CHAPTER 6

Determinants

6.1 DETERMINANTS

At the beginning of this text, reference was made to the ancient Chinese counting board on which colored bamboo rods were manipulated according to prescribed "rules of thumb" in order to solve a system of linear equations. The Chinese counting board is believed to date back to at least 200 B.C., and it was used more or less in the same way for a millennium. The counting board and the "rules of thumb" eventually found their way to Japan where Seki Kowa (1642–1708), a great Japanese mathematician, synthesized the ancient Chinese ideas of array manipulation. Kowa formulated the concept of what we now call the determinant to facilitate solving linear systems—his definition is thought to have been made some time before 1683.

About the same time—somewhere between 1678 and 1693—Gottfried W. Leibniz (1646–1716), a German mathematician, was independently developing his own concept of the determinant together with applications of array manipulation to solve systems of linear equations. It appears that Leibniz's early work dealt with only three equations in three unknowns, whereas Seki Kowa gave a general treatment for n equations in n unknowns. It seems that Kowa and Liebniz both developed what later became known as Cramer's rule (p. 476), but not in the same form or notation. These men had something else in common—their ideas concerning the solution of linear systems were never adopted by the mathematical community of their time, and their discoveries quickly faded into oblivion.

Eventually the determinant was rediscovered, and much was written on the subject between 1750 and 1900. During this era, determinants became the major tool used to analyze and solve linear systems, while the theory of matrices remained relatively undeveloped. But mathematics, like a river, is everchanging

in its course, and major branches can dry up to become minor tributaries while small trickling brooks can develop into raging torrents. This is precisely what occurred with determinants and matrices. The study and use of determinants eventually gave way to Cayley's matrix algebra, and today matrix and linear algebra are in the main stream of applied mathematics, while the role of determinants has been relegated to a minor backwater position. Nevertheless, it is still important to understand what a determinant is and to learn a few of its fundamental properties. Our goal is not to study determinants for their own sake, but rather to explore those properties that are useful in the further development of matrix theory and its applications. Accordingly, many secondary properties are omitted or confined to the exercises, and the details in proofs will be kept to a minimum.

Over the years there have evolved various "slick" ways to define the determinant, but each of these "slick" approaches seems to require at least one "sticky" theorem in order to make the theory sound. We are going to opt for expedience over elegance and proceed with the classical treatment.

A ***permutation*** $p = (p_1, p_2, \ldots, p_n)$ of the numbers $(1, 2, \ldots, n)$ is simply any rearrangement. For example, the set

$$\{(1,2,3) \quad (1,3,2) \quad (2,1,3) \quad (2,3,1) \quad (3,1,2) \quad (3,2,1)\}$$

contains the six distinct permutations of $(1, 2, 3)$. In general, the sequence $(1, 2, \ldots, n)$ has $n! = n(n-1)(n-2)\cdots 1$ different permutations. Given a permutation, consider the problem of restoring it to natural order by a sequence of pairwise interchanges. For example, $(1, 4, 3, 2)$ can be restored to natural order with a single interchange of 2 and 4 or, as indicated in Figure 6.1.1, three *adjacent* interchanges can be used.

FIGURE 6.1.1

The important thing here is that both 1 and 3 are odd. Try to restore $(1, 4, 3, 2)$ to natural order by using an even number of interchanges, and you will discover that it is impossible. This is due to the following general rule that is stated without proof. *The* ***parity*** *of a permutation is unique*—i.e., if a permutation p can be restored to natural order by an even (odd) number of interchanges, then every other sequence of interchanges that restores p to natural order must

also be even (odd). Accordingly, the ***sign of a permutation*** p is defined to be the number

$$\sigma(p) = \begin{cases} +1 & \text{if } p \text{ can be restored to natural order by an } \textit{even} \text{ number of interchanges,} \\ -1 & \text{if } p \text{ can be restored to natural order by an } \textit{odd} \text{ number of interchanges.} \end{cases}$$

For example, if $p = (1,4,3,2)$, then $\sigma(p) = -1$, and if $p = (4,3,2,1)$, then $\sigma(p) = +1$. The sign of the natural order $p = (1,2,3,4)$ is naturally $\sigma(p) = +1$. The general definition of the determinant can now be given.

