Use Givens reduction (i.e., use plane rotations) to reduce the matrix

$$\mathbf{A} = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}$$

to upper-triangular form. Also compute an orthogonal matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{T}$ is upper triangular.

Solution: The plane rotation that uses the (1,1)-entry to annihilate the (2,1)-entry is determined from (5.6.16) to be

$$\mathbf{P}_{12} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{so that} \quad \mathbf{P}_{12}\mathbf{A} = \begin{pmatrix} 3 & 27 & -4 \\ 0 & 20 & 14 \\ 4 & 11 & -2 \end{pmatrix}.$$

Now use the (1,1)-entry in $\mathbf{P}_{12}\mathbf{A}$ to annihilate the (3,1)-entry in $\mathbf{P}_{12}\mathbf{A}$. The plane rotation that does the job is again obtained from (5.6.16) to be

$$\mathbf{P}_{13} = \frac{1}{5}\begin{pmatrix} 3 & 0 & 4 \\ 0 & 5 & 0 \\ -4 & 0 & 3 \end{pmatrix} \quad \text{so that} \quad \mathbf{P}_{13}\mathbf{P}_{12}\mathbf{A} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 20 & 14 \\ 0 & -15 & 2 \end{pmatrix}.$$

Finally, using the (2,2)-entry in $\mathbf{P}_{13}\mathbf{P}_{12}\mathbf{A}$ to annihilate the (3,2)-entry produces

$$\mathbf{P}_{23} = \frac{1}{5}\begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & -3 \\ 0 & 3 & 4 \end{pmatrix} \quad \text{so that} \quad \mathbf{P}_{23}\mathbf{P}_{13}\mathbf{P}_{12}\mathbf{A} = \mathbf{T} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}.$$

Since plane rotation matrices are orthogonal, and since the product of orthogonal matrices is again orthogonal, it must be the case that

$$\mathbf{P} = \mathbf{P}_{23}\mathbf{P}_{13}\mathbf{P}_{12} = \frac{1}{25}\begin{pmatrix} 0 & 15 & 20 \\ -20 & 12 & -9 \\ -15 & -16 & 12 \end{pmatrix}$$

is an orthogonal matrix such that $\mathbf{PA} = \mathbf{T}$.

Householder and Givens reductions are closely related to the results produced by applying the Gram–Schmidt process (p. 307) to the columns of $\mathbf{A}$. When $\mathbf{A}$ is nonsingular, Householder, Givens, and Gram–Schmidt each produce an orthogonal matrix $\mathbf{Q}$ and an upper-triangular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$ ($\mathbf{Q} = \mathbf{P}^T$ in the case of orthogonal reduction). The upper-triangular matrix $\mathbf{R}$ produced by the Gram–Schmidt algorithm has positive diagonal entries, and, as illustrated in Examples 5.7.1 and 5.7.2, we can also force this to be true using the Householder or Givens reduction. This feature makes $\mathbf{Q}$ and $\mathbf{R}$ unique.

QR Factorization

For each nonsingular $\mathbf{A} \in \Re^{n\times n}$, there is a unique orthogonal matrix $\mathbf{Q}$ and a unique upper-triangular matrix $\mathbf{R}$ with positive diagonal entries such that

$$\mathbf{A} = \mathbf{QR}.$$

This "square" QR factorization is a special case of the more general "rectangular" QR factorization discussed on p. 311.

Proof. Only uniqueness needs to be proven. If there are two QR factorizations

$$\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2,$$

let $\mathbf{U} = \mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{R}_2\mathbf{R}_1^{-1}$. The matrix $\mathbf{R}_2\mathbf{R}_1^{-1}$ is upper triangular with positive diagonal entries (Exercises 3.5.8 and 3.7.4) while $\mathbf{Q}_2^T\mathbf{Q}_1$ is an orthogonal matrix (Exercise 5.6.5), and therefore $\mathbf{U}$ is an upper-triangular matrix whose columns are an orthonormal set and whose diagonal entries are positive. Considering the first column of $\mathbf{U}$ we see that

$$\left\|\begin{pmatrix} u_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right\| = 1 \quad\Longrightarrow\quad u_{11} = \pm 1 \quad \text{and} \quad u_{11} > 0 \quad\Longrightarrow\quad u_{11} = 1,$$

so that $\mathbf{U}_{*1} = \mathbf{e}_1$. A similar argument together with the fact that the columns of $\mathbf{U}$ are mutually orthogonal produces

$$\mathbf{U}_{*1}^T\mathbf{U}_{*2} = 0 \quad\Longrightarrow\quad u_{12} = 0 \quad\Longrightarrow\quad u_{22} = 1 \quad\Longrightarrow\quad \mathbf{U}_{*2} = \mathbf{e}_2.$$

Proceeding inductively establishes that $\mathbf{U}_{*k} = \mathbf{e}_k$ for each k (i.e., $\mathbf{U} = \mathbf{I}$), and therefore $\mathbf{Q}_1 = \mathbf{Q}_2$ and $\mathbf{R}_1 = \mathbf{R}_2$. ■

Example 5.7.3

Orthogonal Reduction and Least Squares. Orthogonal reduction can be used to solve the least squares problem associated with an inconsistent system $\mathbf{Ax} = \mathbf{b}$ in which $\mathbf{A} \in \Re^{m\times n}$ and $m \geq n$ (the most common case). If $\boldsymbol{\varepsilon}$ denotes the difference $\boldsymbol{\varepsilon} = \mathbf{Ax} - \mathbf{b}$, then, as described on p. 226, the general least squares problem is to find a vector $\mathbf{x}$ that minimizes the quantity

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \|\boldsymbol{\varepsilon}\|^2,$$

where $\|\star\|$ is the standard euclidean vector norm. Suppose that $\mathbf{A}$ is reduced to an upper-trapezoidal matrix $\mathbf{T}$ by an orthogonal matrix $\mathbf{P}$, and write

$$\mathbf{PA} = \mathbf{T} = \begin{pmatrix} \mathbf{R}_{n\times n} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{Pb} = \begin{pmatrix} \mathbf{c}_{n\times 1} \\ \mathbf{d} \end{pmatrix}$$

in which $\mathbf{R}$ is an upper-triangular matrix. An orthogonal matrix is an isometry—recall (5.6.1)—so that

$$\|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{P}\boldsymbol{\varepsilon}\|^2 = \|\mathbf{P}(\mathbf{Ax}-\mathbf{b})\|^2 = \left\| \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} \mathbf{Rx}-\mathbf{c} \\ \mathbf{d} \end{pmatrix} \right\|^2$$
$$= \|\mathbf{Rx}-\mathbf{c}\|^2 + \|\mathbf{d}\|^2.$$

Consequently, $\|\boldsymbol{\varepsilon}\|^2$ is minimized when $\mathbf{x}$ is a vector such that $\|\mathbf{Rx}-\mathbf{c}\|^2$ is minimal or, in other words, $\mathbf{x}$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$ if and only if $\mathbf{x}$ is a least squares solution for $\mathbf{Rx} = \mathbf{c}$.

Full-Rank Case. In a majority of applications the coefficient matrix $\mathbf{A}$ has linearly independent columns so $rank\,(\mathbf{A}_{m\times n}) = n$. Because multiplication by a nonsingular matrix $\mathbf{P}$ does not change the rank,

$$n = rank\,(\mathbf{A}) = rank\,(\mathbf{PA}) = rank\,(\mathbf{T}) = rank\,(\mathbf{R}_{n\times n}).$$

Thus $\mathbf{R}$ is nonsingular, and we have established the following fact.

- If $\mathbf{A}$ has linearly independent columns, then the (unique) least squares solution for $\mathbf{Ax} = \mathbf{b}$ is obtained by solving the nonsingular triangular system $\mathbf{Rx} = \mathbf{c}$ for $\mathbf{x}$.

As pointed out in Example 4.5.1, computing the matrix product $\mathbf{A}^T\mathbf{A}$ is to be avoided when floating-point computation is used because of the possible loss of significant information. Notice that the method based on orthogonal reduction sidesteps this potential problem because the normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ are avoided and the product $\mathbf{A}^T\mathbf{A}$ is never explicitly computed. Householder reduction (or Givens reduction for sparse problems) is a numerically stable algorithm (see the discussion following this example) for solving the full-rank least squares problem, and, if the computations are properly ordered, it is an attractive alternative to the method of Example 5.5.3 that is based on the modified Gram–Schmidt procedure.

We now have four different ways to reduce a matrix to an upper-triangular (or trapezoidal) form. (1) Gaussian elimination; (2) Gram–Schmidt procedure; (3) Householder reduction; and (4) Givens reduction. It's natural to try to compare them and to sort out the advantages and disadvantages of each.

First consider numerical stability. This is a complicated issue, but you can nevertheless gain an intuitive feel for the situation by considering the effect of applying a sequence of "elementary reduction" matrices to a small perturbation of $\mathbf{A}$. Let $\mathbf{E}$ be a matrix such that $\|\mathbf{E}\|_F$ is small relative to $\|\mathbf{A}\|_F$ (the Frobenius norm was introduced on p. 279), and consider

$$\mathbf{P}_k\cdots\mathbf{P}_2\mathbf{P}_1(\mathbf{A}+\mathbf{E}) = (\mathbf{P}_k\cdots\mathbf{P}_2\mathbf{P}_1\mathbf{A}) + (\mathbf{P}_k\cdots\mathbf{P}_2\mathbf{P}_1\mathbf{E}) = \mathbf{PA}+\mathbf{PE}.$$

If each $\mathbf{P}_i$ is an orthogonal matrix, then the product $\mathbf{P}=\mathbf{P}_k\cdots\mathbf{P}_2\mathbf{P}_1$ is also an orthogonal matrix (Exercise 5.6.5), and consequently $\|\mathbf{PE}\|_F=\|\mathbf{E}\|_F$ (Exercise 5.6.9). In other words, a sequence of orthogonal transformations cannot magnify the magnitude of $\mathbf{E}$, and you might think of $\mathbf{E}$ as representing the effects of roundoff error. This suggests that Householder and Givens reductions should be numerically stable algorithms. On the other hand, if the $\mathbf{P}_i$'s are elementary matrices of Type I, II, or III, then the product $\mathbf{P}=\mathbf{P}_k\cdots\mathbf{P}_2\mathbf{P}_1$ can be any nonsingular matrix—recall (3.9.3). Nonsingular matrices are not generally norm preserving (i.e., it is possible that $\|\mathbf{PE}\|_F>\|\mathbf{E}\|_F$), so the possibility of $\mathbf{E}$ being magnified is generally present in elimination methods, and this suggests the possibility of numerical instability.

Strictly speaking, an algorithm is considered to be ***numerically stable*** if, under floating-point arithmetic, it always returns an answer that is the exact solution of a nearby problem. To give an intuitive argument that the Householder or Givens reduction is a stable algorithm for producing the QR factorization of $\mathbf{A}_{n\times n}$, suppose that $\mathbf{Q}$ and $\mathbf{R}$ are the exact QR factors, and suppose that floating-point arithmetic produces an orthogonal matrix $\mathbf{Q}+\mathbf{E}$ and an upper-triangular matrix $\mathbf{R}+\mathbf{F}$ that are the exact QR factors of a different matrix

$$\tilde{\mathbf{A}} = (\mathbf{Q}+\mathbf{E})(\mathbf{R}+\mathbf{F}) = \mathbf{QR}+\mathbf{QF}+\mathbf{ER}+\mathbf{EF} = \mathbf{A}+\mathbf{QF}+\mathbf{ER}+\mathbf{EF}.$$

If $\mathbf{E}$ and $\mathbf{F}$ account for the roundoff errors, and if their entries are small relative to those in $\mathbf{A}$, then the entries in $\mathbf{EF}$ are negligible, and

$$\tilde{\mathbf{A}} \approx \mathbf{A}+\mathbf{QF}+\mathbf{ER}.$$

But since $\mathbf{Q}$ is orthogonal, $\|\mathbf{QF}\|_F=\|\mathbf{F}\|_F$ and $\|\mathbf{A}\|_F=\|\mathbf{QR}\|_F=\|\mathbf{R}\|_F$, and this means that neither $\mathbf{QF}$ nor $\mathbf{ER}$ can contain entries that are large relative to those in $\mathbf{A}$. Hence $\tilde{\mathbf{A}}\approx\mathbf{A}$, and this is what is required to conclude that the algorithm is stable.

Gaussian elimination is not a stable algorithm because, as alluded to in §1.5, problems arise due to the growth of the magnitude of the numbers that can occur during the process. To see this from a heuristic point of view, consider the LU factorization of $\mathbf{A} = \mathbf{LU}$, and suppose that floating-point Gaussian elimination with no pivoting returns matrices $\mathbf{L} + \mathbf{E}$ and $\mathbf{U} + \mathbf{F}$ that are the exact LU factors of a somewhat different matrix

$$\tilde{\mathbf{A}} = (\mathbf{L} + \mathbf{E})(\mathbf{U} + \mathbf{F}) = \mathbf{LU} + \mathbf{LF} + \mathbf{EU} + \mathbf{EF} = \mathbf{A} + \mathbf{LF} + \mathbf{EU} + \mathbf{EF}.$$

If $\mathbf{E}$ and $\mathbf{F}$ account for the roundoff errors, and if their entries are small relative to those in $\mathbf{A}$, then the entries in $\mathbf{EF}$ are negligible, and

$$\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{LF} + \mathbf{EU} \qquad \text{(using no pivoting)}.$$

However, if $\mathbf{L}$ or $\mathbf{U}$ contains entries that are large relative to those in $\mathbf{A}$ (and this is certainly possible), then $\mathbf{LF}$ or $\mathbf{EU}$ can contain entries that are significant. In other words, Gaussian elimination with no pivoting can return the LU factorization of a matrix $\tilde{\mathbf{A}}$ that is not very close to the original matrix $\mathbf{A}$, and this is what it means to say that an algorithm is unstable. We saw on p. 26 that if partial pivoting is employed, then no multiplier can exceed 1 in magnitude, and hence no entry of $\mathbf{L}$ can be greater than 1 in magnitude (recall that the subdiagonal entries of $\mathbf{L}$ are in fact the multipliers). Consequently, $\mathbf{L}$ cannot greatly magnify the entries of $\mathbf{F}$, so, if the rows of $\mathbf{A}$ have been reordered according to the partial pivoting strategy, then

$$\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{EU} \qquad \text{(using partial pivoting)}.$$

Numerical stability requires that $\tilde{\mathbf{A}} \approx \mathbf{A}$, so the issue boils down to the degree to which $\mathbf{U}$ magnifies the entries in $\mathbf{E}$—i.e., the issue rests on the magnitude of the entries in $\mathbf{U}$. Unfortunately, partial pivoting may not be enough to control the growth of all entries in $\mathbf{U}$. For example, when Gaussian elimination with partial pivoting is applied to

$$\mathbf{W}_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 0 & 1 \\ -1 & -1 & 1 & \ddots & 0 & 0 & 1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \ddots & 1 & 0 & 1 \\ -1 & -1 & -1 & \cdots & -1 & 1 & 1 \\ -1 & -1 & -1 & \cdots & -1 & -1 & 1 \end{pmatrix},$$

the largest entry in $\mathbf{U}$ is $u_{nn} = 2^{n-1}$. However, if complete pivoting is used on $\mathbf{W}_n$, then no entry in the process exceeds 2 in magnitude (Exercises 1.5.7 and 1.5.8). In general, it has been proven that if complete pivoting is used on a well-scaled matrix $\mathbf{A}_{n\times n}$ for which $\max|a_{ij}| = 1$, then no entry of $\mathbf{U}$ can exceed

$\gamma = n^{1/2}\left(2^1 3^{1/2} 4^{1/3} \cdots n^{1/n-1}\right)^{1/2}$ in magnitude. Since γ is a slow growing function of n, the entries in $\mathbf{U}$ won't greatly magnify the entries of $\mathbf{E}$, so

$$\tilde{\mathbf{A}} \approx \mathbf{A} \qquad \text{(using complete pivoting).}$$

In other words, Gaussian elimination with complete pivoting is stable, but Gaussian elimination with partial pivoting is not. Fortunately, in practical work it is rare to encounter problems such as the matrix $\mathbf{W}_n$ in which partial pivoting fails to control the growth in the $\mathbf{U}$ factor, so scaled partial pivoting is generally considered to be a "practically stable" algorithm.

Algorithms based on the Gram–Schmidt procedure are more complicated. First, the Gram–Schmidt algorithms differ from Householder and Givens reductions in that the Gram–Schmidt procedures are not a sequential application of elementary orthogonal transformations. Second, as an algorithm to produce the QR factorization even the modified Gram–Schmidt technique can return a Q factor that is far from being orthogonal, and the intuitive stability argument used earlier is not valid. As an algorithm to return the QR factorization of $\mathbf{A}$, the modified Gram–Schmidt procedure has been proven to be unstable, but as an algorithm used to solve the least squares problem (see Example 5.5.3), it is stable—i.e., stability of modified Gram–Schmidt is problem dependent.

Summary of Numerical Stability

- Gaussian elimination with scaled partial pivoting is theoretically unstable, but it is "practically stable"—i.e., stable for most practical problems.
- Complete pivoting makes Gaussian elimination unconditionally stable.
- For the QR factorization, the Gram–Schmidt procedure (classical or modified) is not stable. However, the modified Gram–Schmidt procedure is a stable algorithm for solving the least squares problem.
- Householder and Givens reductions are unconditionally stable algorithms for computing the QR factorization.

For the algorithms under consideration, the number of multiplicative operations is about the same as the number of additive operations, so computational effort is gauged by counting only multiplicative operations. For the sake of comparison, lower-order terms are not significant, and when they are neglected the following approximations are obtained.

Summary of Computational Effort

The approximate number of multiplications/divisions required to reduce an $n \times n$ matrix to an upper-triangular form is as follows.

- Gaussian elimination (scaled partial pivoting) $\approx n^3/3$.
- Gram–Schmidt procedure (classical and modified) $\approx n^3$.
- Householder reduction $\approx 2n^3/3$.
- Givens reduction $\approx 4n^3/3$.

It's not surprising that the unconditionally stable methods tend to be more costly—there is no free lunch. No one triangularization technique can be considered optimal, and each has found a place in practical work. For example, in solving unstructured linear systems, the probability of Gaussian elimination with scaled partial pivoting failing is not high enough to justify the higher cost of using the safer Householder or Givens reduction, or even complete pivoting. Although much the same is true for the full-rank least squares problem, Householder reduction or modified Gram–Schmidt is frequently used as a safeguard against sensitivities that often accompany least squares problems. For the purpose of computing an orthonormal basis for $R(\mathbf{A})$ in which $\mathbf{A}$ is unstructured and dense (not many zeros), Householder reduction is preferred—the Gram–Schmidt procedures are unstable for this purpose and Givens reduction is too costly. Givens reduction is useful when the matrix being reduced is highly structured or sparse (many zeros).

Example 5.7.4

Reduction to Hessenberg Form. For reasons alluded to in §4.8 and §4.9, it is often desirable to triangularize a square matrix $\mathbf{A}$ by means of a similarity transformation—i.e., find a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{T}$ is upper triangular. But this is a computationally difficult task, so we will try to do the next best thing, which is to find a similarity transformation that will reduce $\mathbf{A}$ to a matrix in which all entries below the first subdiagonal are zero. Such a matrix is said to be in ***upper-Hessenberg form***—illustrated below is a 5×5 Hessenberg form.

$$\mathbf{H} = \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}.$$

Problem: Reduce $\mathbf{A} \in \Re^{n \times n}$ to upper-Hessenberg form by means of an orthogonal similarity transformation—i.e., construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{H}$ is upper Hessenberg.

Solution: At each step, use Householder reduction on entries *below* the main diagonal. Begin by letting $\hat{\mathbf{A}}_{*1}$ denote the entries of the first column that are below the (1,1)-position—this is illustrated below for $n = 5$:

$$\mathbf{A} = \left(\begin{array}{c|cccc} * & * & * & * & * \\ \hline * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{array}\right) = \left(\begin{array}{c|c} a_{11} & \hat{\mathbf{A}}_{1*} \\ \hline \hat{\mathbf{A}}_{*1} & \mathbf{A}_1 \end{array}\right).$$

If $\hat{\mathbf{R}}_1$ is an elementary reflector determined according to (5.7.1) for which $\hat{\mathbf{R}}_1\hat{\mathbf{A}}_{*1} = \begin{pmatrix} * \\ 0 \\ 0 \\ 0 \end{pmatrix}$, then $\mathbf{R}_1 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_1 \end{pmatrix}$ is an orthogonal matrix such that

$$\begin{aligned} \mathbf{R}_1\mathbf{A}\mathbf{R}_1 &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_1 \end{pmatrix} \begin{pmatrix} a_{11} & \hat{\mathbf{A}}_{1*} \\ \hat{\mathbf{A}}_{*1} & \mathbf{A}_1 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_1 \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & \hat{\mathbf{A}}_{1*}\hat{\mathbf{R}}_1 \\ \hat{\mathbf{R}}_1\hat{\mathbf{A}}_{*1} & \hat{\mathbf{R}}_1\mathbf{A}_1\hat{\mathbf{R}}_1 \end{pmatrix} = \left(\begin{array}{c|cccc} * & * & * & * & * \\ \hline * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{array}\right). \end{aligned}$$

At the second step, repeat the process on $\mathbf{A}_2 = \hat{\mathbf{R}}_1\mathbf{A}_1\hat{\mathbf{R}}_1$ to obtain an orthogonal matrix $\hat{\mathbf{R}}_2$ such that $\hat{\mathbf{R}}_2\mathbf{A}_2\hat{\mathbf{R}}_2 = \left(\begin{array}{c|ccc} * & * & * & * \\ \hline * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{array}\right)$. Matrix $\mathbf{R}_2 = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$ is an orthogonal matrix such that

$$\mathbf{R}_2\mathbf{R}_1\mathbf{A}\mathbf{R}_1\mathbf{R}_2 = \left(\begin{array}{c|c|ccc} * & * & * & * & * \\ \hline * & * & * & * & * \\ \hline 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{array}\right).$$

After $n-2$ of these steps, the product $\mathbf{P} = \mathbf{R}_1\mathbf{R}_2 \cdots \mathbf{R}_{n-2}$ is an orthogonal matrix such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{H}$ is in upper-Hessenberg form.

Note: If $\mathbf{A}$ is a symmetric matrix, then $\mathbf{H}^T = (\mathbf{P}^T\mathbf{A}\mathbf{P})^T = \mathbf{P}^T\mathbf{A}^T\mathbf{P} = \mathbf{H}$, so $\mathbf{H}$ is symmetric. But as illustrated below for $n = 5$, a symmetric Hessenberg form is a tridiagonal matrix,

$$\mathbf{H} = \mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ 0 & * & * & * & 0 \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix},$$

so the following useful corollary is obtained.

- Every real-symmetric matrix is orthogonally similar to a tridiagonal matrix, and Householder reduction can be used to compute this tridiagonal matrix. However, the Lanczos technique discussed on p. 651 can be much more efficient.

Example 5.7.5

Problem: Compute the QR factors of a nonsingular upper-Hessenberg matrix $\mathbf{H} \in \Re^{n\times n}$.

Solution: Due to its smaller multiplication count, Householder reduction is generally preferred over Givens reduction. The exception is for matrices that have a zero pattern that can be exploited by the Givens method but not by the Householder method. A Hessenberg matrix $\mathbf{H}$ is such an example. The first step of Householder reduction completely destroys most of the zeros in $\mathbf{H}$, but applying plane rotations does not. This is illustrated below for a 5×5 Hessenberg form—remember that the action of $\mathbf{P}_{k,k+1}$ affects only the k^{th} and $(k+1)^{st}$ rows.

$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} \xrightarrow{\mathbf{P}_{12}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} \xrightarrow{\mathbf{P}_{23}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}$$

$$\xrightarrow{\mathbf{P}_{34}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} \xrightarrow{\mathbf{P}_{45}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * \end{pmatrix}.$$

In general, $\mathbf{P}_{n-1,n}\cdots\mathbf{P}_{23}\mathbf{P}_{12}\mathbf{H} = \mathbf{R}$ is upper triangular in which all diagonal entries, except possibly the last, are positive—the last diagonal can be made positive by the technique illustrated in Example 5.7.2. Thus we obtain an orthogonal matrix $\mathbf{P}$ such that $\mathbf{PH} = \mathbf{R}$, or $\mathbf{H} = \mathbf{QR}$ in which $\mathbf{Q} = \mathbf{P}^T$.

Example 5.7.6

Jacobi Reduction.[49] Given a real-symmetric matrix $\mathbf{A}$, the result of Example 5.7.4 shows that Householder reduction can be used to construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{AP} = \mathbf{T}$ is tridiagonal. Can we do better?—i.e., can we construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{AP} = \mathbf{D}$ is a diagonal matrix? Indeed we can, and much of the material in Chapter 7 concerning eigenvalues and eigenvectors is devoted to this problem. But in the present context, this fact can be constructively established by means of Jacobi's diagonalization algorithm.

Jacobi's Idea. If $\mathbf{A} \in \Re^{n\times n}$ is symmetric, then a plane rotation matrix can be applied to reduce the magnitude of the off-diagonal entries. In particular, suppose that $a_{ij} \neq 0$ is the off-diagonal entry of maximal magnitude, and let $\mathbf{A}'$ denote the matrix obtained by setting each $a_{kk} = 0$. If $\mathbf{P}_{ij}$ is the plane rotation matrix described on p. 333 in which $c = \cos\theta$ and $s = \sin\theta$, where $\cot 2\theta = (a_{ii} - a_{jj})/2a_{ij}$, and if $\mathbf{B} = \mathbf{P}_{ij}^T\mathbf{AP}_{ij}$, then

(1) $b_{ij} = b_{ji} = 0$ (i.e., a_{ij} is annihilated),

(2) $\|\mathbf{B}'\|_F^2 = \|\mathbf{A}'\|_F^2 - 2a_{ij}^2$,

(3) $\|\mathbf{B}'\|_F^2 \leq \left(1 - \dfrac{2}{n^2-n}\right)\|\mathbf{A}'\|_F^2$.

Proof. The entries of $\mathbf{B} = \mathbf{P}_{ij}^T\mathbf{AP}_{ij}$ that lay on the intersection of the i^{th} and j^{th} rows with the i^{th} and j^{th} columns can be described by

$$\hat{\mathbf{B}} = \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \mathbf{P}^T\hat{\mathbf{A}}\mathbf{P}.$$

Use the identities $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\cos\theta\sin\theta$ to verify $b_{ij} = b_{ji} = 0$, and recall that $\|\hat{\mathbf{B}}\|_F = \|\mathbf{P}^T\hat{\mathbf{A}}\mathbf{P}\|_F = \|\hat{\mathbf{A}}\|_F$ (recall Exercise

[49] Karl Gustav Jacob Jacobi (1804–1851) first presented this method in 1846, and it was popular for a time. But the twentieth-century development of electronic computers sparked tremendous interest in numerical algorithms for diagonalizing symmetric matrices, and Jacobi's method quickly fell out of favor because it could not compete with newer procedures—at least on the traditional sequential machines. However, the emergence of multiprocessor parallel computers has resurrected interest in Jacobi's method because of the inherent parallelism in the algorithm. Jacobi was born in Potsdam, Germany, educated at the University of Berlin, and employed as a professor at the University of Königsberg. During his prolific career he made contributions that are still important facets of contemporary mathematics. His accomplishments include the development of elliptic functions; a systematic development and presentation of the theory of determinants; contributions to the theory of rotating liquids; and theorems in the areas of differential equations, calculus of variations, and number theory. In contrast to his great contemporary Gauss, who disliked teaching and was anything but inspiring, Jacobi was regarded as a great teacher (the introduction of the student seminar method is credited to him), and he advocated the view that "the sole end of science is the honor of the human mind, and that under this title a question about numbers is worth as much as a question about the system of the world." Jacobi once defended his excessive devotion to work by saying that "Only cabbages have no nerves, no worries. And what do they get out of their perfect wellbeing?" Jacobi suffered a breakdown from overwork in 1843, and he died at the relatively young age of 46.

5.6.9) to produce the conclusion $b_{ii}^2 + b_{jj}^2 = a_{ii}^2 + 2a_{ij}^2 + a_{jj}^2$. Now use the fact that $b_{kk} = a_{kk}$ for all $k \neq i, j$ together with $\|\mathbf{B}\|_F = \|\mathbf{A}\|_F$ to write

$$\begin{aligned}
\|\mathbf{B}'\|_F^2 &= \|\mathbf{B}\|_F^2 - \sum_k b_{kk}^2 = \|\mathbf{B}\|_F^2 - \sum_{k\neq i,j} b_{kk}^2 - \left(b_{ii}^2 + b_{jj}^2\right) \\
&= \|\mathbf{A}\|_F^2 - \sum_{k\neq i,j} a_{kk}^2 - \left(a_{ii}^2 + 2a_{ij}^2 + a_{jj}^2\right) = \|\mathbf{A}\|_F^2 - \sum_k a_{kk}^2 - 2a_{ij}^2 \\
&= \|\mathbf{A}'\|_F^2 - 2a_{ij}^2.
\end{aligned}$$

Furthermore, since $a_{pq}^2 \leq a_{ij}^2$ for all $p \neq q$,

$$\|\mathbf{A}'\|_F^2 = \sum_{p\neq q} a_{pq}^2 \leq \sum_{p\neq q} a_{ij}^2 = (n^2 - n)a_{ij}^2 \quad \Longrightarrow \quad -a_{ij}^2 \leq -\frac{\|\mathbf{A}'\|_F^2}{n^2 - n},$$

so

$$\|\mathbf{B}'\|_F^2 = \|\mathbf{A}'\|_F^2 - 2a_{ij}^2 \leq \|\mathbf{A}'\|_F^2 - 2\frac{\|\mathbf{A}'\|_F^2}{n^2 - n} = \left(1 - \frac{2}{n^2 - n}\right)\|\mathbf{A}'\|_F^2. \quad \blacksquare$$

Jacobi's Diagonalization Algorithm. Start with $\mathbf{A}_0 = \mathbf{A}$, and produce a sequence of matrices $\mathbf{A}_k = \mathbf{P}_k^T \mathbf{A}_{k-1} \mathbf{P}_k$, where at the k^{th} step $\mathbf{P}_k$ is a plane rotation constructed to annihilate the maximal off-diagonal entry in $\mathbf{A}_{k-1}$. In particular, if a_{ij} is the entry of maximal magnitude in $\mathbf{A}_{k-1}$, then $\mathbf{P}_k$ is the rotator in the (i, j)-plane defined by setting

$$s = \frac{1}{\sqrt{1+\sigma^2}} \quad \text{and} \quad c = \frac{\sigma}{\sqrt{1+\sigma^2}} = \sqrt{1 - s^2}, \quad \text{where} \quad \sigma = \frac{(a_{ii} - a_{jj})}{2a_{ij}}.$$

For $n > 2$ we have

$$\|\mathbf{A}_k'\|_F^2 \leq \left(1 - \frac{2}{n^2 - n}\right)^k \|\mathbf{A}'\|_F^2 \to 0 \quad \text{as} \quad k \to \infty.$$

Therefore, if $\mathbf{P}^{(k)}$ is the orthogonal matrix defined by $\mathbf{P}^{(k)} = \mathbf{P}_1\mathbf{P}_2 \cdots \mathbf{P}_k$, then

$$\lim_{k\to\infty} \mathbf{P}^{(k)^T} \mathbf{A} \mathbf{P}^{(k)} = \lim_{k\to\infty} \mathbf{A}_k = \mathbf{D}$$

is a diagonal matrix.

Exercises for section 5.7

5.7.1. (a) Using Householder reduction, compute the QR factors of

$$\mathbf{A} = \begin{pmatrix} 1 & 19 & -34 \\ -2 & -5 & 20 \\ 2 & 8 & 37 \end{pmatrix}.$$

(b) Repeat part (a) using Givens reduction.

5.7.2. For $\mathbf{A} \in \Re^{m\times n}$, suppose that $rank\,(\mathbf{A}) = n$, and let $\mathbf{P}$ be an orthogonal matrix such that

$$\mathbf{PA} = \mathbf{T} = \begin{pmatrix} \mathbf{R}_{n\times n} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{R}$ is an upper-triangular matrix. If $\mathbf{P}^T$ is partitioned as

$$\mathbf{P}^T = [\mathbf{X}_{m\times n} \,|\, \mathbf{Y}],$$

explain why the columns of $\mathbf{X}$ constitute an orthonormal basis for $R\,(\mathbf{A})$.

5.7.3. By using Householder reduction, find an orthonormal basis for $R\,(\mathbf{A})$, where

$$\mathbf{A} = \begin{pmatrix} 4 & -3 & 4 \\ 2 & -14 & -3 \\ -2 & 14 & 0 \\ 1 & -7 & 15 \end{pmatrix}.$$

5.7.4. Use Householder reduction to compute the least squares solution for $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 4 & -3 & 4 \\ 2 & -14 & -3 \\ -2 & 14 & 0 \\ 1 & -7 & 15 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 5 \\ -15 \\ 0 \\ 30 \end{pmatrix}.$$

Hint: Make use of the factors you computed in Exercise 5.7.3.

5.7.5. If $\mathbf{A} = \mathbf{QR}$ is the QR factorization for $\mathbf{A}$, explain why $\|\mathbf{A}\|_F = \|\mathbf{R}\|_F$, where $\|\star\|_F$ is the Frobenius matrix norm introduced on p. 279.

5.7.6. Find an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{AP} = \mathbf{H}$ is in upper-Hessenberg form, where

$$\mathbf{A} = \begin{pmatrix} -2 & 3 & -4 \\ 3 & -25 & 50 \\ -4 & 50 & 25 \end{pmatrix}.$$

5.7.7. Let $\mathbf{H}$ be an upper-Hessenberg matrix, and suppose that $\mathbf{H} = \mathbf{QR}$, where $\mathbf{R}$ is a nonsingular upper-triangular matrix. Prove that $\mathbf{Q}$ as well as the product $\mathbf{RQ}$ must also be in upper-Hessenberg form.

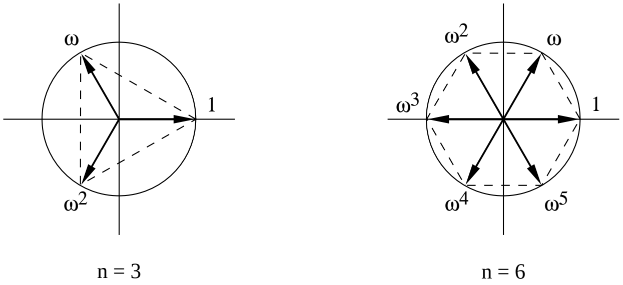
5.7.8. Approximately how many multiplications are needed to reduce an $n \times n$ nonsingular upper-Hessenberg matrix to upper-triangular form by using plane rotations?

5.8 DISCRETE FOURIER TRANSFORM

For a positive integer n, the complex numbers $\left\{1, \omega, \omega^2, \ldots, \omega^{n-1}\right\}$, where

$$\omega = \mathrm{e}^{2\pi \mathrm{i}/n} = \cos\frac{2\pi}{n} + \mathrm{i}\sin\frac{2\pi}{n}$$

are called the ***n*** $\boldsymbol{n^{th}}$ ***roots of unity*** because they represent all solutions to $z^n = 1$. Geometrically, they are the vertices of a regular polygon of n sides as depicted in Figure 5.8.1 for $n = 3$ and $n = 6$.



n = 3 n = 6

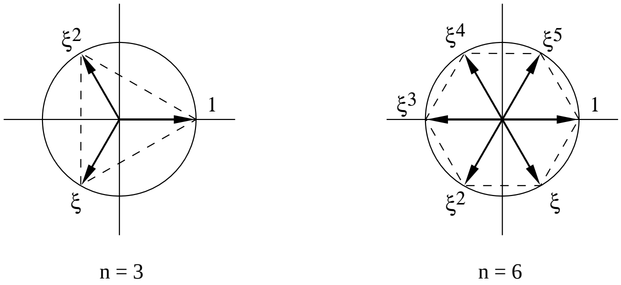
FIGURE 5.8.1

The roots of unity are cyclic in the sense that if $k \geq n$, then $\omega^k = \omega^{k\,(\mathrm{mod}\ n)}$, where $k\,(\mathrm{mod}\ n)$ denotes the remainder when k is divided by n—for example, when $n = 6$, $\omega^6 = 1$, $\omega^7 = \omega$, $\omega^8 = \omega^2$, $\omega^9 = \omega^3, \ldots$.

The numbers $\left\{1, \xi, \xi^2, \ldots, \xi^{n-1}\right\}$, where

$$\xi = \mathrm{e}^{-2\pi \mathrm{i}/n} = \cos\frac{2\pi}{n} - \mathrm{i}\sin\frac{2\pi}{n} = \overline{\omega}$$

are also the n^{th} roots of unity, but, as depicted in Figure 5.8.2 for $n = 3$ and $n = 6$, they are listed in clockwise order around the unit circle rather than counterclockwise.



n = 3 n = 6

FIGURE 5.8.2

The following identities will be useful in our development. If k is an integer, then $1 = |\xi^k|^2 = \xi^k\overline{\xi^k}$ implies that

$$\xi^{-k} = \overline{\xi^k} = \omega^k. \tag{5.8.1}$$

Furthermore, the fact that

$$\xi^k\left(1+\xi^k+\xi^{2k}+\cdots+\xi^{(n-2)k}+\xi^{(n-1)k}\right)=\xi^k+\xi^{2k}+\cdots+\xi^{(n-1)k}+1$$

implies $\left(1+\xi^k+\xi^{2k}+\cdots+\xi^{(n-1)k}\right)\left(1-\xi^k\right)=0$ and, consequently,

$$1+\xi^k+\xi^{2k}+\cdots+\xi^{(n-1)k}=0 \quad \text{whenever} \quad \xi^k\neq 1. \tag{5.8.2}$$

Fourier Matrix

The $n\times n$ matrix whose (j,k)-entry is $\xi^{jk}=\omega^{-jk}$ for $0\leq j,k\leq n-1$ is called the ***Fourier matrix*** of order n, and it has the form

$$\mathbf{F}_n=\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \xi & \xi^2 & \cdots & \xi^{n-1} \\ 1 & \xi^2 & \xi^4 & \cdots & \xi^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi^{n-1} & \xi^{n-2} & \cdots & \xi \end{pmatrix}_{n\times n}.$$

When the context makes it clear, the subscript n on $\mathbf{F}_n$ is omitted.

The Fourier matrix[50] is a special case of the Vandermonde matrix introduced in Example 4.3.4. Using (5.8.1) and (5.8.2), we see that the inner product of any two columns in $\mathbf{F}_n$, say, the r^{th} and s^{th}, is

$$\mathbf{F}_{*r}^{*}\mathbf{F}_{*s}=\sum_{j=0}^{n-1}\overline{\xi^{jr}}\xi^{js}=\sum_{j=0}^{n-1}\xi^{-jr}\xi^{js}=\sum_{j=0}^{n-1}\xi^{j(s-r)}=0.$$

In other words, the columns in $\mathbf{F}_n$ are mutually orthogonal. Furthermore, each column in $\mathbf{F}_n$ has norm $\sqrt{n}$ because

$$\|\mathbf{F}_{*k}\|_2^2=\sum_{j=0}^{n-1}|\xi^{jk}|^2=\sum_{j=0}^{n-1}1=n,$$

[50] Some authors define the Fourier matrix using powers of ω rather than powers of ξ, and some include a scalar multiple $1/n$ or $1/\sqrt{n}$. These differences are superficial, and they do not affect the basic properties. Our definition is the discrete counterpart of the integral operator $F(f)=\int_{-\infty}^{\infty}x(t)\mathrm{e}^{-\mathrm{i}2\pi ft}dt$ that is usually taken as the definition of the continuous Fourier transform.

and consequently every column of $\mathbf{F}_n$ can be normalized by multiplying by the same scalar—namely, $1/\sqrt{n}$. This means that $(1/\sqrt{n}\,)\mathbf{F}_n$ is a unitary matrix. Since it is also true that $\mathbf{F}_n^T = \mathbf{F}_n$, we have

$$\left(\frac{1}{\sqrt{n}}\mathbf{F}_n\right)^{-1} = \left(\frac{1}{\sqrt{n}}\mathbf{F}_n\right)^{*} = \frac{1}{\sqrt{n}}\overline{\mathbf{F}}_n,$$

and therefore $\mathbf{F}_n^{-1} = \overline{\mathbf{F}}_n/n$. But (5.8.1) says that $\overline{\xi^k} = \omega^k$, so it must be the case that

$$\mathbf{F}_n^{-1} = \frac{1}{n}\overline{\mathbf{F}}_n = \frac{1}{n}\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{n-2} & \cdots & \omega \end{pmatrix}_{n\times n}.$$

Example 5.8.1

The Fourier matrices of orders 2 and 4 are given by

$$\mathbf{F}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{F}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix},$$

and their inverses are

$$\mathbf{F}_2^{-1} = \frac{1}{2}\overline{\mathbf{F}}_2 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{F}_4^{-1} = \frac{1}{4}\overline{\mathbf{F}}_4 = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{pmatrix}.$$

Discrete Fourier Transform

Given a vector $\mathbf{x}_{n\times 1}$, the product $\mathbf{F}_n\mathbf{x}$ is called the ***discrete Fourier transform*** of $\mathbf{x}$, and $\mathbf{F}_n^{-1}\mathbf{x}$ is called the ***inverse transform*** of $\mathbf{x}$. The k^{th} entries in $\mathbf{F}_n\mathbf{x}$ and $\mathbf{F}_n^{-1}\mathbf{x}$ are given by

$$[\mathbf{F}_n\mathbf{x}]_k = \sum_{j=0}^{n-1} x_j \xi^{jk} \quad \text{and} \quad [\mathbf{F}_n^{-1}\mathbf{x}]_k = \frac{1}{n}\sum_{j=0}^{n-1} x_j \omega^{jk}. \qquad (5.8.3)$$

Example 5.8.2

Problem: Computing the Inverse Transform. Explain why any algorithm or program designed to compute the discrete Fourier transform of a vector $\mathbf{x}$ can also be used to compute the *inverse* transform of $\mathbf{x}$.

Solution: Call such an algorithm FFT (see p. 373 for a specific example). The fact that

$$\mathbf{F}_n^{-1}\mathbf{x} = \frac{\overline{\mathbf{F}}_n\mathbf{x}}{n} = \frac{\overline{\mathbf{F}_n\overline{\mathbf{x}}}}{n}$$

means that FFT will return the inverse transform of $\mathbf{x}$ by executing the following three steps:

(1) $\mathbf{x} \longleftarrow \overline{\mathbf{x}}$ (compute $\overline{\mathbf{x}}$).

(2) $\mathbf{x} \longleftarrow \text{FFT}(\mathbf{x})$ (compute $\mathbf{F}_n\overline{\mathbf{x}}$).

(3) $\mathbf{x} \longleftarrow (1/n)\overline{\mathbf{x}}$ (compute $n^{-1}\overline{\mathbf{F}_n\overline{\mathbf{x}}} = \mathbf{F}_n^{-1}\mathbf{x}$).

For example, computing the inverse transform of $\mathbf{x} = (\,\mathrm{i} \quad 0 \quad -\mathrm{i} \quad 0\,)^T$ is accomplished as follows—recall that $\mathbf{F}_4$ was given in Example 5.8.1.

$$\overline{\mathbf{x}} = \begin{pmatrix} -\mathrm{i} \\ 0 \\ \mathrm{i} \\ 0 \end{pmatrix}, \quad \mathbf{F}_4\overline{\mathbf{x}} = \begin{pmatrix} 0 \\ -2\mathrm{i} \\ 0 \\ -2\mathrm{i} \end{pmatrix}, \quad \frac{1}{4}\overline{\mathbf{F}_4\overline{\mathbf{x}}} = \frac{1}{4}\begin{pmatrix} 0 \\ 2\mathrm{i} \\ 0 \\ 2\mathrm{i} \end{pmatrix} = \mathbf{F}_4^{-1}\mathbf{x}.$$

You may wish to check that this answer agrees with the result obtained by directly multiplying $\mathbf{F}_4^{-1}$ times $\mathbf{x}$, where $\mathbf{F}_4^{-1}$ is given in Example 5.8.1.

Example 5.8.3

Signal Processing. Suppose that a microphone is placed under a hovering helicopter, and suppose that Figure 5.8.3 represents the sound signal that is recorded during 1 second of time.

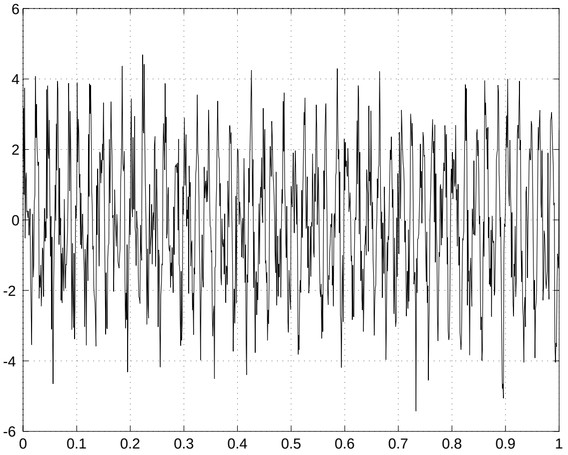


FIGURE 5.8.3

It seems reasonable to expect that the signal should have oscillatory components together with some random noise contamination. That is, we expect the signal to have the form

$$y(\tau) = \left(\sum_k \alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau \right) + \text{Noise.}$$

But due to the noise contamination, the oscillatory nature of the signal is only barely apparent—the characteristic "chop-a chop-a chop-a" is not completely clear. To reveal the oscillatory components, the magic of the Fourier transform is employed. Let $\mathbf{x}$ be the vector obtained by sampling the signal at n equally spaced points between time $\tau = 0$ and $\tau = 1$ ($n = 512$ in our case), and let

$$\mathbf{y} = (2/n)\mathbf{F}_n\mathbf{x} = \mathbf{a} + \mathrm{i}\mathbf{b}, \quad \text{where} \quad \mathbf{a} = (2/n)\mathrm{Re}\,(\mathbf{F}_n\mathbf{x}) \text{ and } \mathbf{b} = (2/n)\mathrm{Im}\,(\mathbf{F}_n\mathbf{x})\,.$$

Using only the first $n/2 = 256$ entries in $\mathbf{a}$ and $\mathrm{i}\mathbf{b}$, we plot the points in

$$\{(0, a_0),\ (1, a_1),\ \ldots,\ (255, a_{255})\} \quad \text{and} \quad \{(0, \mathrm{i}b_0),\ (1, \mathrm{i}b_1),\ \ldots,\ (255, \mathrm{i}b_{255})\}$$

to produce the two graphs shown in Figure 5.8.4.