Definition of Determinant

For an $n \times n$ matrix $\mathbf{A} = [a_{ij}]$, the ***determinant*** of $\mathbf{A}$ is defined to be the scalar

$$\det(\mathbf{A}) = \sum_{p} \sigma(p) a_{1p_1} a_{2p_2} \cdots a_{np_n}, \tag{6.1.1}$$

where the sum is taken over the $n!$ permutations $p = (p_1, p_2, \ldots, p_n)$ of $(1,2,\ldots,n)$. Observe that each term $a_{1p_1}a_{2p_2}\cdots a_{np_n}$ in (6.1.1) contains exactly one entry from each row and each column of $\mathbf{A}$. The determinant of $\mathbf{A}$ can be denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$, whichever is more convenient.

Note: The determinant of a nonsquare matrix is not defined.

For example, when $\mathbf{A}$ is 2×2 there are $2! = 2$ permutations of (1,2), namely, $\{(1,2) \quad (2,1)\}$, so $\det(\mathbf{A})$ contains the two terms

$$\sigma(1,2)a_{11}a_{22} \quad \text{and} \quad \sigma(2,1)a_{12}a_{21}.$$

Since $\sigma(1,2) = +1$ and $\sigma(2,1) = -1$, we obtain the familiar formula

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{6.1.2}$$

Example 6.1.1

Problem: Use the definition to compute $\det(\mathbf{A})$, where $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$.

Solution: The $3! = 6$ permutations of $(1,2,3)$ together with the terms in the expansion of $\det(\mathbf{A})$ are shown in Table 6.1.1.

Table 6.1.1

$p=(p_1,p_2,p_3)$	$\sigma(p)$	$a_{1p_1}a_{2p_2}a_{3p_3}$
$(1,2,3)$	$+$	$1\times 5\times 9=45$
$(1,3,2)$	$-$	$1\times 6\times 8=48$
$(2,1,3)$	$-$	$2\times 4\times 9=72$
$(2,3,1)$	$+$	$2\times 6\times 7=84$
$(3,1,2)$	$+$	$3\times 4\times 8=96$
$(3,2,1)$	$-$	$3\times 5\times 7=105$

Therefore,

$$\det(\mathbf{A})=\sum_p \sigma(p)a_{1p_1}a_{2p_2}a_{3p_3}=45-48-72+84+96-105=0.$$

Perhaps you have seen rules for computing 3×3 determinants that involve running up, down, and around various diagonal lines. These rules do not easily generalize to matrices of order greater than three, and in case you have forgotten (or never knew) them, do not worry about it. Remember the 2×2 rule given in (6.1.2) as well as the following statement concerning triangular matrices and let it go at that.

Triangular Determinants

The determinant of a triangular matrix is the product of its diagonal entries. In other words,

$$\begin{vmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{vmatrix} = t_{11}t_{22}\cdots t_{nn}. \tag{6.1.3}$$

Proof. Recall from the definition (6.1.1) that each term $t_{1p_1}t_{2p_2}\cdots t_{np_n}$ contains exactly one entry from each row and each column. This means that there is only one term in the expansion of the determinant that does not contain an entry below the diagonal, and this term is $t_{11}t_{22}\cdots t_{nn}$. ∎

Transposition Doesn't Alter Determinants

- $\det\left(\mathbf{A}^T\right) = \det(\mathbf{A})$ for all $n \times n$ matrices. (6.1.4)

Proof. As $p = (p_1, p_2, \ldots, p_n)$ varies over all permutations of $(1, 2, \ldots, n)$, the set of all products $\{\sigma(p) a_{1p_1} a_{2p_2} \cdots a_{np_n}\}$ is the same as the set of all products $\{\sigma(p) a_{p_1 1} a_{p_2 2} \cdots a_{p_n n}\}$. Explicitly construct both of these sets for $n = 3$ to convince yourself. ∎

Equation (6.1.4) insures that it's not necessary to distinguish between rows and columns when discussing properties of determinants, so theorems concerning determinants that involve row manipulations will remain true when the word "row" is replaced by "column." For example, it's essential to know how elementary row and column operations alter the determinant of a matrix, but, by virtue of (6.1.4), it suffices to limit the discussion to elementary row operations.