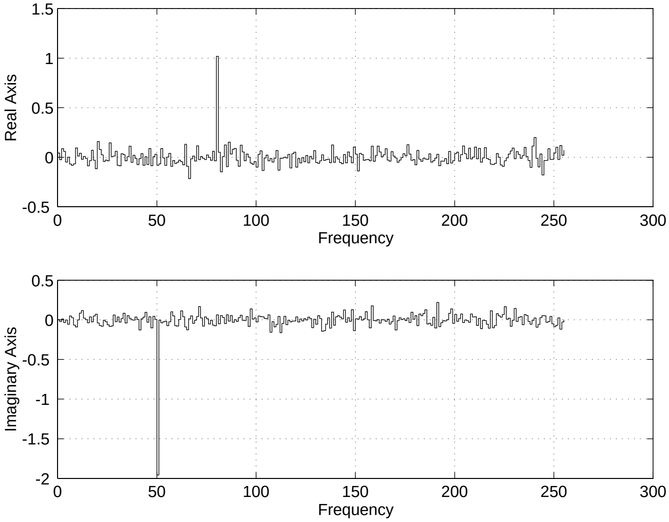


FIGURE 5.8.4

Now there are some obvious characteristics—the plot of $\mathbf{a}$ in the top graph of Figure 5.8.4 has a spike of height approximately 1 at entry 80, and the plot of $\mathrm{i}\mathbf{b}$ in the bottom graph has a spike of height approximately -2 at entry 50. These two spikes indicate that the signal is made up primarily of two oscillatory components—the spike in the real vector $\mathbf{a}$ indicates that one of the oscillatory components is a cosine of frequency 80 Hz (or period $= 1/80$) whose amplitude is approximately 1, and the spike in the imaginary vector $\mathbf{ib}$ indicates there is a sine component with frequency 50 Hz and amplitude of about 2. In other words, the Fourier transform indicates that the signal is

$$y(\tau) = \cos 2\pi(80\tau) + 2\sin 2\pi(50\tau) + \text{Noise}.$$

In truth, the data shown in Figure 5.8.3 was artificially generated by contaminating the function $y(\tau) = \cos 2\pi(80\tau) + 2\sin 2\pi(50\tau)$ with some normally distributed zero-mean noise, and therefore the plot of $(2/n)\mathbf{F}_n\mathbf{x}$ shown in Figure 5.8.4 does indeed accurately reflect the true nature of the signal. To understand why $\mathbf{F}_n$ reveals the hidden frequencies, let $\cos 2\pi f\mathbf{t}$ and $\sin 2\pi f\mathbf{t}$ denote the ***discrete cosine*** and ***discrete sine*** vectors

$$\cos 2\pi f\mathbf{t} = \begin{pmatrix} \cos\left(2\pi f \cdot \frac{0}{n}\right) \\ \cos\left(2\pi f \cdot \frac{1}{n}\right) \\ \cos\left(2\pi f \cdot \frac{2}{n}\right) \\ \vdots \\ \cos\left(2\pi f \cdot \frac{n-1}{n}\right) \end{pmatrix} \quad \text{and} \quad \sin 2\pi f\mathbf{t} = \begin{pmatrix} \sin\left(2\pi f \cdot \frac{0}{n}\right) \\ \sin\left(2\pi f \cdot \frac{1}{n}\right) \\ \sin\left(2\pi f \cdot \frac{2}{n}\right) \\ \vdots \\ \sin\left(2\pi f \cdot \frac{n-1}{n}\right) \end{pmatrix},$$

where $\mathbf{t} = \begin{pmatrix} 0/n & 1/n & 2/n & \cdots & n-1/n \end{pmatrix}^T$ is the ***discrete time vector***. If the ***discrete exponential*** vectors $\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}}$ and $\mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}}$ are defined in the natural way as $\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}} = \cos 2\pi f\mathbf{t} + \mathrm{i}\sin 2\pi f\mathbf{t}$ and $\mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}} = \cos 2\pi f\mathbf{t} - \mathrm{i}\sin 2\pi f\mathbf{t}$, and if $0 \leq f < n$ is an integer frequency,[51] then

$$\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}} = \begin{pmatrix} \omega^{0f} \\ \omega^{1f} \\ \omega^{2f} \\ \vdots \\ \omega^{(n-1)f} \end{pmatrix} = n\left[\mathbf{F}_n^{-1}\right]_{*f} = n\mathbf{F}_n^{-1}\mathbf{e}_f,$$

where $\mathbf{e}_f$ is the $n \times 1$ unit vector with a 1 in the f^{th} component—remember that components of vectors are indexed from 0 to $n-1$ throughout this section. Similarly, the fact that

$$\xi^{kf} = \omega^{-kf} = 1\omega^{-kf} = \omega^{kn}\omega^{-kf} = \omega^{k(n-f)} \quad \text{for} \quad k = 0, 1, 2, \ldots$$

allows us to conclude that if $0 \leq n - f < n$, then

$$\mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}} = \begin{pmatrix} \xi^{0f} \\ \xi^{1f} \\ \xi^{2f} \\ \vdots \\ \xi^{(n-1)f} \end{pmatrix} = \begin{pmatrix} \omega^{0(n-f)} \\ \omega^{1(n-f)} \\ \omega^{2(n-f)} \\ \vdots \\ \omega^{(n-1)(n-f)} \end{pmatrix} = n\left[\mathbf{F}_n^{-1}\right]_{*n-f} = n\mathbf{F}_n^{-1}\mathbf{e}_{n-f}.$$

[51] The assumption that frequencies are integers is not overly harsh because the Fourier series for a periodic function requires only integer frequencies—recall Example 5.4.6.

Therefore, if $0 < f < n$, then

$$\mathbf{F}_n \mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}} = n\mathbf{e}_f \quad \text{and} \quad \mathbf{F}_n \mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}} = n\mathbf{e}_{n-f}. \tag{5.8.4}$$

Because $\cos\theta = (\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{-\mathrm{i}\theta})/2$ and $\sin\theta = (\mathrm{e}^{\mathrm{i}\theta} - \mathrm{e}^{-\mathrm{i}\theta})/2\mathrm{i}$, it follows from (5.8.4) that for any scalars α and β,

$$\mathbf{F}_n(\alpha\cos 2\pi f\mathbf{t}) = \alpha\mathbf{F}_n\left(\frac{\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}} + \mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}}}{2}\right) = \frac{n\alpha}{2}\left(\mathbf{e}_f + \mathbf{e}_{n-f}\right)$$

and

$$\mathbf{F}_n(\beta\sin 2\pi f\mathbf{t}) = \beta\mathbf{F}_n\left(\frac{\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}} - \mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}}}{2\mathrm{i}}\right) = \frac{n\beta}{2\mathrm{i}}\left(\mathbf{e}_f - \mathbf{e}_{n-f}\right),$$

so that

$$\frac{2}{n}\mathbf{F}_n(\alpha\cos 2\pi f\mathbf{t}) = \alpha\mathbf{e}_f + \alpha\mathbf{e}_{n-f} \tag{5.8.5}$$

and

$$\frac{2}{n}\mathbf{F}_n(\beta\sin 2\pi f\mathbf{t}) = -\beta\mathrm{i}\mathbf{e}_f + \beta\mathrm{i}\mathbf{e}_{n-f}. \tag{5.8.6}$$

The trigonometric functions $\alpha\cos 2\pi f\tau$ and $\beta\sin 2\pi f\tau$ have amplitudes α and β, respectively, and their frequency is f (their period is $1/f$). The discrete vectors $\alpha\cos 2\pi f\mathbf{t}$ and $\beta\sin 2\pi f\mathbf{t}$ are obtained by evaluating $\alpha\cos 2\pi f\tau$ and $\beta\sin 2\pi f\tau$ at the discrete points in $\mathbf{t} = (\,0 \quad 1/n \quad 2/n \quad \cdots \quad (n-1)/n\,)^T$. As depicted in Figure 5.8.5 for $n = 32$ and $f = 4$, the vectors $\alpha\mathbf{e}_f$ and $\alpha\mathbf{e}_{n-f}$ are interpreted as two pulses of magnitude α at frequencies f and $n - f$.

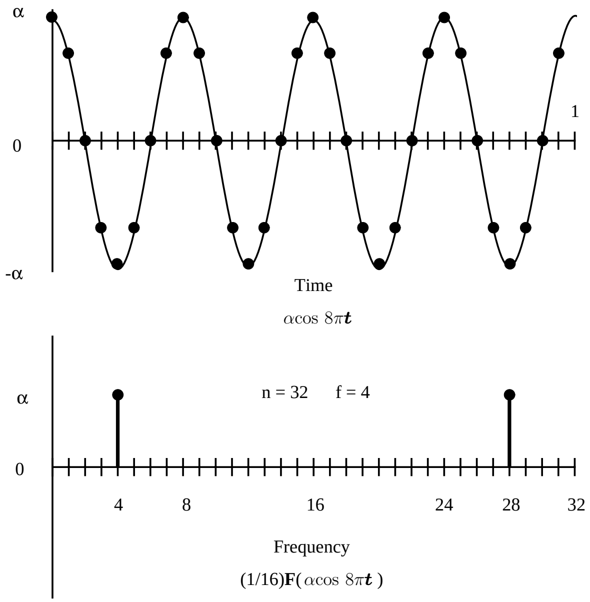
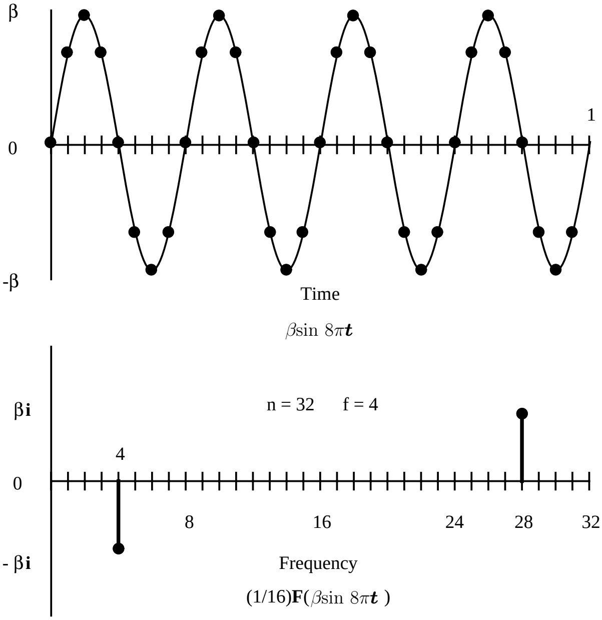


Figure 5.8.5

The vector $\alpha \cos 2\pi f \mathbf{t}$ is said to be in the ***time domain***, while the pulses $\alpha \mathbf{e}_f$ and $\alpha \mathbf{e}_{n-f}$ are said to be in the ***frequency domain***. The situation for $\beta \sin 2\pi f \mathbf{t}$ is similarly depicted in Figure 5.8.6 in which $-\beta \mathrm{i} \mathbf{e}_f$ and $\beta \mathrm{i} \mathbf{e}_{n-f}$ are considered two pulses of height $-\beta$ and β, respectively.



$\beta \sin 8\pi \boldsymbol{t}$

$(1/16)\mathbf{F}(\beta \sin 8\pi \boldsymbol{t})$

Figure 5.8.6

Therefore, if a waveform is given by a finite sum

$$x(\tau) = \sum_k \left(\alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau \right)$$

in which the f_k 's are integers, and if $\mathbf{x}$ is the vector containing the values of $x(\tau)$ at n equally spaced points between time $\tau = 0$ and $\tau = 1$, then, provided that n is sufficiently large,

$$\begin{aligned} \frac{2}{n}\mathbf{F}_n \mathbf{x} &= \frac{2}{n}\mathbf{F}_n \left(\sum_k \alpha_k \cos 2\pi f_k \mathbf{t} + \beta_k \sin 2\pi f_k \mathbf{t} \right) \\ &= \sum_k \frac{2}{n}\mathbf{F}_n \left(\alpha_k \cos 2\pi f_k \mathbf{t} \right) + \sum_k \frac{2}{n}\mathbf{F}_n \left(\beta_k \sin 2\pi f_k \mathbf{t} \right) \qquad (5.8.7) \\ &= \sum_k \alpha_k \left(\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k} \right) + \mathrm{i} \sum_k \beta_k \left(-\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k} \right), \end{aligned}$$

and this exposes the frequency and amplitude of each of the components. If n is chosen so that $\max\{f_k\} < n/2$, then the pulses represented by $\mathbf{e}_f$ and $\mathbf{e}_{n-f}$ are symmetric about the point $n/2$ in the frequency domain, and the information in just the first (or second) half of the frequency domain completely characterizes the original waveform—this is why only 128/2=64 points are plotted in the graphs shown in Figure 5.8.4. In other words, if

$$\mathbf{y} = \frac{2}{n}\mathbf{F}_n\mathbf{x} = \sum_k \alpha_k \left(\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k}\right) + \mathrm{i}\sum_k \beta_k \left(-\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k}\right), \tag{5.8.8}$$

then the information in

$$\mathbf{y}_{n/2} = \sum_k \alpha_k \mathbf{e}_{f_k} - \mathrm{i}\sum_k \beta_k \mathbf{e}_{f_k} \qquad \text{(the first half of } \mathbf{y}\text{)}$$

is enough to reconstruct the original waveform. For example, the equation of the waveform shown in Figure 5.8.7 is

$$x(\tau) = 3\cos 2\pi\tau + 5\sin 2\pi\tau, \tag{5.8.9}$$

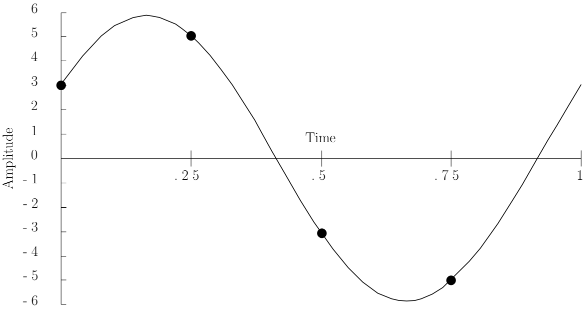


Figure 5.8.7

and it is completely determined by the four values in

$$\mathbf{x} = \begin{pmatrix} x(0) \\ x(1/4) \\ x(1/2) \\ x(3/4) \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ -3 \\ -5 \end{pmatrix}.$$

To capture equation (5.8.9) from these four values, compute the vector $\mathbf{y}$ defined by (5.8.8) to be

$$\mathbf{y} = \frac{2}{4}\mathbf{F}_4\mathbf{x} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix} \begin{pmatrix} 3 \\ 5 \\ -3 \\ -5 \end{pmatrix} = \begin{pmatrix} 0 \\ 3-5\mathrm{i} \\ 0 \\ 3+5\mathrm{i} \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 3 \\ 0 \\ 3 \end{pmatrix} + \mathrm{i}\begin{pmatrix} 0 \\ -5 \\ 0 \\ 5 \end{pmatrix} = 3(\mathbf{e}_1 + \mathbf{e}_3) + 5\mathrm{i}(-\mathbf{e}_1 + \mathbf{e}_3).$$

The real part of $\mathbf{y}$ tells us there is a cosine component with $amplitude = 3$ and $frequency = 1$, while the imaginary part of $\mathbf{y}$ says there is a sine component with $amplitude = 5$ and $frequency = 1$. This is depicted in the frequency domain shown in Figure 5.8.8.

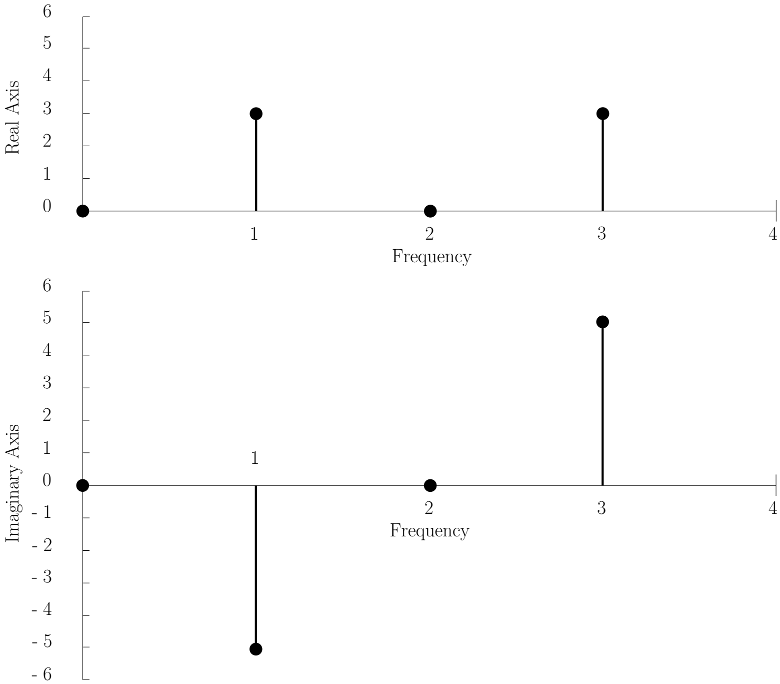


FIGURE 5.8.8

Putting this information together allows us to conclude that the equation of the waveform must be $x(\tau) = 3\cos 2\pi\tau + 5\sin 2\pi\tau$. Since

$$1 = \max\{f_k\} < \frac{n}{2} = \frac{4}{2} = 2,$$

the information in just the first half of $\mathbf{y}$

$$\mathbf{y}_{n/2} = \begin{pmatrix} 0 \\ 3 \end{pmatrix} + \mathrm{i}\begin{pmatrix} 0 \\ -5 \end{pmatrix} = 3\mathbf{e}_1 - 5\mathrm{i}\mathbf{e}_1$$

suffices to completely characterize $x(\tau)$.

These elementary ideas help explain why applying $\mathbf{F}$ to a sample from a signal can reveal the oscillatory components of the signal. But there is still a significant amount of theory that is well beyond the scope of this example. The purpose here is to just hint at how useful the discrete Fourier transform is and why it is so important in analyzing the nature of complicated waveforms.

If

$$\mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}_{n\times 1} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{pmatrix}_{n\times 1},$$

then the vector

$$\mathbf{a}\odot\mathbf{b} = \begin{pmatrix} \alpha_0\beta_0 \\ \alpha_0\beta_1 + \alpha_1\beta_0 \\ \alpha_0\beta_2 + \alpha_1\beta_1 + \alpha_2\beta_0 \\ \vdots \\ \alpha_{n-2}\beta_{n-1} + \alpha_{n-1}\beta_{n-2} \\ \alpha_{n-1}\beta_{n-1} \\ 0 \end{pmatrix}_{2n\times 1} \tag{5.8.10}$$

is called the ***convolution*** of $\mathbf{a}$ and $\mathbf{b}$. The 0 in the last position is for convenience only—it makes the size of the convolution twice the size of the original vectors, and this provides a balance in some of the formulas involving convolution. Furthermore, it is sometimes convenient to pad $\mathbf{a}$ and $\mathbf{b}$ with n additional zeros to consider them to be vectors with $2n$ components. Setting $\alpha_n = \cdots = \alpha_{2n-1} = \beta_n = \cdots = \beta_{2n-1} = 0$ allows us to write the k^{th} entry in $\mathbf{a}\odot\mathbf{b}$ as

$$[\mathbf{a}\odot\mathbf{b}]_k = \sum_{j=0}^{k} \alpha_j\beta_{k-j} \quad \text{for} \quad k = 0, 1, 2, \ldots, 2n-1.$$

A visual way to form $\mathbf{a}\odot\mathbf{b}$ is to "slide" the reversal of $\mathbf{b}$ "against" $\mathbf{a}$ as depicted in Figure 5.8.9, and then sum the resulting products.

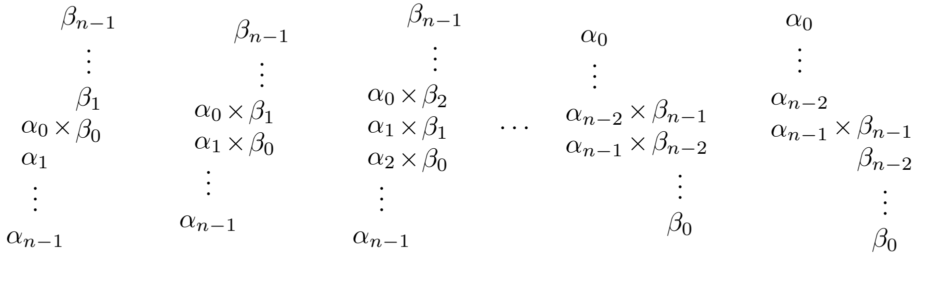


FIGURE 5.8.9

The convolution operation is a natural occurrence in a variety of situations, and polynomial multiplication is one such example.

Example 5.8.4

Polynomial Multiplication. For $p(x)=\sum_{k=0}^{n-1}\alpha_k x^k$, $q(x)=\sum_{k=0}^{n-1}\beta_k x^k$, let $\mathbf{a}=\begin{pmatrix}\alpha_0 & \alpha_1 & \cdots & \alpha_{n-1}\end{pmatrix}^T$ and $\mathbf{b}=\begin{pmatrix}\beta_0 & \beta_1 & \cdots & \beta_{n-1}\end{pmatrix}^T$. The product $p(x)q(x)=\gamma_0+\gamma_1 x+\gamma_2 x^2+\cdots+\gamma_{2n-2}x^{2n-2}$ is a polynomial of degree $2n-2$ in which γ_k is simply the k^{th} component of the convolution $\mathbf{a}\odot\mathbf{b}$ because

$$p(x)q(x)=\sum_{k=0}^{2n-2}\left[\sum_{j=0}^{k}\alpha_j\beta_{k-j}\right]x^k=\sum_{k=0}^{2n-2}[\mathbf{a}\odot\mathbf{b}]_k x^k. \tag{5.8.11}$$

In other words, polynomial multiplication and convolution are equivalent operations, so if we can devise an efficient way to perform a convolution, then we can efficiently multiply two polynomials, and conversely.

There are two facets involved in efficiently performing a convolution. The first is the realization that the discrete Fourier transform has the ability to convert a convolution into an ordinary product, and vice versa. The second is the realization that it's possible to devise a fast algorithm to compute a discrete Fourier transform. These two facets are developed below.

Convolution Theorem

Let $\mathbf{a}\times\mathbf{b}$ denote the entry-by-entry product

$$\mathbf{a}\times\mathbf{b}=\begin{pmatrix}\alpha_0\\ \alpha_1\\ \vdots\\ \alpha_{n-1}\end{pmatrix}\times\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_{n-1}\end{pmatrix}=\begin{pmatrix}\alpha_0\beta_0\\ \alpha_1\beta_1\\ \vdots\\ \alpha_{n-1}\beta_{n-1}\end{pmatrix}_{n\times 1},$$

and let $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ be the padded vectors

$$\hat{\mathbf{a}}=\begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_{n-1}\\ 0\\ \vdots\\ 0\end{pmatrix}_{2n\times 1}\quad\text{and}\quad\hat{\mathbf{b}}=\begin{pmatrix}\beta_0\\ \vdots\\ \beta_{n-1}\\ 0\\ \vdots\\ 0\end{pmatrix}_{2n\times 1}.$$

If $\mathbf{F}=\mathbf{F}_{2n}$ is the Fourier matrix of order $2n$, then

$$\mathbf{F}(\mathbf{a}\odot\mathbf{b})=(\mathbf{F}\hat{\mathbf{a}})\times(\mathbf{F}\hat{\mathbf{b}})\quad\text{and}\quad\mathbf{a}\odot\mathbf{b}=\mathbf{F}^{-1}\big[(\mathbf{F}\hat{\mathbf{a}})\times(\mathbf{F}\hat{\mathbf{b}})\big]. \tag{5.8.12}$$

Proof. Observe that the t^{th} component in $\mathbf{F}_{*j} \times \mathbf{F}_{*k}$ is

$$[\mathbf{F}_{*j} \times \mathbf{F}_{*k}]_t = \xi^{tj}\xi^{tk} = \xi^{t(j+k)} = [\mathbf{F}_{*j+k}]_t,$$

so that the columns of $\mathbf{F}$ have the property that

$$\mathbf{F}_{*j} \times \mathbf{F}_{*k} = \mathbf{F}_{*j+k} \quad \text{for each} \quad j, k = 0, 1, \ldots, (n-1).$$

This means that if $\mathbf{F}\hat{\mathbf{a}}$, $\mathbf{F}\hat{\mathbf{b}}$, and $\mathbf{F}(\mathbf{a} \odot \mathbf{b})$ are expressed as combinations of columns of $\mathbf{F}$ as indicated below,

$$\mathbf{F}\hat{\mathbf{a}} = \sum_{k=0}^{n-1} \alpha_k \mathbf{F}_{*k}, \quad \mathbf{F}\hat{\mathbf{b}} = \sum_{k=0}^{n-1} \beta_k \mathbf{F}_{*k}, \quad \text{and} \quad \mathbf{F}(\mathbf{a} \odot \mathbf{b}) = \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{b}]_k \mathbf{F}_{*k},$$

then the computation of $(\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}})$ is exactly the same as forming the product of two polynomials in the sense that

$$\begin{aligned}(\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}}) &= \left(\sum_{k=0}^{n-1} \alpha_k \mathbf{F}_{*k}\right)\left(\sum_{k=0}^{n-1} \beta_k \mathbf{F}_{*k}\right) = \sum_{k=0}^{2n-2} \left[\sum_{j=0}^{k} \alpha_j \beta_{k-j}\right] \mathbf{F}_{*k} \\ &= \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{b}]_k \mathbf{F}_{*k} = \mathbf{F}(\mathbf{a} \odot \mathbf{b}). \quad \blacksquare\end{aligned}$$

According to the convolution theorem, the convolution of two $n \times 1$ vectors can be computed by executing three discrete Fourier transforms of order $2n$

$$\mathbf{a}_{n\times 1} \odot \mathbf{b}_{n\times 1} = \mathbf{F}_{2n}^{-1}\big[(\mathbf{F}_{2n}\hat{\mathbf{a}}) \times (\mathbf{F}_{2n}\hat{\mathbf{b}})\big]. \tag{5.8.13}$$

The fact that one of them is an inverse transform is not a source of difficulty—recall Example 5.8.2. But it is still not clear that much has been accomplished. Performing a convolution by following the recipe called for in definition (5.8.10) requires n^2 scalar multiplications (you are asked to verify this in the exercises). Performing a discrete Fourier transform of order $2n$ by standard matrix–vector multiplication requires $4n^2$ scalar multiplications, so using matrix–vector multiplication to perform the computations on the right-hand side of (5.8.13) requires at least 12 times the number of scalar multiplications demanded by the definition of convolution. So, if there is an advantage to be gained by using the convolution theorem, then it is necessary to be able to perform a discrete Fourier transform in far fewer scalar multiplications than that required by standard matrix–vector multiplication. It was not until 1965 that this hurdle was overcome. Two Americans, J. W. Cooley and J. W. Tukey, introduced a fast Fourier transform (FFT) algorithm that requires only on the order of $(n/2)\log_2 n$ scalar multiplications to compute $\mathbf{F}_n\mathbf{x}$. Using the FFT together with the convolution theorem requires only about $3n\log_2 n$ multiplications to perform a convolution of two $n\times 1$ vectors, and when n is large, this is significantly less than the n^2 factor demanded by the definition of convolution.

The magic of the fast Fourier transform algorithm emanates from the fact that if n is a power of 2, then a discrete Fourier transform of order n can be executed by performing two transforms of order $n/2$. To appreciate exactly how this comes about, observe that when $n=2^r$ we have $\left(\xi^j\right)^n=\left(\xi^{2j}\right)^{n/2}$, so

$$\begin{gathered}\left\{1,\,\xi,\,\xi^2,\,\xi^3,\,\ldots,\,\xi^{n-1}\right\}=\text{ the } n^{th} \text{ roots of unity}\\ \text{if and only if}\\ \left\{1,\,\xi^2,\,\xi^4,\,\xi^6,\,\ldots,\,\xi^{n-2}\right\}=\text{ the } (n/2)^{th} \text{ roots of unity.}\end{gathered}$$

This means that the (j,k)-entries in the Fourier matrices $\mathbf{F}_n$ and $\mathbf{F}_{n/2}$ are

$$[\mathbf{F}_{n}]_{jk}=\xi^{jk} \quad \text{and} \quad [\mathbf{F}_{n/2}]_{jk}=(\xi^2)^{jk}=\xi^{2jk}. \tag{5.8.14}$$

If the columns of $\mathbf{F}_n$ are permuted so that columns with even subscripts are listed before those with odd subscripts, and if $\mathbf{P}_n^T$ is the corresponding permutation matrix, then we can partition $\mathbf{F}_n\mathbf{P}_n^T$ as

$$\mathbf{F}_n\mathbf{P}_n^T=[\mathbf{F}_{*0}\,\mathbf{F}_{*2}\,\cdots\,\mathbf{F}_{*n-2}\,|\,\mathbf{F}_{*1}\,\mathbf{F}_{*3}\,\cdots\,\mathbf{F}_{*n-1}]=\begin{pmatrix}\mathbf{A}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{B}_{\frac{n}{2}\times\frac{n}{2}}\\ \mathbf{C}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{G}_{\frac{n}{2}\times\frac{n}{2}}\end{pmatrix}.$$

By using (5.8.14) together with the facts that

$$\xi^{nk}=1 \quad \text{and} \quad \xi^{n/2}=\cos\frac{2\pi(n/2)}{n}-\mathrm{i}\sin\frac{2\pi(n/2)}{n}=-1,$$

we see that the entries in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ are

$$\begin{aligned}\mathbf{A}_{jk}&=\mathbf{F}_{j,2k}=\xi^{2jk}=[\mathbf{F}_{n/2}]_{jk},\\ \mathbf{B}_{jk}&=\mathbf{F}_{j,2k+1}=\xi^{j(2k+1)}=\xi^j\xi^{2jk}=\xi^j[\mathbf{F}_{n/2}]_{jk},\\ \mathbf{C}_{jk}&=\mathbf{F}_{\frac{n}{2}+j,\,2k}=\xi^{(\frac{n}{2}+j)2k}=\xi^{nk}\xi^{2jk}=\xi^{2jk}=[\mathbf{F}_{n/2}]_{jk},\\ \mathbf{G}_{jk}&=\mathbf{F}_{\frac{n}{2}+j,\,2k+1}=\xi^{(\frac{n}{2}+j)(2k+1)}=\xi^{nk}\xi^{n/2}\xi^j\xi^{2jk}=-\xi^j\xi^{2jt}=-\xi^j[\mathbf{F}_{n/2}]_{jk}.\end{aligned}$$

In other words, if $\mathbf{D}_{n/2}$ is the diagonal matrix

$$\mathbf{D}_{n/2}=\begin{pmatrix}1 & & & & \\ & \xi & & & \\ & & \xi^2 & & \\ & & & \ddots & \\ & & & & \xi^{\frac{n}{2}-1}\end{pmatrix},$$

then

$$\mathbf{F}_n\mathbf{P}_n^T=\begin{pmatrix}\mathbf{A}_{(n/2)\times(n/2)} & \mathbf{B}_{(n/2)\times(n/2)}\\ \mathbf{C}_{(n/2)\times(n/2)} & \mathbf{G}_{(n/2)\times(n/2)}\end{pmatrix}=\begin{pmatrix}\mathbf{F}_{n/2} & \mathbf{D}_{n/2}\mathbf{F}_{n/2}\\ \mathbf{F}_{n/2} & -\mathbf{D}_{n/2}\mathbf{F}_{n/2}\end{pmatrix}.$$

This fundamental feature of the discrete Fourier transform is summarized below.

Decomposing the Fourier Matrix

If $n = 2^r$, then

$$\mathbf{F}_n = \begin{pmatrix} \mathbf{F}_{n/2} & \mathbf{D}_{n/2}\mathbf{F}_{n/2} \\ \mathbf{F}_{n/2} & -\mathbf{D}_{n/2}\mathbf{F}_{n/2} \end{pmatrix} \mathbf{P}_n, \tag{5.8.15}$$

where

$$\mathbf{D}_{n/2} = \begin{pmatrix} 1 & & & & \\ & \xi & & & \\ & & \xi^2 & & \\ & & & \ddots & \\ & & & & \xi^{\frac{n}{2}-1} \end{pmatrix}$$

contains half of the n^{th} roots of unity and $\mathbf{P}_n$ is the "even–odd" permutation matrix defined by

$$\mathbf{P}_n^T = [\mathbf{e}_0\, \mathbf{e}_2\, \mathbf{e}_4\, \cdots\, \mathbf{e}_{n-2} \,|\, \mathbf{e}_1\, \mathbf{e}_3\, \mathbf{e}_5\, \cdots\, \mathbf{e}_{n-1}] .$$

The decomposition (5.8.15) says that a discrete Fourier transform of order $n = 2^r$ can be accomplished by two Fourier transforms of order $n/2 = 2^{r-1}$, and this leads to the FFT algorithm. To get a feel for how the FFT works, consider the case when $n = 8$, and proceed to "divide and conquer." If

$$\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix}, \quad \text{then} \quad \mathbf{P}_8\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_2 \\ x_4 \\ x_6 \\ \hline x_1 \\ x_3 \\ x_5 \\ x_7 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_4^{(0)} \\ \hline \mathbf{x}_4^{(1)} \end{pmatrix},$$

so

$$\mathbf{F}_8\mathbf{x}_8 = \begin{pmatrix} \mathbf{F}_4 & \mathbf{D}_4\mathbf{F}_4 \\ \mathbf{F}_4 & -\mathbf{D}_4\mathbf{F}_4 \end{pmatrix} \begin{pmatrix} \mathbf{x}_4^{(0)} \\ \mathbf{x}_4^{(1)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_4\mathbf{x}_4^{(0)} + \mathbf{D}_4\mathbf{F}_4\mathbf{x}_4^{(1)} \\ \mathbf{F}_4\mathbf{x}_4^{(0)} - \mathbf{D}_4\mathbf{F}_4\mathbf{x}_4^{(1)} \end{pmatrix}. \tag{5.8.16}$$

But

$$\mathbf{P}_4\mathbf{x}_4^{(0)} = \begin{pmatrix} x_0 \\ x_4 \\ \hline x_2 \\ x_6 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \hline \mathbf{x}_2^{(1)} \end{pmatrix} \quad \text{and} \quad \mathbf{P}_4\mathbf{x}_4^{(1)} = \begin{pmatrix} x_1 \\ x_5 \\ \hline x_3 \\ x_7 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_2^{(2)} \\ \hline \mathbf{x}_2^{(3)} \end{pmatrix},$$

so

$$\mathbf{F}_4\mathbf{x}_4^{(0)} = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \mathbf{x}_2^{(1)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_2\mathbf{x}_2^{(0)} + \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(1)} \\ \mathbf{F}_2\mathbf{x}_2^{(0)} - \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(1)} \end{pmatrix}$$

and (5.8.17)

$$\mathbf{F}_4\mathbf{x}_4^{(1)} = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_2^{(2)} \\ \mathbf{x}_2^{(3)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_2\mathbf{x}_2^{(2)} + \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(3)} \\ \mathbf{F}_2\mathbf{x}_2^{(2)} - \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(3)} \end{pmatrix}.$$

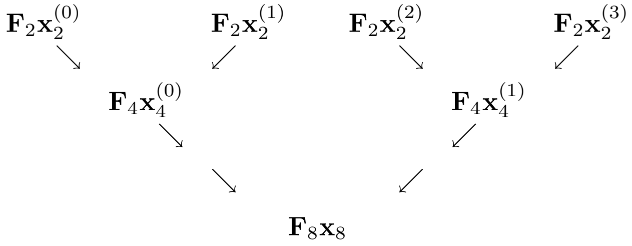
Now, since $\mathbf{F}_2 = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, it is a trivial matter to compute the terms

$$\mathbf{F}_2\mathbf{x}_2^{(0)}, \quad \mathbf{F}_2\mathbf{x}_2^{(1)}, \quad \mathbf{F}_2\mathbf{x}_2^{(2)}, \quad \mathbf{F}_2\mathbf{x}_2^{(3)}.$$

Of course, to actually carry out the computation, we need to work backward through the preceding sequence of steps. That is, we start with

$$\tilde{\mathbf{x}}_8 = \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \hline \mathbf{x}_2^{(1)} \\ \hline \mathbf{x}_2^{(2)} \\ \hline \mathbf{x}_2^{(3)} \end{pmatrix} = \begin{pmatrix} x_0 \\ x_4 \\ \hline x_2 \\ x_6 \\ \hline x_1 \\ x_5 \\ \hline x_3 \\ x_7 \end{pmatrix}, \tag{5.8.18}$$

and use (5.8.17) followed by (5.8.16) to work downward in the following tree.


$\mathbf{F}_2\mathbf{x}_2^{(0)}$
$\mathbf{F}_2\mathbf{x}_2^{(1)}$
$\mathbf{F}_2\mathbf{x}_2^{(2)}$
$\mathbf{F}_2\mathbf{x}_2^{(3)}$
$\mathbf{F}_4\mathbf{x}_4^{(0)}$
$\mathbf{F}_4\mathbf{x}_4^{(1)}$
$\mathbf{F}_8\mathbf{x}_8$

But there appears to be a snag. In order to work downward through this tree, we cannot start directly with $\mathbf{x}_8$—we must start with the permutation $\tilde{\mathbf{x}}_8$ shown in (5.8.18). So how is this initial permutation determined? Looking back reveals that the entries in $\tilde{\mathbf{x}}_8$ were obtained by first sorting the x_j's into two groups—the entries in the even positions were separated from those in the odd positions. Then each group was broken into two more groups by again separating the entries in the even positions from those in the odd positions.

$$\begin{tikzcd}[column sep=small]
 & & & \begin{pmatrix}0 & 1 & 2 & 3 & 4 & 5 & 6 & 7\end{pmatrix} \arrow[dll] \arrow[drr] & & & \\
 & \begin{pmatrix}0 & 2 & 4 & 6\end{pmatrix} \arrow[dl] \arrow[dr] & & & & \begin{pmatrix}1 & 3 & 5 & 7\end{pmatrix} \arrow[dl] \arrow[dr] & \\
 \begin{pmatrix}0 & 4\end{pmatrix} & & \begin{pmatrix}2 & 6\end{pmatrix} & & \begin{pmatrix}1 & 5\end{pmatrix} & & \begin{pmatrix}3 & 7\end{pmatrix}
\end{tikzcd} \tag{5.8.19}$$

In general, this even–odd sorting process (sometimes called a ***perfect shuffle***) produces the permutation necessary to initiate the algorithm. A clever way to perform a perfect shuffle is to use binary representations and observe that the first level of sorting in (5.8.19) is determined according to whether the least significant bit is 0 or 1, the second level of sorting is determined by the second least significant bit, and so on—this is illustrated in Table 5.8.1 for $n = 8$.

Table 5.8.1

Natural order	**First level**	**Second level**
$0 \leftrightarrow 000$	$0 \leftrightarrow 00\mathbf{0}$	$0 \leftrightarrow 0\mathbf{0}0$
$1 \leftrightarrow 001$	$2 \leftrightarrow 01\mathbf{0}$	$4 \leftrightarrow 1\mathbf{0}0$
$2 \leftrightarrow 010$	$4 \leftrightarrow 10\mathbf{0}$	$2 \leftrightarrow 0\mathbf{1}0$
$3 \leftrightarrow 011$	$6 \leftrightarrow 11\mathbf{0}$	$6 \leftrightarrow 1\mathbf{1}0$
$4 \leftrightarrow 100$	$1 \leftrightarrow 00\mathbf{1}$	$1 \leftrightarrow 0\mathbf{0}1$
$5 \leftrightarrow 101$	$3 \leftrightarrow 01\mathbf{1}$	$5 \leftrightarrow 1\mathbf{0}1$
$6 \leftrightarrow 110$	$5 \leftrightarrow 10\mathbf{1}$	$3 \leftrightarrow 0\mathbf{1}1$
$7 \leftrightarrow 111$	$7 \leftrightarrow 11\mathbf{1}$	$7 \leftrightarrow 1\mathbf{1}1$

But all intermediate levels in this sorting process can be eliminated because something very nice occurs. Examination of the last column in Table 5.8.1 reveals that the binary bits in the perfect shuffle ordering are exactly the reversal of the binary bits in the natural ordering. In other words,

- to generate the perfect shuffle of the numbers $0, 1, 2, \ldots, n-1$, simply reverse the bits in the binary representation of each number.

We can summarize the fast Fourier transform by the following implementation that utilizes array operations.[52]

[52] There are a variety of different ways to implement the FFT, and choosing a practical implementation frequently depends on the hardware being used as well as the application under consideration. The FFT ranks high on the list of useful algorithms because it provides an advantage in a large variety of applications, and there are many more facets of the FFT than those presented here (e.g., FFT when n is not a power of 2). In fact, there are entire texts devoted to these issues, so the interested student need only go as far as the nearest library to find more details.

Fast Fourier Transform

For a given input vector $\mathbf{x}$ containing $n = 2^r$ components, the discrete Fourier transform $\mathbf{F}_n\mathbf{x}$ is the result of successively creating the following arrays.

$$\mathbf{X}_{1\times n} \longleftarrow rev(\mathbf{x}) \qquad \text{(bit reverse the subscripts)}$$

For $j = 0,\, 1,\, 2,\, 3,\, \ldots,\, r-1$

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ \mathrm{e}^{-\pi\mathrm{i}/2^j} \\ \mathrm{e}^{-2\pi\mathrm{i}/2^j} \\ \mathrm{e}^{-3\pi\mathrm{i}/2^j} \\ \vdots \\ \mathrm{e}^{-(2^j-1)\pi\mathrm{i}/2^j} \end{pmatrix}_{j+1\times 1} \qquad \begin{array}{l}\text{(Half of the } (2^{j+1})^{th} \text{ roots of 1,} \\ \text{perhaps from a lookup table)}\end{array}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} \mathbf{X}_{*0} & \mathbf{X}_{*2} & \mathbf{X}_{*4} & \cdots & \mathbf{X}_{*2^{r-j}-2} \end{pmatrix}_{2^j\times 2^{r-j-1}}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} \mathbf{X}_{*1} & \mathbf{X}_{*3} & \mathbf{X}_{*5} & \cdots & \mathbf{X}_{*2^{r-j}-1} \end{pmatrix}_{2^j\times 2^{r-j-1}}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix}_{2^{j+1}\times 2^{r-j-1}} \qquad \begin{pmatrix} \times \text{ denotes entry-} \\ \text{by-entry product} \end{pmatrix}$$

Example 5.8.5

Problem: Perform the FFT on $\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

Solution: Start with $\mathbf{X} \longleftarrow rev(\mathbf{x}) = \begin{pmatrix} x_0 & x_2 & x_1 & x_3 \end{pmatrix}$.

For $j = 0$:

$$\mathbf{D} \longleftarrow (1) \qquad \text{(Half of the square roots of 1)}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} x_0 & x_1 \end{pmatrix}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_2 & x_3 \end{pmatrix} \quad \text{and} \quad \mathbf{D} \times \mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_2 & x_3 \end{pmatrix}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix} = \begin{pmatrix} x_0 + x_2 & x_1 + x_3 \\ x_0 - x_2 & x_1 - x_3 \end{pmatrix}$$

For $j = 1$:

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ -\mathrm{i} \end{pmatrix} \qquad \text{(Half of the } 4^{th} \text{ roots of 1)}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} x_0 + x_2 \\ x_0 - x_2 \end{pmatrix}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_1 + x_3 \\ x_1 - x_3 \end{pmatrix} \quad \text{and} \quad \mathbf{D} \times \mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_1 + \ x_3 \\ -\mathrm{i}x_1 + \mathrm{i}x_3 \end{pmatrix}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix} = \begin{pmatrix} x_0 + x_2 + \ x_1 + \ x_3 \\ x_0 - x_2 - \mathrm{i}x_1 + \mathrm{i}x_3 \\ x_0 + x_2 - \ x_1 - \ x_3 \\ x_0 - x_2 + \mathrm{i}x_1 - \mathrm{i}x_3 \end{pmatrix} = \mathbf{F}_4\mathbf{x}$$

Notice that this agrees with the result obtained by using direct matrix–vector multiplication with $\mathbf{F}_4$ given in Example 5.8.1.

To understand why it is called the “fast” Fourier transform, simply count the number of multiplications the FFT requires. Observe that the j^{th} iteration requires 2^j multiplications for each column in $\mathbf{X}^{(1)}$, and there are 2^{r-j-1} columns, so 2^{r-1} multiplications are used for each iteration.[53] Since r iterations are required, the total number of multiplications used by the FFT does not exceed $2^{r-1}r = (n/2)\log_2 n$.

FFT Multiplication Count

If n is a power of 2, then applying the FFT to a vector of n components requires at most $(n/2)\log_2 n$ multiplications.

The $(n/2)\log_2 n$ count represents a tremendous advantage over the n^2 factor demanded by a direct matrix–vector product. To appreciate the magnitude of the difference between n^2 and $(n/2)\log_2 n$, look at Figure 5.8.10.

[53] Actually, we can get by with slightly fewer multiplications if we take advantage of the fact that the first entry in $\mathbf{D}$ is always 1 and if we observe that no multiplications are necessary when $j = 0$. But when n is large, these savings are relatively insignificant, so they are ignored in the multiplication count.

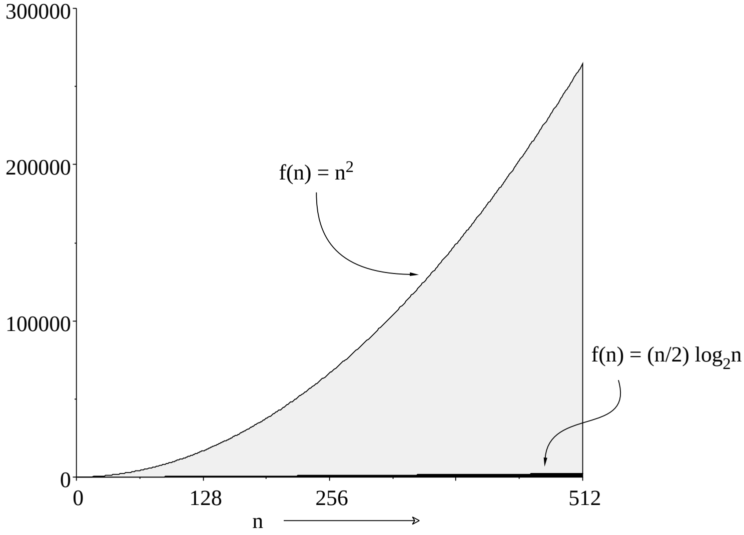


FIGURE 5.8.10

The small dark portion at the bottom of the graph is the area under the curve $f(n) = (n/2)\log_2 n$—it is tiny in comparison to the area under $f(n) = n^2$. For example, if $n = 512$, then $n^2 = 262,144$, but $(n/2)\log_2 n = 2304$. In other words, for $n = 512$, the FFT is on the order of 100 times faster than straightforward matrix–vector multiplication, and for larger values of n, the gap is even wider—Figure 5.8.10 illustrates just how fast the difference between n^2 and $(n/2)\log_2 n$ grows as n increases. Since Cooley and Tukey introduced the FFT in 1965, it has risen to a position of fundamental importance. The FFT and the convolution theorem are extremely powerful tools, and they have been principal components of the computational revolution that now touches our lives countless times each day.

Example 5.8.6

Problem: Fast Integer Multiplication. Consider two positive integers whose base-b representations are

$$c = (\gamma_{n-1}\gamma_{n-2}\cdots\gamma_1\gamma_0)_b \quad \text{and} \quad d = (\delta_{n-1}\delta_{n-2}\cdots\delta_1\delta_0)_b.$$

Use the convolution theorem together with the FFT to compute the product cd.

Solution: If we let

$$p(x) = \sum_{k=0}^{n-1}\gamma_k x^k, \quad q(x) = \sum_{k=0}^{n-1}\delta_k x^k, \quad \mathbf{c} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{n-1} \end{pmatrix}, \quad \text{and} \quad \mathbf{d} = \begin{pmatrix} \delta_0 \\ \delta_1 \\ \vdots \\ \delta_{n-1} \end{pmatrix},$$

then

$$\begin{aligned} c &= \gamma_{n-1}b^{n-1} + \gamma_{n-2}b^{n-2} + \cdots + \gamma_1 b^1 + \gamma_0 b^0 = p(b), \\ d &= \delta_{n-1}b^{n-1} + \delta_{n-2}b^{n-2} + \cdots + \delta_1 b^1 + \delta_0 b^0 = q(b), \end{aligned}$$

and it follows from (5.8.11) that the product cd is given by

$$cd = p(b)q(b) = [\mathbf{c}\odot\mathbf{d}]_{2n-2}b^{2n-2} + [\mathbf{c}\odot\mathbf{d}]_{2n-3}b^{2n-3} + \cdots + [\mathbf{c}\odot\mathbf{d}]_1 b^1 + [\mathbf{c}\odot\mathbf{d}]_0 b^0.$$

It looks as though the convolution $\mathbf{c}\odot\mathbf{d}$ provides the base-b representation for cd, but this is not quite the case because it is possible to have some $[\mathbf{c}\odot\mathbf{d}]_k \geq b$. For example, if $c = 201_{10}$ and $d = 425_{10}$, then

$$\mathbf{c}\odot\mathbf{d} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \odot \begin{pmatrix} 5 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \\ 14 \\ 4 \\ 8 \\ 0 \end{pmatrix},$$

so

$$cd = (8\times10^4) + (4\times10^3) + (14\times10^2) + (2\times10^1) + (5\times10^0). \qquad (5.8.20)$$

But when numbers like 14 (i.e., greater than or equal to the base) appear in $\mathbf{c}\odot\mathbf{d}$, it is relatively easy to decompose them by writing $14 = (1\times10^1) + (4\times10^0)$, so

$$14\times10^2 = \big[(1\times10^1) + (4\times10^0)\big]\times10^2 = (1\times10^3) + (4\times10^2).$$

Substituting this in (5.8.20) and combining coefficients of like powers produces the base-10 representation of the product

$$cd = (8\times10^4) + (5\times10^3) + (4\times10^2) + (2\times10^1) + (5\times10^0) = 85425_{10}.$$

Computing $\mathbf{c}\odot\mathbf{d}$ directly demands n^2 multiplications, but using the FFT in conjunction with the convolution theorem requires only about $3n\log_2 n$ multiplications, which is considerably less than n^2 for large values of n. Thus it is possible to multiply very long base-b integers much faster than by using direct methods. Most digital computers have binary integer multiplication (usually 64-bit multiplication not requiring the FFT) built into their hardware, but for ultra-high-precision multiplication or for more general base-b multiplication, the FFT is a viable tool.

Exercises for section 5.8

5.8.1. Evaluate the following convolutions.

(a) $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \odot \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$. (b) $\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \odot \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$. (c) $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \odot \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$.

5.8.2. (a) Evaluate the discrete Fourier transform of $\begin{pmatrix} 1 \\ -\mathrm{i} \\ -1 \\ \mathrm{i} \end{pmatrix}$.

(b) Evaluate the inverse transform of $\begin{pmatrix} 1 \\ \mathrm{i} \\ -1 \\ -\mathrm{i} \end{pmatrix}$.

5.8.3. Verify directly that $\mathbf{F}_4 = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix}\mathbf{P}_4$, where the $\mathbf{F}_4$, $\mathbf{P}_4$, and $\mathbf{D}_2$, are as defined in (5.8.15).

5.8.4. Use the following vectors to perform the indicated computations:

$$\mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \hat{\mathbf{a}} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ 0 \\ 0 \end{pmatrix}, \quad \hat{\mathbf{b}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ 0 \\ 0 \end{pmatrix}.$$

(a) Compute $\mathbf{a} \odot \mathbf{b}$, $\mathbf{F}_4(\mathbf{a} \odot \mathbf{b})$, and $(\mathbf{F}_4\hat{\mathbf{a}}) \times (\mathbf{F}_4\hat{\mathbf{b}})$.

(b) By using $\mathbf{F}_4^{-1}$ as given in Example 5.8.1, compute

$$\mathbf{F}_4^{-1}\big[(\mathbf{F}_4\hat{\mathbf{a}}) \times (\mathbf{F}_4\hat{\mathbf{b}})\big].$$

Compare this with the results guaranteed by the convolution theorem.

5.8.5. For $p(x) = 2x - 3$ and $q(x) = 3x - 4$, compute the product $p(x)q(x)$ by using the convolution theorem.

5.8.6. Use convolutions to form the following products.

(a) $43_{10} \times 21_{10}$. (b) $123_8 \times 601_8$. (c) $1010_2 \times 1101_2$.

5.8.7. Let $\mathbf{a}$ and $\mathbf{b}$ be $n \times 1$ vectors, where n is a power of 2.

(a) Show that the number of multiplications required to form $\mathbf{a} \odot \mathbf{b}$ by using the definition of convolution is n^2.
Hint: $1 + 2 + \cdots + k = k(k+1)/2$.

(b) Show that the number of multiplications required to form $\mathbf{a} \odot \mathbf{b}$ by using the FFT in conjunction with the convolution theorem is $3n\log_2 n + 7n$. Sketch a graph of $3n\log_2 n$ (the $7n$ factor is dropped because it is not significant), and compare it with the graph of n^2 to illustrate why the FFT in conjunction with the convolution theorem provides such a big advantage.

5.8.8. A waveform given by a finite sum

$$x(\tau) = \sum_k (\alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau)$$

in which the f_k's are integers and $\max\{f_k\} \leq 3$ is sampled at eight equally spaced points between $\tau = 0$ and $\tau = 1$. Let

$$\mathbf{x} = \begin{pmatrix} x(0/8) \\ x(1/8) \\ x(2/8) \\ x(3/8) \\ x(4/8) \\ x(5/8) \\ x(6/8) \\ x(7/8) \end{pmatrix}, \quad \text{and suppose that} \quad \mathbf{y} = \frac{1}{4}\mathbf{F}_8\mathbf{x} = \begin{pmatrix} 0 \\ -5\mathrm{i} \\ 1-3\mathrm{i} \\ 4 \\ 0 \\ 4 \\ 1+3\mathrm{i} \\ 5\mathrm{i} \end{pmatrix}.$$

What is the equation of the waveform?

5.8.9. Prove that $\mathbf{a} \odot \mathbf{b} = \mathbf{b} \odot \mathbf{a}$ for all $\mathbf{a}, \mathbf{b} \in \mathcal{C}^n$—i.e., convolution is a commutative operation.

5.8.10. For $p(x) = \sum_{k=0}^{n-1} \alpha_k x^k$ and the n^{th} roots of unity ξ_k, let

$$\mathbf{a} = (\,\alpha_0\ \alpha_1\ \alpha_2\ \cdots \alpha_{n-1}\,)^T \quad \text{and} \quad \mathbf{p} = (\,p(1)\ p(\xi)\ p(\xi^2)\ \cdots p(\xi^{n-1})\,)^T.$$

Explain why $\mathbf{F}_n\mathbf{a} = \mathbf{p}$ and $\mathbf{a} = \mathbf{F}_n^{-1}\mathbf{p}$. This says that the discrete Fourier transform allows us to go from the representation of a polynomial p in terms of its coefficients α_k to the representation of p in terms of its values $p(\xi^k)$, and the inverse transform takes us in the other direction.

5.8.11. For two polynomials $p(x) = \sum_{k=0}^{n-1} \alpha_k x^k$ and $q(x) = \sum_{k=0}^{n-1} \beta_k x^k$, let

$$\mathbf{p} = \begin{pmatrix} p(1) \\ p(\xi) \\ \vdots \\ p(\xi^{2n-1}) \end{pmatrix} \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} q(1) \\ q(\xi) \\ \vdots \\ q(\xi^{2n-1}) \end{pmatrix},$$

where $\{1, \xi, \xi^2, \ldots, \xi^{2n-1}\}$ are now the $2n^{th}$ roots of unity. Explain why the coefficients in the product

$$p(x)q(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \cdots + \gamma_{2n-2}x^{2n-2}$$

must be given by

$$\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \end{pmatrix} = \mathbf{F}_{2n}^{-1} \begin{pmatrix} p(1)q(1) \\ p(\xi)q(\xi) \\ p(\xi^2)q(\xi^2) \\ \vdots \end{pmatrix}.$$

This says that the product $p(x)q(x)$ is completely determined by the values of $p(x)$ and $q(x)$ at the $2n^{th}$ roots of unity.

5.8.12. A ***circulant*** matrix is defined to be a square matrix that has the form

$$\mathbf{C} = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ c_2 & c_1 & c_0 & \cdots & c_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_0 \end{pmatrix}_{n\times n}.$$

In other words, the entries in each column are the same as the previous column, but they are shifted one position downward and wrapped around at the top—the (j,k)-entry in $\mathbf{C}$ can be described as $c_{jk} = c_{j-k\,(\text{mod } n)}$. (Some authors use $\mathbf{C}^T$ rather than $\mathbf{C}$ as the definition—it doesn't matter.)

(a) If $\mathbf{Q}$ is the circulant matrix defined by

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}_{n\times n},$$

and if $p(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1}$, verify that

$$\mathbf{C} = p(\mathbf{Q}) = c_0\mathbf{I} + c_1\mathbf{Q} + \cdots + c_{n-1}\mathbf{Q}^{n-1}.$$

(b) Explain why the Fourier matrix of order n diagonalizes $\mathbf{Q}$ in the sense that

$$\mathbf{FQF}^{-1} = \mathbf{D} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \xi & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \xi^{n-1} \end{pmatrix},$$

where the ξ^k 's are the n^{th} roots of unity.

(c) Prove that the Fourier matrix of order n diagonalizes every $n \times n$ circulant in the sense that

$$\mathbf{FCF}^{-1} = \begin{pmatrix} p(1) & 0 & \cdots & 0 \\ 0 & p(\xi) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\xi^{n-1}) \end{pmatrix},$$

where $p(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1}$.

(d) If $\mathbf{C}_1$ and $\mathbf{C}_2$ are any pair of $n \times n$ circulants, explain why $\mathbf{C}_1\mathbf{C}_2 = \mathbf{C}_2\mathbf{C}_1$—i.e., all circulants commute with each other.

5.8.13. For a nonsingular circulant $\mathbf{C}_{n\times n}$, explain how to use the FFT algorithm to efficiently perform the following operations.

(a) Solve a system $\mathbf{Cx} = \mathbf{b}$.

(b) Compute $\mathbf{C}^{-1}$.

(c) Multiply two circulants $\mathbf{C}_1\mathbf{C}_2$.

5.8.14. For the vectors

$$\mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{pmatrix}, \quad \hat{\mathbf{a}} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n\times 1}, \text{ and } \hat{\mathbf{b}} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n\times 1},$$

let $\mathbf{C}$ be the $2n \times 2n$ circulant matrix (see Exercise 5.8.12) whose first column is $\hat{\mathbf{a}}$.

(a) Show that the convolution operation can be described as a matrix–vector product by demonstrating that

$$\mathbf{a} \odot \mathbf{b} = \mathbf{C}\hat{\mathbf{b}}.$$

(b) Use this relationship to give an alternate proof of the convolution theorem. **Hint:** Use the diagonalization result of Exercise 5.8.12 together with the result of Exercise 5.8.10.

5.8.15. The ***Kronecker product*** of two matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{p\times q}$ is defined to be the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}.$$

This is also known as the ***tensor product*** or the ***direct product***. Although there is an extensive list of properties that the tensor product satisfies, this exercise requires only the following two elementary facts (which you need not prove unless you feel up to it). The complete list of properties is given in Exercise 7.8.11 (p. 597) along with remarks about Kronecker, and another application appears in Exercise 7.6.10 (p. 573).

$$\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}.$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD} \quad (\text{if } \mathbf{AC} \text{ and } \mathbf{BD} \text{ exist}).$$

(a) If $n = 2^r$, and if $\mathbf{P}_n$ is the even–odd permutation matrix described in (5.8.15), explain why

$$\mathbf{R}_n = (\mathbf{I}_{2^{r-1}} \otimes \mathbf{P}_{2^1})(\mathbf{I}_{2^{r-2}} \otimes \mathbf{P}_{2^2}) \cdots (\mathbf{I}_{2^1} \otimes \mathbf{P}_{2^{r-1}})(\mathbf{I}_{2^0} \otimes \mathbf{P}_{2^r})$$

is the permutation matrix associated with the bit reversing (or perfect shuffle) permutation described in (5.8.19) and Table 5.8.1. **Hint:** Work it out for $n = 8$ by showing

$$\mathbf{R}_8 \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} x_0 \\ x_4 \\ x_2 \\ x_6 \\ x_1 \\ x_5 \\ x_3 \\ x_7 \end{pmatrix},$$

and you will see why it holds in general.

(b) Suppose $n = 2^r$, and set

$$\mathbf{B}_n = \begin{pmatrix} \mathbf{I}_{n/2} & \mathbf{D}_{n/2} \\ \mathbf{I}_{n/2} & -\mathbf{D}_{n/2} \end{pmatrix}.$$

According to (5.8.15), the Fourier matrix can be written as

$$\mathbf{F}_n = \mathbf{B}_n(\mathbf{I}_2 \otimes \mathbf{F}_{n/2})\mathbf{P}_n.$$

Expand on this idea by proving that $\mathbf{F}_n$ can be factored as

$$\mathbf{F}_n = \mathbf{L}_n\mathbf{R}_n$$

in which

$$\mathbf{L}_n = (\mathbf{I}_{2^0} \otimes \mathbf{B}_{2^r})(\mathbf{I}_{2^1} \otimes \mathbf{B}_{2^{r-1}}) \cdots (\mathbf{I}_{2^{r-2}} \otimes \mathbf{B}_{2^2})(\mathbf{I}_{2^{r-1}} \otimes \mathbf{B}_{2^1}),$$

and where $\mathbf{R}_n$ is the bit reversing permutation

$$\mathbf{R}_n = (\mathbf{I}_{2^{r-1}} \otimes \mathbf{P}_{2^1})(\mathbf{I}_{2^{r-2}} \otimes \mathbf{P}_{2^2}) \cdots (\mathbf{I}_{2^1} \otimes \mathbf{P}_{2^{r-1}})(\mathbf{I}_{2^0} \otimes \mathbf{P}_{2^r}).$$

Notice that this says $\mathbf{F}_n\mathbf{x} = \mathbf{L}_n\mathbf{R}_n\mathbf{x}$, so the discrete Fourier transform of $\mathbf{x}$ is obtained by first performing the bit reversing permutation to $\mathbf{x}$ followed by r applications of the terms $(\mathbf{I}_{2^{r-k}} \otimes \mathbf{B}_{2^k})$ from $\mathbf{L}_n$. This in fact is the FFT algorithm in factored form. **Hint:** Define two sequences by the rules

$$\mathbf{L}_{2^k} = (\mathbf{I}_{2^{r-k}} \otimes \mathbf{B}_{2^k})\,\mathbf{L}_{2^{k-1}} \quad \text{and} \quad \mathbf{R}_{2^k} = \mathbf{R}_{2^{k-1}}\,(\mathbf{I}_{2^{r-k}} \otimes \mathbf{P}_{2^k}),$$

where

$$\mathbf{L}_1 = 1, \quad \mathbf{R}_1 = \mathbf{I}_n, \quad \mathbf{B}_2 = \mathbf{F}_2, \quad \mathbf{P}_2 = \mathbf{I}_2,$$

and use induction on k to prove that

$$\mathbf{I}_{2^{r-k}} \otimes \mathbf{F}_{2^k} = \mathbf{L}_{2^k}\mathbf{R}_{2^k} \quad \text{for} \quad k = 1, 2, \ldots, r.$$

5.8.16. For $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{n-1}x^{n-1}$, prove that

$$\frac{1}{n}\sum_{k=0}^{n-1} \left|p(\xi^k)\right|^2 = |\alpha_0|^2 + |\alpha_1|^2 + \cdots + |\alpha_{n-1}|^2,$$

where $\left\{1, \xi, \xi^2, \ldots, \xi^{n-1}\right\}$ are the n^{th} roots of unity.

5.8.17. Consider a waveform that is given by the finite sum

$$x(\tau) = \sum_k (\alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau)$$

in which the f_k 's are distinct integers, and let

$$\mathbf{x} = \sum_k (\alpha_k \cos 2\pi f_k \mathbf{t} + \beta_k \sin 2\pi f_k \mathbf{t})$$

be the vector containing the values of $x(\tau)$ at $n > 2\max\{f_k\}$ equally spaced points between $\tau = 0$ and $\tau = 1$ as described in Example 5.8.3. Use the discrete Fourier transform to prove that

$$\|\mathbf{x}\|_2^2 = \frac{n}{2}\sum_k \left(\alpha_k^2 + \beta_k^2\right).$$

5.8.18. Let η be an arbitrary scalar, and let

$$\mathbf{c} = \begin{pmatrix} 1 \\ \eta \\ \eta^2 \\ \vdots \\ \eta^{2n-1} \end{pmatrix} \quad \text{and} \quad \mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}.$$

Prove that $\mathbf{c}^T(\mathbf{a} \odot \mathbf{a}) = \left(\mathbf{c}^T\hat{\mathbf{a}}\right)^2$.

5.8.19. Apply the FFT algorithm to the vector $\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_7 \end{pmatrix}$, and then verify that your answer agrees with the result obtained by computing $\mathbf{F}_8\mathbf{x}_8$ directly.

5.9 COMPLEMENTARY SUBSPACES

The sum of two subspaces $\mathcal{X}$ and $\mathcal{Y}$ of a vector space $\mathcal{V}$ was defined on p. 166 to be the set $\mathcal{X} + \mathcal{Y} = \{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in \mathcal{X} \text{ and } \mathbf{y} \in \mathcal{Y}\}$, and it was established that $\mathcal{X} + \mathcal{Y}$ is another subspace of $\mathcal{V}$. For example, consider the two subspaces of $\Re^3$ shown in Figure 5.9.1 in which $\mathcal{X}$ is a plane through the origin, and $\mathcal{Y}$ is a line through the origin.

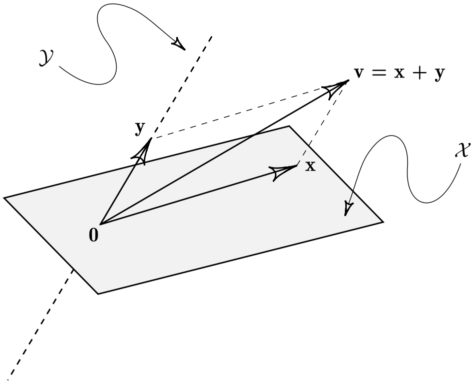


FIGURE 5.9.1

Notice that $\mathcal{X}$ and $\mathcal{Y}$ are ***disjoint*** in the sense that $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$. The parallelogram law for vector addition makes it clear that $\mathcal{X} + \mathcal{Y} = \Re^3$ because each vector in $\Re^3$ can be written as "something from $\mathcal{X}$ plus something from $\mathcal{Y}$." Thus $\Re^3$ is resolved into a pair of disjoint components $\mathcal{X}$ and $\mathcal{Y}$. These ideas generalize as described below.

Complementary Subspaces

Subspaces $\mathcal{X}, \mathcal{Y}$ of a space $\mathcal{V}$ are said to be ***complementary*** whenever

$$\mathcal{V} = \mathcal{X} + \mathcal{Y} \quad \text{and} \quad \mathcal{X} \cap \mathcal{Y} = \mathbf{0}, \tag{5.9.1}$$

in which case $\mathcal{V}$ is said to be the ***direct sum*** of $\mathcal{X}$ and $\mathcal{Y}$, and this is denoted by writing $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$.

- For a vector space $\mathcal{V}$ with subspaces $\mathcal{X}, \mathcal{Y}$ having respective bases $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$, the following statements are equivalent.
  - ▷ $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$. (5.9.2)
  - ▷ For each $\mathbf{v} \in \mathcal{V}$ there are *unique* vectors $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ such that $\mathbf{v} = \mathbf{x} + \mathbf{y}$. (5.9.3)
  - ▷ $\mathcal{B}_{\mathcal{X}} \cap \mathcal{B}_{\mathcal{Y}} = \phi$ and $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\mathcal{V}$. (5.9.4)

Prove these by arguing (5.9.2) $\Longrightarrow$ (5.9.3) $\Longrightarrow$ (5.9.4) $\Longrightarrow$ (5.9.2).
Proof of (5.9.2) $\Longrightarrow$ (5.9.3). First recall from (4.4.19) that

$$\dim \mathcal{V} = \dim (\mathcal{X} + \mathcal{Y}) = \dim \mathcal{X} + \dim \mathcal{Y} - \dim (\mathcal{X} \cap \mathcal{Y}).$$

If $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$, then $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$, and thus $\dim \mathcal{V} = \dim \mathcal{X} + \dim \mathcal{Y}$. To prove (5.9.3), suppose there are two ways to represent a vector $\mathbf{v} \in \mathcal{V}$ as "something from $\mathcal{X}$ plus something from $\mathcal{Y}$." If $\mathbf{v} = \mathbf{x}_1 + \mathbf{y}_1 = \mathbf{x}_2 + \mathbf{y}_2$, where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, then

$$\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{y}_2 - \mathbf{y}_1 \implies \left\{ \begin{array}{c} \mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{X} \\ \text{and} \\ \mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{Y} \end{array} \right\} \implies \mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{X} \cap \mathcal{Y}.$$

But $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$, so $\mathbf{x}_1 = \mathbf{x}_2$ and $\mathbf{y}_1 = \mathbf{y}_2$.
Proof of (5.9.3) $\Longrightarrow$ (5.9.4). The hypothesis insures that $\mathcal{V} = \mathcal{X} + \mathcal{Y}$, and we know from (4.1.2) that $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ spans $\mathcal{X} + \mathcal{Y}$, so $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ must be a spanning set for $\mathcal{V}$. To prove $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ is linearly independent, let $\mathcal{B}_\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ and $\mathcal{B}_\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_s\}$, and suppose that

$$\mathbf{0} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_i + \sum_{j=1}^{s} \beta_j \mathbf{y}_j.$$

This is one way to express $\mathbf{0}$ as "something from $\mathcal{X}$ plus something from $\mathcal{Y}$," while $\mathbf{0} = \mathbf{0} + \mathbf{0}$ is another way. Consequently, (5.9.3) guarantees that

$$\sum_{i=1}^{r} \alpha_i \mathbf{x}_i = \mathbf{0} \quad \text{and} \quad \sum_{j=1}^{s} \beta_j \mathbf{y}_j = \mathbf{0},$$

and hence $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ and $\beta_1 = \beta_2 = \cdots = \beta_s = 0$ because $\mathcal{B}_\mathcal{X}$ and $\mathcal{B}_\mathcal{Y}$ are both linearly independent. Therefore, $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ is linearly independent, and hence it is a basis for $\mathcal{V}$.
Proof of (5.9.4) $\Longrightarrow$ (5.9.2). If $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ is a basis for $\mathcal{V}$, then $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ is a linearly independent set. This together with the fact that $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ always spans $\mathcal{X} + \mathcal{Y}$ means $\mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y}$ is a basis for $\mathcal{X} + \mathcal{Y}$ as well as for $\mathcal{V}$. Consequently, $\mathcal{V} = \mathcal{X} + \mathcal{Y}$, and hence

$$\dim \mathcal{X} + \dim \mathcal{Y} = \dim \mathcal{V} = \dim(\mathcal{X} + \mathcal{Y}) = \dim \mathcal{X} + \dim \mathcal{Y} - \dim (\mathcal{X} \cap \mathcal{Y}),$$

so $\dim (\mathcal{X} \cap \mathcal{Y}) = 0$ or, equivalently, $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$. ■

If $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$, then (5.9.3) says there is one and only one way to resolve each $\mathbf{v} \in \mathcal{V}$ into an "$\mathcal{X}$-component" and a "$\mathcal{Y}$-component" so that $\mathbf{v} = \mathbf{x} + \mathbf{y}$. These two components of $\mathbf{v}$ have a definite geometrical interpretation. Look back at Figure 5.9.1 in which $\Re^3 = \mathcal{X} \oplus \mathcal{Y}$, where $\mathcal{X}$ is a plane and $\mathcal{Y}$ is a line outside the plane, and notice that $\mathbf{x}$ (the $\mathcal{X}$-component of $\mathbf{v}$) is the result of projecting $\mathbf{v}$ onto $\mathcal{X}$ along a line parallel to $\mathcal{Y}$, and $\mathbf{y}$ (the $\mathcal{Y}$-component of $\mathbf{v}$) is obtained by projecting $\mathbf{v}$ onto $\mathcal{Y}$ along a line parallel to $\mathcal{X}$. This leads to the following formal definition of a projection.

Projection

Suppose that $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$ so that for each $\mathbf{v} \in \mathcal{V}$ there are unique vectors $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ such that $\mathbf{v} = \mathbf{x} + \mathbf{y}$.

- The vector $\mathbf{x}$ is called the ***projection*** of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$.
- The vector $\mathbf{y}$ is called the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$.

It's clear that if $\mathcal{X} \perp \mathcal{Y}$ in Figure 5.9.1, then this notion of projection agrees with the concept of orthogonal projection that was discussed on p. 322. The phrase "oblique projection" is sometimes used to emphasize the fact that $\mathcal{X}$ and $\mathcal{Y}$ are not orthogonal subspaces. In this text the word "projection" is synonymous with the term "oblique projection." If it is known that $\mathcal{X} \perp \mathcal{Y}$, then we explicitly say "orthogonal projection." Orthogonal projections are discussed in detail on p. 429.

Given a pair of complementary subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\Re^n$ and an arbitrary vector $\mathbf{v} \in \Re^n = \mathcal{X} \oplus \mathcal{Y}$, how can the projection of $\mathbf{v}$ onto $\mathcal{X}$ be computed? One way is to build a ***projector*** (a projection operator) that is a matrix $\mathbf{P}_{n\times n}$ with the property that for each $\mathbf{v} \in \Re^n$, the product $\mathbf{Pv}$ is the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$. Let $\mathcal{B}_{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ and $\mathcal{B}_{\mathcal{Y}} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-r}\}$ be respective bases for $\mathcal{X}$ and $\mathcal{Y}$ so that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\Re^n$—recall (5.9.4). This guarantees that if the $\mathbf{x}_i$'s and $\mathbf{y}_i$'s are placed as columns in

$$\mathbf{B}_{n\times n} = \big[\mathbf{x}_1\, \mathbf{x}_2 \cdots \mathbf{x}_r \,|\, \mathbf{y}_1\, \mathbf{y}_2 \cdots \mathbf{y}_{n-r}\big] = \big[\mathbf{X}_{n\times r} \,|\, \mathbf{Y}_{n\times(n-r)}\big],$$

then $\mathbf{B}$ is nonsingular. If $\mathbf{P}_{n\times n}$ is to have the property that $\mathbf{Pv}$ is the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$ for every $\mathbf{v} \in \Re^n$, then (5.9.3) implies that $\mathbf{P}\mathbf{x}_i = \mathbf{x}_i,\ \ i = 1, 2, \ldots, r$ and $\mathbf{P}\mathbf{y}_j = \mathbf{0},\ \ j = 1, 2, \ldots, n-r$, so

$$\mathbf{PB} = \mathbf{P}\big[\mathbf{X} \,|\, \mathbf{Y}\big] = \big[\mathbf{PX} \,|\, \mathbf{PY}\big] = \big[\mathbf{X} \,|\, \mathbf{0}\big]$$

and, consequently,

$$\mathbf{P} = \big[\mathbf{X} \,|\, \mathbf{0}\big]\mathbf{B}^{-1} = \mathbf{B}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{B}^{-1}. \tag{5.9.5}$$

To argue that $\mathbf{Pv}$ is indeed the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$, set $\mathbf{x} = \mathbf{Pv}$ and $\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{v}$ and observe that $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where

$$\mathbf{x} = \mathbf{Pv} = \big[\mathbf{X} \,|\, \mathbf{0}\big]\mathbf{B}^{-1}\mathbf{v} \in R\,(\mathbf{X}) = \mathcal{X} \tag{5.9.6}$$

and

$$\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{v} = \mathbf{B}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{B}^{-1}\mathbf{v} = \big[\mathbf{0} \,|\, \mathbf{Y}\big]\mathbf{B}^{-1}\mathbf{v} \in R\,(\mathbf{Y}) = \mathcal{Y}. \tag{5.9.7}$$

Is it possible that there can be more than one projector onto $\mathcal{X}$ along $\mathcal{Y}$? No, $\mathbf{P}$ is unique because if $\mathbf{P}_1$ and $\mathbf{P}_2$ are two such projectors, then for $i = 1, 2,$ we have $\mathbf{P}_i\mathbf{B} = \mathbf{P}_i\big[\mathbf{X}\,|\,\mathbf{Y}\big] = \big[\mathbf{P}_i\mathbf{X}\,|\,\mathbf{P}_i\mathbf{Y}\big] = \big[\mathbf{X}\,|\,\mathbf{0}\big]$, and this implies $\mathbf{P}_1\mathbf{B} = \mathbf{P}_2\mathbf{B}$, which means $\mathbf{P}_1 = \mathbf{P}_2$. Therefore, (5.9.5) is *the* projector onto $\mathcal{X}$ along $\mathcal{Y}$, and this formula for $\mathbf{P}$ is independent of which pair of bases for $\mathcal{X}$ and $\mathcal{Y}$ is selected. Notice that the argument involving (5.9.6) and (5.9.7) also establishes that the ***complementary projector***—the projector onto $\mathcal{Y}$ along $\mathcal{X}$—must be given by

$$\mathbf{Q} = \mathbf{I} - \mathbf{P} = \big[\mathbf{0}\,|\,\mathbf{Y}\big]\mathbf{B}^{-1} = \mathbf{B}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{B}^{-1}.$$

Below is a summary of the basic properties of projectors.

Projectors

Let $\mathcal{X}$ and $\mathcal{Y}$ be complementary subspaces of a vector space $\mathcal{V}$ so that each $\mathbf{v} \in \mathcal{V}$ can be uniquely resolved as $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. The unique linear operator $\mathbf{P}$ defined by $\mathbf{Pv} = \mathbf{x}$ is called the ***projector onto*** $\mathcal{X}$ ***along*** $\mathcal{Y}$, and $\mathbf{P}$ has the following properties.

- $\mathbf{P}^2 = \mathbf{P}$ ($\mathbf{P}$ is idempotent). (5.9.8)
- $\mathbf{I} - \mathbf{P}$ is the complementary projector onto $\mathcal{Y}$ along $\mathcal{X}$. (5.9.9)
- $R(\mathbf{P}) = \{\mathbf{x}\,|\,\mathbf{Px} = \mathbf{x}\}$ (the set of “fixed points” for $\mathbf{P}$). (5.9.10)
- $R(\mathbf{P}) = N(\mathbf{I} - \mathbf{P}) = \mathcal{X}$ and $R(\mathbf{I} - \mathbf{P}) = N(\mathbf{P}) = \mathcal{Y}$. (5.9.11)
- If $\mathcal{V} = \Re^n$ or $\mathcal{C}^n$, then $\mathbf{P}$ is given by

$$\mathbf{P} = \big[\mathbf{X}\,|\,\mathbf{0}\big]\big[\mathbf{X}\,|\,\mathbf{Y}\big]^{-1} = \big[\mathbf{X}\,|\,\mathbf{Y}\big]\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\big[\mathbf{X}\,|\,\mathbf{Y}\big]^{-1}, \tag{5.9.12}$$

where the columns of $\mathbf{X}$ and $\mathbf{Y}$ are respective bases for $\mathcal{X}$ and $\mathcal{Y}$. Other formulas for $\mathbf{P}$ are given on p. 634.

Proof. Some of these properties have already been derived in the context of $\Re^n$. But since the concepts of projections and projectors are valid for all vector spaces, more general arguments that do not rely on properties of $\Re^n$ will be provided. Uniqueness is evident because if $\mathbf{P}_1$ and $\mathbf{P}_2$ both satisfy the defining condition, then $\mathbf{P}_1\mathbf{v} = \mathbf{P}_2\mathbf{v}$ for every $\mathbf{v} \in \mathcal{V}$, and thus $\mathbf{P}_1 = \mathbf{P}_2$. The linearity of $\mathbf{P}$ follows because if $\mathbf{v}_1 = \mathbf{x}_1 + \mathbf{y}_1$ and $\mathbf{v}_2 = \mathbf{x}_2 + \mathbf{y}_2$, where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, then $\mathbf{P}(\alpha\mathbf{v}_1 + \mathbf{v}_2) = \alpha\mathbf{x}_1 + \mathbf{x}_2 = \alpha\mathbf{Pv}_1 + \mathbf{Pv}_2$. To prove that $\mathbf{P}$ is idempotent, write

$$\mathbf{P}^2\mathbf{v} = \mathbf{P}(\mathbf{Pv}) = \mathbf{Px} = \mathbf{x} = \mathbf{Pv} \text{ for every } \mathbf{v} \in \mathcal{V} \quad\Longrightarrow\quad \mathbf{P}^2 = \mathbf{P}.$$

The validity of (5.9.9) is established by observing that $\mathbf{v} = \mathbf{x} + \mathbf{y} = \mathbf{P}\mathbf{v} + \mathbf{y}$ implies $\mathbf{y} = \mathbf{v} - \mathbf{P}\mathbf{v} = (\mathbf{I} - \mathbf{P})\mathbf{v}$. The properties in (5.9.11) and (5.9.10) are immediate consequences of the definition. Formula (5.9.12) is the result of the arguments that culminated in (5.9.5), but it can be more elegantly derived by making use of the material in §4.7 and §4.8. If $\mathcal{B}_\mathcal{X}$ and $\mathcal{B}_\mathcal{Y}$ are bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively, then $\mathcal{B} = \mathcal{B}_\mathcal{X} \cup \mathcal{B}_\mathcal{Y} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-r}\}$ is a basis for $\mathcal{V}$, and (4.7.4) says that the matrix of $\mathbf{P}$ with respect to $\mathcal{B}$ is

$$\begin{aligned}
[\mathbf{P}]_\mathcal{B} &= \Big[[\mathbf{P}\mathbf{x}_1]_\mathcal{B} \,\Big|\, \cdots \,\Big|\, [\mathbf{P}\mathbf{x}_r]_\mathcal{B} \,\Big|\, [\mathbf{P}\mathbf{y}_1]_\mathcal{B} \,\Big|\, \cdots \,\Big|\, [\mathbf{P}\mathbf{y}_{n-r}]_\mathcal{B} \Big] \\
&= \Big[[\mathbf{x}_1]_\mathcal{B} \,\Big|\, \cdots \,\Big|\, [\mathbf{x}_r]_\mathcal{B} \,\Big|\, [\mathbf{0}]_\mathcal{B} \,\Big|\, \cdots \,\Big|\, [\mathbf{0}]_\mathcal{B} \Big] \\
&= \Big[\mathbf{e}_1 \,\Big|\, \cdots \,\Big|\, \mathbf{e}_r \,\Big|\, \mathbf{0} \,\Big|\, \cdots \,\Big|\, \mathbf{0} \Big] = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.
\end{aligned}$$

If $\mathcal{S}$ is the standard basis, then (4.8.5) says that $[\mathbf{P}]_\mathcal{B} = \mathbf{B}^{-1}[\mathbf{P}]_\mathcal{S}\mathbf{B}$ in which

$$\mathbf{B} = [\mathbf{I}]_{\mathcal{B}\mathcal{S}} = \Big[\, [\mathbf{x}_1]_\mathcal{S} \,\Big|\, \cdots \,\Big|\, [\mathbf{x}_r]_\mathcal{S} \,\Big|\, [\mathbf{y}_1]_\mathcal{S} \, \cdots \,\Big|\, [\mathbf{y}_{n-r}]_\mathcal{S} \Big] = \big[\mathbf{X} \,|\, \mathbf{Y}\big],$$

and therefore $[\mathbf{P}]_\mathcal{S} = \mathbf{B}[\mathbf{P}]_\mathcal{B}\mathbf{B}^{-1} = \big[\mathbf{X} \,|\, \mathbf{Y}\big] \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \big[\mathbf{X} \,|\, \mathbf{Y}\big]^{-1}.$ ∎

In the language of §4.8, statement (5.9.12) says that $\mathbf{P}$ is *similar* to the diagonal matrix $\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. In the language of §4.9, this means that $\mathbf{P}$ must be the matrix representation of the linear operator that when restricted to $\mathcal{X}$ is the identity operator and when restricted to $\mathcal{Y}$ is the zero operator.

Statement (5.9.8) says that if $\mathbf{P}$ is a projector, then $\mathbf{P}$ is idempotent ($\mathbf{P}^2 = \mathbf{P}$). But what about the converse—is every idempotent linear operator necessarily a projector? The following theorem says, "Yes."

Projectors and Idempotents

A linear operator $\mathbf{P}$ on $\mathcal{V}$ is a projector if and only if $\mathbf{P}^2 = \mathbf{P}$. (5.9.13)

Proof. The fact that every projector is idempotent was proven in (5.9.8). The proof of the converse rests on the fact that

$$\mathbf{P}^2 = \mathbf{P} \implies R(\mathbf{P}) \text{ and } N(\mathbf{P}) \text{ are complementary subspaces.} \tag{5.9.14}$$

To prove this, observe that $\mathcal{V} = R(\mathbf{P}) + N(\mathbf{P})$ because for each $\mathbf{v} \in \mathcal{V}$,

$$\mathbf{v} = \mathbf{P}\mathbf{v} + (\mathbf{I} - \mathbf{P})\mathbf{v}, \quad \text{where} \quad \mathbf{P}\mathbf{v} \in R(\mathbf{P}) \text{ and } (\mathbf{I} - \mathbf{P})\mathbf{v} \in N(\mathbf{P}). \tag{5.9.15}$$

Furthermore, $R(\mathbf{P}) \cap N(\mathbf{P}) = \mathbf{0}$ because

$$\mathbf{x} \in R(\mathbf{P}) \cap N(\mathbf{P}) \implies \mathbf{x} = \mathbf{Py} \text{ and } \mathbf{Px} = \mathbf{0} \implies \mathbf{x} = \mathbf{Py} = \mathbf{P}^2\mathbf{y} = \mathbf{0},$$

and thus (5.9.14) is established. Now that we know $R(\mathbf{P})$ and $N(\mathbf{P})$ are complementary, we can conclude that $\mathbf{P}$ is a projector because each $\mathbf{v} \in \mathcal{V}$ can be uniquely written as $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in R(\mathbf{P})$ and $\mathbf{y} \in N(\mathbf{P})$, and (5.9.15) guarantees $\mathbf{Pv} = \mathbf{x}$. ∎

Notice that there is a one-to-one correspondence between the set of idempotents (or projectors) defined on a vector space $\mathcal{V}$ and the set of all pairs of complementary subspaces of $\mathcal{V}$ in the following sense.

- Each idempotent $\mathbf{P}$ defines a pair of complementary spaces—namely, $R(\mathbf{P})$ and $N(\mathbf{P})$.
- Every pair of complementary subspaces $\mathcal{X}$ and $\mathcal{Y}$ defines an idempotent—namely, the projector onto $\mathcal{X}$ along $\mathcal{Y}$.

Example 5.9.1

Problem: Let $\mathcal{X}$ and $\mathcal{Y}$ be the subspaces of $\Re^3$ that are spanned by

$$\mathcal{B}_{\mathcal{X}} = \left\{ \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix} \right\} \quad \text{and} \quad \mathcal{B}_{\mathcal{Y}} = \left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right\},$$

respectively. Explain why $\mathcal{X}$ and $\mathcal{Y}$ are complementary, and then determine the projector onto $\mathcal{X}$ along $\mathcal{Y}$. What is the projection of $\mathbf{v} = (\,-2 \quad 1 \quad 3\,)^T$ onto $\mathcal{X}$ along $\mathcal{Y}$? What is the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$?

Solution: $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$ are linearly independent, so they are bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are complementary because

$$rank\,[\mathbf{X}\,|\,\mathbf{Y}] = rank \left(\begin{array}{rr|r} 1 & 0 & 1 \\ -1 & 1 & -1 \\ -1 & -2 & 0 \end{array} \right) = 3$$

insures that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\Re^3$—recall (5.9.4). The projector onto $\mathcal{X}$ along $\mathcal{Y}$ is obtained from (5.9.12) as

$$\mathbf{P} = \left[\mathbf{X}\,|\,\mathbf{0}\right]\left[\mathbf{X}\,|\,\mathbf{Y}\right]^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -2 & 0 \end{pmatrix} \begin{pmatrix} -2 & -2 & -1 \\ 1 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & -2 & -1 \\ 3 & 3 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

You may wish to verify that $\mathbf{P}$ is indeed idempotent. The projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$ is $\mathbf{Pv}$, and, according to (5.9.9), the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$ is $(\mathbf{I} - \mathbf{P})\mathbf{v}$.

Example 5.9.2

Angle between Complementary Subspaces. The angle between nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ in $\Re^n$ was defined on p. 295 to be the number $0 \leq \theta \leq \pi/2$ such that $\cos\theta = \mathbf{v}^T\mathbf{u}/\left\|\mathbf{v}\right\|_2\left\|\mathbf{u}\right\|_2$. It's natural to try to extend this idea to somehow make sense of angles between subspaces of $\Re^n$. Angles between completely general subspaces are presently out of our reach—they are discussed in §5.15—but the angle between a pair of *complementary* subspaces is within our grasp. When $\Re^n = \mathcal{R} \oplus \mathcal{N}$ with $\mathcal{R} \neq \mathbf{0} \neq \mathcal{N}$, the ***angle*** (also known as the ***minimal angle***) between $\mathcal{R}$ and $\mathcal{N}$ is defined to be the number $0 < \theta \leq \pi/2$ that satisfies

$$\cos\theta = \max_{\substack{\mathbf{u}\in\mathcal{R}\\ \mathbf{v}\in\mathcal{N}}} \frac{\mathbf{v}^T\mathbf{u}}{\left\|\mathbf{v}\right\|_2\left\|\mathbf{u}\right\|_2} = \max_{\substack{\mathbf{u}\in\mathcal{R},\, \mathbf{v}\in\mathcal{N}\\ \left\|\mathbf{u}\right\|_2 = \left\|\mathbf{v}\right\|_2 = 1}} \mathbf{v}^T\mathbf{u}. \tag{5.9.16}$$

While this is a good definition, it's not easy to use—especially if one wants to compute the numerical value of $\cos\theta$. The trick in making θ more accessible is to think in terms of projections and $\sin\theta = (1 - \cos^2\theta)^{1/2}$. Let $\mathbf{P}$ be the projector such that $R(\mathbf{P}) = \mathcal{R}$ and $N(\mathbf{P}) = \mathcal{N}$, and recall that the matrix 2-norm (p. 281) of $\mathbf{P}$ is

$$\left\|\mathbf{P}\right\|_2 = \max_{\left\|\mathbf{x}\right\|_2 = 1} \left\|\mathbf{P}\mathbf{x}\right\|_2. \tag{5.9.17}$$

In other words, $\left\|\mathbf{P}\right\|_2$ is the length of a longest vector in the image of the unit sphere under transformation by $\mathbf{P}$. To understand how $\sin\theta$ is related to $\left\|\mathbf{P}\right\|_2$, consider the situation in $\Re^3$. The image of the unit sphere under $\mathbf{P}$ is obtained by projecting the sphere onto $\mathcal{R}$ along lines parallel to $\mathcal{N}$. As depicted in Figure 5.9.2, the result is an ellipse in $\mathcal{R}$.

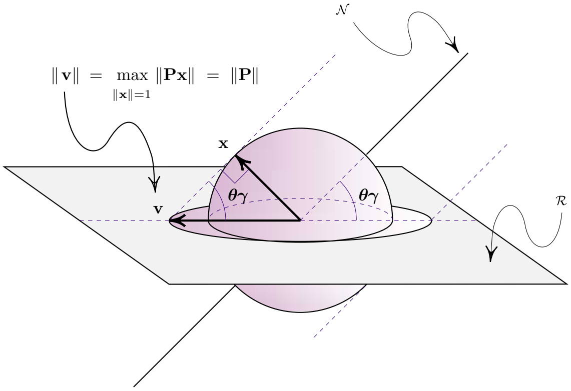

$\mathcal{N}$
$\|\mathbf{v}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{P}\mathbf{x}\| = \|\mathbf{P}\|$
$\mathbf{x}$
$\mathbf{v}$
$\theta\gamma$
$\theta\gamma$
$\mathcal{R}$

FIGURE 5.9.2

The norm of a longest vector $\mathbf{v}$ on this ellipse equals the norm of $\mathbf{P}$. That is, $\left\|\mathbf{v}\right\|_2 = \max_{\left\|\mathbf{x}\right\|_2=1}\left\|\mathbf{P}\mathbf{x}\right\|_2 = \left\|\mathbf{P}\right\|_2$, and it is apparent from the right triangle in

Figure 5.9.2 that

$$\sin\theta = \frac{\|\mathbf{x}\|_2}{\|\mathbf{v}\|_2} = \frac{1}{\|\mathbf{v}\|_2} = \frac{1}{\|\mathbf{P}\|_2}. \tag{5.9.18}$$

A little reflection on the geometry associated with Figure 5.9.2 should convince you that in $\Re^3$ a number θ satisfies (5.9.16) if and only if θ satisfies (5.9.18)—a completely rigorous proof validating this fact in $\Re^n$ is given in §5.15.

Note: Recall from p. 281 that $\|\mathbf{P}\|_2 = \sqrt{\lambda_{max}}$, where λ_{max} is the largest number λ such that $\mathbf{P}^T\mathbf{P} - \lambda\mathbf{I}$ is a singular matrix. Consequently,

$$\sin\theta = \frac{1}{\|\mathbf{P}\|_2} = \frac{1}{\sqrt{\lambda_{max}}}.$$

Numbers λ such that $\mathbf{P}^T\mathbf{P} - \lambda\mathbf{I}$ is singular are called *eigenvalues* of $\mathbf{P}^T\mathbf{P}$ (they are the main topic of discussion in Chapter 7, p. 489), and the numbers $\sqrt{\lambda}$ are the *singular values* of $\mathbf{P}$ discussed on p. 411.

Exercises for section 5.9

5.9.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be subspaces of $\Re^3$ whose respective bases are

$$\mathcal{B}_{\mathcal{X}} = \left\{ \begin{pmatrix}1\\1\\1\end{pmatrix}, \begin{pmatrix}1\\2\\2\end{pmatrix} \right\} \quad \text{and} \quad \mathcal{B}_{\mathcal{Y}} = \left\{ \begin{pmatrix}1\\2\\3\end{pmatrix} \right\}.$$

(a) Explain why $\mathcal{X}$ and $\mathcal{Y}$ are complementary subspaces of $\Re^3$.
(b) Determine the projector $\mathbf{P}$ onto $\mathcal{X}$ along $\mathcal{Y}$ as well as the complementary projector $\mathbf{Q}$ onto $\mathcal{Y}$ along $\mathcal{X}$.
(c) Determine the projection of $\mathbf{v} = \begin{pmatrix}2\\-1\\1\end{pmatrix}$ onto $\mathcal{Y}$ along $\mathcal{X}$.
(d) Verify that $\mathbf{P}$ and $\mathbf{Q}$ are both idempotent.
(e) Verify that $R(\mathbf{P}) = \mathcal{X} = N(\mathbf{Q})$ and $N(\mathbf{P}) = \mathcal{Y} = R(\mathbf{Q})$.

5.9.2. Construct an example of a pair of nontrivial complementary subspaces of $\Re^5$, and explain why your example is valid.

5.9.3. Construct an example to show that if $\mathcal{V} = \mathcal{X} + \mathcal{Y}$ but $\mathcal{X} \cap \mathcal{Y} \neq \mathbf{0}$, then a vector $\mathbf{v} \in \mathcal{V}$ can have two different representations as

$$\mathbf{v} = \mathbf{x}_1 + \mathbf{y}_1 \quad \text{and} \quad \mathbf{v} = \mathbf{x}_2 + \mathbf{y}_2,$$

where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, but $\mathbf{x}_1 \neq \mathbf{x}_2$ and $\mathbf{y}_1 \neq \mathbf{y}_2$.

5.9.4. Explain why $\Re^{n\times n} = \mathcal{S} \oplus \mathcal{K}$, where $\mathcal{S}$ and $\mathcal{K}$ are the subspaces of $n \times n$ symmetric and skew-symmetric matrices, respectively. What is the projection of $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ onto $\mathcal{S}$ along $\mathcal{K}$? **Hint:** Recall Exercise 3.2.6.

5.9.5. For a general vector space, let $\mathcal{X}$ and $\mathcal{Y}$ be two subspaces with respective bases $\mathcal{B}_{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and $\mathcal{B}_{\mathcal{Y}} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$.

(a) Prove that $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$ if and only if $\{\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{y}_1, \ldots, \mathbf{y}_n\}$ is a linearly independent set.

(b) Does $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ being linear independent imply $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$?

(c) If $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a linearly independent set, does it follow that $\mathcal{X}$ and $\mathcal{Y}$ are complementary subspaces? Why?

5.9.6. Let $\mathbf{P}$ be a projector defined on a vector space $\mathcal{V}$. Prove that (5.9.10) is true—i.e., prove that the range of a projector is the set of its "fixed points" in the sense that $R(\mathbf{P}) = \{\mathbf{x} \in \mathcal{V} \mid \mathbf{Px} = \mathbf{x}\}$.