Effects of Row Operations

Let $\mathbf{B}$ be the matrix obtained from $\mathbf{A}_{n\times n}$ by one of the three elementary row operations:

Type I: Interchange rows i and j.

Type II: Multiply row i by $\alpha \neq 0$.

Type III: Add α times row i to row j.

The value of $\det(\mathbf{B})$ is as follows:

- $\det(\mathbf{B}) = -\det(\mathbf{A})$ for Type I operations. (6.1.5)
- $\det(\mathbf{B}) = \alpha \det(\mathbf{A})$ for Type II operations. (6.1.6)
- $\det(\mathbf{B}) = \det(\mathbf{A})$ for Type III operations. (6.1.7)

Proof of (6.1.5). If $\mathbf{B}$ agrees with $\mathbf{A}$ except that $\mathbf{B}_{i*} = \mathbf{A}_{j*}$ and $\mathbf{B}_{j*} = \mathbf{A}_{i*}$, then for each permutation $p = (p_1, p_2, \ldots, p_n)$ of $(1, 2, \ldots, n)$,

$$\begin{aligned} b_{1p_1} \cdots b_{ip_i} \cdots b_{jp_j} \cdots b_{np_n} &= a_{1p_1} \cdots a_{jp_i} \cdots a_{ip_j} \cdots a_{np_n} \\ &= a_{1p_1} \cdots a_{ip_j} \cdots a_{jp_i} \cdots a_{np_n}. \end{aligned}$$

Furthermore, $\sigma(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) = -\sigma(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_n)$ because the two permutations differ only by one interchange. Consequently, definition (6.1.1) of the determinant guarantees that $\det(\mathbf{B}) = -\det(\mathbf{A})$.

Proof of (6.1.6). If $\mathbf{B}$ agrees with $\mathbf{A}$ except that $\mathbf{B}_{i*} = \alpha\mathbf{A}_{i*}$, then for each permutation $p = (p_1, p_2, \ldots, p_n)$,

$$b_{1p_1}\cdots b_{ip_i}\cdots b_{np_n} = a_{1p_1}\cdots \alpha a_{ip_i}\cdots a_{np_n} = \alpha(a_{1p_1}\cdots a_{ip_i}\cdots a_{np_n}),$$

and therefore the expansion (6.1.1) yields $\det(\mathbf{B}) = \alpha\det(\mathbf{A})$.

Proof of (6.1.7). If $\mathbf{B}$ agrees with $\mathbf{A}$ except that $\mathbf{B}_{j*} = \mathbf{A}_{j*} + \alpha\mathbf{A}_{i*}$, then for each permutation $p = (p_1, p_2, \ldots, p_n)$,

$$\begin{aligned} b_{1p_1}\cdots b_{ip_i}\cdots b_{jp_j}\cdots b_{np_n} &= a_{1p_1}\cdots a_{ip_i}\cdots (a_{jp_j} + \alpha a_{ip_j})\cdots a_{np_n} \\ &= a_{1p_1}\cdots a_{ip_i}\cdots a_{jp_j}\cdots a_{np_n} + \alpha(a_{1p_1}\cdots a_{ip_i}\cdots a_{ip_j}\cdots a_{np_n}), \end{aligned}$$

so that

$$\begin{aligned} \det(\mathbf{B}) = &\sum_p \sigma(p) a_{1p_1}\cdots a_{ip_i}\cdots a_{jp_j}\cdots a_{np_n} \\ &+\alpha\sum_p \sigma(p) a_{1p_1}\cdots a_{ip_i}\cdots a_{ip_j}\cdots a_{np_n}. \end{aligned} \tag{6.1.8}$$

The first sum on the right-hand side of (6.1.8) is $\det(\mathbf{A})$, while the second sum is the expansion of the determinant of a matrix $\tilde{\mathbf{A}}$ in which the i^{th} and j^{th} rows are identical. For such a matrix, $\det(\tilde{\mathbf{A}}) = 0$ because (6.1.5) says that the sign of the determinant is reversed whenever the i^{th} and j^{th} rows are interchanged, so $\det(\tilde{\mathbf{A}}) = -\det(\tilde{\mathbf{A}})$. Consequently, the second sum on the right-hand side of (6.1.8) is zero, and thus $\det(\mathbf{B}) = \det(\mathbf{A})$. ∎