5.9.7. Suppose that $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$, and let $\mathbf{P}$ be the projector onto $\mathcal{X}$ along $\mathcal{Y}$. Prove that (5.9.11) is true—i.e., prove

$$R(\mathbf{P}) = N(\mathbf{I} - \mathbf{P}) = \mathcal{X} \quad \text{and} \quad R(\mathbf{I} - \mathbf{P}) = N(\mathbf{P}) = \mathcal{Y}.$$

5.9.8. Explain why $\|\mathbf{P}\|_2 \geq 1$ for every projector $\mathbf{P} \neq \mathbf{0}$. When is $\|\mathbf{P}\|_2 = 1$?

5.9.9. Explain why $\|\mathbf{I} - \mathbf{P}\|_2 = \|\mathbf{P}\|_2$ for all projectors that are not zero and not equal to the identity.

5.9.10. Prove that if $\mathbf{u}, \mathbf{v} \in \Re^{n\times 1}$ are vectors such that $\mathbf{v}^T\mathbf{u} = 1$, then

$$\left\|\mathbf{I} - \mathbf{uv}^T\right\|_2 = \left\|\mathbf{uv}^T\right\|_2 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 = \left\|\mathbf{uv}^T\right\|_F,$$

where $\|\star\|_F$ is the Frobenius matrix norm defined in (5.2.1) on p. 279.

5.9.11. Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are complementary subspaces of $\Re^n$, and let $\mathbf{B} = [\mathbf{X} \mid \mathbf{Y}]$ be a nonsingular matrix in which the columns of $\mathbf{X}$ and $\mathbf{Y}$ constitute respective bases for $\mathcal{X}$ and $\mathcal{Y}$. For an arbitrary vector $\mathbf{v} \in \Re^{n\times 1}$, explain why the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$ can be obtained by the following two-step process.

(1) Solve the system $\mathbf{Bz} = \mathbf{v}$ for $\mathbf{z}$.

(2) Partition $\mathbf{z}$ as $\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{pmatrix}$, and set $\mathbf{p} = \mathbf{Xz}_1$.

5.9.12. Let $\mathbf{P}$ and $\mathbf{Q}$ be projectors.

(a) Prove $R(\mathbf{P}) = R(\mathbf{Q})$ if and only if $\mathbf{PQ} = \mathbf{Q}$ and $\mathbf{QP} = \mathbf{P}$.

(b) Prove $N(\mathbf{P}) = N(\mathbf{Q})$ if and only if $\mathbf{PQ} = \mathbf{P}$ and $\mathbf{QP} = \mathbf{Q}$.

(c) Prove that if $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_k$ are projectors with the same range, and if $\alpha_1, \alpha_2, \ldots, \alpha_k$ are scalars such that $\sum_j \alpha_j = 1$, then $\sum_j \alpha_j \mathbf{E}_j$ is a projector.

5.9.13. Prove that $\mathit{rank}(\mathbf{P}) = \mathit{trace}(\mathbf{P})$ for every projector $\mathbf{P}$ defined on $\Re^n$. **Hint:** Recall Example 3.6.5 (p. 110).

5.9.14. Let $\{\mathcal{X}_i\}_{i=1}^{k}$ be a collection of subspaces from a vector space $\mathcal{V}$, and let $\mathcal{B}_i$ denote a basis for $\mathcal{X}_i$. Prove that the following statements are equivalent.

(i) $\mathcal{V} = \mathcal{X}_1 + \mathcal{X}_2 + \cdots + \mathcal{X}_k$ and $\mathcal{X}_j \cap (\mathcal{X}_1 + \cdots + \mathcal{X}_{j-1}) = \mathbf{0}$ for each $j = 2, 3, \ldots, k$.

(ii) For each vector $\mathbf{v} \in \mathcal{V}$, there is one and only one way to write $\mathbf{v} = \mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_k$, where $\mathbf{x}_i \in \mathcal{X}_i$.

(iii) $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ with $\mathcal{B}_i \cap \mathcal{B}_j = \phi$ for $i \neq j$ is a basis for $\mathcal{V}$.

Whenever any one of the above statements is true, $\mathcal{V}$ is said to be the ***direct sum*** of the $\mathcal{X}_i$'s, and we write $\mathcal{V} = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \cdots \oplus \mathcal{X}_k$. Notice that for $k = 2$, (i) and (5.9.1) say the same thing, and (ii) and (iii) reduce to (5.9.3) and (5.9.4), respectively.

5.9.15. For complementary subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\Re^n$, let $\mathbf{P}$ be the projector onto $\mathcal{X}$ along $\mathcal{Y}$, and let $\mathbf{Q} = [\mathbf{X} \,|\, \mathbf{Y}]$ in which the columns of $\mathbf{X}$ and $\mathbf{Y}$ constitute bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively. Prove that if $\mathbf{Q}^{-1}\mathbf{A}_{n\times n}\mathbf{Q}$ is partitioned as $\mathbf{Q}^{-1}\mathbf{AQ} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$, then

$$\mathbf{Q}\begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{PAP}, \qquad \mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{PA}(\mathbf{I}-\mathbf{P}),$$

$$\mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = (\mathbf{I}-\mathbf{P})\mathbf{AP}, \ \mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{12} \end{pmatrix}\mathbf{Q}^{-1} = (\mathbf{I}-\mathbf{P})\mathbf{A}(\mathbf{I}-\mathbf{P}).$$

This means that if $\mathbf{A}$ is considered as a linear operator on $\Re^n$, and if $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$, where $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$ are the respective bases for $\mathcal{X}$ and $\mathcal{Y}$ defined by the columns of $\mathbf{X}$ and $\mathbf{Y}$, then, in the context of §4.8, the matrix representation of $\mathbf{A}$ with respect to $\mathcal{B}$ is $[\mathbf{A}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$

in which the blocks are matrix representations of restricted operators as shown below.

$$\mathbf{A}_{11} = \Big[\mathbf{PAP}_{/\mathcal{X}}\Big]_{\mathcal{B}_{\mathcal{X}}}. \qquad \mathbf{A}_{12} = \Big[\mathbf{PA}(\mathbf{I}-\mathbf{P})_{/\mathcal{Y}}\Big]_{\mathcal{B}_{\mathcal{Y}}\mathcal{B}_{\mathcal{X}}}.$$

$$\mathbf{A}_{21} = \Big[(\mathbf{I}-\mathbf{P})\mathbf{AP}_{/\mathcal{X}}\Big]_{\mathcal{B}_{\mathcal{X}}\mathcal{B}_{\mathcal{Y}}}. \quad \mathbf{A}_{22} = \Big[(\mathbf{I}-\mathbf{P})\mathbf{A}(\mathbf{I}-\mathbf{P})_{/\mathcal{Y}}\Big]_{\mathcal{B}_{\mathcal{Y}}}.$$

5.9.16. Suppose that $\Re^n = \mathcal{X} \oplus \mathcal{Y}$, where $\dim \mathcal{X} = r$, and let $\mathbf{P}$ be the projector onto $\mathcal{X}$ along $\mathcal{Y}$. Explain why there exist matrices $\mathbf{X}_{n\times r}$ and $\mathbf{A}_{r\times n}$ such that $\mathbf{P} = \mathbf{XA}$, where $rank\,(\mathbf{X}) = rank\,(\mathbf{A}) = r$ and $\mathbf{AX} = \mathbf{I}_r$. This is a *full-rank factorization* for $\mathbf{P}$ (recall Exercise 3.9.8).

5.9.17. For either a real or complex vector space, let $\mathbf{E}$ be the projector onto $\mathcal{X}_1$ along $\mathcal{Y}_1$, and let $\mathbf{F}$ be the projector onto $\mathcal{X}_2$ along $\mathcal{Y}_2$. Prove that $\mathbf{E}+\mathbf{F}$ is a projector if and only if $\mathbf{EF} = \mathbf{FE} = \mathbf{0}$, and under this condition, prove that $R\,(\mathbf{E}+\mathbf{F}) = \mathcal{X}_1 \oplus \mathcal{X}_2$ and $N\,(\mathbf{E}+\mathbf{F}) = \mathcal{Y}_1 \cap \mathcal{Y}_2$.

5.9.18. For either a real or complex vector space, let $\mathbf{E}$ be the projector onto $\mathcal{X}_1$ along $\mathcal{Y}_1$, and let $\mathbf{F}$ be the projector onto $\mathcal{X}_2$ along $\mathcal{Y}_2$. Prove that $\mathbf{E}-\mathbf{F}$ is a projector if and only if $\mathbf{EF} = \mathbf{FE} = \mathbf{F}$, and under this condition, prove that $R\,(\mathbf{E}-\mathbf{F}) = \mathcal{X}_1 \cap \mathcal{Y}_2$ and $N\,(\mathbf{E}-\mathbf{F}) = \mathcal{Y}_1 \oplus \mathcal{X}_2$.
Hint: $\mathbf{P}$ is a projector if and only if $\mathbf{I}-\mathbf{P}$ is a projector.

5.9.19. For either a real or complex vector space, let $\mathbf{E}$ be the projector onto $\mathcal{X}_1$ along $\mathcal{Y}_1$, and let $\mathbf{F}$ be the projector onto $\mathcal{X}_2$ along $\mathcal{Y}_2$. Prove that if $\mathbf{EF} = \mathbf{P} = \mathbf{FE}$, then $\mathbf{P}$ is the projector onto $\mathcal{X}_1 \cap \mathcal{X}_2$ along $\mathcal{Y}_1 + \mathcal{Y}_2$.

5.9.20. An ***inner pseudoinverse*** for $\mathbf{A}_{m\times n}$ is a matrix $\mathbf{X}_{n\times m}$ such that $\mathbf{AXA} = \mathbf{A}$, and an ***outer pseudoinverse*** for $\mathbf{A}$ is a matrix $\mathbf{X}$ satisfying $\mathbf{XAX} = \mathbf{X}$. When $\mathbf{X}$ is both an inner and outer pseudoinverse, $\mathbf{X}$ is called a ***reflexive pseudoinverse.***

(a) If $\mathbf{Ax} = \mathbf{b}$ is a consistent system of m equations in n unknowns, and if $\mathbf{A}^-$ is any inner pseudoinverse for $\mathbf{A}$, explain why the set of all solutions to $\mathbf{Ax} = \mathbf{b}$ can be expressed as

$$\mathbf{A}^-\mathbf{b} + R\left(\mathbf{I} - \mathbf{A}^-\mathbf{A}\right) = \{\mathbf{A}^-\mathbf{b} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{h} \,|\, \mathbf{h} \in \Re^n\}.$$

(b) Let $\mathcal{M}$ and $\mathcal{L}$ be respective complements of $R\,(\mathbf{A})$ and $N\,(\mathbf{A})$ so that $\mathcal{C}^m = R\,(\mathbf{A}) \oplus \mathcal{M}$ and $\mathcal{C}^n = \mathcal{L} \oplus N\,(\mathbf{A})$. Prove that there is a unique reflexive pseudoinverse $\mathbf{X}$ for $\mathbf{A}$ such that $R\,(\mathbf{X}) = \mathcal{L}$ and $N\,(\mathbf{X}) = \mathcal{M}$. Show that $\mathbf{X} = \mathbf{QA}^-\mathbf{P}$, where $\mathbf{A}^-$ is any inner pseudoinverse for $\mathbf{A}$, $\mathbf{P}$ is the projector onto $R\,(\mathbf{A})$ along $\mathcal{M}$, and $\mathbf{Q}$ is the projector onto $\mathcal{L}$ along $N\,(\mathbf{A})$.

5.10 RANGE-NULLSPACE DECOMPOSITION

Since there are infinitely many different pairs of complementary subspaces in $\Re^n$ (or $\mathcal{C}^n$),[54] is some pair more "natural" than the rest? Without reference to anything else the question is hard to answer. But if we start with a given matrix $\mathbf{A}_{n\times n}$, then there is a very natural direct sum decomposition of $\Re^n$ defined by fundamental subspaces associated with powers of $\mathbf{A}$. The rank plus nullity theorem on p. 199 says that $\dim R(\mathbf{A})+\dim N(\mathbf{A})=n$, so it's reasonable to ask about the possibility of $R(\mathbf{A})$ and $N(\mathbf{A})$ being complementary subspaces. If $\mathbf{A}$ is nonsingular, then it's trivially true that $R(\mathbf{A})$ and $N(\mathbf{A})$ are complementary, but when $\mathbf{A}$ is singular, this need not be the case because $R(\mathbf{A})$ and $N(\mathbf{A})$ need not be disjoint. For example,

$$\mathbf{A}=\begin{pmatrix}0&1\\0&0\end{pmatrix} \implies \begin{pmatrix}1\\0\end{pmatrix}\in R(\mathbf{A})\cap N(\mathbf{A}).$$

But all is not lost if we are willing to consider powers of $\mathbf{A}$.

Range-Nullspace Decomposition

For every singular matrix $\mathbf{A}_{n\times n}$, there exists a positive integer k such that $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$ are complementary subspaces. That is,

$$\Re^n = R\left(\mathbf{A}^k\right)\oplus N\left(\mathbf{A}^k\right). \tag{5.10.1}$$

The smallest positive integer k for which (5.10.1) holds is called the ***index*** of $\mathbf{A}$. For nonsingular matrices we define $index(\mathbf{A})=0$.

Proof. First observe that as $\mathbf{A}$ is powered the nullspaces grow and the ranges shrink—recall Exercise 4.2.12.

$$\begin{aligned}&N\left(\mathbf{A}^0\right)\subseteq N(\mathbf{A})\subseteq N\left(\mathbf{A}^2\right)\subseteq\cdots\subseteq N\left(\mathbf{A}^k\right)\subseteq N\left(\mathbf{A}^{k+1}\right)\subseteq\cdots\\&R\left(\mathbf{A}^0\right)\supseteq R(\mathbf{A})\supseteq R\left(\mathbf{A}^2\right)\supseteq\cdots\supseteq R\left(\mathbf{A}^k\right)\supseteq R\left(\mathbf{A}^{k+1}\right)\supseteq\cdots.\end{aligned} \tag{5.10.2}$$

The proof of (5.10.1) is attained by combining the four following properties.

Property 1. *There is equality at some point in each of the chains* (5.10.2).

Proof. If there is strict containment at each link in the nullspace chain in (5.10.2), then the sequence of inequalities

$$\dim N\left(\mathbf{A}^0\right)<\dim N(\mathbf{A})<\dim N\left(\mathbf{A}^2\right)<\dim N\left(\mathbf{A}^3\right)<\cdots$$

[54] All statements and arguments in this section are phrased in terms of $\Re^n$, but everything we say has a trivial extension to $\mathcal{C}^n$.

holds, and this forces $n < \dim N\left(\mathbf{A}^{n+1}\right)$, which is impossible. A similar argument proves equality exists somewhere in the range chain.

Property 2. *Once equality is attained, it is maintained throughout the rest of both chains in* (5.10.2). *In other words,*

$$\begin{aligned} &N\left(\mathbf{A}^0\right) \subset N(\mathbf{A}) \subset \cdots \subset N\left(\mathbf{A}^k\right) = N\left(\mathbf{A}^{k+1}\right) = N\left(\mathbf{A}^{k+2}\right) = \cdots \\ &R\left(\mathbf{A}^0\right) \supset R(\mathbf{A}) \supset \cdots \supset R\left(\mathbf{A}^k\right) = R\left(\mathbf{A}^{k+1}\right) = R\left(\mathbf{A}^{k+2}\right) = \cdots . \end{aligned} \tag{5.10.3}$$

To prove this for the range chain, observe that if k is the smallest nonnegative integer such that $R\left(\mathbf{A}^k\right) = R\left(\mathbf{A}^{k+1}\right)$, then for all $i \geq 1$,

$$R\left(\mathbf{A}^{i+k}\right) = R\left(\mathbf{A}^i\mathbf{A}^k\right) = \mathbf{A}^iR\left(\mathbf{A}^k\right) = \mathbf{A}^iR\left(\mathbf{A}^{k+1}\right) = R\left(\mathbf{A}^{i+k+1}\right).$$

The nullspace chain stops growing at exactly the same place the ranges stop shrinking because the rank plus nullity theorem (p. 199) insures that $\dim N\left(\mathbf{A}^p\right) = n - \dim R\left(\mathbf{A}^p\right)$.

Property 3. *If* k *is the value at which the ranges stop shrinking and the nullspaces stop growing in* (5.10.3), *then* $R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right) = \mathbf{0}$.

Proof. If $\mathbf{x} \in R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right)$, then $\mathbf{A}^k\mathbf{y} = \mathbf{x}$ for some $\mathbf{y} \in \Re^n$, and $\mathbf{A}^k\mathbf{x} = \mathbf{0}$. Hence $\mathbf{A}^{2k}\mathbf{y} = \mathbf{A}^k\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{y} \in N\left(\mathbf{A}^{2k}\right) = N\left(\mathbf{A}^k\right) \Rightarrow \mathbf{x} = \mathbf{0}$.

Property 4. *If* k *is the value at which the ranges stop shrinking and the nullspaces stop growing in* (5.10.3), *then* $R\left(\mathbf{A}^k\right) + N\left(\mathbf{A}^k\right) = \Re^n$.

Proof. Use Property 3 along with (4.4.19), (4.4.15), and (4.4.6), to write

$$\begin{aligned} \dim\left[R\left(\mathbf{A}^k\right) + N\left(\mathbf{A}^k\right)\right] &= \dim R\left(\mathbf{A}^k\right) + \dim N\left(\mathbf{A}^k\right) - \dim R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right) \\ &= \dim R\left(\mathbf{A}^k\right) + \dim N\left(\mathbf{A}^k\right) = n \\ &\Longrightarrow \quad R\left(\mathbf{A}^k\right) + N\left(\mathbf{A}^k\right) = \Re^n. \quad \blacksquare \end{aligned}$$

Below is a summary of our observations concerning the index of a square matrix.

Index

The index of a square matrix $\mathbf{A}$ is the smallest nonnegative integer k such that any one of the three following statements is true.

- $rank\left(\mathbf{A}^k\right) = rank\left(\mathbf{A}^{k+1}\right)$.
- $R\left(\mathbf{A}^k\right) = R\left(\mathbf{A}^{k+1}\right)$—i.e., the point where $R\left(\mathbf{A}^k\right)$ stops shrinking.
- $N\left(\mathbf{A}^k\right) = N\left(\mathbf{A}^{k+1}\right)$—i.e., the point where $N\left(\mathbf{A}^k\right)$ stops growing.

For nonsingular matrices, $index(\mathbf{A}) = 0$. For singular matrices, $index(\mathbf{A})$ is the smallest *positive* integer k such that either of the following two statements is true.

- $R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right) = \mathbf{0}$. (5.10.4)
- $\Re^n = R\left(\mathbf{A}^k\right) \oplus N\left(\mathbf{A}^k\right)$.

Example 5.10.1

Problem: Determine the index of $\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{pmatrix}$.

Solution: $\mathbf{A}$ is singular (because $rank(\mathbf{A}) = 2$), so $index(\mathbf{A}) > 0$. Since

$$\mathbf{A}^2 = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{A}^3 = \begin{pmatrix} 8 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

we see that $rank(\mathbf{A}) > rank\left(\mathbf{A}^2\right) = rank\left(\mathbf{A}^3\right)$, so $index(\mathbf{A}) = 2$. Alternately,

$$R(\mathbf{A}) = span\left\{\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}\right\}, \; R\left(\mathbf{A}^2\right) = span\begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}, \; R\left(\mathbf{A}^3\right) = span\begin{pmatrix} 8 \\ 0 \\ 0 \end{pmatrix},$$

so $R(\mathbf{A}) \supset R\left(\mathbf{A}^2\right) = R\left(\mathbf{A}^3\right)$ implies $index(\mathbf{A}) = 2$.

Nilpotent Matrices

- $\mathbf{N}_{n\times n}$ is said to be ***nilpotent*** whenever $\mathbf{N}^k = \mathbf{0}$ for some positive integer k.
- $k = index(\mathbf{N})$ is the smallest positive integer such that $\mathbf{N}^k = \mathbf{0}$. (Some authors refer to $index(\mathbf{N})$ as the ***index of nilpotency***.)

Proof. To prove that $k = index(\mathbf{N})$ is the smallest positive integer such that $\mathbf{N}^k = \mathbf{0}$, suppose p is a positive integer such that $\mathbf{N}^p = \mathbf{0}$, but $\mathbf{N}^{p-1} \neq \mathbf{0}$. We know from (5.10.3) that $R\left(\mathbf{N}^0\right) \supset R(\mathbf{N}) \supset \cdots \supset R\left(\mathbf{N}^k\right) = R\left(\mathbf{N}^{k+1}\right) = R\left(\mathbf{N}^{k+2}\right) = \cdots$, and this makes it clear that it's impossible to have $p < k$ or $p > k$, so $p = k$ is the only choice. ■

Example 5.10.2

Problem: Verify that

$$\mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

is a nilpotent matrix, and determine its index.

Solution: Computing the powers

$$\mathbf{N}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{N}^3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

reveals that $\mathbf{N}$ is indeed nilpotent, and it shows that $index(\mathbf{N}) = 3$ because $\mathbf{N}^3 = \mathbf{0}$, but $\mathbf{N}^2 \neq \mathbf{0}$.

Anytime $\Re^n$ can be written as the direct sum of two complementary subspaces such that one of them is an invariant subspace for a given square matrix $\mathbf{A}$ we have a block-triangular representation for $\mathbf{A}$ according to formula (4.9.9) on p. 263. And if both complementary spaces are invariant under $\mathbf{A}$, then (4.9.10) says that this block-triangular representation is actually block diagonal.

Herein lies the true value of the range-nullspace decomposition (5.10.1) because it turns out that if $k = index(\mathbf{A})$, then $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$ are both invariant subspaces under $\mathbf{A}$. $R\left(\mathbf{A}^k\right)$ is invariant under $\mathbf{A}$ because

$$\mathbf{A}\big(R\left(\mathbf{A}^k\right)\big) = R\left(\mathbf{A}^{k+1}\right) = R\left(\mathbf{A}^k\right),$$

and $N\left(\mathbf{A}^k\right)$ is invariant because

$$\begin{aligned}\mathbf{x} \in \mathbf{A}\big(N\left(\mathbf{A}^k\right)\big) &\implies \mathbf{x} = \mathbf{A}\mathbf{w} \text{ for some } \mathbf{w} \in N\left(\mathbf{A}^k\right) = N\left(\mathbf{A}^{k+1}\right)\\ &\implies \mathbf{A}^k\mathbf{x} = \mathbf{A}^{k+1}\mathbf{w} = \mathbf{0} \implies \mathbf{x} \in N\left(\mathbf{A}^k\right)\\ &\implies \mathbf{A}\big(N\left(\mathbf{A}^k\right)\big) \subseteq N\left(\mathbf{A}^k\right).\end{aligned}$$

This brings us to a matrix decomposition that is an important building block for developments that culminate in the Jordan form on p. 590.

Core-Nilpotent Decomposition

If $\mathbf{A}$ is an $n \times n$ singular matrix of index k such that $rank\left(\mathbf{A}^k\right) = r$, then there exists a nonsingular matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix} \tag{5.10.5}$$

in which $\mathbf{C}$ is nonsingular, and $\mathbf{N}$ is nilpotent of index k. In other words, $\mathbf{A}$ is *similar* to a 2×2 block-diagonal matrix containing a nonsingular "core" and a nilpotent component. The block-diagonal matrix in (5.10.5) is called a ***core-nilpotent decomposition*** of $\mathbf{A}$.

Note: When $\mathbf{A}$ is nonsingular, $k = 0$ and $r = n$, so $\mathbf{N}$ is not present, and we can set $\mathbf{Q} = \mathbf{I}$ and $\mathbf{C} = \mathbf{A}$ (the nonsingular core is everything). So (5.10.5) says absolutely nothing about nonsingular matrices.

Proof. Let $\mathbf{Q} = \left(\mathbf{X} \,|\, \mathbf{Y}\right)$, where the columns of $\mathbf{X}_{n\times r}$ and $\mathbf{Y}_{n\times n-r}$ constitute bases for $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$, respectively. Equation (4.9.10) guarantees that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ must be block diagonal in form, and thus (5.10.5) is established. To see that $\mathbf{N}$ is nilpotent, let

$$\mathbf{Q}^{-1} = \left(\frac{\mathbf{U}}{\mathbf{V}}\right),$$

and write

$$\begin{pmatrix} \mathbf{C}^k & \mathbf{0} \\ \mathbf{0} & \mathbf{N}^k \end{pmatrix} = \mathbf{Q}^{-1}\mathbf{A}^k\mathbf{Q} = \left(\frac{\mathbf{U}}{\mathbf{V}}\right)\mathbf{A}^k(\mathbf{X}\,|\,\mathbf{Y}) = \begin{pmatrix} \mathbf{U}\mathbf{A}^k\mathbf{X} & \mathbf{0} \\ \mathbf{V}\mathbf{A}^k\mathbf{X} & \mathbf{0} \end{pmatrix}.$$

Therefore, $\mathbf{N}^k = \mathbf{0}$ and $\mathbf{Q}^{-1}\mathbf{A}^k\mathbf{Q} = \begin{pmatrix} \mathbf{C}^k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. Since $\mathbf{C}^k$ is $r \times r$ and $r = \mathit{rank}\left(\mathbf{A}^k\right) = \mathit{rank}\left(\mathbf{Q}^{-1}\mathbf{A}^k\mathbf{Q}\right) = \mathit{rank}\left(\mathbf{C}^k\right)$, it must be the case that $\mathbf{C}^k$ is nonsingular, and hence $\mathbf{C}$ is nonsingular. Finally, notice that $\mathit{index}(\mathbf{N}) = k$ because if $\mathit{index}(\mathbf{N}) \neq k$, then $\mathbf{N}^{k-1} = \mathbf{0}$, so

$$\begin{aligned} \mathit{rank}\left(\mathbf{A}^{k-1}\right) &= \mathit{rank}\left(\mathbf{Q}^{-1}\mathbf{A}^{k-1}\mathbf{Q}\right) = \mathit{rank}\begin{pmatrix} \mathbf{C}^{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}^{k-1} \end{pmatrix} = \mathit{rank}\begin{pmatrix} \mathbf{C}^{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= \mathit{rank}\left(\mathbf{C}^{k-1}\right) = r = \mathit{rank}\left(\mathbf{A}^k\right), \end{aligned}$$

which is impossible because $\mathit{index}(\mathbf{A}) = k$ is the smallest integer for which there is equality in ranks of powers. ■

Example 5.10.3

Problem: Let $\mathbf{A}_{n\times n}$ be a rank-r matrix of index k, and let

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix} \quad \text{with} \quad \mathbf{Q} = \left(\mathbf{X}_{n\times r}\,|\,\mathbf{Y}\right) \text{ and } \mathbf{Q}^{-1} = \left(\frac{\mathbf{U}_{r\times n}}{\mathbf{V}}\right)$$

be the core-nilpotent decomposition described in (5.10.5). Explain why

$$\mathbf{Q}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{X}\mathbf{U} = \text{the projector onto } R\left(\mathbf{A}^k\right) \text{ along } N\left(\mathbf{A}^k\right)$$

and

$$\mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{Y}\mathbf{V} = \text{the projector onto } N\left(\mathbf{A}^k\right) \text{ along } R\left(\mathbf{A}^k\right).$$

Solution: Because $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$ are complementary subspaces, and because the columns of $\mathbf{X}$ and $\mathbf{Y}$ constitute respective bases for these spaces, it follows from the discussion concerning projectors on p. 386 that

$$\mathbf{P} = (\mathbf{X}\,|\,\mathbf{Y})\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}(\mathbf{X}\,|\,\mathbf{Y})^{-1} = \mathbf{Q}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{X}\mathbf{U}$$

must be the projector onto $R\left(\mathbf{A}^k\right)$ along $N\left(\mathbf{A}^k\right)$, and

$$\mathbf{I} - \mathbf{P} = (\mathbf{X}\,|\,\mathbf{Y})\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}(\mathbf{X}\,|\,\mathbf{Y})^{-1} = \mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{Y}\mathbf{V}$$

is the complementary projector onto $N\left(\mathbf{A}^k\right)$ along $R\left(\mathbf{A}^k\right)$.

Example 5.10.4

Problem: Explain how each noninvertible linear operator defined on an n-dimensional vector space $\mathcal{V}$ can be decomposed as the "direct sum" of an invertible operator and a nilpotent operator.

Solution: Let $\mathbf{T}$ be a linear operator of index k defined on $\mathcal{V} = \mathcal{R} \oplus \mathcal{N}$, where $\mathcal{R} = R\left(\mathbf{T}^k\right)$ and $\mathcal{N} = N\left(\mathbf{T}^k\right)$, and let $\mathbf{E} = \mathbf{T}_{/\mathcal{R}}$ and $\mathbf{F} = \mathbf{T}_{/\mathcal{N}}$ be the restriction operators as described in §4.9. Since $\mathcal{R}$ and $\mathcal{N}$ are invariant subspaces for $\mathbf{T}$, we know from the discussion of matrix representations on p. 263 that the right-hand side of the core-nilpotent decomposition in (5.10.5) must be the matrix representation of $\mathbf{T}$ with respect to a basis $\mathcal{B}_{\mathcal{R}} \cup \mathcal{B}_{\mathcal{N}}$, where $\mathcal{B}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{N}}$ are respective bases for $\mathcal{R}$ and $\mathcal{N}$. Furthermore, the nonsingular matrix $\mathbf{C}$ and the nilpotent matrix $\mathbf{N}$ are the matrix representations of $\mathbf{E}$ and $\mathbf{F}$ with respect to $\mathcal{B}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{N}}$, respectively. Consequently, $\mathbf{E}$ is an invertible operator on $\mathcal{R}$, and $\mathbf{F}$ is a nilpotent operator on $\mathcal{N}$. Since $\mathcal{V} = \mathcal{R} \oplus \mathcal{N}$, each $\mathbf{x} \in \mathcal{V}$ can be expressed as $\mathbf{x} = \mathbf{r} + \mathbf{n}$ with $\mathbf{r} \in \mathcal{R}$ and $\mathbf{n} \in \mathcal{N}$. This allows us to formulate the concept of the ***direct sum*** of $\mathbf{E}$ and $\mathbf{F}$ by defining $\mathbf{E} \oplus \mathbf{F}$ to be the linear operator on $\mathcal{V}$ such that $(\mathbf{E} \oplus \mathbf{F})(\mathbf{x}) = \mathbf{E}(\mathbf{r}) + \mathbf{F}(\mathbf{n})$ for each $\mathbf{x} \in \mathcal{V}$. Therefore,

$$\begin{aligned}\mathbf{T}(\mathbf{x}) &= \mathbf{T}(\mathbf{r}+\mathbf{n}) = \mathbf{T}(\mathbf{r}) + \mathbf{T}(\mathbf{n}) = (\mathbf{T}_{/\mathcal{R}})(\mathbf{r}) + (\mathbf{T}_{/\mathcal{N}})(\mathbf{n}) \\ &= \mathbf{E}(\mathbf{r}) + \mathbf{F}(\mathbf{n}) = (\mathbf{E} \oplus \mathbf{F})(\mathbf{x}) \quad \text{for each} \quad \mathbf{x} \in \mathcal{V}.\end{aligned}$$

In other words, $\mathbf{T} = \mathbf{E} \oplus \mathbf{F}$ in which $\mathbf{E} = \mathbf{T}_{/\mathcal{R}}$ is invertible and $\mathbf{F} = \mathbf{T}_{/\mathcal{N}}$ is nilpotent.

Example 5.10.5

Drazin Inverse. Inverting the nonsingular core $\mathbf{C}$ and neglecting the nilpotent part $\mathbf{N}$ in the core-nilpotent decomposition (5.10.5) produces a natural generalization of matrix inversion. More precisely, if

$$\mathbf{A} = \mathbf{Q}\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix}\mathbf{Q}^{-1}, \quad \text{then} \quad \mathbf{A}^D = \mathbf{Q}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} \tag{5.10.6}$$

defines the ***Drazin inverse*** of $\mathbf{A}$. Even though the components in a core-nilpotent decomposition are not uniquely defined by $\mathbf{A}$, it can be proven that $\mathbf{A}^D$ is unique and has the following properties.

- $\mathbf{A}^D = \mathbf{A}^{-1}$ when $\mathbf{A}$ is nonsingular (the nilpotent part is not present).
- $\mathbf{A}^D\mathbf{A}\mathbf{A}^D = \mathbf{A}^D$, $\mathbf{A}\mathbf{A}^D = \mathbf{A}^D\mathbf{A}$, $\mathbf{A}^{k+1}\mathbf{A}^D = \mathbf{A}^k$, where $k = index(\mathbf{A})$.[55]

[55] These three properties served as Michael P. Drazin's original definition in 1968. Initially,

- If $\mathbf{Ax} = \mathbf{b}$ is a consistent system of linear equations in which $\mathbf{b} \in R\left(\mathbf{A}^k\right)$, then $\mathbf{x} = \mathbf{A}^D\mathbf{b}$ is the unique solution that belongs to $R\left(\mathbf{A}^k\right)$ (Exercise 5.10.9).

- $\mathbf{AA}^D$ is the projector onto $R\left(\mathbf{A}^k\right)$ along $N\left(\mathbf{A}^k\right)$, and $\mathbf{I} - \mathbf{AA}^D$ is the complementary projector onto $N\left(\mathbf{A}^k\right)$ along $R\left(\mathbf{A}^k\right)$ (Exercise 5.10.10).

- If $\mathbf{A}$ is considered as a linear operator on $\Re^n$, then, with respect to a basis $\mathcal{B}_{\mathcal{R}}$ for $R\left(\mathbf{A}^k\right)$, $\mathbf{C}$ is the matrix representation for the restricted operator $\mathbf{A}_{/R(\mathbf{A}^k)}$ (see p. 263). Thus $\mathbf{A}_{/R(\mathbf{A}^k)}$ is invertible. Moreover,

$$\left[\mathbf{A}^D{}_{/R(\mathbf{A}^k)}\right]_{\mathcal{B}_{\mathcal{R}}} = \mathbf{C}^{-1} = \left[\left(\mathbf{A}_{/R(\mathbf{A}^k)}\right)^{-1}\right]_{\mathcal{B}_{\mathcal{R}}}, \quad \text{so} \quad \mathbf{A}^D{}_{/R(\mathbf{A}^k)} = \left(\mathbf{A}_{/R(\mathbf{A}^k)}\right)^{-1}.$$

In other words, $\mathbf{A}^D$ is the inverse of $\mathbf{A}$ on $R\left(\mathbf{A}^k\right)$, and $\mathbf{A}^D$ is the zero operator on $N\left(\mathbf{A}^k\right)$, so, in the context of Example 5.10.4,

$$\mathbf{A} = \mathbf{A}_{/R(\mathbf{A}^k)} \oplus \mathbf{A}_{/N(\mathbf{A}^k)} \quad \text{and} \quad \mathbf{A}^D = \left(\mathbf{A}_{/R(\mathbf{A}^k)}\right)^{-1} \oplus \mathbf{0}_{/N(\mathbf{A}^k)}.$$

Exercises for section 5.10

5.10.1. If $\mathbf{A}$ is a square matrix of index $k > 0$, prove that $index(\mathbf{A}^k) = 1$.

5.10.2. If $\mathbf{A}$ is a nilpotent matrix of index k, describe the components in a core-nilpotent decomposition of $\mathbf{A}$.

5.10.3. Prove that if $\mathbf{A}$ is a symmetric matrix, then $index(\mathbf{A}) \leq 1$.

5.10.4. $\mathbf{A} \in \mathcal{C}^{n\times n}$ is said to be a ***normal matrix*** whenever $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A}$. Prove that if $\mathbf{A}$ is normal, then $index(\mathbf{A}) \leq 1$.
Note: All symmetric matrices are normal, so the result of this exercise includes the result of Exercise 5.10.3 as a special case.

Drazin's concept attracted little interest—perhaps due to Drazin's abstract algebraic presentation. But eventually Drazin's generalized inverse was recognized to be a useful tool for analyzing nonorthogonal types of problems involving singular matrices. In this respect, the Drazin inverse is complementary to the Moore–Penrose pseudoinverse discussed in Exercise 4.5.20 and on p. 423 because the Moore–Penrose pseudoinverse is more useful in applications where orthogonality is somehow wired in (e.g., least squares).

5.10.5. Find a core-nilpotent decomposition and the Drazin inverse of

$$\mathbf{A} = \begin{pmatrix} -2 & 0 & -4 \\ 4 & 2 & 4 \\ 3 & 2 & 2 \end{pmatrix}.$$

5.10.6. For a square matrix $\mathbf{A}$, any scalar λ that makes $\mathbf{A} - \lambda\mathbf{I}$ singular is called an ***eigenvalue*** for $\mathbf{A}$. The *index of an eigenvalue* λ is defined to be the index of the associated matrix $\mathbf{A} - \lambda\mathbf{I}$. In other words, $index(\lambda) = index(\mathbf{A} - \lambda\mathbf{I})$. Determine the eigenvalues and the index of each eigenvalue for the following matrices:

$$\text{(a)} \quad \mathbf{J} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}. \qquad \text{(b)} \quad \mathbf{J} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

5.10.7. Let $\mathbf{P}$ be a projector different from the identity.

(a) Explain why $index(\mathbf{P}) = 1$. What is the index of $\mathbf{I}$?

(b) Determine the core-nilpotent decomposition for $\mathbf{P}$.

5.10.8. Let $\mathbf{N}$ be a nilpotent matrix of index k, and suppose that $\mathbf{x}$ is a vector such that $\mathbf{N}^{k-1}\mathbf{x} \neq \mathbf{0}$. Prove that the set

$$\mathcal{C} = \{\mathbf{x}, \mathbf{N}\mathbf{x}, \mathbf{N}^2\mathbf{x}, \ldots, \mathbf{N}^{k-1}\mathbf{x}\}$$

is a linearly independent set. $\mathcal{C}$ is sometimes called a ***Jordan chain*** or a ***Krylov sequence***.

5.10.9. Let $\mathbf{A}$ be a square matrix of index k, and let $\mathbf{b} \in R\left(\mathbf{A}^k\right)$.

(a) Explain why the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ must be consistent.

(b) Explain why $\mathbf{x} = \mathbf{A}^D\mathbf{b}$ is the unique solution in $R\left(\mathbf{A}^k\right)$.

(c) Explain why the general solution is given by $\mathbf{A}^D\mathbf{b} + N(\mathbf{A})$.

5.10.10. Suppose that $\mathbf{A}$ is a square matrix of index k, and let $\mathbf{A}^D$ be the Drazin inverse of $\mathbf{A}$ as defined in Example 5.10.5. Explain why $\mathbf{A}\mathbf{A}^D$ is the projector onto $R\left(\mathbf{A}^k\right)$ along $N\left(\mathbf{A}^k\right)$. What does $\mathbf{I} - \mathbf{A}\mathbf{A}^D$ project onto and along?

5.10.11. An ***algebraic group*** is a set $\mathcal{G}$ together with an associative operation between its elements such that $\mathcal{G}$ is closed with respect to this operation; $\mathcal{G}$ possesses an identity element $\mathbf{E}$ (which can be proven to be unique); and every member $\mathbf{A} \in \mathcal{G}$ has an inverse $\mathbf{A}^{\#}$ (which can be proven to be unique). These are essentially the axioms (A1), (A2), (A4), and (A5) in the definition of a vector space given on p. 160. A ***matrix group*** is a set of square matrices that forms an algebraic group under ordinary matrix multiplication.

(a) Show that the set of $n \times n$ nonsingular matrices is a matrix group.

(b) Show that the set of $n \times n$ unitary matrices is a *subgroup* of the $n \times n$ nonsingular matrices.

(c) Show that the set $\mathcal{G} = \left\{ \begin{pmatrix} \alpha & \alpha \\ \alpha & \alpha \end{pmatrix} \,\middle|\, \alpha \neq 0 \right\}$ is a matrix group. In particular, what does the identity element $\mathbf{E} \in \mathcal{G}$ look like, and what does the inverse $\mathbf{A}^{\#}$ of $\mathbf{A} \in \mathcal{G}$ look like?

5.10.12. For singular matrices, prove that the following statements are equivalent.

(a) $\mathbf{A}$ is a group matrix (i.e., $\mathbf{A}$ belongs to a matrix group).

(b) $R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0}$.

(c) $R(\mathbf{A})$ and $N(\mathbf{A})$ are complementary subspaces.

(d) $index(\mathbf{A}) = 1$.

(e) There are nonsingular matrices $\mathbf{Q}_{n\times n}$ and $\mathbf{C}_{r\times r}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \text{where} \quad r = rank(\mathbf{A}).$$

5.10.13. Let $\mathbf{A} \in \mathcal{G}$ for some matrix group $\mathcal{G}$.

(a) Show that the identity element $\mathbf{E} \in \mathcal{G}$ is the projector onto $R(\mathbf{A})$ along $N(\mathbf{A})$ by arguing that $\mathbf{E}$ must be of the form

$$\mathbf{E} = \mathbf{Q} \begin{pmatrix} \mathbf{I}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}^{-1}.$$

(b) Show that the ***group inverse*** of $\mathbf{A}$ (the inverse of $\mathbf{A}$ in $\mathcal{G}$) must be of the form

$$\mathbf{A}^{\#} = \mathbf{Q} \begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}^{-1}.$$

5.11 ORTHOGONAL DECOMPOSITION

The orthogonal complement of a single vector $\mathbf{x}$ was defined on p. 322 to be the set of all vectors orthogonal to $\mathbf{x}$. Below is the natural extension of this idea.

Orthogonal Complement

For a subset $\mathcal{M}$ of an inner-product space $\mathcal{V}$, the ***orthogonal complement*** $\mathcal{M}^{\perp}$ (pronounced "$\mathcal{M}$ perp") of $\mathcal{M}$ is defined to be the set of all vectors in $\mathcal{V}$ that are orthogonal to every vector in $\mathcal{M}$. That is,

$$\mathcal{M}^{\perp} = \left\{ \mathbf{x} \in \mathcal{V} \;\middle|\; \langle \mathbf{m} | \mathbf{x} \rangle = 0 \text{ for all } \mathbf{m} \in \mathcal{M} \right\}.$$

For example, if $\mathcal{M} = \{\mathbf{x}\}$ is a single vector in $\Re^2$, then, as illustrated in Figure 5.11.1, $\mathcal{M}^{\perp}$ is the line through the origin that is perpendicular to $\mathbf{x}$. If $\mathcal{M}$ is a plane through the origin in $\Re^3$, then $\mathcal{M}^{\perp}$ is the line through the origin that is perpendicular to the plane.

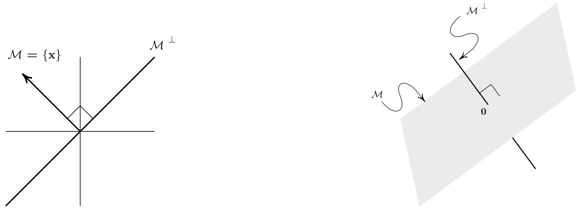


Figure 5.11.1

Notice that $\mathcal{M}^{\perp}$ is a subspace of $\mathcal{V}$ even if $\mathcal{M}$ is not a subspace because $\mathcal{M}^{\perp}$ is closed with respect to vector addition and scalar multiplication (Exercise 5.11.4). But if $\mathcal{M}$ is a subspace, then $\mathcal{M}$ and $\mathcal{M}^{\perp}$ decompose $\mathcal{V}$ as described below.

Orthogonal Complementary Subspaces

If $\mathcal{M}$ is a subspace of a finite-dimensional inner-product space $\mathcal{V}$, then

$$\mathcal{V} = \mathcal{M} \oplus \mathcal{M}^{\perp}. \tag{5.11.1}$$

Furthermore, if $\mathcal{N}$ is a subspace such that $\mathcal{V} = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{N} \perp \mathcal{M}$ (every vector in $\mathcal{N}$ is orthogonal to every vector in $\mathcal{M}$), then

$$\mathcal{N} = \mathcal{M}^{\perp}. \tag{5.11.2}$$

Proof. Observe that $\mathcal{M} \cap \mathcal{M}^{\perp} = \mathbf{0}$ because if $\mathbf{x} \in \mathcal{M}$ and $\mathbf{x} \in \mathcal{M}^{\perp}$, then $\mathbf{x}$ must be orthogonal to itself, and $\langle \mathbf{x}|\mathbf{x} \rangle = 0$ implies $\mathbf{x} = 0$. To prove that $\mathcal{M} \oplus \mathcal{M}^{\perp} = \mathcal{V}$, suppose that $\mathcal{B}_{\mathcal{M}}$ and $\mathcal{B}_{\mathcal{M}^{\perp}}$ are orthonormal bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively. Since $\mathcal{M}$ and $\mathcal{M}^{\perp}$ are disjoint, $\mathcal{B}_{\mathcal{M}} \cup \mathcal{B}_{\mathcal{M}^{\perp}}$ is an orthonormal basis for some subspace $\mathcal{S} = \mathcal{M} \oplus \mathcal{M}^{\perp} \subseteq \mathcal{V}$. If $\mathcal{S} \neq \mathcal{V}$, then the basis extension technique of Example 4.4.5 followed by the Gram–Schmidt orthogonalization procedure of §5.5 yields a nonempty set of vectors $\mathcal{E}$ such that $\mathcal{B}_{\mathcal{M}} \cup \mathcal{B}_{\mathcal{M}^{\perp}} \cup \mathcal{E}$ is an orthonormal basis for $\mathcal{V}$. Consequently,

$$\mathcal{E} \perp \mathcal{B}_{\mathcal{M}} \implies \mathcal{E} \perp \mathcal{M} \implies \mathcal{E} \subseteq \mathcal{M}^{\perp} \implies \mathcal{E} \subseteq span\,(\mathcal{B}_{\mathcal{M}^{\perp}})\,.$$

But this is impossible because $\mathcal{B}_{\mathcal{M}} \cup \mathcal{B}_{\mathcal{M}^{\perp}} \cup \mathcal{E}$ is linearly independent. Therefore, $\mathcal{E}$ is the empty set, and thus $\mathcal{V} = \mathcal{M} \oplus \mathcal{M}^{\perp}$. To prove statement (5.11.2), note that $\mathcal{N} \perp \mathcal{M}$ implies $\mathcal{N} \subseteq \mathcal{M}^{\perp}$, and coupling this with the fact that $\mathcal{M} \oplus \mathcal{M}^{\perp} = \mathcal{V} = \mathcal{M} \oplus \mathcal{N}$ together with (4.4.19) insures

$$\dim \mathcal{N} = \dim \mathcal{V} - \dim \mathcal{M} = \dim \mathcal{M}^{\perp}. \quad \blacksquare$$

Example 5.11.1

Problem: Let $\mathbf{U}_{m \times m} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ be a partitioned orthogonal matrix. Explain why $R\,(\mathbf{U}_1)$ and $R\,(\mathbf{U}_2)$ must be orthogonal complements of each other.

Solution: Statement (5.9.4) insures that $\Re^m = R\,(\mathbf{U}_1) \oplus R\,(\mathbf{U}_2)$, and we know that $R\,(\mathbf{U}_1) \perp R\,(\mathbf{U}_2)$ because the columns of $\mathbf{U}$ are an orthonormal set. Therefore, (5.11.2) guarantees that $R\,(\mathbf{U}_2) = R\,(\mathbf{U}_1)^{\perp}$.

Perp Operation

If $\mathcal{M}$ is a subspace of an n-dimensional inner-product space, then the following statements are true.