It is now possible to evaluate the determinant of an elementary matrix associated with any of the three types of elementary operations. Let $\mathbf{E}$, $\mathbf{F}$, and $\mathbf{G}$ be elementary matrices of Types I, II, and III, respectively, and recall from the discussion in §3.9 that each of these elementary matrices can be obtained by performing the associated row (or column) operation to an identity matrix of appropriate size. The result concerning triangular determinants (6.1.3) guarantees that $\det(\mathbf{I}) = 1$ regardless of the size of $\mathbf{I}$, so if $\mathbf{E}$ is obtained by interchanging any two rows (or columns) in $\mathbf{I}$, then (6.1.5) insures that

$$\det(\mathbf{E}) = -\det(\mathbf{I}) = -1. \tag{6.1.9}$$

Similarly, if $\mathbf{F}$ is obtained by multiplying any row (or column) in $\mathbf{I}$ by $\alpha \neq 0$, then (6.1.6) implies that

$$\det(\mathbf{F}) = \alpha\det(\mathbf{I}) = \alpha, \tag{6.1.10}$$

and if $\mathbf{G}$ is the result of adding a multiple of one row (or column) in $\mathbf{I}$ to another row (or column) in $\mathbf{I}$, then (6.1.7) guarantees that

$$\det(\mathbf{G}) = \det(\mathbf{I}) = 1. \tag{6.1.11}$$

In particular, (6.1.9)–(6.1.11) guarantee that the determinants of elementary matrices of Types I, II, and III are nonzero.

As discussed in §3.9, if $\mathbf{P}$ is an elementary matrix of Type I, II, or III, and if $\mathbf{A}$ is any other matrix, then the product $\mathbf{PA}$ is the matrix obtained by performing the elementary operation associated with $\mathbf{P}$ to the rows of $\mathbf{A}$. This, together with the observations (6.1.5)–(6.1.7) and (6.1.9)–(6.1.11), leads to the conclusion that for every square matrix $\mathbf{A}$,

$$\begin{aligned}
\det(\mathbf{EA}) &= -\det(\mathbf{A}) &&= \det(\mathbf{E})\det(\mathbf{A}),\\
\det(\mathbf{FA}) &= \alpha\det(\mathbf{A}) &&= \det(\mathbf{F})\det(\mathbf{A}),\\
\det(\mathbf{GA}) &= \det(\mathbf{A}) &&= \det(\mathbf{G})\det(\mathbf{A}).
\end{aligned}$$

In other words, $\det(\mathbf{PA}) = \det(\mathbf{P})\det(\mathbf{A})$ whenever $\mathbf{P}$ is an elementary matrix of Type I, II, or III. It's easy to extend this observation to any number of these elementary matrices, $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k$, by writing

$$\begin{aligned}
\det(\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k\mathbf{A}) &= \det(\mathbf{P}_1)\det(\mathbf{P}_2\cdots\mathbf{P}_k\mathbf{A})\\
&= \det(\mathbf{P}_1)\det(\mathbf{P}_2)\det(\mathbf{P}_3\cdots\mathbf{P}_k\mathbf{A})\\
&\;\;\vdots\\
&= \det(\mathbf{P}_1)\det(\mathbf{P}_2)\cdots\det(\mathbf{P}_k)\det(\mathbf{A}).
\end{aligned} \tag{6.1.12}$$

This leads to a characterization of invertibility in terms of determinants.

Invertibility and Determinants

- $\mathbf{A}_{n\times n}$ is nonsingular if and only if $\det(\mathbf{A}) \neq 0$ (6.1.13)

  or, equivalently,
- $\mathbf{A}_{n\times n}$ is singular if and only if $\det(\mathbf{A}) = 0$. (6.1.14)

Proof. Let $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k$ be a sequence of elementary matrices of Type I, II, or III such that $\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k\mathbf{A} = \mathbf{E}_\mathbf{A}$, and apply (6.1.12) to conclude

$$\det(\mathbf{P}_1)\det(\mathbf{P}_2)\cdots\det(\mathbf{P}_k)\det(\mathbf{A}) = \det(\mathbf{E}_\mathbf{A}).$$