- $\dim \mathcal{M}^{\perp} = n - \dim \mathcal{M}.$ (5.11.3)
- $\mathcal{M}^{\perp^{\perp}} = \mathcal{M}.$ (5.11.4)

Proof. Property (5.11.3) follows from the fact that $\mathcal{M}$ and $\mathcal{M}^{\perp}$ are complementary subspaces—recall (4.4.19). To prove (5.11.4), first show that $\mathcal{M}^{\perp^{\perp}} \subseteq \mathcal{M}$. If $\mathbf{x} \in \mathcal{M}^{\perp^{\perp}}$, then (5.11.1) implies $\mathbf{x} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in \mathcal{M}$ and $\mathbf{n} \in \mathcal{M}^{\perp}$, so

$$0 = \langle \mathbf{n}|\mathbf{x} \rangle = \langle \mathbf{n}|\mathbf{m} + \mathbf{n} \rangle = \langle \mathbf{n}|\mathbf{m} \rangle + \langle \mathbf{n}|\mathbf{n} \rangle = \langle \mathbf{n}|\mathbf{n} \rangle \implies \mathbf{n} = 0 \implies \mathbf{x} \in \mathcal{M},$$

and thus $\mathcal{M}^{\perp^{\perp}} \subseteq \mathcal{M}$. We know from (5.11.3) that $\dim \mathcal{M}^{\perp} = n - \dim \mathcal{M}$ and $\dim \mathcal{M}^{\perp^{\perp}} = n - \dim \mathcal{M}^{\perp}$, so $\dim \mathcal{M}^{\perp^{\perp}} = \dim \mathcal{M}$. Therefore, (4.4.6) guarantees that $\mathcal{M}^{\perp^{\perp}} = \mathcal{M}$. $\blacksquare$

We are now in a position to understand why the four fundamental subspaces associated with a matrix $\mathbf{A} \in \Re^{m\times n}$ are indeed "fundamental." First consider $R(\mathbf{A})^{\perp}$, and observe that for all $\mathbf{y} \in \Re^{n}$,

$$\begin{aligned} \mathbf{x} \in R(\mathbf{A})^{\perp} &\iff \langle \mathbf{Ay}|\mathbf{x}\rangle = 0 \iff \mathbf{y}^T\mathbf{A}^T\mathbf{x} = 0 \\ &\iff \left\langle \mathbf{y}\middle|\mathbf{A}^T\mathbf{x}\right\rangle = 0 \iff \mathbf{A}^T\mathbf{x} = \mathbf{0} \quad \text{(Exercise 5.3.2)} \\ &\iff \mathbf{x} \in N\left(\mathbf{A}^T\right). \end{aligned}$$

Therefore, $R(\mathbf{A})^{\perp} = N\left(\mathbf{A}^T\right)$. Perping both sides of this equation and replacing[56] $\mathbf{A}$ by $\mathbf{A}^T$ produces $R\left(\mathbf{A}^T\right) = N(\mathbf{A})^{\perp}$. Combining these observations produces one of the fundamental theorems of linear algebra.

Orthogonal Decomposition Theorem

For every $\mathbf{A} \in \Re^{m\times n}$,

$$R(\mathbf{A})^{\perp} = N\left(\mathbf{A}^T\right) \quad \text{and} \quad N(\mathbf{A})^{\perp} = R\left(\mathbf{A}^T\right). \tag{5.11.5}$$

In light of (5.11.1), this means that every matrix $\mathbf{A} \in \Re^{m\times n}$ produces an orthogonal decomposition of $\Re^m$ and $\Re^n$ in the sense that

$$\Re^m = R(\mathbf{A}) \oplus R(\mathbf{A})^{\perp} = R(\mathbf{A}) \oplus N\left(\mathbf{A}^T\right), \tag{5.11.6}$$

and

$$\Re^n = N(\mathbf{A}) \oplus N(\mathbf{A})^{\perp} = N(\mathbf{A}) \oplus R\left(\mathbf{A}^T\right). \tag{5.11.7}$$

Theorems without hypotheses tend to be extreme in the sense that they either say very little or they reveal a lot. The orthogonal decomposition theorem has no hypothesis—it holds for all matrices—so, does it really say something significant? Yes, it does, and here's part of the reason why.

In addition to telling us how to decompose $\Re^m$ and $\Re^n$ in terms of the four fundamental subspaces of $\mathbf{A}$, the orthogonal decomposition theorem also tells us how to decompose $\mathbf{A}$ itself into more basic components. Suppose that $rank(\mathbf{A}) = r$, and let

$$\mathcal{B}_{R(\mathbf{A})} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\} \quad \text{and} \quad \mathcal{B}_{N\left(\mathbf{A}^T\right)} = \{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$$

be orthonormal bases for $R(\mathbf{A})$ and $N\left(\mathbf{A}^T\right)$, respectively, and let

$$\mathcal{B}_{R\left(\mathbf{A}^T\right)} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\} \quad \text{and} \quad \mathcal{B}_{N(\mathbf{A})} = \{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n\}$$

[56] Here, as well as throughout the rest of this section, $(\star)^T$ can be replaced by $(\star)^*$ whenever $\Re^{m\times n}$ is replaced by $\mathcal{C}^{m\times n}$.

be orthonormal bases for $R\left(\mathbf{A}^T\right)$ and $N(\mathbf{A})$, respectively. It follows that $\mathcal{B}_{R(\mathbf{A})} \cup \mathcal{B}_{N\left(\mathbf{A}^T\right)}$ and $\mathcal{B}_{R\left(\mathbf{A}^T\right)} \cup \mathcal{B}_{N(\mathbf{A})}$ are orthonormal bases for $\Re^m$ and $\Re^n$, respectively, and hence

$$\mathbf{U}_{m\times m} = \Big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_m\Big) \quad \text{and} \quad \mathbf{V}_{n\times n} = \Big(\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n\Big) \tag{5.11.8}$$

are orthogonal matrices. Now consider the product $\mathbf{R} = \mathbf{U}^T\mathbf{A}\mathbf{V}$, and notice that $r_{ij} = \mathbf{u}_i^T\mathbf{A}\mathbf{v}_j$. However, $\mathbf{u}_i^T\mathbf{A} = \mathbf{0}$ for $i = r+1, \ldots, m$ and $\mathbf{A}\mathbf{v}_j = \mathbf{0}$ for $j = r+1, \ldots, n$, so

$$\mathbf{R} = \mathbf{U}^T\mathbf{A}\mathbf{V} = \begin{pmatrix} \mathbf{u}_1^T\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{u}_1^T\mathbf{A}\mathbf{v}_r & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \mathbf{u}_r^T\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{u}_r^T\mathbf{A}\mathbf{v}_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{5.11.9}$$

In other words, $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{V}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T. \tag{5.11.10}$$

Moreover, $\mathbf{C}$ is nonsingular because it is $r \times r$ and

$$rank\,(\mathbf{C}) = rank\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = rank\left(\mathbf{U}^T\mathbf{A}\mathbf{V}\right) = rank\,(\mathbf{A}) = r.$$

For lack of a better name, we will refer to (5.11.10) as a ***URV factorization.***

We have just observed that every set of orthonormal bases for the four fundamental subspaces defines a URV factorization. The situation is also reversible in the sense that every URV factorization of $\mathbf{A}$ defines an orthonormal basis for each fundamental subspace. Starting with orthogonal matrices $\mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$ and $\mathbf{V} = \left(\mathbf{V}_1 \,|\, \mathbf{V}_2\right)$ together with a nonsingular matrix $\mathbf{C}_{r\times r}$ such that (5.11.10) holds, use the fact that right-hand multiplication by a nonsingular matrix does not alter the range (Exercise 4.5.12) to observe

$$R(\mathbf{A}) = R(\mathbf{U}\mathbf{R}) = R(\mathbf{U}_1\mathbf{C} \,|\, \mathbf{0}) = R(\mathbf{U}_1\mathbf{C}) = R(\mathbf{U}_1).$$

By (5.11.5) and Example 5.11.1, $N\left(\mathbf{A}^T\right) = R(\mathbf{A})^{\perp} = R(\mathbf{U}_1)^{\perp} = R(\mathbf{U}_2)$. Similarly, left-hand multiplication by a nonsingular matrix does not change the nullspace, so the second equation in (5.11.5) along with Example 5.11.1 yields

$$N(\mathbf{A}) = N\left(\mathbf{R}\mathbf{V}^T\right) = N\begin{pmatrix} \mathbf{C}\mathbf{V}_1^T \\ \mathbf{0} \end{pmatrix} = N\left(\mathbf{C}\mathbf{V}_1^T\right) = N\left(\mathbf{V}_1^T\right) = R(\mathbf{V}_1)^{\perp} = R(\mathbf{V}_2),$$

and $R\left(\mathbf{A}^T\right) = N(\mathbf{A})^{\perp} = R(\mathbf{V}_2)^{\perp} = R(\mathbf{V}_1)$. A summary is given below.

URV Factorization

For each $\mathbf{A} \in \Re^{m\times n}$ of rank r, there are orthogonal matrices $\mathbf{U}_{m\times m}$ and $\mathbf{V}_{n\times n}$ and a nonsingular matrix $\mathbf{C}_{r\times r}$ such that

$$\mathbf{A} = \mathbf{URV}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n} \mathbf{V}^T. \tag{5.11.11}$$

- The first r columns in $\mathbf{U}$ are an orthonormal basis for $R(\mathbf{A})$.
- The last $m-r$ columns of $\mathbf{U}$ are an orthonormal basis for $N(\mathbf{A}^T)$.
- The first r columns in $\mathbf{V}$ are an orthonormal basis for $R(\mathbf{A}^T)$.
- The last $n-r$ columns of $\mathbf{V}$ are an orthonormal basis for $N(\mathbf{A})$.

Each different collection of orthonormal bases for the four fundamental subspaces of $\mathbf{A}$ produces a different URV factorization of $\mathbf{A}$. In the complex case, replace $(\star)^T$ by $(\star)^*$ and “orthogonal” by “unitary.”

Example 5.11.2

Problem: Explain how to make $\mathbf{C}$ lower triangular in (5.11.11).

Solution: Apply Householder (or Givens) reduction to produce an orthogonal matrix $\mathbf{P}_{m\times m}$ such that $\mathbf{PA} = \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix}$, where $\mathbf{B}$ is $r \times n$ of rank r. Householder (or Givens) reduction applied to $\mathbf{B}^T$ results in an orthogonal matrix $\mathbf{Q}_{n\times n}$ and a nonsingular upper-triangular matrix $\mathbf{T}$ such that

$$\mathbf{QB}^T = \begin{pmatrix} \mathbf{T}_{r\times r} \\ \mathbf{0} \end{pmatrix} \implies \mathbf{B} = (\mathbf{T}^T \,|\, \mathbf{0})\mathbf{Q} \implies \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{T}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q},$$

so $\mathbf{A} = \mathbf{P}^T \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \mathbf{P}^T \begin{pmatrix} \mathbf{T}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}$ is a URV factorization.

Note: $\mathbf{C}$ can in fact be made diagonal—see (p. 412).

Have you noticed the duality that has emerged concerning the use of fundamental subspaces of $\mathbf{A}$ to decompose $\Re^n$ (or $\mathcal{C}^n$)? On one hand there is the range-nullspace decomposition (p. 394), and on the other is the orthogonal decomposition theorem (p. 405). Each produces a decomposition of $\mathbf{A}$. The range-nullspace decomposition of $\Re^n$ produces the core-nilpotent decomposition of $\mathbf{A}$ (p. 397), and the orthogonal decomposition theorem produces the URV factorization. In the next section, the URV factorization specializes to become the singular value decomposition (p. 412), and in a somewhat parallel manner, the core-nilpotent decomposition paves the way to the Jordan form (p. 590). These two parallel tracks constitute the backbone for the theory of modern linear algebra, so it's worthwhile to take a moment and reflect on them.

The range-nullspace decomposition decomposes $\Re^n$ with *square* matrices while the orthogonal decomposition theorem does it with *rectangular* matrices. So does this mean that the range-nullspace decomposition is a special case of, or somehow weaker than, the orthogonal decomposition theorem? No! Even for square matrices they are not very comparable because each says something that the other doesn't. The core-nilpotent decomposition (and eventually the Jordan form) is obtained by a similarity transformation, and, as discussed in §§4.8–4.9, similarity is the primary mechanism for revealing characteristics of $\mathbf{A}$ that are independent of bases or coordinate systems. The URV factorization has little to say about such things because it is generally not a similarity transformation. Orthogonal decomposition has the advantage whenever orthogonality is naturally built into a problem—such as least squares applications. And, as discussed in §5.7, orthogonal methods often produce numerically stable algorithms for floating-point computation, whereas similarity transformations are generally not well suited for numerical computations. The value of similarity is mainly on the theoretical side of the coin.

So when do we get the best of both worlds—i.e., when is a URV factorization also a core-nilpotent decomposition? First, $\mathbf{A}$ must be square and, second, (5.11.11) must be a similarity transformation, so $\mathbf{U} = \mathbf{V}$. Surprisingly, this happens for a rather large class of matrices described below.

Range Perpendicular to Nullspace

For $rank\,(\mathbf{A}_{n\times n}) = r,$ the following statements are equivalent:

- $R\,(\mathbf{A}) \perp N\,(\mathbf{A}),$ (5.11.12)
- $R\,(\mathbf{A}) = R\left(\mathbf{A}^T\right),$ (5.11.13)
- $N\,(\mathbf{A}) = N\left(\mathbf{A}^T\right),$ (5.11.14)
- $\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T$ (5.11.15)

in which $\mathbf{U}$ is orthogonal and $\mathbf{C}$ is nonsingular. Such matrices will be called ***RPN matrices,*** short for"range perpendicular to nullspace." Some authors call them *range-symmetric* or *EP* matrices. Nonsingular matrices are trivially RPN because they have a zero nullspace. For complex matrices, replace $(\star)^T$ by $(\star)^*$ and "orthogonal" by "unitary."

Proof. The fact that (5.11.12) $\Longleftrightarrow$ (5.11.13) $\Longleftrightarrow$ (5.11.14) is a direct consequence of (5.11.5). It suffices to prove (5.11.15) $\Longleftrightarrow$ (5.11.13). If (5.11.15) is a

URV factorization with $\mathbf{V} = \mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$, then $R(\mathbf{A}) = R(\mathbf{U}_1) = R(\mathbf{V}_1) = R\left(\mathbf{A}^T\right)$. Conversely, if $R(\mathbf{A}) = R\left(\mathbf{A}^T\right)$, perping both sides and using equation (5.11.5) produces $N(\mathbf{A}) = N\left(\mathbf{A}^T\right)$, so (5.11.8) yields a URV factorization with $\mathbf{U} = \mathbf{V}$. ∎

Example 5.11.3

$\mathbf{A} \in \mathcal{C}^{n\times n}$ is called a ***normal matrix*** whenever $\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A}$. As illustrated in Figure 5.11.2, normal matrices fill the niche between hermitian and (complex) RPN matrices in the sense that real-symmetric ⇒ hermitian ⇒ normal ⇒ RPN, with no implication being reversible—details are called for in Exercise 5.11.13.

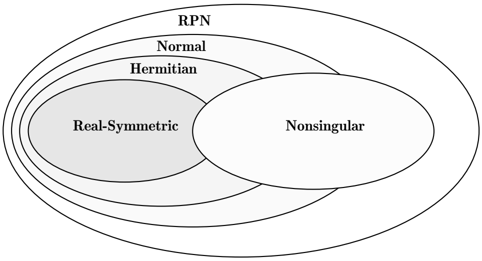


FIGURE 5.11.2

Exercises for section 5.11

5.11.1. Verify the orthogonal decomposition theorem for $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ -1 & -1 & 0 \\ -2 & -1 & -1 \end{pmatrix}$.

5.11.2. For an inner-product space $\mathcal{V}$, what is $\mathcal{V}^\perp$? What is $\mathbf{0}^\perp$?

5.11.3. Find a basis for the orthogonal complement of $\mathcal{M} = span\left\{\begin{pmatrix} 1 \\ 2 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 1 \\ 6 \end{pmatrix}\right\}$.

5.11.4. For every inner-product space $\mathcal{V}$, prove that if $\mathcal{M} \subseteq \mathcal{V}$, then $\mathcal{M}^\perp$ is a subspace of $\mathcal{V}$.

5.11.5. If $\mathcal{M}$ and $\mathcal{N}$ are subspaces of an n-dimensional inner-product space, prove that the following statements are true.

(a) $\mathcal{M} \subseteq \mathcal{N} \implies \mathcal{N}^\perp \subseteq \mathcal{M}^\perp$.
(b) $(\mathcal{M} + \mathcal{N})^\perp = \mathcal{M}^\perp \cap \mathcal{N}^\perp$.
(c) $(\mathcal{M} \cap \mathcal{N})^\perp = \mathcal{M}^\perp + \mathcal{N}^\perp$.

5.11.6. Explain why the rank plus nullity theorem on p. 199 is a corollary of the orthogonal decomposition theorem.

5.11.7. Suppose $\mathbf{A} = \mathbf{URV}^T$ is a URV factorization of an $m \times n$ matrix of rank r, and suppose $\mathbf{U}$ is partitioned as $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$, where $\mathbf{U}_1$ is $m \times r$. Prove that $\mathbf{P} = \mathbf{U}_1\mathbf{U}_1^T$ is the projector onto $R(\mathbf{A})$ along $N\big(\mathbf{A}^T\big)$. In this case, $\mathbf{P}$ is said to be an *orthogonal projector* because its range is orthogonal to its nullspace. What is the orthogonal projector onto $N\big(\mathbf{A}^T\big)$ along $R(\mathbf{A})$? (Orthogonal projectors are discussed in more detail on p. 429.)

5.11.8. Use the Householder reduction method as described in Example 5.11.2 to compute a URV factorization as well as orthonormal bases for the four fundamental subspaces of $\mathbf{A} = \begin{pmatrix} -4 & -2 & -4 & -2 \\ 2 & -2 & 2 & 1 \\ -4 & 1 & -4 & -2 \end{pmatrix}$.

5.11.9. Compute a URV factorization for the matrix given in Exercise 5.11.8 by using elementary row operations together with Gram–Schmidt orthogonalization. Are the results the same as those of Exercise 5.11.8?

5.11.10. For the matrix $\mathbf{A}$ of Exercise 5.11.8, find vectors $\mathbf{x} \in R(\mathbf{A})$ and $\mathbf{y} \in N\big(\mathbf{A}^T\big)$ such that $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where $\mathbf{v} = (\,3 \quad 3 \quad 3\,)^T$. Is there more than one choice for $\mathbf{x}$ and $\mathbf{y}$?

5.11.11. Construct a square matrix such that $R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0}$, but $R(\mathbf{A})$ is not orthogonal to $N(\mathbf{A})$.

5.11.12. For $\mathbf{A}_{n \times n}$ singular, explain why $R(\mathbf{A}) \perp N(\mathbf{A})$ implies $index(\mathbf{A}) = 1$, but not conversely.

5.11.13. Prove that real-symmetric matrix $\Rightarrow$ hermitian $\Rightarrow$ normal $\Rightarrow$ (complex) RPN. Construct examples to show that none of the implications is reversible.

5.11.14. Let $\mathbf{A}$ be a normal matrix.

(a) Prove that $R(\mathbf{A} - \lambda\mathbf{I}) \perp N(\mathbf{A} - \lambda\mathbf{I})$ for every scalar λ.

(b) Let λ and μ be scalars such that $\mathbf{A} - \lambda\mathbf{I}$ and $\mathbf{A} - \mu\mathbf{I}$ are singular matrices—such scalars are called ***eigenvalues*** of $\mathbf{A}$. Prove that if $\lambda \neq \mu$, then $N(\mathbf{A} - \lambda\mathbf{I}) \perp N(\mathbf{A} - \mu\mathbf{I})$.

5.12 SINGULAR VALUE DECOMPOSITION

For an $m \times n$ matrix $\mathbf{A}$ of rank r, Example 5.11.2 shows how to build a URV factorization

$$\mathbf{A} = \mathbf{URV}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n}\mathbf{V}^T$$

in which $\mathbf{C}$ is triangular. The purpose of this section is to prove that it's possible to do even better by showing that $\mathbf{C}$ can be made to be *diagonal*. To see how, let $\sigma_1 = \|\mathbf{A}\|_2 = \|\mathbf{C}\|_2$ (Exercise 5.6.9), and recall from the proof of (5.2.7) on p. 281 that $\|\mathbf{C}\|_2 = \|\mathbf{Cx}\|_2$ for some vector $\mathbf{x}$ such that

$$(\mathbf{C}^T\mathbf{C} - \lambda\mathbf{I})\mathbf{x} = 0, \quad \text{where} \quad \|\mathbf{x}\|_2 = 1 \text{ and } \lambda = \mathbf{x}^T\mathbf{C}^T\mathbf{Cx} = \sigma_1^2. \tag{5.12.1}$$

Set $\mathbf{y} = \mathbf{Cx}/\|\mathbf{Cx}\|_2 = \mathbf{Cx}/\sigma_1$, and let $\mathbf{R}_y = \big(\mathbf{y} \,|\, \mathbf{Y}\big)$ and $\mathbf{R}_x = \big(\mathbf{x} \,|\, \mathbf{X}\big)$ be elementary reflectors having $\mathbf{y}$ and $\mathbf{x}$ as their first columns, respectively—recall Example 5.6.3. Reflectors are orthogonal matrices, so $\mathbf{x}^T\mathbf{X} = \mathbf{0}$ and $\mathbf{Y}^T\mathbf{y} = \mathbf{0}$, and these together with (5.12.1) yield

$$\mathbf{y}^T\mathbf{CX} = \frac{\mathbf{x}^T\mathbf{C}^T\mathbf{CX}}{\sigma_1} = \frac{\lambda\mathbf{x}^T\mathbf{X}}{\sigma_1} = \mathbf{0} \quad \text{and} \quad \mathbf{Y}^T\mathbf{Cx} = \sigma_1\mathbf{Y}^T\mathbf{y} = \mathbf{0}.$$

Coupling these facts with $\mathbf{y}^T\mathbf{Cx} = \mathbf{y}^T(\sigma_1\mathbf{y}) = \sigma_1$ and $\mathbf{R}_y = \mathbf{R}_y^T$ produces

$$\mathbf{R}_y\mathbf{CR}_x = \begin{pmatrix} \mathbf{y}^T \\ \hline \mathbf{Y}^T \end{pmatrix}\mathbf{C}\big(\mathbf{x} \,|\, \mathbf{X}\big) = \begin{pmatrix} \mathbf{y}^T\mathbf{Cx} & \mathbf{y}^T\mathbf{CX} \\ \mathbf{Y}^T\mathbf{Cx} & \mathbf{Y}^T\mathbf{CX} \end{pmatrix} = \begin{pmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 \end{pmatrix}$$

with $\sigma_1 \geq \|\mathbf{C}_2\|_2$ (because $\sigma_1 = \|\mathbf{C}\|_2 = \max\{\sigma_1, \|\mathbf{C}_2\|\}$ by (5.2.12)). Repeating the process on $\mathbf{C}_2$ yields reflectors $\mathbf{S}_y$, $\mathbf{S}_x$ such that

$$\mathbf{S}_y\mathbf{C}_2\mathbf{S}_x = \begin{pmatrix} \sigma_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_3 \end{pmatrix}, \quad \text{where } \sigma_2 \geq \|\mathbf{C}_3\|_2\,.$$

If $\mathbf{P}_2$ and $\mathbf{Q}_2$ are the orthogonal matrices

$$\mathbf{P}_2 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{S}_y \end{pmatrix}\mathbf{R}_y, \quad \mathbf{Q}_2 = \mathbf{R}_x\begin{pmatrix} 1 & 0 \\ 0 & \mathbf{S}_x \end{pmatrix}, \text{ then } \mathbf{P}_2\mathbf{CQ}_2 = \begin{pmatrix} \sigma_1 & 0 & \mathbf{0} \\ 0 & \sigma_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_3 \end{pmatrix}$$

in which $\sigma_1 \geq \sigma_2 \geq \|\mathbf{C}_3\|_2$. Continuing for $r-1$ times produces orthogonal matrices $\mathbf{P}_{r-1}$ and $\mathbf{Q}_{r-1}$ such that $\mathbf{P}_{r-1}\mathbf{CQ}_{r-1} = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_r) = \mathbf{D}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. If $\tilde{\mathbf{U}}^T$ and $\tilde{\mathbf{V}}$ are the orthogonal matrices

$$\tilde{\mathbf{U}}^T = \begin{pmatrix} \mathbf{P}_{r-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\mathbf{U}^T \text{ and } \tilde{\mathbf{V}} = \mathbf{V}\begin{pmatrix} \mathbf{Q}_{r-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \text{ then } \tilde{\mathbf{U}}^T\mathbf{A}\tilde{\mathbf{V}} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

and thus the *singular value decomposition* (SVD) is derived.[57]

[57] The SVD has been independently discovered and rediscovered several times. Those credited with the early developments include Eugenio Beltrami (1835–1899) in 1873, M. E. Camille Jordan (1838–1922) in 1875, James J. Sylvester (1814–1897) in 1889, L. Autonne in 1913, and C. Eckart and G. Young in 1936.

Singular Value Decomposition

For each $\mathbf{A} \in \Re^{m\times n}$ of rank r, there are orthogonal matrices $\mathbf{U}_{m\times m}$, $\mathbf{V}_{n\times n}$ and a diagonal matrix $\mathbf{D}_{r\times r} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ such that

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n}\mathbf{V}^T \quad \text{with} \quad \sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0. \tag{5.12.2}$$

The σ_i 's are called the nonzero ***singular values*** of $\mathbf{A}$. When $r < p = \min\{m, n\}$, $\mathbf{A}$ is said to have $p - r$ additional zero singular values. The factorization in (5.12.2) is called a ***singular value decomposition*** of $\mathbf{A}$, and the columns in $\mathbf{U}$ and $\mathbf{V}$ are called left-hand and right-hand ***singular vectors*** for $\mathbf{A}$, respectively.

While the constructive method used to derive the SVD can be used as an algorithm, more sophisticated techniques exist, and all good matrix computation packages contain numerically stable SVD implementations. However, the details of a practical SVD algorithm are too complicated to be discussed at this point.

The SVD is valid for complex matrices when $(\star)^T$ is replaced by $(\star)^*$, and it can be shown that the singular values are unique, but the singular vectors are not. In the language of Chapter 7, the σ_i^2 's are the eigenvalues of $\mathbf{A}^T\mathbf{A}$, and the singular vectors are specialized sets of eigenvectors for $\mathbf{A}^T\mathbf{A}$—see the summary on p. 555. In fact, the practical algorithm for computing the SVD is an implementation of the QR iteration (p. 535) that is cleverly applied to $\mathbf{A}^T\mathbf{A}$ without ever explicitly computing $\mathbf{A}^T\mathbf{A}$.

Singular values reveal something about the geometry of linear transformations because the singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$ of a matrix $\mathbf{A}$ tell us how much distortion can occur under transformation by $\mathbf{A}$. They do so by giving us an explicit picture of how $\mathbf{A}$ distorts the unit sphere. To develop this, suppose that $\mathbf{A} \in \Re^{n\times n}$ is nonsingular (Exercise 5.12.5 treats the singular and rectangular case), and let $\mathcal{S}_2 = \{\mathbf{x} \mid \|\mathbf{x}\|_2 = 1\}$ be the unit 2-sphere in $\Re^n$. The nature of the image $\mathbf{A}(\mathcal{S}_2)$ is revealed by considering the singular value decompositions

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad \text{and} \quad \mathbf{A}^{-1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T \quad \text{with} \quad \mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. For each $\mathbf{y} \in \mathbf{A}(\mathcal{S}_2)$ there is an $\mathbf{x} \in \mathcal{S}_2$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$, so, with $\mathbf{w} = \mathbf{U}^T\mathbf{y}$,

$$\begin{aligned} 1 = \|\mathbf{x}\|_2^2 &= \left\|\mathbf{A}^{-1}\mathbf{A}\mathbf{x}\right\|_2^2 = \left\|\mathbf{A}^{-1}\mathbf{y}\right\|_2^2 = \left\|\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}\right\|_2^2 = \left\|\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}\right\|_2^2 \\ &= \left\|\mathbf{D}^{-1}\mathbf{w}\right\|_2^2 = \frac{w_1^2}{\sigma_1^2} + \frac{w_2^2}{\sigma_2^2} + \cdots + \frac{w_r^2}{\sigma_r^2}. \end{aligned} \tag{5.12.3}$$

This means that $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ is an ellipsoid whose k^{th} semiaxis has length σ_k. Because orthogonal transformations are isometries (length preserving transformations), $\mathbf{U}^T$ can only affect the orientation of $\mathbf{A}(\mathcal{S}_2)$, so $\mathbf{A}(\mathcal{S}_2)$ is also an ellipsoid whose k^{th} semiaxis has length σ_k. Furthermore, (5.12.3) implies that the ellipsoid $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ is in standard position—i.e., its axes are directed along the standard basis vectors $\mathbf{e}_k$. Since $\mathbf{U}$ maps $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ to $\mathbf{A}(\mathcal{S}_2)$, and since $\mathbf{U}\mathbf{e}_k = \mathbf{U}_{*k}$, it follows that the axes of $\mathbf{A}(\mathcal{S}_2)$ are directed along the left-hand singular vectors defined by the columns of $\mathbf{U}$. Therefore, the k^{th} semiaxis of $\mathbf{A}(\mathcal{S}_2)$ is $\sigma_k\mathbf{U}_{*k}$. Finally, since $\mathbf{AV} = \mathbf{UD}$ implies $\mathbf{AV}_{*k} = \sigma_k\mathbf{U}_{*k}$, the right-hand singular vector $\mathbf{V}_{*k}$ is a point on $\mathcal{S}_2$ that is mapped to the k^{th} semiaxis vector on the ellipsoid $\mathbf{A}(\mathcal{S}_2)$. The picture in $\Re^3$ looks like Figure 5.12.1.

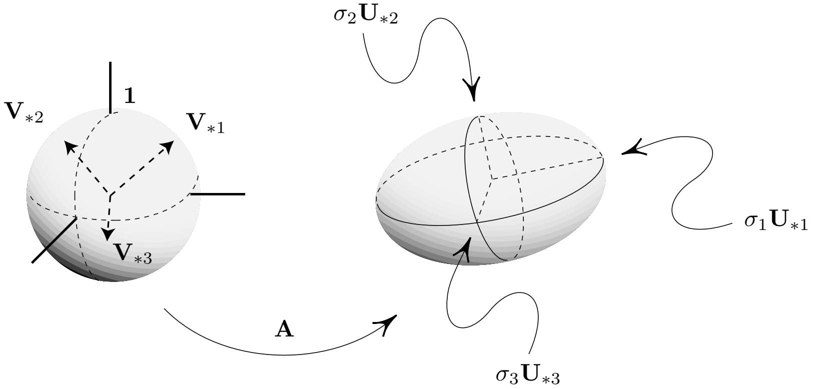

$\sigma_2\mathbf{U}_{*2}$
$\mathbf{V}_{*2}$
$\mathbf{1}$
$\mathbf{V}_{*1}$
$\mathbf{V}_{*3}$
$\sigma_1\mathbf{U}_{*1}$
$\mathbf{A}$
$\sigma_3\mathbf{U}_{*3}$

Figure 5.12.1

The degree of distortion of the unit sphere under transformation by $\mathbf{A}$ is therefore measured by $\kappa_2 = \sigma_1/\sigma_n$, the ratio of the largest singular value to the smallest singular value. Moreover, from the discussion of induced matrix norms (p. 280) and the unitary invariance of the 2-norm (Exercise 5.6.9),

$$\max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \|\mathbf{A}\|_2 = \left\|\mathbf{UDV}^T\right\|_2 = \|\mathbf{D}\|_2 = \sigma_1$$

and

$$\min_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \frac{1}{\|\mathbf{A}^{-1}\|_2} = \frac{1}{\|\mathbf{VD}^{-1}\mathbf{U}^T\|_2} = \frac{1}{\|\mathbf{D}^{-1}\|_2} = \sigma_n.$$

In other words, longest and shortest vectors on $\mathbf{A}(\mathcal{S}_2)$ have respective lengths $\sigma_1 = \|\mathbf{A}\|_2$ and $\sigma_n = 1/\left\|\mathbf{A}^{-1}\right\|_2$ (this justifies Figure 5.2.1 on p. 281), so $\kappa_2 = \|\mathbf{A}\|_2\left\|\mathbf{A}^{-1}\right\|_2$. This is called the 2-*norm condition number* of $\mathbf{A}$. Different norms result in condition numbers with different values but with more or less the same order of magnitude as κ_2 (see Exercise 5.12.3), so the qualitative information about distortion is the same. Below is a summary.

Image of the Unit Sphere

For a nonsingular $\mathbf{A}_{n\times n}$ having singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ and an SVD $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with $\mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, the image of the unit 2-sphere is an ellipsoid whose k^{th} semiaxis is given by $\sigma_k \mathbf{U}_{*k}$ (see Figure 5.12.1). Furthermore, $\mathbf{V}_{*k}$ is a point on the unit sphere such that $\mathbf{A}\mathbf{V}_{*k} = \sigma_k \mathbf{U}_{*k}$. In particular,

- $\sigma_1 = \|\mathbf{A}\mathbf{V}_{*1}\|_2 = \max\limits_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\|_2,$ (5.12.4)
- $\sigma_n = \|\mathbf{A}\mathbf{V}_{*n}\|_2 = \min\limits_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = 1/\|\mathbf{A}^{-1}\|_2.$ (5.12.5)

The degree of distortion of the unit sphere under transformation by $\mathbf{A}$ is measured by the 2-norm ***condition number***

- $\kappa_2 = \dfrac{\sigma_1}{\sigma_n} = \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2 \geq 1.$ (5.12.6)

Notice that $\kappa_2 = 1$ if and only if $\mathbf{A}$ is an orthogonal matrix.

The amount of distortion of the unit sphere under transformation by $\mathbf{A}$ determines the degree to which uncertainties in a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be magnified. This is explained in the following example.

Example 5.12.1

Uncertainties in Linear Systems. Systems of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ arising in practical work almost always come with built-in uncertainties due to modeling errors (because assumptions are almost always necessary), data collection errors (because infinitely precise gauges don't exist), and data entry errors (because numbers like $\sqrt{2}$, π, and $2/3$ can't be entered exactly). In addition, roundoff error in floating-point computation is a prevalent source of uncertainty. In all cases it's important to estimate the degree of uncertainty in the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$. This is not difficult when $\mathbf{A}$ is known exactly and all uncertainty resides in the right-hand side. Even if this is not the case, it's sometimes possible to aggregate uncertainties and shift all of them to the right-hand side.

Problem: Let $\mathbf{A}\mathbf{x} = \mathbf{b}$ be a nonsingular system in which $\mathbf{A}$ is known exactly but $\mathbf{b}$ is subject to an uncertainty $\mathbf{e}$, and consider $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} - \mathbf{e} = \tilde{\mathbf{b}}$. Estimate the *relative uncertainty*[58] $\|\mathbf{x} - \tilde{\mathbf{x}}\| / \|\mathbf{x}\|$ in $\mathbf{x}$ in terms of the relative uncertainty $\|\mathbf{b} - \tilde{\mathbf{b}}\| / \|\mathbf{b}\| = \|\mathbf{e}\| / \|\mathbf{b}\|$ in $\mathbf{b}$. Use any vector norm and its induced matrix norm (p. 280).

[58] Knowing the *absolute* uncertainty $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ by itself may not be meaningful. For example, an absolute uncertainty of a half of an inch might be fine when measuring the distance between the earth and the moon, but it's not good in the practice of eye surgery.

Solution: Use $\|\mathbf{b}\| = \|\mathbf{Ax}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\|$ with $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{e}$ to write

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} = \frac{\left\|\mathbf{A}^{-1}\mathbf{e}\right\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \|\mathbf{e}\|}{\|\mathbf{b}\|} = \kappa \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}, \tag{5.12.7}$$

where $\kappa = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|$ is a *condition number* as discussed earlier ($\kappa = \sigma_1/\sigma_n$ if the 2-norm is used). Furthermore, $\|\mathbf{e}\| = \|\mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})\| \leq \|\mathbf{A}\| \, \|(\mathbf{x} - \tilde{\mathbf{x}})\|$ and $\|\mathbf{x}\| \leq \left\|\mathbf{A}^{-1}\right\| \|\mathbf{b}\|$ imply

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \geq \frac{\|\mathbf{e}\|}{\|\mathbf{A}\| \, \|\mathbf{x}\|} \geq \frac{\|\mathbf{e}\|}{\|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \|\mathbf{b}\|} = \frac{1}{\kappa} \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}.$$

This with (5.12.7) yields the following bounds on the relative uncertainty:

$$\kappa^{-1} \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} \leq \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}, \quad \text{where} \quad \kappa = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|. \tag{5.12.8}$$

In other words, when $\mathbf{A}$ is *well conditioned* (i.e., when κ is small—see the rule of thumb in Example 3.8.2 to get a feeling of what “small” and “large” might mean), (5.12.8) insures that small relative uncertainties in $\mathbf{b}$ cannot greatly affect the solution, but when $\mathbf{A}$ is *ill conditioned* (i.e., when κ is large), a relatively small uncertainty in $\mathbf{b}$ *might* result in a relatively large uncertainty in $\mathbf{x}$. To be more sure, the following problem needs to be addressed.

Problem: Can equality be realized in each bound in (5.12.8) for every nonsingular $\mathbf{A}$, and if so, how?

Solution: Use the 2-norm, and let $\mathbf{A} = \mathbf{UDV}^T$ be an SVD so $\mathbf{AV}_{*k} = \sigma_k \mathbf{U}_{*k}$ for each k. If $\mathbf{b}$ and $\mathbf{e}$ are directed along left-hand singular vectors associated with σ_1 and σ_n, respectively—say, $\mathbf{b} = \beta \mathbf{U}_{*1}$ and $\mathbf{e} = \epsilon \mathbf{U}_{*n}$, then

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{A}^{-1}(\beta \mathbf{U}_{*1}) = \frac{\beta \mathbf{V}_{*1}}{\sigma_1} \quad \text{and} \quad \mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{e} = \mathbf{A}^{-1}(\epsilon \mathbf{U}_{*n}) = \frac{\epsilon \mathbf{V}_{*n}}{\sigma_n},$$

so

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} = \left(\frac{\sigma_1}{\sigma_n}\right) \frac{|\epsilon|}{|\beta|} = \kappa_2 \frac{\|\mathbf{e}\|_2}{\|\mathbf{b}\|_2} \quad \text{when } \mathbf{b} = \beta \mathbf{U}_{*1} \text{ and } \mathbf{e} = \epsilon \mathbf{U}_{*n}.$$

Thus the upper bound (the worst case) in (5.12.8) is attainable for all $\mathbf{A}$. The lower bound (the best case) is realized in the opposite situation when $\mathbf{b}$ and $\mathbf{e}$ are directed along $\mathbf{U}_{*n}$ and $\mathbf{U}_{*1}$, respectively. If $\mathbf{b} = \beta \mathbf{U}_{*n}$ and $\mathbf{e} = \epsilon \mathbf{U}_{*1}$, then the same argument yields $\mathbf{x} = \sigma_n^{-1} \beta \mathbf{V}_{*n}$ and $\mathbf{x} - \tilde{\mathbf{x}} = \sigma_1^{-1} \epsilon \mathbf{V}_{*1}$, so

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} = \left(\frac{\sigma_n}{\sigma_1}\right) \frac{|\epsilon|}{|\beta|} = \kappa_2^{-1} \frac{\|\mathbf{e}\|_2}{\|\mathbf{b}\|_2} \quad \text{when } \mathbf{b} = \beta \mathbf{U}_{*n} \text{ and } \mathbf{e} = \epsilon \mathbf{U}_{*1}.$$

Therefore, if $\mathbf{A}$ is well conditioned, then relatively small uncertainties in $\mathbf{b}$ can't produce relatively large uncertainties in $\mathbf{x}$. But when $\mathbf{A}$ is ill conditioned, it's possible for relatively small uncertainties in $\mathbf{b}$ to have relatively large effects on $\mathbf{x}$, and it's also possible for large uncertainties in $\mathbf{b}$ to have almost no effect on $\mathbf{x}$. Since the direction of $\mathbf{e}$ is almost always unknown, we must guard against the worst case and proceed with caution when dealing with ill-conditioned matrices.

Problem: What if there are uncertainties in both sides of $\mathbf{Ax} = \mathbf{b}$?

Solution: Use calculus to analyze the situation by considering the entries of $\mathbf{A} = \mathbf{A}(t)$ and $\mathbf{b} = \mathbf{b}(t)$ to be differentiable functions of a variable t, and compute the relative size of the derivative of $\mathbf{x} = \mathbf{x}(t)$ by differentiating $\mathbf{b} = \mathbf{Ax}$ to obtain $\mathbf{b}' = (\mathbf{Ax})' = \mathbf{A}'\mathbf{x} + \mathbf{Ax}'$ (with $\star'$ denoting $d\star/dt$), so

$$\begin{aligned}\|\mathbf{x}'\| &= \left\|\mathbf{A}^{-1}\mathbf{b}' - \mathbf{A}^{-1}\mathbf{A}'\mathbf{x}\right\| \le \left\|\mathbf{A}^{-1}\mathbf{b}'\right\| + \left\|\mathbf{A}^{-1}\mathbf{A}'\mathbf{x}\right\| \\ &\le \left\|\mathbf{A}^{-1}\right\| \|\mathbf{b}'\| + \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}'\| \|\mathbf{x}\|.\end{aligned}$$

Consequently,

$$\begin{aligned}\frac{\|\mathbf{x}'\|}{\|\mathbf{x}\|} &\le \frac{\left\|\mathbf{A}^{-1}\right\| \|\mathbf{b}'\|}{\|\mathbf{x}\|} + \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}'\| \\ &\le \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\mathbf{b}'\|}{\|\mathbf{A}\| \|\mathbf{x}\|} + \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\mathbf{A}'\|}{\|\mathbf{A}\|} \\ &\le \kappa\frac{\|\mathbf{b}'\|}{\|\mathbf{b}\|} + \kappa\frac{\|\mathbf{A}'\|}{\|\mathbf{A}\|} = \kappa\left(\frac{\|\mathbf{b}'\|}{\|\mathbf{b}\|} + \frac{\|\mathbf{A}'\|}{\|\mathbf{A}\|}\right).\end{aligned}$$

In other words, the relative sensitivity of the solution is the sum of the relative sensitivities of $\mathbf{A}$ and $\mathbf{b}$ magnified by $\kappa = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|$. A discrete analog of the above inequality is developed in Exercise 5.12.12.

Conclusion: In all cases, the credibility of the solution to $\mathbf{Ax} = \mathbf{b}$ in the face of uncertainties must be gauged in relation to the condition of $\mathbf{A}$.

As the next example shows, the condition number is pivotal also in determining whether or not the residual $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ is a reliable indicator of the accuracy of an approximate solution $\tilde{\mathbf{x}}$.

Example 5.12.2

Checking an Answer. Suppose that $\tilde{\mathbf{x}}$ is a computed (or otherwise approximate) solution for a nonsingular system $\mathbf{Ax} = \mathbf{b}$, and suppose the accuracy of $\tilde{\mathbf{x}}$ is "checked" by computing the ***residual*** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$. If $\mathbf{r} = \mathbf{0}$, *exactly,* then $\tilde{\mathbf{x}}$ must be the exact solution. But if $\mathbf{r}$ is not exactly zero—say, $\|\mathbf{r}\|_2$ is zero to t significant digits—are we guaranteed that $\tilde{\mathbf{x}}$ is accurate to roughly t significant figures? This question was briefly examined in Example 1.6.3, but it's worth another look.

Problem: To what extent does the size of the residual reflect the accuracy of an approximate solution?

Solution: Without realizing it, we answered this question in Example 5.12.1. To bound the accuracy of $\tilde{\mathbf{x}}$ relative to the exact solution $\mathbf{x}$, write $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ as $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} - \mathbf{r}$, and apply (5.12.8) with $\mathbf{e} = \mathbf{r}$ to obtain

$$\kappa^{-1}\frac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2} \le \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \kappa\frac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2}, \quad \text{where} \quad \kappa = \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2. \tag{5.12.9}$$

Therefore, for a well-conditioned $\mathbf{A}$, the residual $\mathbf{r}$ is relatively small if and only if $\tilde{\mathbf{x}}$ is relatively accurate. However, as demonstrated in Example 5.12.1, equality on either side of (5.12.9) is possible, so, when $\mathbf{A}$ is ill conditioned, a very inaccurate approximation $\tilde{\mathbf{x}}$ can produce a small residual $\mathbf{r}$, and a very accurate approximation can produce a large residual.

Conclusion: Residuals are reliable indicators of accuracy only when $\mathbf{A}$ is well conditioned—if $\mathbf{A}$ is ill conditioned, residuals are nearly meaningless.

In addition to measuring the distortion of the unit sphere and gauging the sensitivity of linear systems, singular values provide a measure of how close $\mathbf{A}$ is to a matrix of lower rank.

Distance to Lower-Rank Matrices

If $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$ are the nonzero singular values of $\mathbf{A}_{m\times n}$, then for each $k < r$, the distance from $\mathbf{A}$ to the closest matrix of rank k is

$$\sigma_{k+1} = \min_{rank(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2. \tag{5.12.10}$$

Proof. Suppose $rank\,(\mathbf{B}_{m\times n}) = k$, and let $\mathbf{A} = \mathbf{U}\begin{pmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T$ be an SVD for $\mathbf{A}$ with $\mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$. Define $\mathbf{S} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{k+1})$, and partition $\mathbf{V} = \left(\mathbf{F}_{n\times k+1} \,|\, \mathbf{G}\right)$. Since $rank\,(\mathbf{BF}) \le rank\,(\mathbf{B}) = k$ (by (4.5.2)), $\dim N(\mathbf{BF}) = k{+}1{-}rank\,(\mathbf{BF}) \ge 1$, so there is an $\mathbf{x} \in N(\mathbf{BF})$ with $\|\mathbf{x}\|_2 = 1$. Consequently, $\mathbf{BFx} = \mathbf{0}$ and

$$\mathbf{AFx} = \mathbf{U}\begin{pmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T\mathbf{Fx} = \mathbf{U}\begin{pmatrix}\mathbf{S} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \star & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} = \mathbf{U}\begin{pmatrix}\mathbf{Sx}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}.$$

Since $\|\mathbf{A} - \mathbf{B}\|_2 = \max_{\|\mathbf{y}\|_2=1} \|(\mathbf{A} - \mathbf{B})\mathbf{y}\|_2$, and since $\|\mathbf{Fx}\|_2 = \|\mathbf{x}\|_2 = 1$ (recall (5.2.4), p. 280, and (5.2.13), p. 283),

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \ge \|(\mathbf{A} - \mathbf{B})\mathbf{Fx}\|_2^2 = \|\mathbf{Sx}\|_2^2 = \sum_{i=1}^{k+1}\sigma_i^2 x_i^2 \ge \sigma_{k+1}^2\sum_{i=1}^{k+1} x_i^2 = \sigma_{k+1}^2.$$

Equality holds for $\mathbf{B}_k = \mathbf{U}\begin{pmatrix}\mathbf{D}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T$ with $\mathbf{D}_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$, and thus (5.12.10) is proven. ∎

Example 5.12.3

Filtering Noisy Data. The SVD can be a useful tool in applications involving the need to sort through noisy data and lift out relevant information. Suppose that $\mathbf{A}_{m\times n}$ is a matrix containing data that are contaminated with a certain level of noise—e.g., the entries $\mathbf{A}$ might be digital samples of a noisy video or audio signal such as that in Example 5.8.3 (p. 359). The SVD resolves the data in $\mathbf{A}$ into r mutually orthogonal components by writing

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i, \tag{5.12.11}$$

where $\mathbf{Z}_i = \mathbf{u}_i\mathbf{v}_i^T$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. The matrices $\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_r\}$ constitute an orthonormal set because

$$\langle \mathbf{Z}_i | \mathbf{Z}_j \rangle = trace\left(\mathbf{Z}_i^{\mathbf{T}}\mathbf{Z}_j\right) = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

In other words, the SVD (5.12.11) can be regarded as a Fourier expansion as described on p. 299 and, consequently, $\sigma_i = \langle \mathbf{Z}_i | \mathbf{A} \rangle$ can be interpreted as the proportion of $\mathbf{A}$ lying in the "direction" of $\mathbf{Z}_i$. In many applications the noise contamination in $\mathbf{A}$ is random (or nondirectional) in the sense that the noise is distributed more or less uniformly across the $\mathbf{Z}_i$'s. That is, there is about as much noise in the "direction" of one $\mathbf{Z}_i$ as there is in the "direction" of any other. Consequently, we expect each term $\sigma_i\mathbf{Z}_i$ to contain approximately the same level of noise. This means that if $\text{SNR}(\sigma_i\mathbf{Z}_i)$ denotes the *signal-to-noise ratio* in $\sigma_i\mathbf{Z}_i$, then

$$\text{SNR}(\sigma_1\mathbf{Z}_1) \geq \text{SNR}(\sigma_2\mathbf{Z}_2) \geq \cdots \geq \text{SNR}(\sigma_r\mathbf{Z}_r),$$

more or less. If some of the singular values, say, $\sigma_{k+1}, \ldots, \sigma_r$, are small relative to (total noise)$/r$, then the terms $\sigma_{k+1}\mathbf{Z}_{k+1}, \ldots, \sigma_r\mathbf{Z}_r$ have small signal-to-noise ratios. Therefore, if we delete these terms from (5.12.11), then we lose a small part of the total signal, but we remove a disproportionately large component of the total noise in $\mathbf{A}$. This explains why a *truncated* SVD $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i\mathbf{Z}_i$ can, in many instances, filter out some of the noise without losing significant information about the signal in $\mathbf{A}$. Determining the best value of k often requires empirical techniques that vary from application to application, but looking for obvious gaps between large and small singular values is usually a good place to start. The next example presents an interesting application of this idea to building an Internet search engine.

Example 5.12.4

Search Engines. The filtering idea presented in Example 5.12.3 is widely used, but a particularly novel application is the method of *latent semantic indexing* used in the areas of information retrieval and text mining. You can think of this in terms of building an Internet search engine. Start with a dictionary of terms $T_1, T_2, \ldots, T_m$. Terms are usually single words, but sometimes a term may contain more that one word such as "landing gear." It's up to you to decide how extensive your dictionary should be, but even if you use the entire English language, you probably won't be using more than a few hundred-thousand terms, and this is within the capacity of existing computer technology. Each document (or web page) D_j of interest is scanned for key terms (this is called *indexing* the document), and an associated *document vector* $\mathbf{d}_j = (\text{freq}_{1j}, \text{freq}_{2j}, \ldots, \text{freq}_{mj})^T$ is created in which

$$\text{freq}_{ij} = \text{number of times term } T_i \text{ occurs in document } D_j.$$

(More sophisticated search engines use weighted frequency strategies.) After a collection of documents $D_1, D_2, \ldots, D_n$ has been indexed, the associated document vectors $\mathbf{d}_j$ are placed as columns in a *term-by-document matrix*

$$\mathbf{A}_{m\times n} = \big(\mathbf{d}_1 \,|\, \mathbf{d}_2 \,\cdots\, |\, \mathbf{d}_n\big) = \begin{array}{c} \\ T_1 \\ T_2 \\ \vdots \\ T_m \end{array} \begin{array}{c} \begin{array}{cccc} D_1 & D_2 & \cdots & D_n \end{array} \\ \begin{pmatrix} \text{freq}_{11} & \text{freq}_{12} & \cdots & \text{freq}_{1n} \\ \text{freq}_{21} & \text{freq}_{22} & \cdots & \text{freq}_{2n} \\ \vdots & \vdots & & \vdots \\ \text{freq}_{m1} & \text{freq}_{m2} & \cdots & \text{freq}_{mn} \end{pmatrix} \end{array}.$$

Naturally, most entries in each document vector $\mathbf{d}_j$ will be zero, so $\mathbf{A}$ is a sparse matrix—this is good because it means that sparse matrix technology can be applied. When a query composed of a few terms is submitted to the search engine, a *query vector* $\mathbf{q}^T = (q_1, q_2, \ldots, q_n)$ is formed in which

$$q_i = \begin{cases} 1 & \text{if term } T_i \text{ appears in the query,} \\ 0 & \text{otherwise.} \end{cases}$$

(The q_i 's might also be weighted.) To measure how well a query $\mathbf{q}$ matches a document D_j, we check how close $\mathbf{q}$ is to $\mathbf{d}_j$ by computing the magnitude of

$$\cos\theta_j = \frac{\mathbf{q}^T\mathbf{d}_j}{\|\mathbf{q}\|_2\,\|\mathbf{d}_j\|_2} = \frac{\mathbf{q}^T\mathbf{A}\mathbf{e}_j}{\|\mathbf{q}\|_2\,\|\mathbf{A}\mathbf{e}_j\|_2}. \tag{5.12.12}$$

If $|\cos\theta_j| \geq \tau$ for some threshold tolerance τ, then document D_j is considered relevant and is returned to the user. Selecting τ is part art and part science that's based on experimentation and desired performance criteria. If the columns of $\mathbf{A}$ along with $\mathbf{q}$ are initially normalized to have unit length, then

$|\mathbf{q}^T\mathbf{A}| = \big(|\cos\theta_1|,\, |\cos\theta_2|,\, \ldots,\, |\cos\theta_n|\big)$ provides the information that allows the search engine to rank the relevance of each document relative to the query. However, due to things like variation and ambiguity in the use of vocabulary, presentation style, and even the indexing process, there is a lot of "noise" in $\mathbf{A}$, so the results in $|\mathbf{q}^T\mathbf{A}|$ are nowhere near being an exact measure of how well query $\mathbf{q}$ matches the various documents. To filter out some of this noise, the techniques of Example 5.12.3 are employed. An SVD $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ is judiciously truncated, and

$$\mathbf{A}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T = \big(\mathbf{u}_1 \,|\, \cdots \,|\, \mathbf{u}_k\big) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{pmatrix} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

is used in place of $\mathbf{A}$ in (5.12.12). In other words, instead of using $\cos\theta_j$, query $\mathbf{q}$ is compared with document D_j by using the magnitude of

$$\cos\phi_j = \frac{\mathbf{q}^T\mathbf{A}_k\mathbf{e}_j}{\|\mathbf{q}\|_2\, \|\mathbf{A}_k\mathbf{e}_j\|_2}.$$

To make this more suitable for computation, set $\mathbf{S}_k = \mathbf{D}_k\mathbf{V}_k^T = \big(\mathbf{s}_1 \,|\, \mathbf{s}_2 \,|\, \cdots \,|\, \mathbf{s}_k\big)$, and use

$$\|\mathbf{A}_k\mathbf{e}_j\|_2 = \left\|\mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T\mathbf{e}_j\right\|_2 = \|\mathbf{U}_k\mathbf{s}_j\|_2 = \|\mathbf{s}_j\|_2$$

to write

$$\cos\phi_j = \frac{\mathbf{q}^T\mathbf{U}_k\mathbf{s}_j}{\|\mathbf{q}\|_2\, \|\mathbf{s}_j\|_2}. \tag{5.12.13}$$

The vectors in $\mathbf{U}_k$ and $\mathbf{S}_k$ only need to be computed once (and they can be determined without computing the entire SVD), so (5.12.13) requires very little computation to process each new query. Furthermore, we can be generous in the number of SVD components that are dropped because variation in the use of vocabulary and the ambiguity of many words produces significant noise in $\mathbf{A}$. Coupling this with the fact that numerical accuracy is not an important issue (knowing a cosine to two or three significant digits is sufficient) means that we are more than happy to replace the SVD of $\mathbf{A}$ by a low-rank truncation $\mathbf{A}_k$, where k is *significantly* less than r.

Alternate Query Matching Strategy. An alternate way to measuring how close a given query $\mathbf{q}$ is to a document vector $\mathbf{d}_j$ is to replace the query vector $\mathbf{q}$ in (5.12.12) by the *projected query* $\widetilde{\mathbf{q}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{q}$, where $\mathbf{P}_{R(\mathbf{A})} = \mathbf{U}_r\mathbf{U}_r^T$ is the orthogonal projector onto $R(\mathbf{A})$ along $R(\mathbf{A})^{\perp}$ (Exercise 5.12.15) to produce

$$\cos\widetilde{\theta}_j = \frac{\widetilde{\mathbf{q}}^T\mathbf{A}\mathbf{e}_j}{\|\widetilde{\mathbf{q}}\|_2\, \|\mathbf{A}\mathbf{e}_j\|_2}. \tag{5.12.14}$$

It's proven on p. 435 that $\widetilde{\mathbf{q}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{q}$ is the vector in $R(\mathbf{A})$ (the *document space*) that is closest to $\mathbf{q}$, so using $\widetilde{\mathbf{q}}$ in place of $\mathbf{q}$ has the effect of using the best approximation to $\mathbf{q}$ that is a linear combination of the document vectors $\mathbf{d}_i$. Since $\widetilde{\mathbf{q}}^T\mathbf{A} = \mathbf{q}^T\mathbf{A}$ and $\|\widetilde{\mathbf{q}}\|_2 \leq \|\mathbf{q}\|_2$, it follows that $\cos\widetilde{\theta}_j \geq \cos\theta_j$, so more documents are deemed relevant when the projected query is used. Just as in the unprojected query matching strategy, the noise is filtered out by replacing $\mathbf{A}$ in (5.12.14) with a truncated SVD $\mathbf{A}_k = \sum_{i=1}^k \sigma_i\mathbf{u}_i\mathbf{v}_i^T$. The result is

$$\cos\widetilde{\phi}_j = \frac{\mathbf{q}^T\mathbf{U}_k\mathbf{s}_j}{\left\|\mathbf{U}_k^T\mathbf{q}\right\|_2 \|\mathbf{s}_j\|_2}$$

and, just as in (5.12.13), $\cos\widetilde{\phi}_j$ is easily and quickly computed for each new query $\mathbf{q}$ because $\mathbf{U}_k$ and $\mathbf{s}_j$ need only be computed once.

The next example shows why singular values are the primary mechanism for numerically determining the rank of a matrix.

Example 5.12.5

Perturbations and Numerical Rank. For $\mathbf{A} \in \Re^{m\times n}$ with $p = \min\{m,n\}$, let $\{\sigma_1,\sigma_2,\ldots,\sigma_p\}$ and $\{\beta_1,\beta_2,\ldots,\beta_p\}$ be all singular values (nonzero as well as any zero ones) for $\mathbf{A}$ and $\mathbf{A}+\mathbf{E}$, respectively.

Problem: Prove that

$$|\sigma_k - \beta_k| \leq \|\mathbf{E}\|_2 \quad \text{for each} \quad k = 1,2,\ldots,p. \tag{5.12.15}$$

Solution: If the SVD for $\mathbf{A}$ given in (5.12.2) is written in the form

$$\mathbf{A} = \sum_{i=1}^{p} \sigma_i\mathbf{u}_i\mathbf{v}_i^T, \quad \text{and if we set} \quad \mathbf{A}_{k-1} = \sum_{i=1}^{k-1} \sigma_i\mathbf{u}_i\mathbf{v}_i^T,$$

then

$$\begin{aligned}\sigma_k &= \|\mathbf{A}-\mathbf{A}_{k-1}\|_2 = \|\mathbf{A}+\mathbf{E}-\mathbf{A}_{k-1}-\mathbf{E}\|_2 \\ &\geq \|\mathbf{A}+\mathbf{E}-\mathbf{A}_{k-1}\|_2 - \|\mathbf{E}\|_2 \quad \text{(recall (5.1.6) on p. 273)} \\ &\geq \beta_k - \|\mathbf{E}\|_2 \quad \text{by (5.12.10).}\end{aligned}$$

Couple this with the observation that

$$\begin{aligned}\sigma_k &= \min_{rank(\mathbf{B})=k-1} \|\mathbf{A}-\mathbf{B}\|_2 = \min_{rank(\mathbf{B})=k-1} \|\mathbf{A}+\mathbf{E}-\mathbf{B}-\mathbf{E}\|_2 \\ &\leq \min_{rank(\mathbf{B})=k-1} \|\mathbf{A}+\mathbf{E}-\mathbf{B}\|_2 + \|\mathbf{E}\|_2 = \beta_k + \|\mathbf{E}\|_2\end{aligned}$$

to conclude that $|\sigma_k - \beta_k| \leq \|\mathbf{E}\|_2$.

Problem: Explain why this means that computing the singular values of $\mathbf{A}$ with any stable algorithm (one that returns the exact singular values β_k of a nearby matrix $\mathbf{A}+\mathbf{E}$) is a good way to compute $rank(\mathbf{A})$.

Solution: If $rank(\mathbf{A}) = r$, then $p-r$ of the σ_k 's are exactly zero, so the perturbation result (5.12.15) guarantees that $p-r$ of the computed β_k 's cannot be larger than $\|\mathbf{E}\|_2$. So if

$$\beta_1 \geq \cdots \geq \beta_{\tilde{r}} > \|\mathbf{E}\|_2 \geq \beta_{\tilde{r}+1} \geq \cdots \geq \beta_p,$$

then it's reasonable to consider $\tilde{r}$ to be the ***numerical rank*** of $\mathbf{A}$. For most algorithms, $\|\mathbf{E}\|_2$ is not known exactly, but adequate estimates of $\|\mathbf{E}\|_2$ often can be derived. Considerable effort has gone into the development of stable algorithms for computing singular values, but such algorithms are too involved to discuss here—consult an advanced book on matrix computations. Generally speaking, good SVD algorithms have $\|\mathbf{E}\|_2 \approx 5 \times 10^{-t}\|\mathbf{A}\|_2$ when t-digit floating-point arithmetic is used.

Just as the range-nullspace decomposition was used in Example 5.10.5 to define the Drazin inverse of a square matrix, a URV factorization or an SVD can be used to define a generalized inverse for rectangular matrices. For a URV factorization

$$\mathbf{A}_{m\times n} = \mathbf{U}\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n}\mathbf{V}^T, \quad \text{we define} \quad \mathbf{A}^{\dagger}_{n\times m} = \mathbf{V}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{n\times m}\mathbf{U}^T$$

to be the ***Moore–Penrose inverse*** (or the ***pseudoinverse***) of $\mathbf{A}$. (Replace $(\star)^T$ by $(\star)^*$ when $\mathbf{A} \in \mathcal{C}^{m\times n}$.) Although the URV factors are not uniquely defined by $\mathbf{A}$, it can be proven that $\mathbf{A}^{\dagger}$ is unique by arguing that $\mathbf{A}^{\dagger}$ is the unique solution to the four Penrose equations

$$\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}, \qquad \mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger},$$

$$\left(\mathbf{A}\mathbf{A}^{\dagger}\right)^T = \mathbf{A}\mathbf{A}^{\dagger}, \qquad \left(\mathbf{A}^{\dagger}\mathbf{A}\right)^T = \mathbf{A}^{\dagger}\mathbf{A},$$

so $\mathbf{A}^{\dagger}$ is the same matrix defined in Exercise 4.5.20. Since it doesn't matter which URV factorization is used, we can use the SVD (5.12.2), in which case $\mathbf{C} = \mathbf{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. Some "inverselike" properties that relate $\mathbf{A}^{\dagger}$ to solutions and least squares solutions for linear systems are given in the following summary. Other useful properties appear in the exercises.

Moore–Penrose Pseudoinverse

- In terms of URV factors, the Moore–Penrose pseudoinverse of

$$\mathbf{A}_{m\times n}=\mathbf{U}\begin{pmatrix}\mathbf{C}_{r\times r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T \quad\text{is}\quad \mathbf{A}^\dagger_{n\times m}=\mathbf{V}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T. \qquad (5.12.16)$$

- When $\mathbf{Ax}=\mathbf{b}$ is consistent, $\mathbf{x}=\mathbf{A}^\dagger\mathbf{b}$ is the solution of minimal euclidean norm. (5.12.17)
- When $\mathbf{Ax}=\mathbf{b}$ is inconsistent, $\mathbf{x}=\mathbf{A}^\dagger\mathbf{b}$ is the least squares solution of minimal euclidean norm. (5.12.18)
- When an SVD is used, $\mathbf{C}=\mathbf{D}=\operatorname{diag}(\sigma_1,\ldots,\sigma_r)$, so

$$\mathbf{A}^\dagger=\mathbf{V}\begin{pmatrix}\mathbf{D}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T=\sum_{i=1}^{r}\frac{\mathbf{v}_i\mathbf{u}_i^T}{\sigma_i} \quad\text{and}\quad \mathbf{A}^\dagger\mathbf{b}=\sum_{i=1}^{r}\frac{\left(\mathbf{u}_i^T\mathbf{b}\right)}{\sigma_i}\mathbf{v}_i.$$

Proof. To prove (5.12.17), suppose $\mathbf{Ax}_0=\mathbf{b}$, and replace $\mathbf{A}$ by $\mathbf{AA}^\dagger\mathbf{A}$ to write $\mathbf{b}=\mathbf{Ax}_0=\mathbf{AA}^\dagger\mathbf{Ax}_0=\mathbf{AA}^\dagger\mathbf{b}$. Thus $\mathbf{A}^\dagger\mathbf{b}$ solves $\mathbf{Ax}=\mathbf{b}$ when it is consistent. To see that $\mathbf{A}^\dagger\mathbf{b}$ is the solution of minimal norm, observe that the general solution is $\mathbf{A}^\dagger\mathbf{b}+N(\mathbf{A})$ (a particular solution plus the general solution of the homogeneous equation), so every solution has the form $\mathbf{z}=\mathbf{A}^\dagger\mathbf{b}+\mathbf{n}$, where $\mathbf{n}\in N(\mathbf{A})$. It's not difficult to see that $\mathbf{A}^\dagger\mathbf{b}\in R\left(\mathbf{A}^\dagger\right)=R\left(\mathbf{A}^T\right)$ (Exercise 5.12.16), so $\mathbf{A}^\dagger\mathbf{b}\perp\mathbf{n}$. Therefore, by the Pythagorean theorem (Exercise 5.4.14),

$$\|\mathbf{z}\|_2^2=\left\|\mathbf{A}^\dagger\mathbf{b}+\mathbf{n}\right\|_2^2=\left\|\mathbf{A}^\dagger\mathbf{b}\right\|_2^2+\|\mathbf{n}\|_2^2\geq\left\|\mathbf{A}^\dagger\mathbf{b}\right\|_2^2.$$

Equality is possible if and only if $\mathbf{n}=\mathbf{0}$, so $\mathbf{A}^\dagger\mathbf{b}$ is the *unique* minimum norm solution. When $\mathbf{Ax}=\mathbf{b}$ is inconsistent, the least squares solutions are the solutions of the normal equations $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$, and it's straightforward to verify that $\mathbf{A}^\dagger\mathbf{b}$ is one such solution (Exercise 5.12.16(c)). To prove that $\mathbf{A}^\dagger\mathbf{b}$ is the least squares solution of minimal norm, apply the same argument used in the consistent case to the normal equations. ■

Caution! Generalized inverses are useful in formulating theoretical statements such as those above, but, just as in the case of the ordinary inverse, generalized inverses are not practical computational tools. In addition to being computationally inefficient, serious numerical problems result from the fact that $\mathbf{A}^\dagger$ need not be a continuous function of the entries of $\mathbf{A}$. For example,

$$\mathbf{A}(x) = \begin{pmatrix} 1 & 0 \\ 0 & x \end{pmatrix} \implies \mathbf{A}^{\dagger}(x) = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1/x \end{pmatrix} & \text{for } x \neq 0, \\ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} & \text{for } x = 0. \end{cases}$$

Not only is $\mathbf{A}^{\dagger}(x)$ discontinuous in the sense that $\lim_{x\to 0} \mathbf{A}^{\dagger}(x) \neq \mathbf{A}^{\dagger}(0)$, but it is discontinuous in the worst way because as $\mathbf{A}(x)$ comes closer to $\mathbf{A}(0)$ the matrix $\mathbf{A}^{\dagger}(x)$ moves farther away from $\mathbf{A}^{\dagger}(0)$. This type of behavior translates into insurmountable computational difficulties because small errors due to roundoff (or anything else) can produce enormous errors in the computed $\mathbf{A}^{\dagger}$, and as errors in $\mathbf{A}$ become smaller the resulting errors in $\mathbf{A}^{\dagger}$ can become greater. This diabolical fact is also true for the Drazin inverse (p. 399). The inherent numerical problems coupled with the fact that it's extremely rare for an application to require explicit knowledge of the entries of $\mathbf{A}^{\dagger}$ or $\mathbf{A}^{D}$ constrains them to being theoretical or notational tools. But don't underestimate this role—go back and read Laplace's statement quoted in the footnote on p. 81.

Example 5.12.6

Another way to view the URV or SVD factorizations in relation to the Moore–Penrose inverse is to consider $\mathbf{A}_{/R(\mathbf{A}^T)}$ and $\mathbf{A}^{\dagger}_{/R(\mathbf{A})}$, the restrictions of $\mathbf{A}$ and $\mathbf{A}^{\dagger}$ to $R\left(\mathbf{A}^T\right)$ and $R\left(\mathbf{A}\right)$, respectively. Begin by making the straightforward observations that $R\left(\mathbf{A}^{\dagger}\right) = R\left(\mathbf{A}^T\right)$ and $N\left(\mathbf{A}^{\dagger}\right) = N\left(\mathbf{A}^T\right)$ (Exercise 5.12.16). Since $\Re^n = R\left(\mathbf{A}^T\right) \oplus N\left(\mathbf{A}\right)$ and $\Re^m = R\left(\mathbf{A}\right) \oplus N\left(\mathbf{A}^T\right)$, it follows that $R\left(\mathbf{A}\right) = \mathbf{A}(\Re^n) = \mathbf{A}(R\left(\mathbf{A}^T\right))$ and $R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^{\dagger}\right) = \mathbf{A}^{\dagger}(\Re^m) = \mathbf{A}^{\dagger}(R\left(\mathbf{A}\right))$. In other words, $\mathbf{A}_{/R(\mathbf{A}^T)}$ and $\mathbf{A}^{\dagger}_{/R(\mathbf{A})}$ are linear transformations such that

$$\mathbf{A}_{/R(\mathbf{A}^T)} : R\left(\mathbf{A}^T\right) \to R\left(\mathbf{A}\right) \quad \text{and} \quad \mathbf{A}^{\dagger}_{/R(\mathbf{A})} : R\left(\mathbf{A}\right) \to R\left(\mathbf{A}^T\right).$$

If $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ and $\mathcal{B}' = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ are the first r columns from $\mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$ and $\mathbf{V} = \left(\mathbf{V}_1 \,|\, \mathbf{V}_2\right)$ in (5.11.11), then $\mathbf{A}\mathbf{V}_1 = \mathbf{U}_1\mathbf{C}$ and $\mathbf{A}^{\dagger}\mathbf{U}_1 = \mathbf{V}_1\mathbf{C}^{-1}$ implies (recall (4.7.4)) that

$$\left[\mathbf{A}_{/R(\mathbf{A}^T)}\right]_{\mathcal{B}'\mathcal{B}} = \mathbf{C} \quad \text{and} \quad \left[\mathbf{A}^{\dagger}_{/R(\mathbf{A})}\right]_{\mathcal{B}\mathcal{B}'} = \mathbf{C}^{-1}. \tag{5.12.19}$$

If left-hand and right-hand singular vectors from the SVD (5.12.2) are used in $\mathcal{B}$ and $\mathcal{B}'$, respectively, then $\mathbf{C} = \mathbf{D} = \operatorname{diag}\left(\sigma_1, \ldots, \sigma_r\right)$. Thus (5.12.19) reveals the exact sense in which $\mathbf{A}$ and $\mathbf{A}^{\dagger}$ are "inverses." Compare these results with the analogous statements for the Drazin inverse in Example 5.10.5 on p. 399.

Exercises for section 5.12

5.12.1. Following the derivation in the text, find an SVD for

$$\mathbf{C} = \begin{pmatrix} -4 & -6 \\ 3 & -8 \end{pmatrix}.$$

5.12.2. If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the nonzero singular values of $\mathbf{A}$, then it can be shown that the function $\nu_k(\mathbf{A}) = \left(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2\right)^{1/2}$ defines a unitarily invariant norm (recall Exercise 5.6.9) for $\Re^{m\times n}$ (or $\mathcal{C}^{m\times n}$) for each $k = 1, 2, \ldots, r$. Explain why the 2-norm and the Frobenius norm (p. 279) are the extreme cases in the sense that $\|\mathbf{A}\|_2^2 = \sigma_1^2$ and $\|\mathbf{A}\|_F^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2$.

5.12.3. Each of the four common matrix norms can be bounded above and below by a constant multiple of each of the other matrix norms. To be precise, $\|\mathbf{A}\|_i \leq \alpha \|\mathbf{A}\|_j$, where α is the (i, j)-entry in the following matrix.

$$\begin{array}{c} \\ 1 \\ 2 \\ \infty \\ F \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \infty & F \end{array} \\ \begin{pmatrix} * & \sqrt{n} & n & \sqrt{n} \\ \sqrt{n} & * & \sqrt{n} & 1 \\ n & \sqrt{n} & * & \sqrt{n} \\ \sqrt{n} & \sqrt{n} & \sqrt{n} & * \end{pmatrix} \end{array}.$$

For analyzing limiting behavior, it therefore makes no difference which of these norms is used, so they are said to be *equivalent matrix norms.* (A similar statement for vector norms was given in Exercise 5.1.8.) Explain why the $(2, F)$ and the $(F, 2)$ entries are correct.

5.12.4. Prove that if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the nonzero singular values of a rank r matrix $\mathbf{A}$, and if $\|\mathbf{E}\|_2 < \sigma_r$, then $rank\,(\mathbf{A} + \mathbf{E}) \geq rank\,(\mathbf{A})$. **Note:** This clarifies the meaning of the term "sufficiently small" in the assertion on p. 216 that small perturbations can't reduce rank.

5.12.5. **Image of the Unit Sphere.** Extend the result on p. 414 concerning the image of the unit sphere to include singular and rectangular matrices by showing that if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the nonzero singular values of $\mathbf{A}_{m\times n}$, then the image $\mathbf{A}(\mathcal{S}_2) \subset \Re^m$ of the unit 2-sphere $\mathcal{S}_2 \subset \Re^n$ is an ellipsoid (possibly degenerate) in which the k^{th} semiaxis is $\sigma_k \mathbf{U}_{*k} = \mathbf{A}\mathbf{V}_{*k}$, where $\mathbf{U}_{*k}$ and $\mathbf{V}_{*k}$ are respective left-hand and right-hand singular vectors for $\mathbf{A}$.

5.12.6. Prove that if σ_r is the smallest nonzero singular value of $\mathbf{A}_{m\times n}$, then

$$\sigma_r = \min_{\substack{\|\mathbf{x}\|_2=1 \\ \mathbf{x}\in R(\mathbf{A}^T)}} \|\mathbf{Ax}\|_2 = 1/\left\|\mathbf{A}^\dagger\right\|_2,$$

which is the generalization of (5.12.5).

5.12.7. Generalized Condition Number. Extend the bound in (5.12.8) to include singular and rectangular matrices by showing that if $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the respective minimum 2-norm solutions of consistent systems $\mathbf{Ax}=\mathbf{b}$ and $\mathbf{A}\tilde{\mathbf{x}}=\tilde{\mathbf{b}}=\mathbf{b}-\mathbf{e}$, then

$$\kappa^{-1}\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} \le \frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \kappa\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}, \quad \text{where} \quad \kappa = \|\mathbf{A}\|\left\|\mathbf{A}^\dagger\right\|.$$

Can the same reasoning given in Example 5.12.1 be used to argue that for $\|\star\|_2$, the upper and lower bounds are attainable for every $\mathbf{A}$?

5.12.8. Prove that if $|\epsilon| < \sigma_r^2$ for the smallest nonzero singular value of $\mathbf{A}_{m\times n}$, then $(\mathbf{A}^T\mathbf{A}+\epsilon\mathbf{I})^{-1}$ exists, and $\lim_{\epsilon\to 0}(\mathbf{A}^T\mathbf{A}+\epsilon\mathbf{I})^{-1}\mathbf{A}^T = \mathbf{A}^\dagger$.

5.12.9. Consider a system $\mathbf{Ax}=\mathbf{b}$ in which

$$\mathbf{A} = \begin{pmatrix} .835 & .667 \\ .333 & .266 \end{pmatrix},$$

and suppose $\mathbf{b}$ is subject to an uncertainty $\mathbf{e}$. Using ∞-norms, determine the directions of $\mathbf{b}$ and $\mathbf{e}$ that give rise to the worst-case scenario in (5.12.8) in the sense that $\|\mathbf{x}-\tilde{\mathbf{x}}\|_\infty / \|\mathbf{x}\|_\infty = \kappa_\infty\|\mathbf{e}\|_\infty / \|\mathbf{b}\|_\infty$.

5.12.10. An ill-conditioned matrix is suspected when a small pivot u_{ii} emerges during the LU factorization of $\mathbf{A}$ because $\left[\mathbf{U}^{-1}\right]_{ii} = 1/u_{ii}$ is then large, and this opens the possibility of $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$ having large entries. Unfortunately, this is not an absolute test, and no guarantees about conditioning can be made from the pivots alone.

(a) Construct an example of a matrix that is well conditioned but has a small pivot.

(b) Construct an example of a matrix that is ill conditioned but has no small pivots.

5.12.11. Bound the relative uncertainty in the solution of a nonsingular system $\mathbf{Ax} = \mathbf{b}$ for which there is some uncertainty in $\mathbf{A}$ but not in $\mathbf{b}$ by showing that if $(\mathbf{A}-\mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}$, where $\alpha = \left\|\mathbf{A}^{-1}\mathbf{E}\right\| < 1$ for any matrix norm such that $\|\mathbf{I}\| = 1$, then

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\kappa}{1-\alpha}\frac{\|\mathbf{E}\|}{\|\mathbf{A}\|}, \quad \text{where} \quad \kappa = \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|.$$

Note: If the 2-norm is used, then $\|\mathbf{E}\|_2 < \sigma_n$ insures $\alpha < 1$.
Hint: If $\mathbf{B} = \mathbf{A}^{-1}\mathbf{E}$, then $\mathbf{A}-\mathbf{E} = \mathbf{A}(\mathbf{I}-\mathbf{B})$, and $\alpha = \|\mathbf{B}\| < 1$ $\Longrightarrow \left\|\mathbf{B}^k\right\| \leq \|\mathbf{B}\|^k \to 0 \Longrightarrow \mathbf{B}^k \to \mathbf{0}$, so the Neumann series expansion (p. 126) yields $(\mathbf{I}-\mathbf{B})^{-1} = \sum_{i=0}^{\infty}\mathbf{B}^i$.

5.12.12. Now bound the relative uncertainty in the solution of a nonsingular system $\mathbf{Ax} = \mathbf{b}$ for which there is some uncertainty in both $\mathbf{A}$ and $\mathbf{b}$ by showing that if $(\mathbf{A}-\mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}-\mathbf{e}$, where $\alpha = \left\|\mathbf{A}^{-1}\mathbf{E}\right\| < 1$ for any matrix norm such that $\|\mathbf{I}\| = 1$, then

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\kappa}{1-\kappa\|\mathbf{E}\|/\|\mathbf{A}\|}\left(\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}+\frac{\|\mathbf{E}\|}{\|\mathbf{A}\|}\right), \quad \text{where} \quad \kappa = \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|.$$

Note: If the 2-norm is used, then $\|\mathbf{E}\|_2 < \sigma_n$ insures $\alpha < 1$. This exercise underscores the conclusion of Example 5.12.1 stating that if $\mathbf{A}$ is well conditioned, and if the relative uncertainties in $\mathbf{A}$ and $\mathbf{b}$ are small, then the relative uncertainty in $\mathbf{x}$ must be small.

5.12.13. Consider the matrix $\mathbf{A} = \begin{pmatrix} -4 & -2 & -4 & -2 \\ 2 & -2 & 2 & 1 \\ -4 & 1 & -4 & -2 \end{pmatrix}$.

(a) Use the URV factorization you computed in Exercise 5.11.8 to determine $\mathbf{A}^{\dagger}$.

(b) Now use the URV factorization you obtained in Exercise 5.11.9 to determine $\mathbf{A}^{\dagger}$. Do your results agree with those of part (a)?

5.12.14. For matrix $\mathbf{A}$ in Exercise 5.11.8, and for $\mathbf{b} = (-12 \quad 3 \quad -9)^T$, find the solution of $\mathbf{Ax} = \mathbf{b}$ that has minimum euclidean norm.

5.12.15. Suppose $\mathbf{A} = \mathbf{URV}^T$ is a URV factorization (so it could be an SVD) of an $m \times n$ matrix of rank r, and suppose $\mathbf{U}$ is partitioned as $\mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$, where $\mathbf{U}_1$ is $m \times r$. Prove that $\mathbf{P} = \mathbf{U}_1\mathbf{U}_1^T = \mathbf{AA}^{\dagger}$ is the projector onto $R(\mathbf{A})$ along $N\left(\mathbf{A}^T\right)$. In this case, $\mathbf{P}$ is said to be an *orthogonal projector* because its range is orthogonal to its nullspace. What is the orthogonal projector onto $N\left(\mathbf{A}^T\right)$ along $R(\mathbf{A})$? (Orthogonal projectors are discussed in more detail on p. 429.)

5.12.16. Establish the following properties of $\mathbf{A}^\dagger$.

(a) $\mathbf{A}^\dagger = \mathbf{A}^{-1}$ when $\mathbf{A}$ is nonsingular.

(b) $(\mathbf{A}^\dagger)^\dagger = \mathbf{A}$.

(c) $(\mathbf{A}^\dagger)^T = (\mathbf{A}^T)^\dagger$.

(d) $\mathbf{A}^\dagger = \begin{cases} (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T & \text{when } rank\,(\mathbf{A}_{m\times n}) = n, \\ \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1} & \text{when } rank\,(\mathbf{A}_{m\times n}) = m. \end{cases}$

(e) $\mathbf{A}^T = \mathbf{A}^T\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger\mathbf{A}\mathbf{A}^T$ for all $\mathbf{A} \in \Re^{m\times n}$.

(f) $\mathbf{A}^\dagger = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^\dagger = (\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T$ for all $\mathbf{A} \in \Re^{m\times n}$.

(g) $R\left(\mathbf{A}^\dagger\right) = R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^\dagger\mathbf{A}\right)$, and
$N\left(\mathbf{A}^\dagger\right) = N\left(\mathbf{A}^T\right) = N\left(\mathbf{A}\mathbf{A}^\dagger\right)$.

(h) $(\mathbf{PAQ})^\dagger = \mathbf{Q}^T\mathbf{A}^\dagger\mathbf{P}^T$ when $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal matrices, but in general $(\mathbf{AB})^\dagger \neq \mathbf{B}^\dagger\mathbf{A}^\dagger$ (the reverse-order law fails).

(i) $(\mathbf{A}^T\mathbf{A})^\dagger = \mathbf{A}^\dagger(\mathbf{A}^T)^\dagger$ and $(\mathbf{A}\mathbf{A}^T)^\dagger = (\mathbf{A}^T)^\dagger\mathbf{A}^\dagger$.

5.12.17. Explain why $\mathbf{A}^\dagger = \mathbf{A}^D$ if and only if $\mathbf{A}$ is an RPN matrix.

5.12.18. Let $\mathbf{X}$, $\mathbf{Y} \in \Re^{m\times n}$ be such that $R(\mathbf{X}) \perp R(\mathbf{Y})$.

(a) Establish the Pythagorean theorem for matrices by proving

$$\|\mathbf{X}+\mathbf{Y}\|_F^2 = \|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2 .$$

(b) Give an example to show that the result of part (a) does not hold for the matrix 2-norm.

(c) Demonstrate that $\mathbf{A}^\dagger$ is the ***best approximate inverse*** for $\mathbf{A}$ in the sense that $\mathbf{A}^\dagger$ is the matrix of smallest Frobenius norm that minimizes $\|\mathbf{I}-\mathbf{AX}\|_F$.

5.13 ORTHOGONAL PROJECTION

As discussed in §5.9, every pair of complementary subspaces defines a projector. But when the complementary subspaces happen to be *orthogonal* complements, the resulting projector has some particularly nice properties, and the purpose of this section is to develop this special case in more detail. Discussions are in the context of real spaces, but generalizations to complex spaces are straightforward by replacing $(\star)^T$ by $(\star)^*$ and "orthogonal matrix" by "unitary matrix."

If $\mathcal{M}$ is a subspace of an inner-product space $\mathcal{V}$, then $\mathcal{V} = \mathcal{M} \oplus \mathcal{M}^\perp$ by (5.11.1), and each $\mathbf{v} \in \mathcal{V}$ can be written uniquely as $\mathbf{v} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in \mathcal{M}$ and $\mathbf{n} \in \mathcal{M}^\perp$ by (5.9.3). The vector $\mathbf{m}$ was defined on p. 385 to be the projection of $\mathbf{v}$ onto $\mathcal{M}$ along $\mathcal{M}^\perp$, so the following definitions are natural.

Orthogonal Projection

For $\mathbf{v} \in \mathcal{V}$, let $\mathbf{v} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in \mathcal{M}$ and $\mathbf{n} \in \mathcal{M}^\perp$.

- $\mathbf{m}$ is called the ***orthogonal projection*** of $\mathbf{v}$ onto $\mathcal{M}$.
- The projector $\mathbf{P}_{\mathcal{M}}$ onto $\mathcal{M}$ along $\mathcal{M}^\perp$ is called the ***orthogonal projector*** onto $\mathcal{M}$.
- $\mathbf{P}_{\mathcal{M}}$ is the unique linear operator such that $\mathbf{P}_{\mathcal{M}}\mathbf{v} = \mathbf{m}$ (see p. 386).

These ideas are illustrated illustrated in Figure 5.13.1 for $\mathcal{V} = \Re^3$.

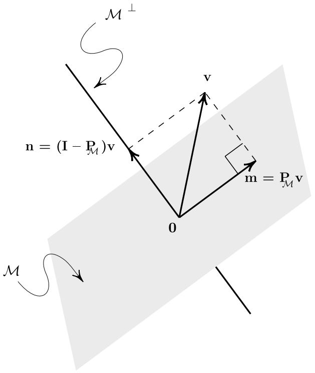

$\mathcal{M}^\perp$
$\mathbf{v}$
$\mathbf{n} = (\mathbf{I} - \mathbf{P}_{\mathcal{M}})\mathbf{v}$
$\mathbf{m} = \mathbf{P}_{\mathcal{M}}\mathbf{v}$
$\mathbf{0}$
$\mathcal{M}$

Figure 5.13.1

Given an arbitrary pair of complementary subspaces $\mathcal{M}$, $\mathcal{N}$ of $\Re^n$, formula (5.9.12) on p. 386 says that the projector $\mathbf{P}$ onto $\mathcal{M}$ along $\mathcal{N}$ is given by

$$\mathbf{P} = \left(\mathbf{M} \,|\, \mathbf{N}\right) \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \left(\mathbf{M} \,|\, \mathbf{N}\right)^{-1} = \left(\mathbf{M} \,|\, \mathbf{0}\right)\left(\mathbf{M} \,|\, \mathbf{N}\right)^{-1}, \tag{5.13.1}$$

where the columns of $\mathbf{M}$ and $\mathbf{N}$ constitute bases for $\mathcal{M}$ and $\mathcal{N}$, respectively. So, how does this expression simplify when $\mathcal{N} = \mathcal{M}^\perp$? To answer the question, observe that if $\mathcal{N} = \mathcal{M}^\perp$, then $\mathbf{N}^T\mathbf{M} = \mathbf{0}$ and $\mathbf{M}^T\mathbf{N} = \mathbf{0}$. Furthermore, if $\dim \mathcal{M} = r$, then $\mathbf{M}^T\mathbf{M}$ is $r \times r$, and $rank\left(\mathbf{M}^T\mathbf{M}\right) = rank\,(\mathbf{M}) = r$ by (4.5.4), so $\mathbf{M}^T\mathbf{M}$ is nonsingular. Therefore, if the columns of $\mathbf{N}$ are chosen to be an orthonormal basis for $\mathcal{M}^\perp$, then

$$\left(\begin{array}{c}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\\ \hline \mathbf{N}^T\end{array}\right)\left(\mathbf{M}\,|\,\mathbf{N}\right)=\begin{pmatrix}\mathbf{I}&\mathbf{0}\\ \mathbf{0}&\mathbf{I}\end{pmatrix} \implies \left(\mathbf{M}\,|\,\mathbf{N}\right)^{-1}=\left(\begin{array}{c}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\\ \hline \mathbf{N}^T\end{array}\right).$$

This together with (5.13.1) says the orthogonal projector onto $\mathcal{M}$ is given by

$$\mathbf{P}_{\mathcal{M}} = \left(\mathbf{M}\,|\,\mathbf{0}\right)\left(\begin{array}{c}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\\ \hline \mathbf{N}^T\end{array}\right) = \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T. \qquad (5.13.2)$$

As discussed in §5.9, the projector associated with any given pair of complementary subspaces is unique, and it doesn't matter which bases are used to form $\mathbf{M}$ and $\mathbf{N}$ in (5.13.1). Consequently, formula $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T$ is independent of the choice of $\mathbf{M}$—just as long as its columns constitute some basis for $\mathcal{M}$. In particular, the columns of $\mathbf{M}$ need not be an *orthonormal* basis for $\mathcal{M}$. But if they are, then $\mathbf{M}^T\mathbf{M} = \mathbf{I}$, and (5.13.2) becomes $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\mathbf{M}^T$. Moreover, if the columns of $\mathbf{M}$ and $\mathbf{N}$ constitute orthonormal bases for $\mathcal{M}$ and $\mathcal{M}^\perp$, respectively, then $\mathbf{U} = \left(\mathbf{M}\,|\,\mathbf{N}\right)$ is an orthogonal matrix, and (5.13.1) becomes

$$\mathbf{P}_{\mathcal{M}} = \mathbf{U}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T.$$

In other words, every orthogonal projector is orthogonally similar to a diagonal matrix in which the diagonal entries are 1's and 0's.

Below is a summary of the formulas used to build orthogonal projectors.

Constructing Orthogonal Projectors

Let $\mathcal{M}$ be an r-dimensional subspace of $\Re^n$, and let the columns of $\mathbf{M}_{n\times r}$ and $\mathbf{N}_{n\times n-r}$ be bases for $\mathcal{M}$ and $\mathcal{M}^\perp$, respectively. The orthogonal projectors onto $\mathcal{M}$ and $\mathcal{M}^\perp$ are

- $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T$ and $\mathbf{P}_{\mathcal{M}^\perp} = \mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T$. (5.13.3)

If $\mathbf{M}$ and $\mathbf{N}$ contain *orthonormal* bases for $\mathcal{M}$ and $\mathcal{M}^\perp$, then

- $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\mathbf{M}^T$ and $\mathbf{P}_{\mathcal{M}^\perp} = \mathbf{N}\mathbf{N}^T$. (5.13.4)
- $\mathbf{P}_{\mathcal{M}} = \mathbf{U}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T$, where $\mathbf{U} = \left(\mathbf{M}\,|\,\mathbf{N}\right)$. (5.13.5)
- $\mathbf{P}_{\mathcal{M}^\perp} = \mathbf{I} - \mathbf{P}_{\mathcal{M}}$ in all cases. (5.13.6)

Note: Extensions of (5.13.3) appear on p. 634.

Example 5.13.1

Problem: Let $\mathbf{u}_{n\times 1} \neq \mathbf{0}$, and consider the line $\mathcal{L} = span\{\mathbf{u}\}$. Construct the orthogonal projector onto $\mathcal{L}$, and then determine the orthogonal projection of a vector $\mathbf{x}_{n\times 1}$ onto $\mathcal{L}$.

Solution: The vector $\mathbf{u}$ by itself is a basis for $\mathcal{L}$, so, according to (5.13.3),

$$\mathbf{P}_{\mathcal{L}} = \mathbf{u}\left(\mathbf{u}^T\mathbf{u}\right)^{-1}\mathbf{u}^T = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}$$

is the orthogonal projector onto $\mathcal{L}$. The orthogonal projection of a vector $\mathbf{x}$ onto $\mathcal{L}$ is therefore given by

$$\mathbf{P}_{\mathcal{L}}\mathbf{x} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\mathbf{x} = \left(\frac{\mathbf{u}^T\mathbf{x}}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u}.$$

Note: If $\|\mathbf{u}\|_2 = 1$, then $\mathbf{P}_{\mathcal{L}} = \mathbf{u}\mathbf{u}^T$, so $\mathbf{P}_{\mathcal{L}}\mathbf{x} = \mathbf{u}\mathbf{u}^T\mathbf{x} = (\mathbf{u}^T\mathbf{x})\mathbf{u}$, and

$$\|\mathbf{P}_{\mathcal{L}}\mathbf{x}\|_2 = |\mathbf{u}^T\mathbf{x}|\ \|\mathbf{u}\|_2 = |\mathbf{u}^T\mathbf{x}|.$$

This yields a geometrical interpretation for the magnitude of the standard inner product. It says that if $\mathbf{u}$ is a vector of unit length in $\mathcal{L}$, then, as illustrated in Figure 5.13.2, $|\mathbf{u}^T\mathbf{x}|$ is the length of the orthogonal projection of $\mathbf{x}$ onto the line spanned by $\mathbf{u}$.

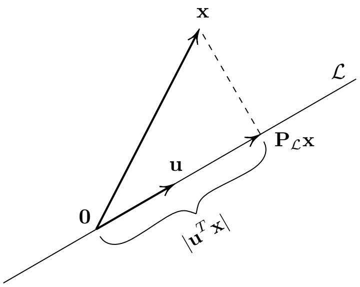

x
$\mathcal{L}$
$\mathbf{P}_{\mathcal{L}}\mathbf{x}$
u
0
$|\mathbf{u}^T\mathbf{x}|$

FIGURE 5.13.2

Finally, notice that since $\mathbf{P}_{\mathcal{L}} = \mathbf{u}\mathbf{u}^T$ is the orthogonal projector onto $\mathcal{L}$, it must be the case that $\mathbf{P}_{\mathcal{L}^\perp} = \mathbf{I} - \mathbf{P}_{\mathcal{L}} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$ is the orthogonal projection onto $\mathcal{L}^\perp$. This was called an ***elementary orthogonal projector*** on p. 322—go back and reexamine Figure 5.6.1.

Example 5.13.2

Volume, Gram–Schmidt, and QR. A solid in $\Re^m$ with parallel opposing faces whose adjacent sides are defined by vectors from a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is called an n-dimensional ***parallelepiped***. As shown in the shaded portions of Figure 5.13.3, a two-dimensional parallelepiped is a parallelogram, and a three-dimensional parallelepiped is a skewed rectangular box.

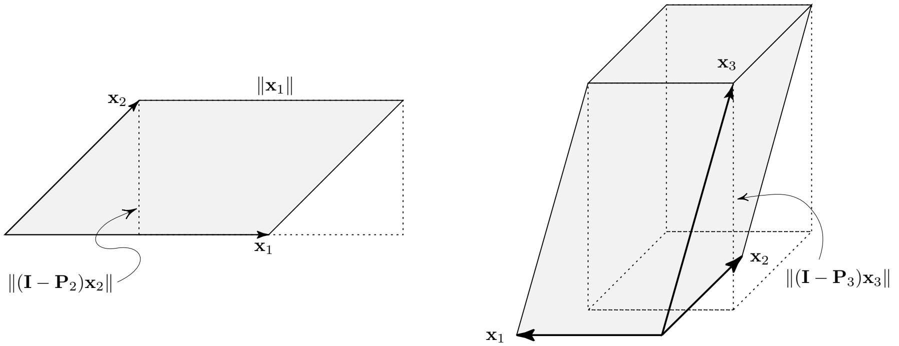


Figure 5.13.3

Problem: Determine the volumes of a two-dimensional and a three-dimensional parallelepiped, and then make the natural extension to define the volume of an n-dimensional parallelepiped.