Since elementary matrices have nonzero determinants,

$$\begin{aligned}
\det(\mathbf{A}) \neq 0 &\iff \det(\mathbf{E}_\mathbf{A}) \neq 0 \iff \text{there are no zero pivots}\\
&\iff \text{every column in } \mathbf{E}_\mathbf{A} \text{ (and in } \mathbf{A}) \text{ is basic}\\
&\iff \mathbf{A} \text{ is nonsingular.} \;\blacksquare
\end{aligned}$$

Example 6.1.2

Caution! Small Determinants $\not\Longleftrightarrow$ Near Singularity. Because of (6.1.13) and (6.1.14), it might be easy to get the idea that $\det(\mathbf{A})$ is somehow a measure of how close $\mathbf{A}$ is to being singular, but this is not necessarily the case. Nearly singular matrices need not have determinants of small magnitude. For example, $\mathbf{A}_n = \begin{pmatrix} n & 0 \\ 0 & 1/n \end{pmatrix}$ is nearly singular when n is large, but $\det(\mathbf{A}_n) = 1$ for all n. Furthermore, small determinants do not necessarily signal nearly singular matrices. For example,

$$\mathbf{A}_n = \begin{pmatrix} .1 & 0 & \cdots & 0 \\ 0 & .1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & .1 \end{pmatrix}_{n\times n}$$

is not close to any singular matrix—see (5.12.10) on p. 417—but $\det(\mathbf{A}_n) = (.1)^n$ is extremely small for large n.

A ***minor determinant*** (or simply a ***minor***) of $\mathbf{A}_{m\times n}$ is defined to be the determinant of any $k \times k$ submatrix of $\mathbf{A}$. For example,

$$\begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} = -3 \text{ and } \begin{vmatrix} 2 & 3 \\ 8 & 9 \end{vmatrix} = -6 \text{ are } 2\times 2 \text{ minors of } \mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

An individual entry of $\mathbf{A}$ can be regarded as a 1×1 minor, and $\det(\mathbf{A})$ itself is considered to be a 3×3 minor of $\mathbf{A}$.

We already know that the rank of any matrix $\mathbf{A}$ is the size of the largest nonsingular submatrix in $\mathbf{A}$ (p. 215). But (6.1.13) guarantees that the nonsingular submatrices of $\mathbf{A}$ are simply those submatrices with nonzero determinants, so we have the following characterization of rank.

Rank and Determinants

- $rank(\mathbf{A})$ = the size of the largest nonzero minor of $\mathbf{A}$.

Example 6.1.3

Problem: Use determinants to compute the rank of $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 1 \\ 4 & 5 & 6 & 1 \\ 7 & 8 & 9 & 1 \end{pmatrix}$.

Solution: Clearly, there are 1×1 and 2×2 minors that are nonzero, so $rank(\mathbf{A}) \geq 2$. In order to decide if the rank is three, we must see if there

are any 3×3 nonzero minors. There are exactly four 3×3 minors, and they are

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = 0, \quad \begin{vmatrix} 1 & 2 & 1 \\ 4 & 5 & 1 \\ 7 & 8 & 1 \end{vmatrix} = 0, \quad \begin{vmatrix} 1 & 3 & 1 \\ 4 & 6 & 1 \\ 7 & 9 & 1 \end{vmatrix} = 0, \quad \begin{vmatrix} 2 & 3 & 1 \\ 5 & 6 & 1 \\ 8 & 9 & 1 \end{vmatrix} = 0.$$

Since all 3×3 minors are 0, we conclude that $rank\,(\mathbf{A}) = 2$. You should be able to see from this example that using determinants is generally not a good way to compute the rank of a matrix.

In (6.1.12) we observed that the determinant of a product of elementary matrices is the product of their respective determinants. We are now in a position to extend this observation.