Solution: In the two-dimensional case, volume is area, and it's evident from Figure 5.13.3 that the area of the shaded parallelogram is the same as the area of the dotted rectangle. The width of the dotted rectangle is $\nu_1 = \|\mathbf{x}_1\|_2$, and the height is $\nu_2 = \|(\mathbf{I} - \mathbf{P}_2)\mathbf{x}_2\|_2$, where $\mathbf{P}_2$ is the orthogonal projector onto the space (line) spanned by $\mathbf{x}_1$, and $\mathbf{I} - \mathbf{P}_2$ is the orthogonal projector onto $span\{\mathbf{x}_1\}^{\perp}$. In other words, the area, V_2, of the parallelogram is the length of its base times its *projected height*, ν_2, so

$$V_2 = \|\mathbf{x}_1\|_2 \, \|(\mathbf{I} - \mathbf{P}_2)\mathbf{x}_2\|_2 = \nu_1\nu_2.$$

Similarly, the volume of a three-dimensional parallelepiped is the area of its base times its projected height. The area of the base was just determined to be $V_2 = \|\mathbf{x}_1\|_2 \, \|(\mathbf{I} - \mathbf{P}_2)\mathbf{x}_2\|_2 = \nu_1\nu_2$, and it's evident from Figure 5.13.3 that the projected height is $\nu_3 = \|(\mathbf{I} - \mathbf{P}_3)\mathbf{x}_3\|_2$, where $\mathbf{P}_3$ is the orthogonal projector onto $span\{\mathbf{x}_1, \mathbf{x}_2\}$. Therefore, the volume of the parallelepiped generated by $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is

$$V_3 = \|\mathbf{x}_1\|_2 \, \|(\mathbf{I} - \mathbf{P}_2)\mathbf{x}_2\|_2 \, \|(\mathbf{I} - \mathbf{P}_3)\mathbf{x}_3\|_2 = \nu_1\nu_2\nu_3.$$

It's now clear how to inductively define V_4, V_5, etc. In general, the volume of the parallelepiped generated by a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is

$$V_n = \|\mathbf{x}_1\|_2 \, \|(\mathbf{I} - \mathbf{P}_2)\mathbf{x}_2\|_2 \, \|(\mathbf{I} - \mathbf{P}_3)\mathbf{x}_3\|_2 \cdots \|(\mathbf{I} - \mathbf{P}_n)\mathbf{x}_n\|_2 = \nu_1\nu_2 \cdots \nu_n,$$

where $\mathbf{P}_k$ is the orthogonal projector onto $span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$, and where

$$\nu_1 = \|\mathbf{x}_1\|_2 \quad \text{and} \quad \nu_k = \|(\mathbf{I} - \mathbf{P}_k)\mathbf{x}_k\|_2 \quad \text{for} \quad k > 1. \tag{5.13.7}$$

Note that if $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is an *orthogonal* set, $V_n = \|\mathbf{x}_1\|_2 \, \|\mathbf{x}_2\|_2 \cdots \|\mathbf{x}_n\|_2$, which is what we would expect.

Connections with Gram–Schmidt and QR. Recall from (5.5.4) on p. 309 that the vectors in the Gram–Schmidt sequence generated from a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \Re^m$ are $\mathbf{u}_1 = \mathbf{x}_1 / \|\mathbf{x}_1\|_2$ and

$$\mathbf{u}_k = \frac{\left(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T\right)\mathbf{x}_k}{\left\|\left(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T\right)\mathbf{x}_k\right\|_2}, \quad \text{where} \quad \mathbf{U}_k = \left[\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\right] \quad \text{for } k > 1.$$

Since $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k-1}\}$ is an *orthonormal* basis for $span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$, it follows from (5.13.4) that $\mathbf{U}_k\mathbf{U}_k^T$ must be the orthogonal projector onto $span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$. Hence $\mathbf{U}_k\mathbf{U}_k^T = \mathbf{P}_k$ and $(\mathbf{I}-\mathbf{P}_k)\mathbf{x}_k = (\mathbf{I}-\mathbf{U}_k\mathbf{U}_k^T)\mathbf{x}_k$, so $\left\|\left(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T\right)\mathbf{x}_k\right\|_2 = \nu_k$ is the k^{th} projected height in (5.13.7). This means that when the Gram–Schmidt equations are written in the form of a QR factorization as explained on p. 311, the diagonal elements of the upper-triangular matrix $\mathbf{R}$ are the ν_k's. Consequently, the product of the diagonal entries in $\mathbf{R}$ is the volume of the parallelepiped generated by the $\mathbf{x}_k$'s. But the QR factorization of $\mathbf{A} = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n\right]$ is unique (Exercise 5.5.8), so it doesn't matter whether Gram–Schmidt or another method is used to determine the QR factors. Therefore, we arrive at the following conclusion.

- If $\mathbf{A}_{m\times n} = \mathbf{Q}_{m\times n}\mathbf{R}_{n\times n}$ is the (rectangular) QR factorization of a matrix with linearly independent columns, then the volume of the n-dimensional parallelepiped generated by the columns of $\mathbf{A}$ is $V_n = \nu_1\nu_2\cdots\nu_n$, where the ν_k's are the diagonal elements of $\mathbf{R}$. We will see on p. 468 what this means in terms of determinants.

Of course, not all projectors are *orthogonal* projectors, so a natural question to ask is, "What characteristic features distinguish orthogonal projectors from more general oblique projectors?" Some answers are given below.

Orthogonal Projectors

Suppose that $\mathbf{P} \in \Re^{n\times n}$ is a projector—i.e., $\mathbf{P}^2 = \mathbf{P}$. The following statements are equivalent to saying that $\mathbf{P}$ is an *orthogonal* projector.

- $R(\mathbf{P}) \perp N(\mathbf{P})$. (5.13.8)
- $\mathbf{P}^T = \mathbf{P}$ (i.e., orthogonal projector $\Longleftrightarrow \mathbf{P}^2 = \mathbf{P} = \mathbf{P}^T$). (5.13.9)
- $\|\mathbf{P}\|_2 = 1$ for the matrix 2-norm (p. 281). (5.13.10)

Proof. Every projector projects vectors onto its range along (parallel to) its nullspace, so statement (5.13.8) is essentially a restatement of the definition of an orthogonal projector. To prove (5.13.9), note that if $\mathbf{P}$ is an orthogonal projector, then (5.13.3) insures that $\mathbf{P}$ is symmetric. Conversely, if a projector

$\mathbf{P}$ is symmetric, then it must be an orthogonal projector because (5.11.5) on p. 405 allows us to write

$$\mathbf{P}=\mathbf{P}^T \implies R(\mathbf{P})=R\left(\mathbf{P}^T\right) \implies R(\mathbf{P})\perp N(\mathbf{P}).$$

To see why (5.13.10) characterizes projectors that are orthogonal, refer back to Example 5.9.2 on p. 389 (or look ahead to (5.15.3)) and note that $\|\mathbf{P}\|_2 = 1/\sin\theta$, where θ is the angle between $R(\mathbf{P})$ and $N(\mathbf{P})$. This makes it clear that $\|\mathbf{P}\|_2 \geq 1$ for all projectors, and $\|\mathbf{P}\|_2 = 1$ if and only if $\theta=\pi/2$, (i.e., if and only if $R(\mathbf{P})\perp N(\mathbf{P})$). ■

Example 5.13.3

Problem: For $\mathbf{A}\in\Re^{m\times n}$ such that $rank(\mathbf{A})=r$, describe the orthogonal projectors onto each of the four fundamental subspaces of $\mathbf{A}$.

Solution 1: Let $\mathbf{B}_{m\times r}$ and $\mathbf{N}_{n\times n-r}$ be matrices whose columns are bases for $R(\mathbf{A})$ and $N(\mathbf{A})$, respectively—e.g., $\mathbf{B}$ might contain the basic columns of $\mathbf{A}$. The orthogonal decomposition theorem on p. 405 says $R(\mathbf{A})^{\perp}=N\left(\mathbf{A}^T\right)$ and $N(\mathbf{A})^{\perp}=R\left(\mathbf{A}^T\right)$, so, by making use of (5.13.3) and (5.13.6), we can write

$$\begin{aligned}
\mathbf{P}_{R(\mathbf{A})} &= \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T,\\
\mathbf{P}_{N(\mathbf{A}^T)} &= \mathbf{P}_{R(\mathbf{A})^{\perp}} = \mathbf{I}-\mathbf{P}_{R(\mathbf{A})} = \mathbf{I}-\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T,\\
\mathbf{P}_{N(\mathbf{A})} &= \mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T,\\
\mathbf{P}_{R(\mathbf{A}^T)} &= \mathbf{P}_{N(\mathbf{A})^{\perp}} = \mathbf{I}-\mathbf{P}_{N(\mathbf{A})} = \mathbf{I}-\mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T.
\end{aligned}$$

Note: If $rank(\mathbf{A})=n$, then all columns of $\mathbf{A}$ are basic and

$$\mathbf{P}_{R(\mathbf{A})}=\mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T. \tag{5.13.11}$$

Solution 2: Another way to describe these projectors is to make use of the Moore–Penrose pseudoinverse $\mathbf{A}^{\dagger}$ (p. 423). Recall that if $\mathbf{A}$ has a URV factorization

$$\mathbf{A}=\mathbf{U}\begin{pmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T, \quad \text{then} \quad \mathbf{A}^{\dagger}=\mathbf{V}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T,$$

where $\mathbf{U}=\left(\mathbf{U}_1\,|\,\mathbf{U}_2\right)$ and $\mathbf{V}=\left(\mathbf{V}_1\,|\,\mathbf{V}_2\right)$ are orthogonal matrices in which the columns of $\mathbf{U}_1$ and $\mathbf{V}_1$ constitute orthonormal bases for $R(\mathbf{A})$ and $R\left(\mathbf{A}^T\right)$, respectively, and the columns of $\mathbf{U}_2$ and $\mathbf{V}_2$ are orthonormal bases for $N\left(\mathbf{A}^T\right)$ and $N(\mathbf{A})$, respectively. Computing the products $\mathbf{A}\mathbf{A}^{\dagger}$ and $\mathbf{A}^{\dagger}\mathbf{A}$ reveals

$$\mathbf{A}\mathbf{A}^{\dagger}=\mathbf{U}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T=\mathbf{U}_1\mathbf{U}_1^T \quad \text{and} \quad \mathbf{A}^{\dagger}\mathbf{A}=\mathbf{V}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T=\mathbf{V}_1\mathbf{V}_1^T,$$

so, according to (5.13.4),

$$\begin{aligned} \mathbf{P}_{R(\mathbf{A})} &= \mathbf{U}_1\mathbf{U}_1^T = \mathbf{A}\mathbf{A}^\dagger, & \mathbf{P}_{N(\mathbf{A}^T)} &= \mathbf{I} - \mathbf{P}_{R(\mathbf{A})} = \mathbf{I} - \mathbf{A}\mathbf{A}^\dagger, \\ \mathbf{P}_{R(\mathbf{A}^T)} &= \mathbf{V}_1\mathbf{V}_1^T = \mathbf{A}^\dagger\mathbf{A}, & \mathbf{P}_{N(\mathbf{A})} &= \mathbf{I} - \mathbf{P}_{R(\mathbf{A}^T)} = \mathbf{I} - \mathbf{A}^\dagger\mathbf{A}. \end{aligned} \tag{5.13.12}$$

The notion of orthogonal projection in higher-dimensional spaces is consistent with the visual geometry in $\Re^2$ and $\Re^3$. In particular, it is visually evident from Figure 5.13.4 that if $\mathcal{M}$ is a subspace of $\Re^3$, and if $\mathbf{b}$ is a vector outside of $\mathcal{M}$, then the point in $\mathcal{M}$ that is closest to $\mathbf{b}$ is $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$.

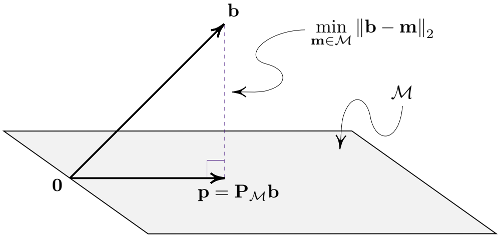

$\mathbf{b}$
$\min_{\mathbf{m}\in\mathcal{M}} \|\mathbf{b}-\mathbf{m}\|_2$
$\mathcal{M}$
$\mathbf{0}$
$\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$

FIGURE 5.13.4

The situation is exactly the same in higher dimensions. But rather than using our eyes to understand why, we use mathematics—it's surprising just how easy it is to "see" such things in abstract spaces.

Closest Point Theorem

Let $\mathcal{M}$ be a subspace of an inner-product space $\mathcal{V}$, and let $\mathbf{b}$ be a vector in $\mathcal{V}$. The unique vector in $\mathcal{M}$ that is closest to $\mathbf{b}$ is $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$. In other words,

$$\min_{\mathbf{m}\in\mathcal{M}} \|\mathbf{b}-\mathbf{m}\|_2 = \|\mathbf{b}-\mathbf{P}_{\mathcal{M}}\mathbf{b}\|_2 = \mathit{dist}\,(\mathbf{b},\mathcal{M}). \tag{5.13.13}$$

This is called the ***orthogonal distance*** between $\mathbf{b}$ and $\mathcal{M}$.

Proof. If $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, then $\mathbf{p}-\mathbf{m} \in \mathcal{M}$ for all $\mathbf{m} \in \mathcal{M}$, and

$$\mathbf{b}-\mathbf{p} = (\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{b} \in \mathcal{M}^\perp,$$

so $(\mathbf{p}-\mathbf{m}) \perp (\mathbf{b}-\mathbf{p})$. The Pythagorean theorem says $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$ whenever $\mathbf{x} \perp \mathbf{y}$ (recall Exercise 5.4.14), and hence

$$\|\mathbf{b}-\mathbf{m}\|_2^2 = \|\mathbf{b}-\mathbf{p}+\mathbf{p}-\mathbf{m}\|_2^2 = \|\mathbf{b}-\mathbf{p}\|_2^2 + \|\mathbf{p}-\mathbf{m}\|_2^2 \geq \|\mathbf{p}-\mathbf{m}\|_2^2\,.$$

In other words, $\min_{\mathbf{m}\in\mathcal{M}} \|\mathbf{b}-\mathbf{m}\|_2 = \|\mathbf{b}-\mathbf{p}\|_2$. Now argue that there is not another point in $\mathcal{M}$ that is as close to $\mathbf{b}$ as $\mathbf{p}$ is. If $\widehat{\mathbf{m}} \in \mathcal{M}$ such that $\|\mathbf{b}-\widehat{\mathbf{m}}\|_2 = \|\mathbf{b}-\mathbf{p}\|_2$, then by using the Pythagorean theorem again we see

$$\|\mathbf{b}-\widehat{\mathbf{m}}\|_2^2 = \|\mathbf{b}-\mathbf{p}+\mathbf{p}-\widehat{\mathbf{m}}\|_2^2 = \|\mathbf{b}-\mathbf{p}\|_2^2 + \|\mathbf{p}-\widehat{\mathbf{m}}\|_2^2 \quad \Longrightarrow \quad \|\mathbf{p}-\widehat{\mathbf{m}}\|_2 = 0,$$

and thus $\widehat{\mathbf{m}} = \mathbf{p}$. ■

Example 5.13.4

To illustrate some of the previous ideas, consider $\Re^{n\times n}$ with the inner product $\langle \mathbf{A}|\mathbf{B}\rangle = trace\left(\mathbf{A}^T\mathbf{B}\right)$. If $\mathcal{S}_n$ is the subspace of $n\times n$ real-symmetric matrices, then each of the following statements is true.

- $\mathcal{S}_n^\perp =$ the subspace $\mathcal{K}_n$ of $n\times n$ skew-symmetric matrices.
  - ▷ $\mathcal{S}_n \perp \mathcal{K}_n$ because for all $\mathbf{S}\in\mathcal{S}_n$ and $\mathbf{K}\in\mathcal{K}_n$,

$$\begin{aligned}\langle \mathbf{S}|\mathbf{K}\rangle &= trace\left(\mathbf{S}^T\mathbf{K}\right) = -trace\left(\mathbf{S}\mathbf{K}^T\right) = -trace\left(\mathbf{S}\mathbf{K}^T\right)^T \\ &= -trace\left(\mathbf{K}\mathbf{S}^T\right) = -trace\left(\mathbf{S}^T\mathbf{K}\right) = -\langle \mathbf{S}|\mathbf{K}\rangle \\ \Longrightarrow \quad & \langle \mathbf{S}|\mathbf{K}\rangle = 0.\end{aligned}$$

  - ▷ $\Re^{n\times n} = \mathcal{S}_n \oplus \mathcal{K}_n$ because every $\mathbf{A}\in\Re^{n\times n}$ can be uniquely expressed as the sum of a symmetric and a skew-symmetric matrix by writing

$$\mathbf{A} = \frac{\mathbf{A}+\mathbf{A}^T}{2} + \frac{\mathbf{A}-\mathbf{A}^T}{2} \qquad \text{(recall (5.9.3) and Exercise 3.2.6).}$$

- The orthogonal projection of $\mathbf{A}\in\Re^{n\times n}$ onto $\mathcal{S}_n$ is $\mathbf{P}(\mathbf{A}) = (\mathbf{A}+\mathbf{A}^T)/2$.
- The closest symmetric matrix to $\mathbf{A}\in\Re^{n\times n}$ is $\mathbf{P}(\mathbf{A}) = (\mathbf{A}+\mathbf{A}^T)/2$.
- The distance from $\mathbf{A}\in\Re^{n\times n}$ to $\mathcal{S}_n$ (the deviation from symmetry) is

$$dist(\mathbf{A},\mathcal{S}_n) = \|\mathbf{A}-\mathbf{P}(\mathbf{A})\|_F = \left\|(\mathbf{A}-\mathbf{A}^T)/2\right\|_F = \sqrt{\frac{trace\left(\mathbf{A}^T\mathbf{A}\right) - trace\left(\mathbf{A}^2\right)}{2}}.$$

Example 5.13.5

Affine Projections. If $\mathbf{v}\neq\mathbf{0}$ is a vector in a space $\mathcal{V}$, and if $\mathcal{M}$ is a subspace of $\mathcal{V}$, then the set of points $\mathcal{A} = \mathbf{v}+\mathcal{M}$ is called an ***affine space*** in $\mathcal{V}$. Strictly speaking, $\mathcal{A}$ is not a subspace (e.g., it doesn't contain $\mathbf{0}$), but, as depicted in Figure 5.13.5, $\mathcal{A}$ is the translate of a subspace—i.e., $\mathcal{A}$ is just a copy of $\mathcal{M}$ that has been translated away from the origin through $\mathbf{v}$. Consequently, notions such as projection onto $\mathcal{A}$ and points closest to $\mathcal{A}$ are analogous to the corresponding concepts for subspaces.

Problem: For $\mathbf{b} \in \mathcal{V}$, determine the point $\mathbf{p}$ in $\mathcal{A} = \mathbf{v} + \mathcal{M}$ that is closest to $\mathbf{b}$. In other words, explain how to project $\mathbf{b}$ orthogonally onto $\mathcal{A}$.

Solution: The trick is to subtract $\mathbf{v}$ from $\mathbf{b}$ as well as from everything in $\mathcal{A}$ to put things back into the context of subspaces where we already know the answers. As illustrated in Figure 5.13.5, this moves $\mathcal{A}$ back down to $\mathcal{M}$, and it translates $\mathbf{v} \to \mathbf{0}$, $\mathbf{b} \to (\mathbf{b} - \mathbf{v})$, and $\mathbf{p} \to (\mathbf{p} - \mathbf{v})$.

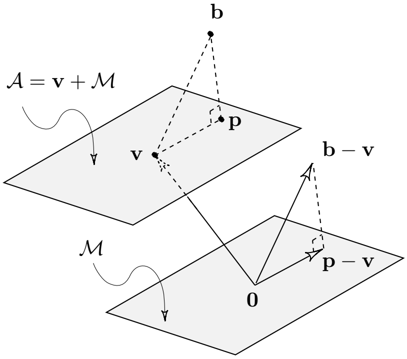


Figure 5.13.5

If $\mathbf{p}$ is to be the orthogonal projection of $\mathbf{b}$ onto $\mathcal{A}$, then $\mathbf{p} - \mathbf{v}$ must be the orthogonal projection of $\mathbf{b} - \mathbf{v}$ onto $\mathcal{M}$, so

$$\mathbf{p} - \mathbf{v} = \mathbf{P}_{\mathcal{M}}(\mathbf{b} - \mathbf{v}) \quad \Longrightarrow \quad \mathbf{p} = \mathbf{v} + \mathbf{P}_{\mathcal{M}}(\mathbf{b} - \mathbf{v}), \tag{5.13.14}$$

and thus $\mathbf{p}$ is the point in $\mathcal{A}$ that is closest to $\mathbf{b}$. Applications to the solution of linear systems are developed in Exercises 5.13.17–5.13.22.

We are now in a position to replace the classical calculus-based theory of least squares presented in §4.6 with a more modern vector space development. In addition to being straightforward, the modern geometrical approach puts the entire least squares picture in much sharper focus. Viewing concepts from more than one perspective generally produces deeper understanding, and this is particularly true for the theory of least squares.

Recall from p. 226 that for an inconsistent system $\mathbf{A}_{m\times n}\mathbf{x} = \mathbf{b}$, the object of the least squares problem is to find vectors $\mathbf{x}$ that minimize the quantity

$$(\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 . \tag{5.13.15}$$

The classical development in §4.6 relies on calculus to argue that the set of vectors $\mathbf{x}$ that minimize (5.13.15) is exactly the set that solves the (always consistent) system of normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. In the context of the closest point theorem the least squares problem asks for vectors $\mathbf{x}$ such that $\mathbf{A}\mathbf{x}$ is as close to $\mathbf{b}$ as possible. But $\mathbf{Ax}$ is always a vector in $R(\mathbf{A})$, and the closest point theorem says that the vector in $R(\mathbf{A})$ that is closest to $\mathbf{b}$ is $\mathbf{P}_{R(\mathbf{A})}\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $R(\mathbf{A})$. Figure 5.13.6 illustrates the situation in $\Re^3$.

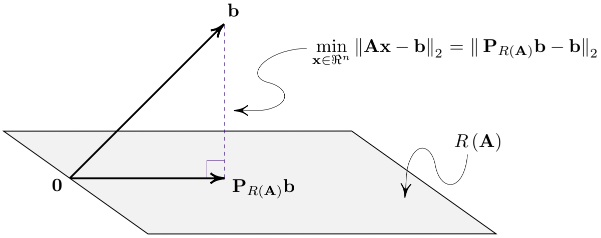

$\mathbf{b}$
$\min_{\mathbf{x}\in\Re^n} \|\mathbf{Ax}-\mathbf{b}\|_2 = \|\mathbf{P}_{R(\mathbf{A})}\mathbf{b}-\mathbf{b}\|_2$
$R(\mathbf{A})$
$\mathbf{0}$
$\mathbf{P}_{R(\mathbf{A})}\mathbf{b}$

Figure 5.13.6

So the least squares problem boils down to finding vectors $\mathbf{x}$ such that

$$\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}.$$

But this system is equivalent to the system of normal equations because

$$\begin{aligned}\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b} &\iff \mathbf{P}_{R(\mathbf{A})}\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}\\ &\iff \mathbf{P}_{R(\mathbf{A})}(\mathbf{Ax}-\mathbf{b}) = \mathbf{0}\\ &\iff (\mathbf{Ax}-\mathbf{b}) \in N\left(\mathbf{P}_{R(\mathbf{A})}\right) = R(\mathbf{A})^{\perp} = N\left(\mathbf{A}^T\right)\\ &\iff \mathbf{A}^T(\mathbf{Ax}-\mathbf{b}) = \mathbf{0}\\ &\iff \mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}.\end{aligned}$$

Characterizing the set of least squares solutions as the solutions to $\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}$ makes it obvious that $\mathbf{x} = \mathbf{A}^{\dagger}\mathbf{b}$ is a particular least squares solution because (5.13.12) insures $\mathbf{AA}^{\dagger} = \mathbf{P}_{R(\mathbf{A})}$, and thus

$$\mathbf{A}(\mathbf{A}^{\dagger}\mathbf{b}) = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}.$$

Furthermore, since $\mathbf{A}^{\dagger}\mathbf{b}$ is a particular solution of $\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}$, the general solution—i.e., the set of all least squares solutions—must be the affine space $\mathcal{S} = \mathbf{A}^{\dagger}\mathbf{b} + N(\mathbf{A})$. Finally, the fact that $\mathbf{A}^{\dagger}\mathbf{b}$ is the least squares solution of minimal norm follows from Example 5.13.5 together with

$$R\left(\mathbf{A}^{\dagger}\right) = R\left(\mathbf{A}^T\right) = N(\mathbf{A})^{\perp} \quad \text{(see part (g) of Exercise 5.12.16)}$$

because (5.13.14) insures that the point in $\mathcal{S}$ that is closest to the origin is

$$\mathbf{p} = \mathbf{A}^{\dagger}\mathbf{b} + \mathbf{P}_{N(\mathbf{A})}(\mathbf{0} - \mathbf{A}^{\dagger}\mathbf{b}) = \mathbf{A}^{\dagger}\mathbf{b}.$$

The classical development in §4.6 based on partial differentiation is not easily generalized to cover the case of complex matrices, but the vector space approach given in this example trivially extends to complex matrices by simply replacing $(\star)^T$ by $(\star)^*$.

Below is a summary of some of the major points concerning the theory of least squares.

Least Squares Solutions

Each of the following four statements is equivalent to saying that $\widehat{\mathbf{x}}$ is a least squares solution for a possibly inconsistent linear system $\mathbf{Ax} = \mathbf{b}$.

- $\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}\|_2 = \min\limits_{\mathbf{x}\in\Re^n} \|\mathbf{Ax} - \mathbf{b}\|_2 .$ (5.13.16)
- $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}.$ (5.13.17)
- $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$ $\quad (\mathbf{A}^*\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^*\mathbf{b}$ when $\mathbf{A} \in \mathcal{C}^{m\times n}$). (5.13.18)
- $\widehat{\mathbf{x}} \in \mathbf{A}^\dagger\mathbf{b} + N(\mathbf{A})$ ($\mathbf{A}^\dagger\mathbf{b}$ is the minimal 2-norm LSS). (5.13.19)

Caution! These are valuable theoretical characterizations, but none is recommended for floating-point computation. Directly solving (5.13.17) or (5.13.18) or explicitly computing $\mathbf{A}^\dagger$ can be inefficient and numerically unstable. Computational issues are discussed in Example 4.5.1 on p. 214; Example 5.5.3 on p. 313; and Example 5.7.3 on p. 346.

The least squares story will not be complete until the following fundamental question is answered: "Why is the method of least squares the best way to make estimates of physical phenomena in the face of uncertainty?" This is the focal point of the next section.

Exercises for section 5.13

5.13.1. Find the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M} = span\{\mathbf{u}\}$, and then determine the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}^\perp$, where $\mathbf{b} = (4 \quad 8)^T$ and $\mathbf{u} = (3 \quad 1)^T$.

5.13.2. Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 1 & 2 & 0 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

(a) Compute the orthogonal projectors onto each of the four fundamental subspaces associated with $\mathbf{A}$.

(b) Find the point in $N(\mathbf{A})^\perp$ that is closest to $\mathbf{b}$.

5.13.3. For an orthogonal projector $\mathbf{P}$, prove that $\|\mathbf{Px}\|_2 = \|\mathbf{x}\|_2$ if and only if $\mathbf{x} \in R(\mathbf{P})$.

5.13.4. Explain why $\mathbf{A}^T\mathbf{P}_{R(\mathbf{A})} = \mathbf{A}^T$ for all $\mathbf{A} \in \Re^{m\times n}$.

5.13.5. Explain why $\mathbf{P}_{\mathcal{M}} = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i{}^T$ whenever $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ is an orthonormal basis for $\mathcal{M} \subseteq \Re^{n\times 1}$.

5.13.6. Explain how to use orthogonal reduction techniques to compute the orthogonal projectors onto each of the four fundamental subspaces of a matrix $\mathbf{A} \in \Re^{m\times n}$.

5.13.7. (a) Describe all 2×2 orthogonal projectors in $\Re^{2\times 2}$.
(b) Describe all 2×2 projectors in $\Re^{2\times 2}$.

5.13.8. The line $\mathcal{L}$ in $\Re^n$ passing through two distinct points $\mathbf{u}$ and $\mathbf{v}$ is $\mathcal{L} = \mathbf{u} + \mathit{span}\{\mathbf{u} - \mathbf{v}\}$. If $\mathbf{u} \neq \mathbf{0}$ and $\mathbf{v} \neq \alpha\mathbf{u}$, then $\mathcal{L}$ is a line not passing through the origin—i.e., $\mathcal{L}$ is not a subspace. Sketch a picture in $\Re^2$ or $\Re^3$ to visualize this, and then explain how to project a vector $\mathbf{b}$ orthogonally onto $\mathcal{L}$.

5.13.9. Explain why $\widehat{\mathbf{x}}$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$ if and only if $\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}\|_2 = \left\|\mathbf{P}_{N(\mathbf{A}^T)}\mathbf{b}\right\|_2$.

5.13.10. Prove that if $\boldsymbol{\varepsilon} = \mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}$, where $\widehat{\mathbf{x}}$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$, then $\|\boldsymbol{\varepsilon}\|_2^2 = \|\mathbf{b}\|_2^2 - \left\|\mathbf{P}_{R(\mathbf{A})}\mathbf{b}\right\|_2^2$.

5.13.11. Let $\mathcal{M}$ be an r-dimensional subspace of $\Re^n$. We know from (5.4.3) that if $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ is an orthonormal basis for $\mathcal{M}$, and if $\mathbf{x} \in \mathcal{M}$, then $\mathbf{x}$ is equal to its Fourier expansion with respect to $\mathcal{B}$. That is, $\mathbf{x} = \sum_{i=1}^{r} (\mathbf{u}_i{}^T\mathbf{x})\mathbf{u}_i$. However, if $\mathbf{x} \notin \mathcal{M}$, then equality is not possible (why?), so the question that arises is, "What does the Fourier expansion on the right-hand side of this expression represent?" Answer this question by showing that the Fourier expansion $\sum_{i=1}^{r} (\mathbf{u}_i{}^T\mathbf{x})\mathbf{u}_i$ is the point in $\mathcal{M}$ that is closest to $\mathbf{x}$ in the euclidean norm. In other words, show that $\sum_{i=1}^{r} (\mathbf{u}_i{}^T\mathbf{x})\mathbf{u}_i = \mathbf{P}_{\mathcal{M}}\mathbf{x}$.

5.13.12. Determine the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$, where

$$\mathbf{b} = \begin{pmatrix} 5 \\ 2 \\ 5 \\ 3 \end{pmatrix} \quad \text{and} \quad \mathcal{M} = \mathit{span}\left\{ \begin{pmatrix} -3/5 \\ 0 \\ 4/5 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 4/5 \\ 0 \\ 3/5 \\ 0 \end{pmatrix} \right\}.$$

Hint: Is this spanning set in fact an orthonormal basis?

5.13.13. Let $\mathcal{M}$ and $\mathcal{N}$ be subspaces of a vector space $\mathcal{V}$, and consider the associated orthogonal projectors $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$.

(a) Prove that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$ if and only if $\mathcal{M} \perp \mathcal{N}$.

(b) Is it true that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$ if and only if $\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}} = \mathbf{0}$? Why?

5.13.14. Let $\mathcal{M}$ and $\mathcal{N}$ be subspaces of the same vector space, and let $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ be orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively.

(a) Prove that $R(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}) = R(\mathbf{P}_{\mathcal{M}}) + R(\mathbf{P}_{\mathcal{N}}) = \mathcal{M} + \mathcal{N}$. **Hint:** Use Exercise 4.2.9 along with (4.5.5).

(b) Explain why $\mathcal{M} \perp \mathcal{N}$ if and only if $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$.

(c) Explain why $\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}$ is an orthogonal projector if and only if $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$, in which case $R(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}) = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{M} \perp \mathcal{N}$. **Hint:** Recall Exercise 5.9.17.

5.13.15. Anderson–Duffin Formula.[59] Prove that if $\mathcal{M}$ and $\mathcal{N}$ are subspaces of the same vector space, then the orthogonal projector onto $\mathcal{M} \cap \mathcal{N}$ is given by $\mathbf{P}_{\mathcal{M}\cap\mathcal{N}} = 2\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}}$. **Hint:** Use (5.13.12) and Exercise 5.13.14 to show $\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}$. Argue that if $\mathbf{Z} = 2\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}$, then $\mathbf{Z} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\mathbf{Z} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}$.

5.13.16. Given a square matrix $\mathbf{X}$, the ***matrix exponential*** $\mathrm{e}^{\mathbf{X}}$ is defined as

$$\mathrm{e}^{\mathbf{X}} = \mathbf{I} + \mathbf{X} + \frac{\mathbf{X}^2}{2!} + \frac{\mathbf{X}^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{\mathbf{X}^n}{n!}.$$

It can be shown that this series converges for all $\mathbf{X}$, and it is legitimate to differentiate and integrate it term by term to produce the statements $d\mathrm{e}^{\mathbf{A}t}/dt = \mathbf{A}\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\mathbf{A}t}\mathbf{A}$ and $\int \mathrm{e}^{\mathbf{A}t}\mathbf{A}\,dt = \mathrm{e}^{\mathbf{A}t}$.

(a) Use the fact that $\lim_{t\to\infty} \mathrm{e}^{-\mathbf{A}^T\mathbf{A}t} = \mathbf{0}$ for all $\mathbf{A} \in \Re^{m\times n}$ to show $\mathbf{A}^{\dagger} = \int_0^{\infty} \mathrm{e}^{-\mathbf{A}^T\mathbf{A}t}\mathbf{A}^T dt$.

(b) If $\lim_{t\to\infty} \mathrm{e}^{-\mathbf{A}^{k+1}t} = \mathbf{0}$, show $\mathbf{A}^D = \int_0^{\infty} \mathrm{e}^{-\mathbf{A}^{k+1}t}\mathbf{A}^k dt$, where $k = index(\mathbf{A})$.[60]

(c) For nonsingular matrices, show that if $\lim_{t\to\infty} \mathrm{e}^{-\mathbf{A}t} = \mathbf{0}$, then $\mathbf{A}^{-1} = \int_0^{\infty} \mathrm{e}^{-\mathbf{A}t} dt$.

[59] W. N. Anderson, Jr., and R. J. Duffin discovered this formula for the orthogonal projector onto an intersection in 1969. They called $\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}}$ the *parallel sum* of $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ because it is the matrix generalization of the scalar function $r_1r_2/(r_1 + r_2) = r_1(r_1 + r_2)^{-1}r_2$ that is the resistance of a circuit composed of two resistors r_1 and r_2 connected in parallel. The simple elegance of the Anderson–Duffin formula makes it one of the innumerable little sparkling facets in the jewel that is linear algebra.

[60] A more useful integral representation for $\mathbf{A}^D$ is given in Exercise 7.9.22 (p. 615).

5.13.17. An affine space $\mathbf{v} + \mathcal{M} \subseteq \Re^n$ for which $\dim \mathcal{M} = n - 1$ is called a ***hyperplane.*** For example, a hyperplane in $\Re^2$ is a line (not necessarily through the origin), and a hyperplane in $\Re^3$ is a plane (not necessarily through the origin). The i^{th} equation $\mathbf{A}_{i*}\mathbf{x} = b_i$ in a linear system $\mathbf{A}_{m\times n}\mathbf{x} = \mathbf{b}$ is a hyperplane in $\Re^n$, so the solutions of $\mathbf{Ax} = \mathbf{b}$ occur at the intersection of the m hyperplanes defined by the rows of $\mathbf{A}$.

(a) Prove that for a given scalar β and a nonzero vector $\mathbf{u} \in \Re^n$, the set $\mathcal{H} = \{\mathbf{x} \,|\, \mathbf{u}^T\mathbf{x} = \beta\}$ is a hyperplane in $\Re^n$.

(b) Explain why the orthogonal projection of $\mathbf{b} \in \Re^n$ onto $\mathcal{H}$ is $\mathbf{p} = \mathbf{b} - \big(\mathbf{u}^T\mathbf{b} - \beta/\mathbf{u}^T\mathbf{u}\big)\,\mathbf{u}$.

5.13.18. For $\mathbf{u}, \mathbf{w} \in \Re^n$ such that $\mathbf{u}^T\mathbf{w} \neq 0$, let $\mathcal{M} = \mathbf{u}^{\perp}$ and $\mathcal{W} = span\,\{\mathbf{w}\}$.

(a) Explain why $\Re^n = \mathcal{M} \oplus \mathcal{W}$.

(b) For $\mathbf{b} \in \Re^{n\times 1}$, explain why the *oblique* projection of $\mathbf{b}$ onto $\mathcal{M}$ along $\mathcal{W}$ is given by $\mathbf{p} = \mathbf{b} - \mathbf{u}^T\mathbf{b}/\mathbf{u}^T\mathbf{w}\mathbf{w}$.

(c) For a given scalar β, let $\mathcal{H}$ be the hyperplane in $\Re^n$ defined by $\mathcal{H} = \{\mathbf{x} \,|\, \mathbf{u}^T\mathbf{x} = \beta\}$—see Exercise 5.13.17. Explain why the *oblique* projection of $\mathbf{b}$ onto $\mathcal{H}$ along $\mathcal{W}$ should be given by $\mathbf{p} = \mathbf{b} - \big(\mathbf{u}^T\mathbf{b} - \beta/\mathbf{u}^T\mathbf{w}\big)\,\mathbf{w}$.

5.13.19. Kaczmarz's[61] Projection Method. The solution of a nonsingular system

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

is the intersection of the two hyperplanes (lines in this case) defined by

$$\mathcal{H}_1 = \{(x_1, x_2) \,|\, a_{11}x_1 + a_{12}x_2 = b_1\},\ \mathcal{H}_2 = \{(x_1, x_2) \,|\, a_{21}x_1 + a_{22}x_2 = b_2\}.$$

It's visually evident that by starting with an arbitrary point $\mathbf{p}_0$ and alternately projecting orthogonally onto $\mathcal{H}_1$ and $\mathcal{H}_2$ as depicted in Figure 5.13.7, the resulting sequence of projections $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4, \ldots\}$ converges to $\mathcal{H}_1 \cap \mathcal{H}_2$, the solution of $\mathbf{Ax} = \mathbf{b}$.

[61] Although this idea has probably occurred to many people down through the ages, credit is usually given to Stefan Kaczmarz, who published his results in 1937. Kaczmarz was among a school of bright young Polish mathematicians who were beginning to flower in the first part of the twentieth century. Tragically, this group was decimated by Hitler's invasion of Poland, and Kaczmarz himself was killed in military action while trying to defend his country.

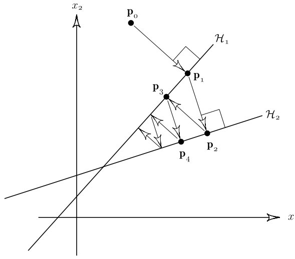


FIGURE 5.13.7

This idea can be generalized by using Exercise 5.13.17. For a consistent system $\mathbf{A}_{n\times r}\mathbf{x} = \mathbf{b}$ with $rank\,(\mathbf{A}) = r$, scale the rows so that $\|\mathbf{A}_{i*}\|_2 = 1$ for each i, and let $\mathcal{H}_i = \{\mathbf{x} \,|\, \mathbf{A}_{i*}\mathbf{x} = b_i\}$ be the hyperplane defined by the i^{th} equation. Begin with an arbitrary vector $\mathbf{p}_0 \in \Re^{r\times 1}$, and successively perform orthogonal projections onto each hyperplane to generate the following sequence:

$$\begin{aligned}
\mathbf{p}_1 &= \mathbf{p}_0 - (\mathbf{A}_{1*}\mathbf{p}_0 - b_1)\,(\mathbf{A}_{1*})^T & &\text{(project } \mathbf{p}_0 \text{ onto } \mathcal{H}_1\text{),}\\
\mathbf{p}_2 &= \mathbf{p}_1 - (\mathbf{A}_{2*}\mathbf{p}_1 - b_2)\,(\mathbf{A}_{2*})^T & &\text{(project } \mathbf{p}_1 \text{ onto } \mathcal{H}_2\text{),}\\
&\vdots & &\qquad\vdots\\
\mathbf{p}_n &= \mathbf{p}_{n-1} - (\mathbf{A}_{n*}\mathbf{p}_{n-1} - b_n)\,(\mathbf{A}_{n*})^T & &\text{(project } \mathbf{p}_{n-1} \text{ onto } \mathcal{H}_n\text{).}
\end{aligned}$$

When all n hyperplanes have been used, continue by repeating the process. For example, on the second pass project $\mathbf{p}_n$ onto $\mathcal{H}_1$; then project $\mathbf{p}_{n+1}$ onto $\mathcal{H}_2$, etc. For an arbitrary $\mathbf{p}_0$, the entire Kaczmarz sequence is generated by executing the following double loop:

```
For $k = 0, 1, 2, 3, \ldots$
    For $i = 1, 2, \ldots, n$
        $\mathbf{p}_{kn+i} = \mathbf{p}_{kn+i-1} - (\mathbf{A}_{i*}\mathbf{p}_{kn+i-1} - b_i)\,(\mathbf{A}_{i*})^T$
```

Prove that the Kaczmarz sequence converges to the solution of $\mathbf{Ax} = \mathbf{b}$ by showing $\|\mathbf{p}_{kn+i} - \mathbf{x}\|_2^2 = \|\mathbf{p}_{kn+i-1} - \mathbf{x}\|_2^2 - (\mathbf{A}_{i*}\mathbf{p}_{kn+i-1} - b_i)^2$.

5.13.20. Oblique Projection Method. Assume that a nonsingular system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$ has been row scaled so that $\|\mathbf{A}_{i*}\|_2 = 1$ for each i, and let $\mathcal{H}_i = \{\mathbf{x} \,|\, \mathbf{A}_{i*}\mathbf{x} = b_i\}$ be the hyperplane defined by the i^{th} equation—see Exercise 5.13.17. In theory, the system can be solved by making $n-1$ oblique projections of the type described in Exercise 5.13.18 because if an arbitrary point $\mathbf{p}_1$ in $\mathcal{H}_1$ is projected obliquely onto $\mathcal{H}_2$ along $\mathcal{H}_1$ to produce $\mathbf{p}_2$, then $\mathbf{p}_2$ is in $\mathcal{H}_1 \cap \mathcal{H}_2$. If $\mathbf{p}_2$ is projected onto $\mathcal{H}_3$ along $\mathcal{H}_1 \cap \mathcal{H}_2$ to produce $\mathbf{p}_3$, then $\mathbf{p}_3 \in \mathcal{H}_1 \cap \mathcal{H}_2 \cap \mathcal{H}_3$, and so forth until $\mathbf{p}_n \in \cap_{i=1}^{n}\mathcal{H}_i$. This is similar to Kaczmarz's method given in Exercise 5.13.19, but here we are projecting obliquely instead of orthogonally. However, projecting $\mathbf{p}_k$ onto $\mathcal{H}_{k+1}$ along $\cap_{i=1}^{k}\mathcal{H}_i$ is difficult because

$\cap_{i=1}^{k}\mathcal{H}_i$ is generally unknown. This problem is overcome by modifying the procedure as follows—use Figure 5.13.8 with $n=3$ as a guide.

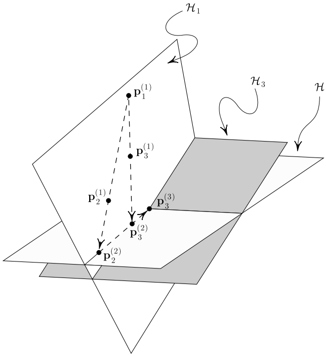


FIGURE 5.13.8

Step 0. Begin with any set $\{\mathbf{p}_1^{(1)},\mathbf{p}_2^{(1)},\ldots,\mathbf{p}_n^{(1)}\}\subset\mathcal{H}_1$ such that $\{(\mathbf{p}_1^{(1)}-\mathbf{p}_2^{(1)}),\,(\mathbf{p}_1^{(1)}-\mathbf{p}_3^{(1)}),\,\ldots,\,(\mathbf{p}_1^{(1)}-\mathbf{p}_n^{(1)})\}$ is linearly independent and $\mathbf{A}_{2*}(\mathbf{p}_1^{(1)}-\mathbf{p}_k^{(1)})\neq 0$ for $k=2,3,\ldots,n$.

Step 1. In turn, project $\mathbf{p}_1^{(1)}$ onto $\mathcal{H}_2$ through $\mathbf{p}_2^{(1)},\mathbf{p}_3^{(1)},\ldots,\mathbf{p}_n^{(1)}$ to produce $\{\mathbf{p}_2^{(2)},\,\mathbf{p}_3^{(2)},\,\ldots,\,\mathbf{p}_n^{(2)}\}\subset\mathcal{H}_1\cap\mathcal{H}_2$ (see Figure 5.13.8).

Step 2. Project $\mathbf{p}_2^{(2)}$ onto $\mathcal{H}_3$ through $\mathbf{p}_3^{(2)},\mathbf{p}_4^{(2)},\ldots,\mathbf{p}_n^{(2)}$ to produce $\{\mathbf{p}_3^{(3)},\,\mathbf{p}_4^{(3)},\,\ldots,\,\mathbf{p}_n^{(3)}\}\subset\mathcal{H}_1\cap\mathcal{H}_2\cap\mathcal{H}_3$. And so the process continues.

Step n−1. Project $\mathbf{p}_{n-1}^{(n-1)}$ through $\mathbf{p}_n^{(n-1)}$ to produce $\mathbf{p}_n^{(n)}\in\cap_{i=1}^{n}\mathcal{H}_i$. Of course, $\mathbf{x}=\mathbf{p}_n^{(n)}$ is the solution of the system.

For any initial set $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}\subset\mathcal{H}_1$ satisfying the properties described in Step 0, explain why the following algorithm performs the computations described in Steps $1,2,\ldots,n-1$.

```
For $i = 2$ to $n$
    For $j = i$ to $n$
        $\mathbf{x}_j \leftarrow \mathbf{x}_j - \dfrac{(\mathbf{A}_{i*}\mathbf{x}_{i-1} - b_i)}{\mathbf{A}_{i*}(\mathbf{x}_{i-1} - \mathbf{x}_j)}(\mathbf{x}_{i-1} - \mathbf{x}_j)$
$\mathbf{x} \leftarrow \mathbf{x}_n$    (the solution of the system)
```

5.13.21. Let $\mathcal{M}$ be a subspace of $\Re^n$, and let $\mathbf{R}=\mathbf{I}-2\mathbf{P}_{\mathcal{M}}$. Prove that the orthogonal distance between any point $\mathbf{x}\in\Re^n$ and $\mathcal{M}^\perp$ is the same as the orthogonal distance between $\mathbf{Rx}$ and $\mathcal{M}^\perp$. In other words, prove that $\mathbf{R}$ reflects everything in $\Re^n$ about $\mathcal{M}^\perp$. Naturally, $\mathbf{R}$ is called the ***reflector*** about $\mathcal{M}^\perp$. The *elementary* reflectors $\mathbf{I}-2\mathbf{u}\mathbf{u}^T/\mathbf{u}^T\mathbf{u}$ discussed on p. 324 are special cases—go back and look at Figure 5.6.2.

5.13.22. Cimmino's Reflection Method. In 1938 the Italian mathematician Gianfranco Cimmino used the following elementary observation to construct an iterative algorithm for solving linear systems. For a 2×2 system $\mathbf{Ax} = \mathbf{b}$, let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the two lines (hyperplanes) defined by the two equations. For an arbitrary guess $\mathbf{r}_0$, let $\mathbf{r}_1$ be the reflection of $\mathbf{r}_0$ about the line $\mathcal{H}_1$, and let $\mathbf{r}_2$ be the reflection of $\mathbf{r}_0$ about the line $\mathcal{H}_2$. As illustrated in Figure 5.13.9, the three points $\mathbf{r}_0$, $\mathbf{r}_1$, and $\mathbf{r}_2$ lie on a circle whose center is $\mathcal{H}_1 \cap \mathcal{H}_2$ (the solution of the system).

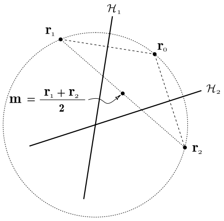


FIGURE 5.13.9

The mean value $\mathbf{m} = (\mathbf{r}_1 + \mathbf{r}_2)/2$ is strictly inside the circle, so $\mathbf{m}$ is a better approximation to the solution than $\mathbf{r}_0$. It's visually evident that iteration produces a sequence that converges to the solution of $\mathbf{Ax} = \mathbf{b}$. Prove this in general by using the following blueprint.

(a) For a scalar β and a vector $\mathbf{u} \in \Re^n$ such that $\|\mathbf{u}\|_2 = 1$, consider the hyperplane $\mathcal{H} = \{\mathbf{x} \mid \mathbf{u}^T\mathbf{x} = \beta\}$ (Exercise 5.13.17). Use (5.6.8) to show that the reflection of a vector $\mathbf{b}$ about $\mathcal{H}$ is $\mathbf{r} = \mathbf{b} - 2(\mathbf{u}^T\mathbf{b} - \beta)\mathbf{u}$.

(b) For a system $\mathbf{Ax} = \mathbf{b}$ in which the rows of $\mathbf{A} \in \Re^{n \times r}$ have been scaled so that $\|\mathbf{A}_{i*}\|_2 = 1$ for each i, let $\mathcal{H}_i = \{\mathbf{x} \mid \mathbf{A}_{i*}\mathbf{x} = b_i\}$ be the hyperplane defined by the i^{th} equation. If $\mathbf{r}_0 \in \Re^{r \times 1}$ is an arbitrary vector, and if $\mathbf{r}_i$ is the reflection of $\mathbf{r}_0$ about $\mathcal{H}_i$, explain why the mean value of the reflections $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ is $\mathbf{m} = \mathbf{r}_0 - (2/n)\mathbf{A}^T\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = \mathbf{Ar}_0 - \mathbf{b}$.

(c) Iterating part (b) produces $\mathbf{m}_k = \mathbf{m}_{k-1} - (2/n)\mathbf{A}^T\boldsymbol{\varepsilon}_{k-1}$, where $\boldsymbol{\varepsilon}_{k-1} = \mathbf{Am}_{k-1} - \mathbf{b}$. Show that if $\mathbf{A}$ is nonsingular, and if $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, then $\mathbf{x} - \mathbf{m}_k = \left(\mathbf{I} - (2/n)\mathbf{A}^T\mathbf{A}\right)^k (\mathbf{x} - \mathbf{m}_0)$. **Note:** It can be proven that $\left(\mathbf{I} - (2/n)\mathbf{A}^T\mathbf{A}\right)^k \to \mathbf{0}$ as $k \to \infty$, so $\mathbf{m}_k \to \mathbf{x}$ for all $\mathbf{m}_0$. In fact, $\mathbf{m}_k$ converges even if $\mathbf{A}$ is rank deficient—if consistent, it converges to a solution, and, if inconsistent, the limit is a least squares solution. Cimmino's method also works with weighted means. If $\mathbf{W} = \operatorname{diag}(w_1, w_2, \ldots, w_n)$, where $w_i > 0$ and $\sum w_i = 1$, then $\mathbf{m}_k = \mathbf{m}_{k-1} - \omega\mathbf{A}^T\mathbf{W}\boldsymbol{\varepsilon}_{k-1}$ is a convergent sequence in which $0 < \omega < 2$ is a "relaxation parameter" that can be adjusted to alter the rate of convergence.

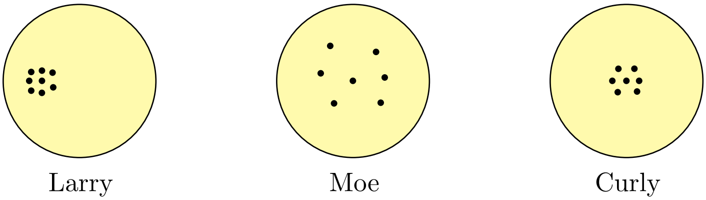
5.14 WHY LEAST SQUARES?

Drawing inferences about natural phenomena based upon physical observations and estimating characteristics of large populations by examining small samples are fundamental concerns of applied science. Numerical characteristics of a phenomenon or population are often called *parameters,* and the goal is to design functions or rules called *estimators* that use observations or samples to estimate parameters of interest. For example, the mean height h of all people is a parameter of the world's population, and one way of estimating h is to observe the mean height of a sample of k people. In other words, if h_i is the height of the i^{th} person in a sample, the function $\hat{h}$ defined by

$$\hat{h}(h_1, h_2, \ldots, h_k) = \frac{1}{k}\left(\sum_{i=1}^{k} h_i\right)$$

is an estimator for h. Moreover, $\hat{h}$ is a *linear estimator* because $\hat{h}$ is a linear function of the observations.

Good estimators should possess at least two properties—they should be *unbiased* and they should have *minimal variance.* For example, consider estimating the center of a circle drawn on a wall by asking Larry, Moe, and Curly to each throw one dart at the circle. To decide which estimator is best, we need to know more about each thrower's style. While being able to throw a tight pattern, it is known that Larry tends to have a left-hand bias in his style. Moe doesn't suffer from a bias, but he tends to throw a rather large pattern. However, Curly can throw a tight pattern without a bias. Typical patterns are shown below.



Although Larry has a small variance, he is an unacceptable estimator because he is biased in the sense that his average is significantly different than the center. Moe and Curly are each unbiased estimators because they have an average that is the center, but Curly is clearly the preferred estimator because his variance is much smaller than Moe's. In other words, Curly is the unbiased estimator of minimal variance.

To make these ideas more formal, let's adopt the following standard notation and terminology from elementary probability theory concerning random variables X and Y.

- $E[X] = \mu_X$ denotes the ***mean*** (or expected value) of X.
- $\text{Var}[X] = E\left[(X-\mu_X)^2\right] = E[X^2] - \mu_X^2$ is the ***variance*** of X.
- $\text{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$ is the ***covariance*** of X and Y.

Minimum Variance Unbiased Estimators

An estimator $\hat{\theta}$ (consider as a random variable) for a parameter θ is said to be ***unbiased*** when $E[\hat{\theta}] = \theta$, and $\hat{\theta}$ is called a ***minimum variance unbiased estimator*** for θ whenever $\text{Var}[\hat{\theta}] \leq \text{Var}[\hat{\phi}]$ for all unbiased estimators $\hat{\phi}$ of θ.

These ideas make it possible to precisely articulate why the method of least squares is the best way to fit observed data. Let Y be a variable that is known (or assumed) to be linearly related to other variables $X_1, X_2, \ldots, X_n$ according to the equation[62]

$$Y = \beta_1 X_1 + \cdots + \beta_n X_n, \qquad (5.14.1),$$

where the β_i's are unknown constants (parameters). Suppose that the values assumed by the X_i's are not subject to error or variation and can be exactly observed or specified, but, due perhaps to measurement error, the values of Y cannot be exactly observed. Instead, we observe

$$y = Y + \varepsilon = \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon, \qquad (5.14.2)$$

where ε is a random variable accounting for the measurement error. For example, consider the problem of determining the velocity v of a moving object by measuring the distance D it has traveled at various points in time T by using the linear relation $D = vT$. Time can be prescribed at exact values such as $T_1 = 1$ second, $T_2 = 2$ seconds, etc., but observing the distance traveled at the prescribed values of T will almost certainly involve small measurement errors so that in reality the observed distances satisfy $d = D + \varepsilon = vT + \varepsilon$. Now consider the general problem of determining the parameters β_k in (5.14.1) by observing (or measuring) values of Y at m different points $\mathbf{X}_{i*} = (x_{i1}, x_{i2}, \ldots, x_{in}) \in \Re^n$, where x_{ij} is the value of X_j to be used when making the i^{th} observation. If y_i denotes the random variable that represents the outcome of the i^{th} observation of Y, then according to (5.14.2),

$$y_i = \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \varepsilon_i, \quad i = 1, 2, \ldots, m, \qquad (5.14.3)$$

[62] Equation (5.14.1) is called a ***no-intercept model,*** whereas the slightly more general equation $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$ is known as an ***intercept model.*** Since the analysis for an intercept model is not significantly different from the analysis of the no-intercept case, we deal only with the no-intercept case and leave the intercept model for the reader to develop.

where ε_i is a random variable accounting for the i^{th} observation (or measurement) error.[63] It is generally valid to assume that observation errors are not correlated with each other but have a common variance (not necessarily known) and a zero mean. In other words, we assume that

$$E[\varepsilon_i] = 0 \text{ for each } i \quad \text{and} \quad \mathrm{Cov}[\varepsilon_i, \varepsilon_j] = \begin{cases} \sigma^2 & \text{when } i = j, \\ 0 & \text{when } i \neq j. \end{cases}$$

If $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}$, then the equations in (5.14.3) can be written as $\mathbf{y} = \mathbf{X}_{m\times n}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In practice, the points $\mathbf{X}_{i*}$ at which observations y_i are made can almost always be selected to insure that $rank\,(\mathbf{X}_{m\times n}) = n$, so the complete statement of the ***standard linear model*** is

$$\mathbf{y} = \mathbf{X}_{m\times n}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{such that} \quad \begin{cases} rank\,(\mathbf{X}) = n, \\ E[\boldsymbol{\varepsilon}] = \mathbf{0}, \\ \mathrm{Cov}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}, \end{cases} \tag{5.14.4}$$

where we have adopted the conventions

$$E[\boldsymbol{\varepsilon}] = \begin{pmatrix} E[\varepsilon_1] \\ E[\varepsilon_2] \\ \vdots \\ E[\varepsilon_m] \end{pmatrix} \text{ and } \mathrm{Cov}[\boldsymbol{\varepsilon}] = \begin{pmatrix} \mathrm{Cov}[\varepsilon_1,\varepsilon_1] & \mathrm{Cov}[\varepsilon_1,\varepsilon_2] & \cdots & \mathrm{Cov}[\varepsilon_1,\varepsilon_m] \\ \mathrm{Cov}[\varepsilon_2,\varepsilon_1] & \mathrm{Cov}[\varepsilon_2,\varepsilon_2] & \cdots & \mathrm{Cov}[\varepsilon_2,\varepsilon_m] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[\varepsilon_m,\varepsilon_1] & \mathrm{Cov}[\varepsilon_m,\varepsilon_2] & \cdots & \mathrm{Cov}[\varepsilon_m,\varepsilon_m] \end{pmatrix}.$$

The problem is to determine the best (minimum variance) linear (linear function of the y_i 's) unbiased estimators for the components of $\boldsymbol{\beta}$. Gauss realized in 1821 that this is precisely what the least squares solution provides.

Gauss–Markov Theorem

For the standard linear model (5.14.4), the minimum variance linear unbiased estimator for β_i is given by the i^{th} component $\hat{\beta}_i$ in the vector $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^\dagger\mathbf{y}$. In other words, the best linear unbiased estimator for $\boldsymbol{\beta}$ is the least squares solution of $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}$.

[63] In addition to observation and measurement errors, other errors such as modeling errors or those induced by imposing simplifying assumptions produce the same kind of equation—recall the discussion of ice cream on p. 228.

Proof. It is clear that $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{y}$ is a linear estimator of $\boldsymbol{\beta}$ because each component $\hat{\beta}_i = \sum_k [\mathbf{X}^\dagger]_{ik}\, y_k$ is a linear function of the observations. The fact that $\hat{\boldsymbol{\beta}}$ is unbiased follows by using the linear nature of expected value to write

$$E[\mathbf{y}] = E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = E[\mathbf{X}\boldsymbol{\beta}] + E[\boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta},$$

so that

$$E\big[\hat{\boldsymbol{\beta}}\big] = E\big[\mathbf{X}^\dagger \mathbf{y}\big] = \mathbf{X}^\dagger E[\mathbf{y}] = \mathbf{X}^\dagger \mathbf{X}\boldsymbol{\beta} = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

To argue that $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{y}$ has minimal variance among all linear unbiased estimators for $\boldsymbol{\beta}$, let $\boldsymbol{\beta}^*$ be an arbitrary linear unbiased estimator for $\boldsymbol{\beta}$. Linearity of $\boldsymbol{\beta}^*$ implies the existence of a matrix $\mathbf{L}_{n\times m}$ such that $\boldsymbol{\beta}^* = \mathbf{L}\mathbf{y}$, and unbiasedness insures $\boldsymbol{\beta} = E[\boldsymbol{\beta}^*] = E[\mathbf{L}\mathbf{y}] = \mathbf{L}E[\mathbf{y}] = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$. We want $\boldsymbol{\beta} = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$ to hold irrespective of the values of the components in $\boldsymbol{\beta}$, so it must be the case that $\mathbf{L}\mathbf{X} = \mathbf{I}_n$ (recall Exercise 3.5.5). For $i \neq j$ we have

$$0 = \mathrm{Cov}[\varepsilon_i, \varepsilon_j] = E[\varepsilon_i\varepsilon_j] - \mu_{\varepsilon_i}\mu_{\varepsilon_j} \quad \Longrightarrow \quad E[\varepsilon_i\varepsilon_j] = E[\varepsilon_i]E[\varepsilon_j] = 0,$$

so that

$$\mathrm{Cov}[y_i, y_j] = \begin{cases} E[(y_i - \mu_{y_i})^2] = E[\varepsilon_i^2] = \mathrm{Var}[\varepsilon_i] = \sigma^2 & \text{when } i = j, \\ E[(y_i - \mu_{y_i})(y_j - \mu_{y_j})] = E[\varepsilon_i\varepsilon_j] = 0 & \text{when } i \neq j. \end{cases} \tag{5.14.5}$$

This together with the fact that $\mathrm{Var}[aW + bZ] = a^2\mathrm{Var}[W] + b^2\mathrm{Var}[Z]$ whenever $\mathrm{Cov}[W, Z] = 0$ allows us to write

$$\mathrm{Var}[\beta_i^*] = \mathrm{Var}[\mathbf{L}_{i*}\mathbf{y}] = \mathrm{Var}\left[\sum_{k=1}^{m} l_{ik}y_k\right] = \sigma^2 \sum_{k=1}^{m} l_{ik}^2 = \sigma^2 \left\|\mathbf{L}_{i*}\right\|_2^2.$$

Since $\mathbf{L}\mathbf{X} = \mathbf{I}$, it follows that $\mathrm{Var}[\beta_i^*]$ is minimal if and only if $\mathbf{L}_{i*}$ is the minimum norm solution of the system $\mathbf{z}^T\mathbf{X} = \mathbf{e}_i^T$. We know from (5.12.17) that the (unique) minimum norm solution is given by $\mathbf{z}^T = \mathbf{e}_i^T\mathbf{X}^\dagger = \mathbf{X}_{i*}^\dagger$, so $\mathrm{Var}[\beta_i^*]$ is minimal if and only if $\mathbf{L}_{i*} = \mathbf{X}_{i*}^\dagger$. Since this holds for $i = 1, 2, \ldots, m$, it follows that $\mathbf{L} = \mathbf{X}^\dagger$. In other words, the components of $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{y}$ are the (unique) minimal variance linear unbiased estimators for the parameters in $\boldsymbol{\beta}$. ■

Exercises for section 5.14

5.14.1. For a matrix $\mathbf{Z}_{m\times n} = [z_{ij}]$, of random variables, $E[\mathbf{Z}]$ is defined to be the $m \times n$ matrix whose (i,j)-entry is $E[z_{ij}]$. Consider the standard linear model described in (5.14.4), and let $\hat{\mathbf{e}}$ denote the vector of random variables defined by $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ in which $\hat{\boldsymbol{\beta}} = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^\dagger\mathbf{y}$. Demonstrate that

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{m - n}$$

is an unbiased estimator for σ^2. **Hint:** $\mathbf{d}^T\mathbf{c} = \mathit{trace}(\mathbf{c}\mathbf{d}^T)$ for column vectors $\mathbf{c}$ and $\mathbf{d}$, and, by virtue of Exercise 5.9.13,

$$\mathit{trace}\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger\right) = m - \mathit{trace}\left(\mathbf{X}\mathbf{X}^\dagger\right) = m - \mathit{rank}\left(\mathbf{X}\mathbf{X}^\dagger\right) = m - n.$$

5.15 ANGLES BETWEEN SUBSPACES

Consider the problem of somehow gauging the separation between a pair of nontrivial but otherwise general subspaces $\mathcal{M}$ and $\mathcal{N}$ of $\Re^n$. Perhaps the first thing that comes to mind is to measure the angle between them. But defining the "angle" between subspaces in $\Re^n$ is not as straightforward as the visual geometry of $\Re^2$ or $\Re^3$ might suggest. There is just too much "wiggle room" in higher dimensions to make any one definition completely satisfying, and the "correct" definition usually varies with the specific application under consideration.

Before exploring general angles, recall what has already been said about some special cases beginning with the angle between a pair of one-dimensional subspaces. If $\mathcal{M}$ and $\mathcal{N}$ are spanned by vectors $\mathbf{u}$ and $\mathbf{v}$, respectively, and if $\|\mathbf{u}\| = 1 = \|\mathbf{v}\|$, then the angle between $\mathcal{M}$ and $\mathcal{N}$ is defined by the expression $\cos\theta = \mathbf{v}^T\mathbf{u}$ (p. 295). This idea was carried one step further on p. 389 to define the angle between two *complementary* subspaces, and an intuitive connection to norms of projectors was presented. These intuitive ideas are now made rigorous.

Minimal Angle

The ***minimal angle*** between nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$ is defined to be the number $0 \le \theta_{min} \le \pi/2$ for which

$$\cos\theta_{min} = \max_{\substack{\mathbf{u}\in\mathcal{M},\,\mathbf{v}\in\mathcal{N}\\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u}. \tag{5.15.1}$$

- If $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ are the orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively, then

$$\cos\theta_{min} = \|\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\|_2\,. \tag{5.15.2}$$

- If $\mathcal{M}$ and $\mathcal{N}$ are *complementary* subspaces, and if $\mathbf{P}_{\mathcal{MN}}$ is the *oblique* projector onto $\mathcal{M}$ along $\mathcal{N}$, then

$$\sin\theta_{min} = \frac{1}{\|\mathbf{P}_{\mathcal{MN}}\|_2}. \tag{5.15.3}$$

- $\mathcal{M}$ and $\mathcal{N}$ are complementary subspaces if and only if $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is invertible, and in this case

$$\sin\theta_{min} = \frac{1}{\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\|_2}. \tag{5.15.4}$$

Proof of (5.15.2). If $f : \mathcal{V} \to \Re$ is a function defined on a space $\mathcal{V}$ such that $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$ for all scalars $\alpha \ge 0$, then

$$\max_{\|\mathbf{x}\|=1} f(\mathbf{x}) = \max_{\|\mathbf{x}\|\le 1} f(\mathbf{x}) \quad \text{(see Exercise 5.15.8).} \tag{5.15.5}$$

This together with (5.2.9) and the fact that $\mathbf{P}_{\mathcal{M}}\mathbf{x} \in \mathcal{M}$ and $\mathbf{P}_{\mathcal{N}}\mathbf{y} \in \mathcal{N}$ means

$$\begin{aligned}\cos\theta_{min} &= \max_{\substack{\mathbf{u}\in\mathcal{M},\, \mathbf{v}\in\mathcal{N} \\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u} = \max_{\substack{\mathbf{u}\in\mathcal{M},\, \mathbf{v}\in\mathcal{N} \\ \|\mathbf{u}\|_2\le 1,\, \|\mathbf{v}\|_2\le 1}} \mathbf{v}^T\mathbf{u} \\ &= \max_{\|\mathbf{x}\|_2\le 1,\, \|\mathbf{y}\|_2\le 1} \mathbf{y}^T\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\mathbf{x} = \|\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\|_2 . \quad \blacksquare\end{aligned}$$

Proof of (5.15.3). Let $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ be orthogonal matrices in which the columns of $\mathbf{U}_1$ and $\mathbf{U}_2$ constitute orthonormal bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively, and $\mathbf{V}_1$ and $\mathbf{V}_2$ are orthonormal bases for $\mathcal{N}^{\perp}$ and $\mathcal{N}$, respectively, so that $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}$ and $\mathbf{V}_i^T\mathbf{V}_i = \mathbf{I}$ for $i = 1, 2,$ and

$$\mathbf{P}_{\mathcal{M}} = \mathbf{U}_1\mathbf{U}_1^T,\ \ \mathbf{I} - \mathbf{P}_{\mathcal{M}} = \mathbf{U}_2\mathbf{U}_2^T,\ \ \mathbf{P}_{\mathcal{N}} = \mathbf{V}_2\mathbf{V}_2^T,\ \ \mathbf{I} - \mathbf{P}_{\mathcal{N}} = \mathbf{V}_1\mathbf{V}_1^T.$$

As discussed on p. 407, there is a nonsingular matrix $\mathbf{C}$ such that

$$\mathbf{P}_{\mathcal{M}\mathcal{N}} = \mathbf{U}\begin{pmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T = \mathbf{U}_1\mathbf{C}\mathbf{V}_1^T. \tag{5.15.6}$$

Notice that $\mathbf{P}_{\mathcal{M}\mathcal{N}}^2 = \mathbf{P}_{\mathcal{M}\mathcal{N}}$ implies $\mathbf{C} = \mathbf{C}\mathbf{V}_1^T\mathbf{U}_1\mathbf{C}$, which in turn insures $\mathbf{C}^{-1} = \mathbf{V}_1^T\mathbf{U}_1$. Recall that $\|\mathbf{XAY}\|_2 = \|\mathbf{A}\|_2$ whenever $\mathbf{X}$ has orthonormal columns and $\mathbf{Y}$ has orthonormal rows (Exercise 5.6.9). Consequently,

$$\|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2 = \|\mathbf{C}\|_2 = \frac{1}{\displaystyle\min_{\|\mathbf{x}\|_2=1}\big\|\mathbf{C}^{-1}\mathbf{x}\big\|_2} = \frac{1}{\displaystyle\min_{\|\mathbf{x}\|_2=1}\big\|\mathbf{V}_1^T\mathbf{U}_1\mathbf{x}\big\|_2} \quad \text{(recall (5.2.6)).}$$

Combining this with (5.15.2) produces (5.15.3) by writing

$$\begin{aligned}\sin^2\theta_{min} &= 1 - \cos^2\theta_{min} = 1 - \|\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\|_2^2 = 1 - \big\|\mathbf{V}_2\mathbf{V}_2^T\mathbf{U}_1\mathbf{U}_1^T\big\|_2^2 \\ &= 1 - \big\|(\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_1\big\|_2^2 = 1 - \max_{\|\mathbf{x}\|_2=1}\big\|(\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_1\mathbf{x}\big\|_2^2 \\ &= 1 - \max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{U}_1^T(\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_1\mathbf{x} = 1 - \max_{\|\mathbf{x}\|_2=1}\Big(1 - \big\|\mathbf{V}_1^T\mathbf{U}_1\mathbf{x}\big\|_2^2\Big) \\ &= 1 - \Big(1 - \min_{\|\mathbf{x}\|_2=1}\big\|\mathbf{V}_1^T\mathbf{U}_1\mathbf{x}\big\|_2^2\Big) = \frac{1}{\|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2^2}. \quad \blacksquare\end{aligned}$$

Proof of (5.15.4). Observe that

$$\begin{aligned}\mathbf{U}^T(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})\mathbf{V} &= \begin{pmatrix}\mathbf{U}_1^T \\ \hline \mathbf{U}_2^T\end{pmatrix}(\mathbf{U}_1\mathbf{U}_1^T - \mathbf{V}_2\mathbf{V}_2^T)\big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big) \\ &= \begin{pmatrix}\mathbf{U}_1^T\mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & -\mathbf{U}_2^T\mathbf{V}_2\end{pmatrix},\end{aligned} \tag{5.15.7}$$

where $\mathbf{U}_1^T\mathbf{V}_1 = (\mathbf{C}^{-1})^T$ is nonsingular. To see that $\mathbf{U}_2^T\mathbf{V}_2$ is also nonsingular, suppose $\dim \mathcal{M} = r$ so that $\dim \mathcal{N} = n - r$ and $\mathbf{U}_2^T\mathbf{V}_2$ is $n - r \times n - r$. Use the formula for the rank of a product (4.5.1) to write

$$rank\left(\mathbf{U}_2^T\mathbf{V}_2\right) = rank\left(\mathbf{U}_2^T\right) - \dim N\left(\mathbf{U}_2^T\right) \cap R\left(\mathbf{V}_2\right) = n - r - \dim \mathcal{M} \cap \mathcal{N} = n - r.$$

It now follows from (5.15.7) that $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular, and

$$\mathbf{V}^T(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\mathbf{U} = \begin{pmatrix} (\mathbf{U}_1^T\mathbf{V}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & -(\mathbf{U}_2^T\mathbf{V}_2)^{-1} \end{pmatrix}.$$

(Showing that $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular implies $\mathcal{M} \oplus \mathcal{N} = \Re^n$ is Exercise 5.15.6.) Formula (5.2.12) on p. 283 for the 2-norm of a block-diagonal matrix can now be applied to yield

$$\left\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\right\|_2 = \max\left\{ \left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2, \left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2 \right\}. \tag{5.15.8}$$

But $\left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2 = \left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2$ because we can again use (5.2.6) to write

$$\begin{aligned} \frac{1}{\left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2^2} &= \min_{\|\mathbf{x}\|_2=1} \left\|\mathbf{U}_1^T\mathbf{V}_1\mathbf{x}\right\|_2^2 = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{V}_1^T\mathbf{U}_1\mathbf{U}_1^T\mathbf{V}_1\mathbf{x} \\ &= \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{V}_1^T(\mathbf{I} - \mathbf{U}_2\mathbf{U}_2^T)\mathbf{V}_1\mathbf{x} \\ &= \min_{\|\mathbf{x}\|_2=1} (1 - \mathbf{x}^T\mathbf{V}_1^T\mathbf{U}_2\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}) \\ &= 1 - \max_{\|\mathbf{x}\|_2=1} \left\|\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}\right\|_2^2 = 1 - \left\|\mathbf{U}_2^T\mathbf{V}_1\right\|_2^2. \end{aligned}$$

By a similar argument, $1/\left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2^2 = 1 - \left\|\mathbf{U}_2^T\mathbf{V}_1\right\|_2^2$ (Exercise 5.15.11(a)). Therefore,

$$\left\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\right\|_2 = \left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2 = \left\|\mathbf{C}^T\right\|_2 = \|\mathbf{C}\|_2 = \|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2. \quad \blacksquare$$

While the minimal angle works fine for complementary spaces, it may not convey much information about the separation between noncomplementary subspaces. For example, $\theta_{min} = 0$ whenever $\mathcal{M}$ and $\mathcal{N}$ have a nontrivial intersection, but there nevertheless might be a nontrivial "gap" between $\mathcal{M}$ and $\mathcal{N}$—look at Figure 5.15.1. Rather than thinking about angles to measure such a gap, consider orthogonal distances as discussed in (5.13.13). Define

$$\delta(\mathcal{M}, \mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M} \\ \|\mathbf{m}\|_2=1}} dist\,(\mathbf{m}, \mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M} \\ \|\mathbf{m}\|_2=1}} \|(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2$$

to be the ***directed distance*** from $\mathcal{M}$ to $\mathcal{N}$, and notice that $\delta(\mathcal{M},\mathcal{N}) \le 1$ because (5.2.5) and (5.13.10) can be combined to produce

$$dist\,(\mathbf{m},\mathcal{N}) = \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2 = \|\mathbf{P}_{\mathcal{N}^\perp}\mathbf{m}\|_2 \le \|\mathbf{P}_{\mathcal{N}^\perp}\|_2\,\|\mathbf{m}\|_2 = 1.$$

Figure 5.15.1 illustrates $\delta(\mathcal{M},\mathcal{N})$ for two planes in $\Re^3$.

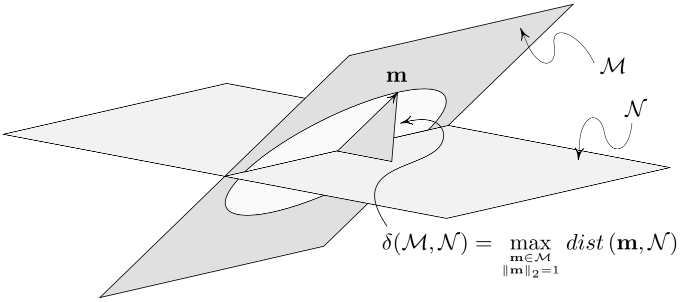

$\mathbf{m}$
$\mathcal{M}$
$\mathcal{N}$
$\delta(\mathcal{M},\mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} dist\,(\mathbf{m},\mathcal{N})$

FIGURE 5.15.1

This picture is a bit misleading because $\delta(\mathcal{M},\mathcal{N}) = \delta(\mathcal{N},\mathcal{M})$ for this particular situation. However, $\delta(\mathcal{M},\mathcal{N})$ and $\delta(\mathcal{N},\mathcal{M})$ need not always agree—that's why the phrase *directed* distance is used. For example, if $\mathcal{M}$ is the xy-plane in $\Re^3$ and $\mathcal{N} = span\,\{(0,1,1)\}$, then $\delta(\mathcal{N},\mathcal{M}) = 1/\sqrt{2}$ while $\delta(\mathcal{M},\mathcal{N}) = 1$. Consequently, using orthogonal distance to gauge the degree of maximal separation between an arbitrary pair of subspaces requires that both values of δ be taken into account. Hence we make the following definition.

Gap Between Subspaces

The ***gap*** between subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$ is defined to be

$$\text{gap}\,(\mathcal{M},\mathcal{N}) = \max\big\{\delta(\mathcal{M},\mathcal{N}),\ \delta(\mathcal{N},\mathcal{M})\big\}, \tag{5.15.9}$$

where $\delta(\mathcal{M},\mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} dist\,(\mathbf{m},\mathcal{N})$.

Evaluating the gap between a given pair of subspaces requires knowing some properties of directed distance. Observe that (5.15.5) together with the fact that $\|\mathbf{A}^T\|_2 = \|\mathbf{A}\|_2$ can be used to write

$$\begin{aligned}\delta(\mathcal{M},\mathcal{N}) &= \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} dist\,(\mathbf{m},\mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2\\ &= \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2\le 1}} \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\mathbf{x}\|_2 \qquad (5.15.10)\\ &= \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\|_2 = \|\mathbf{P}_{\mathcal{M}}(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\|_2\,.\end{aligned}$$

Similarly, $\delta(\mathcal{N},\mathcal{M}) = \|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2 = \|\mathbf{P}_{\mathcal{N}}(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\|_2$. If $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ are the orthogonal matrices introduced on p. 451, then

$$\delta(\mathcal{M},\mathcal{N}) = \|\mathbf{P}_{\mathcal{M}}(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\|_2 = \left\|\mathbf{U}_1\mathbf{U}_1^T\mathbf{V}_1\mathbf{V}_1^T\right\|_2 = \left\|\mathbf{U}_1^T\mathbf{V}_1\right\|_2$$

and (5.15.11)

$$\delta(\mathcal{N},\mathcal{M}) = \|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2 = \left\|\mathbf{U}_2\mathbf{U}_2^T\mathbf{V}_2\mathbf{V}_2^T\right\|_2 = \left\|\mathbf{U}_2^T\mathbf{V}_2\right\|_2.$$

Combining these observations with (5.15.7) leads us to conclude that

$$\begin{aligned} \|\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}\|_2 &= \max\left\{ \left\|\mathbf{U}_1^T\mathbf{V}_1\right\|_2,\ \left\|\mathbf{U}_2^T\mathbf{V}_2\right\|_2 \right\} \\ &= \max\left\{\delta(\mathcal{M},\mathcal{N}),\ \delta(\mathcal{N},\mathcal{M})\right\} \qquad (5.15.12) \\ &= \operatorname{gap}(\mathcal{M},\mathcal{N}). \end{aligned}$$

Below is a summary of these and other properties of the gap measure.

Gap Properties

The following statements are true for subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$.

- $\operatorname{gap}(\mathcal{M},\mathcal{N}) = \|\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}\|_2$.
- $\operatorname{gap}(\mathcal{M},\mathcal{N}) = \max\left\{ \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\|_2,\ \|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2 \right\}$.
- $\operatorname{gap}(\mathcal{M},\mathcal{N}) = 1$ whenever $\dim\mathcal{M} \neq \dim\mathcal{N}$. (5.15.13)
- If $\dim\mathcal{M} = \dim\mathcal{N}$, then $\delta(\mathcal{M},\mathcal{N}) = \delta(\mathcal{N},\mathcal{M})$, and
  - ▷ $\operatorname{gap}(\mathcal{M},\mathcal{N}) = 1$ when $\mathcal{M}^\perp \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^\perp$) $\neq \mathbf{0}$, (5.15.14)
  - ▷ $\operatorname{gap}(\mathcal{M},\mathcal{N}) < 1$ when $\mathcal{M}^\perp \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^\perp$) $= \mathbf{0}$. (5.15.15)

Proof of (5.15.13). Suppose that $\dim\mathcal{M} = r$ and $\dim\mathcal{N} = k$, where $r < k$. Notice that this implies that $\mathcal{M}^\perp \cap \mathcal{N} \neq \mathbf{0}$, for otherwise the formula for the dimension of a sum (4.4.19) yields

$$n \geq \dim(\mathcal{M}^\perp + \mathcal{N}) = \dim\mathcal{M}^\perp + \dim\mathcal{N} = n - r + k > n,$$

which is impossible. Thus there exists a nonzero vector $\mathbf{x} \in \mathcal{M}^\perp \cap \mathcal{N}$, and by normalization we can take $\|\mathbf{x}\|_2 = 1$. Consequently, $(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{x} = \mathbf{x} = \mathbf{P}_{\mathcal{N}}\mathbf{x}$, so $\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\mathbf{x}\|_2 = 1$. This insures that $\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2 = 1$, which implies $\delta(\mathcal{N},\mathcal{M}) = 1$. ■

Proof of (5.15.14). Assume $\dim\mathcal{M} = \dim\mathcal{N} = r$, and use the formula for the dimension of a sum along with $(\mathcal{M} \cap \mathcal{N}^\perp)^\perp = \mathcal{M}^\perp + \mathcal{N}$ (Exercise 5.11.5) to conclude that

$$\begin{aligned} \dim\left(\mathcal{M}^\perp \cap \mathcal{N}\right) &= \dim\mathcal{M}^\perp + \dim\mathcal{N} - \dim\left(\mathcal{M}^\perp + \mathcal{N}\right) \\ &= (n-r) + r - \dim\left(\mathcal{M} \cap \mathcal{N}^\perp\right)^\perp = \dim\left(\mathcal{M} \cap \mathcal{N}^\perp\right). \end{aligned}$$

When $\dim\left(\mathcal{M}\cap\mathcal{N}^{\perp}\right)=\dim\left(\mathcal{M}^{\perp}\cap\mathcal{N}\right)>0$, there are vectors $\mathbf{x}\in\mathcal{M}^{\perp}\cap\mathcal{N}$ and $\mathbf{y}\in\mathcal{M}\cap\mathcal{N}^{\perp}$ such that $\|\mathbf{x}\|_2=1=\|\mathbf{y}\|_2$. Hence, $\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\mathbf{x}\|_2=\|\mathbf{x}\|_2=1$, and $\|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\mathbf{y}\|_2=\|\mathbf{y}\|_2=1$, so

$$\delta(\mathcal{N},\mathcal{M})=\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2=1=\|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\|_2=\delta(\mathcal{M},\mathcal{N}).\quad\blacksquare$$

Proof of (5.15.15). If $\dim\left(\mathcal{M}\cap\mathcal{N}^{\perp}\right)=\dim\left(\mathcal{M}^{\perp}\cap\mathcal{N}\right)=0$, then $\mathbf{U}_2^T\mathbf{V}_1$ is nonsingular because it is $r\times r$ and has rank r—apply the formula (4.5.1) for the rank of a product. From (5.15.11) we have

$$\begin{aligned}\delta^2(\mathcal{M},\mathcal{N})&=\left\|\mathbf{U}_1^T\mathbf{V}_1\right\|_2^2=\left\|\mathbf{U}_1\mathbf{U}_1^T\mathbf{V}_1\right\|_2^2=\left\|(\mathbf{I}-\mathbf{U}_2\mathbf{U}_2^T)\mathbf{V}_1\right\|_2^2\\&=\max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{V}_1^T(\mathbf{I}-\mathbf{U}_2\mathbf{U}_2^T)\mathbf{V}_1\mathbf{x}=\max_{\|\mathbf{x}\|_2=1}\left(1-\left\|\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}\right\|_2^2\right)\\&=1-\min_{\|\mathbf{x}\|_2=1}\left\|\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}\right\|_2^2=1-\frac{1}{\left\|(\mathbf{U}_2^T\mathbf{V}_1)^{-1}\right\|_2^2}<1\text{ (recall (5.2.6)).}\end{aligned}$$

A similar argument shows $\delta^2(\mathcal{N},\mathcal{M})=\left\|\mathbf{U}_2^T\mathbf{V}_2\right\|_2^2=1-1/\left\|(\mathbf{U}_2^T\mathbf{V}_1)^{-1}\right\|_2^2$ (Exercise 5.15.11(b)), so $\delta(\mathcal{N},\mathcal{M})=\delta(\mathcal{M},\mathcal{N})<1$. $\blacksquare$

Because $0\le\operatorname{gap}(\mathcal{M},\mathcal{N})\le1$, the gap measure defines another angle between $\mathcal{M}$ and $\mathcal{N}$.

Maximal Angle

The ***maximal angle*** between subspaces $\mathcal{M},\mathcal{N}\subseteq\Re^n$ is defined to be the number $0\le\theta_{max}\le\pi/2$ for which

$$\sin\theta_{max}=\operatorname{gap}(\mathcal{M},\mathcal{N})=\|\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{N}}\|_2.\qquad(5.15.16)$$

For applications requiring knowledge of the degree of separation between a pair of nontrivial complementary subspaces, the minimal angle does the job. Similarly, the maximal angle adequately handles the task for subspaces of equal dimension. However, neither the minimal nor maximal angle may be of much help for more general subspaces. For example, if $\mathcal{M}$ and $\mathcal{N}$ are subspaces of unequal dimension that have a nontrivial intersection, then $\theta_{min}=0$ and $\theta_{max}=\pi/2$, but neither of these numbers might convey the desired information. Consequently, it seems natural to try to formulate definitions of “intermediate” angles between θ_{min} and θ_{max}. There are a host of such angles known as the ***principal*** or ***canonical angles***, and they are derived as follows.

Let $k = \min\{\dim \mathcal{M}, \dim \mathcal{N}\}$, and set $\mathcal{M}_1 = \mathcal{M}$, $\mathcal{N}_1 = \mathcal{N}$, and $\theta_1 = \theta_{min}$. Let $\mathbf{u}_1$ and $\mathbf{v}_1$ be vectors of unit 2-norm such that the following maximum is attained when $\mathbf{u} = \mathbf{u}_1$ and $\mathbf{v} = \mathbf{v}_1$:

$$\cos\theta_{min} = \max_{\substack{\mathbf{u}\in\mathcal{M},\, \mathbf{v}\in\mathcal{N} \\ \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1}} \mathbf{v}^T\mathbf{u} = \mathbf{v}_1^T\mathbf{u}_1.$$

Set

$$\mathcal{M}_2 = \mathbf{u}_1^{\perp} \cap \mathcal{M}_1 \quad \text{and} \quad \mathcal{N}_2 = \mathbf{v}_1^{\perp} \cap \mathcal{N}_1,$$

and define the second principal angle θ_2 to be the minimal angle between $\mathcal{M}_2$ and $\mathcal{N}_2$. Continue in this manner—e.g., if $\mathbf{u}_2$ and $\mathbf{v}_2$ are vectors such that $\|\mathbf{u}_2\|_2 = 1 = \|\mathbf{v}_2\|_2$ and

$$\cos\theta_2 = \max_{\substack{\mathbf{u}\in\mathcal{M}_2,\, \mathbf{v}\in\mathcal{N}_2 \\ \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1}} \mathbf{v}^T\mathbf{u} = \mathbf{v}_2^T\mathbf{u}_2,$$

set

$$\mathcal{M}_3 = \mathbf{u}_2^{\perp} \cap \mathcal{M}_2 \quad \text{and} \quad \mathcal{N}_3 = \mathbf{v}_2^{\perp} \cap \mathcal{N}_2,$$

and define the third principal angle θ_3 to be the minimal angle between $\mathcal{M}_3$ and $\mathcal{N}_3$. This process is repeated k times, at which point one of the subspaces is zero. Below is a summary.

Principal Angles

For nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$ with $k = \min\{\dim \mathcal{M}, \dim \mathcal{N}\}$, the principal angles between $\mathcal{M} = \mathcal{M}_1$ and $\mathcal{N} = \mathcal{N}_1$ are recursively defined to be the numbers $0 \le \theta_i \le \pi/2$ such that

$$\cos\theta_i = \max_{\substack{\mathbf{u}\in\mathcal{M}_i,\, \mathbf{v}\in\mathcal{N}_i \\ \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1}} \mathbf{v}^T\mathbf{u} = \mathbf{v}_i^T\mathbf{u}_i, \quad i = 1, 2, \ldots, k,$$

where $\|\mathbf{u}_i\|_2 = 1 = \|\mathbf{v}_i\|_2$, $\mathcal{M}_i = \mathbf{u}_{i-1}^{\perp} \cap \mathcal{M}_{i-1}$, and $\mathcal{N}_i = \mathbf{v}_{i-1}^{\perp} \cap \mathcal{N}_{i-1}$.

- It's possible to prove that $\theta_{min} = \theta_1 \le \theta_2 \le \cdots \le \theta_k \le \theta_{max}$, where $\theta_k = \theta_{max}$ when $\dim \mathcal{M} = \dim \mathcal{N}$.
- The vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are not uniquely defined, but the θ_i 's are unique. In fact, it can be proven that the $\sin\theta_i$'s are singular values (p. 412) for $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$. Furthermore, if $\dim \mathcal{M} \ge \dim \mathcal{N} = k$, then the $\cos\theta_i$'s are the singular values of $\mathbf{V}_2^T\mathbf{U}_1$, and the $\sin\theta_i$'s are the singular values of $\mathbf{V}_2^T\mathbf{U}_2\mathbf{U}_2^T$, where $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ are the orthogonal matrices from p. 451.

Exercises for section 5.15

5.15.1. Determine the angles θ_{min} and θ_{max} between the following subspaces of $\Re^3$.

(a) $\mathcal{M} = \text{xy-plane}, \quad \mathcal{N} = \mathit{span}\left\{(1,\,0,\,0),\ (0,\,1,\,1)\right\}.$

(b) $\mathcal{M} = \text{xy-plane}, \quad \mathcal{N} = \mathit{span}\left\{(0,\,1,\,1)\right\}.$

5.15.2. Determine the principal angles between the following subspaces of $\Re^3$.

(a) $\mathcal{M} = \text{xy-plane}, \quad \mathcal{N} = \mathit{span}\left\{(1,\,0,\,0),\ (0,\,1,\,1)\right\}.$

(b) $\mathcal{M} = \text{xy-plane}, \quad \mathcal{N} = \mathit{span}\left\{(0,\,1,\,1)\right\}.$

5.15.3. Let θ_{min} be the minimal angle between nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$.

(a) Explain why $\theta_{max} = 0$ if and only if $\mathcal{M} = \mathcal{N}$.

(b) Explain why $\theta_{min} = 0$ if and only if $\mathcal{M} \cap \mathcal{N} \neq \mathbf{0}$.

(c) Explain why $\theta_{min} = \pi/2$ if and only if $\mathcal{M} \perp \mathcal{N}$.

5.15.4. Let θ_{min} be the minimal angle between nonzero subspaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, and let $\theta^{\perp}_{min}$ denote the minimal angle between $\mathcal{M}^{\perp}$ and $\mathcal{N}^{\perp}$. Prove that if $\mathcal{M} \oplus \mathcal{N} = \Re^n$, then $\theta_{min} = \theta^{\perp}_{min}$.

5.15.5. For nonzero subspaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, let $\tilde{\theta}_{min}$ denote the minimal angle between $\mathcal{M}$ and $\mathcal{N}^{\perp}$, and let θ_{max} be the maximal angle between $\mathcal{M}$ and $\mathcal{N}$. Prove that if $\mathcal{M} \oplus \mathcal{N}^{\perp} = \Re^n$, then $\cos\tilde{\theta}_{min} = \sin\theta_{max}$.

5.15.6. For subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$, prove that $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular if and only if $\mathcal{M}$ and $\mathcal{N}$ are complementary.

5.15.7. For complementary spaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, let $\mathbf{P} = \mathbf{P}_{\mathcal{M}\mathcal{N}}$ be the oblique projector onto $\mathcal{M}$ along $\mathcal{N}$, and let $\mathbf{Q} = \mathbf{P}_{\mathcal{M}^{\perp}\mathcal{N}^{\perp}}$ be the oblique projector onto $\mathcal{M}^{\perp}$ along $\mathcal{N}^{\perp}$.

(a) Prove that $(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1} = \mathbf{P} - \mathbf{Q}$.

(b) If θ_{min} is the minimal angle between $\mathcal{M}$ and $\mathcal{N}$, explain why

$$\sin\theta_{min} = \frac{1}{\left\|\mathbf{P} - \mathbf{Q}\right\|_2}.$$

(c) Explain why $\left\|\mathbf{P} - \mathbf{Q}\right\|_2 = \left\|\mathbf{P}\right\|_2.$

5.15.8. Prove that if $f : \mathcal{V} \rightarrow \Re$ is a function defined on a space $\mathcal{V}$ such that $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$ for scalars $\alpha \geq 0$, then

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

5.15.9. Let $\mathcal{M}$ and $\mathcal{N}$ be nonzero complementary subspaces of $\Re^n$.

(a) Explain why $\mathbf{P}_{\mathcal{MN}} = \big[(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\big]^{\dagger}$, where $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ are the orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively, and $\mathbf{P}_{\mathcal{MN}}$ is the *oblique* projector onto $\mathcal{M}$ along $\mathcal{N}$.

(b) If θ_{min} is the minimal angle between $\mathcal{M}$ and $\mathcal{N}$, explain why

$$\begin{aligned}\sin\theta_{min} &= \Big\| \big[(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\big]^{\dagger} \Big\|_2^{-1} = \Big\| \big[\mathbf{P}_{\mathcal{M}}(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\big]^{\dagger} \Big\|_2^{-1} \\ &= \Big\| \big[(\mathbf{I} - \mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\big]^{\dagger} \Big\|_2^{-1} = \Big\| \big[\mathbf{P}_{\mathcal{N}}(\mathbf{I} - \mathbf{P}_{\mathcal{M}})\big]^{\dagger} \Big\|_2^{-1}.\end{aligned}$$

5.15.10. For complementary subspaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, let θ_{min} be the minimal angle between $\mathcal{M}$ and $\mathcal{N}$, and let $\bar{\theta}_{min}$ denote the minimal angle between $\mathcal{M}$ and $\mathcal{N}^{\perp}$.

(a) If $\mathbf{P}_{\mathcal{MN}}$ is the oblique projector onto $\mathcal{M}$ along $\mathcal{N}$, prove that

$$\cos\bar{\theta}_{min} = \Big\| \mathbf{P}_{\mathcal{MN}}^{\dagger} \Big\|_2.$$

(b) Explain why $\sin\theta_{min} \leq \cos\bar{\theta}_{min}$.

5.15.11. Let $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ be the orthogonal matrices defined on p. 451.

(a) Prove that if $\mathbf{U}_2^T\mathbf{V}_2$ is nonsingular, then

$$\frac{1}{\big\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\big\|_2^2} = 1 - \big\|\mathbf{U}_2^T\mathbf{V}_1\big\|_2^2.$$

(b) Prove that if $\mathbf{U}_2^T\mathbf{V}_1$ is nonsingular, then

$$\big\|\mathbf{U}_2^T\mathbf{V}_2\big\|_2^2 = 1 - \frac{1}{\big\|(\mathbf{U}_2^T\mathbf{V}_1)^{-1}\big\|_2^2}.$$