Product Rules

- $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ for all $n \times n$ matrices. (6.1.15)
- $\det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} = \det(\mathbf{A})\det(\mathbf{D})$ if $\mathbf{A}$ and $\mathbf{D}$ are square. (6.1.16)

Proof of (6.1.15). If $\mathbf{A}$ is singular, then $\mathbf{AB}$ is also singular because (4.5.2) says that $rank\,(\mathbf{AB}) \leq rank\,(\mathbf{A})$. Consequently, (6.1.14) implies that

$$\det(\mathbf{AB}) = 0 = \det(\mathbf{A})\det(\mathbf{B}),$$

so (6.1.15) is trivially true when $\mathbf{A}$ is singular. If $\mathbf{A}$ is nonsingular, then $\mathbf{A}$ can be written as a product of elementary matrices $\mathbf{A} = \mathbf{P}_1\mathbf{P}_2 \cdots \mathbf{P}_k$ that are of Type I, II, or III—recall (3.9.3). Therefore, (6.1.12) can be applied to produce

$$\begin{aligned} \det(\mathbf{AB}) &= \det(\mathbf{P}_1\mathbf{P}_2 \cdots \mathbf{P}_k\mathbf{B}) = \det(\mathbf{P}_1)\det(\mathbf{P}_2) \cdots \det(\mathbf{P}_k)\det(\mathbf{B}) \\ &= \det(\mathbf{P}_1\mathbf{P}_2 \cdots \mathbf{P}_k)\det(\mathbf{B}) = \det(\mathbf{A})\det(\mathbf{B}). \end{aligned}$$

Proof of (6.1.16). First consider the special case $\mathbf{X} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, and use the definition to write $\det(\mathbf{X}) = \sum_{\sigma(p)} x_{1j_1}x_{2j_2} \cdots x_{rj_r}x_{r+1,j_{r+1}} \cdots x_{n,j_n}$. But

$$x_{rj_r}x_{r+1,j_{r+1}} \cdots x_{n,j_n} = \begin{cases} 1 & \text{when } p = \begin{pmatrix} 1 & \cdots & r & r+1 & \cdots & n \\ j_1 & \cdots & j_r & r+1 & \cdots & n \end{pmatrix}, \\ 0 & \text{for all other permutations,} \end{cases}$$

so, if p_r denotes permutations of only the first r positive integers, then

$$\det(\mathbf{X}) = \sum_{\sigma(p)} x_{1j_1}x_{2j_2} \cdots x_{rj_r}x_{r+1,j_{r+1}} \cdots x_{n,j_n} = \sum_{\sigma(p_r)} x_{1j_1}x_{2j_2} \cdots x_{rj_r} = \det(\mathbf{A}).$$

Thus $\begin{vmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{vmatrix} = \det(\mathbf{A})$. Similarly, $\begin{vmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} = \det(\mathbf{D})$, so, by (6.1.15),

$$\begin{vmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} = \det\left\{ \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \right\} = \begin{vmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{vmatrix} \begin{vmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} = \det(\mathbf{A})\det(\mathbf{D}).$$

If $\mathbf{A} = \mathbf{Q}_A\mathbf{R}_A$ and $\mathbf{D} = \mathbf{Q}_D\mathbf{R}_D$ are the respective QR factorizations (p. 345) of $\mathbf{A}$ and $\mathbf{D}$, then $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_D \end{pmatrix} \begin{pmatrix} \mathbf{R}_A & \mathbf{Q}_A^T\mathbf{B} \\ \mathbf{0} & \mathbf{R}_D \end{pmatrix}$ is also a QR factorization. By (6.1.3), the determinant of a triangular matrix is the product of its diagonal entries, and this together with the previous results yield

$$\begin{aligned} \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} &= \begin{vmatrix} \mathbf{Q}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_D \end{vmatrix} \begin{vmatrix} \mathbf{R}_A & \mathbf{Q}_A^T\mathbf{B} \\ \mathbf{0} & \mathbf{R}_D \end{vmatrix} = \det(\mathbf{Q}_A)\det(\mathbf{Q}_D)\det(\mathbf{R}_A)\det(\mathbf{R}_D) \\ &= \det(\mathbf{Q}_A\mathbf{R}_A)\det(\mathbf{Q}_D\mathbf{R}_D) = \det(\mathbf{A})\det(\mathbf{D}). \quad \blacksquare \end{aligned}$$

Example 6.1.4

Volume and Determinants. The definition of a determinant is purely algebraic, but there is a concrete geometrical interpretation. 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 depicted in Figure 6.1.2, a two-dimensional parallelepiped is a parallelogram, and a three-dimensional parallelepiped is a skewed rectangular box.

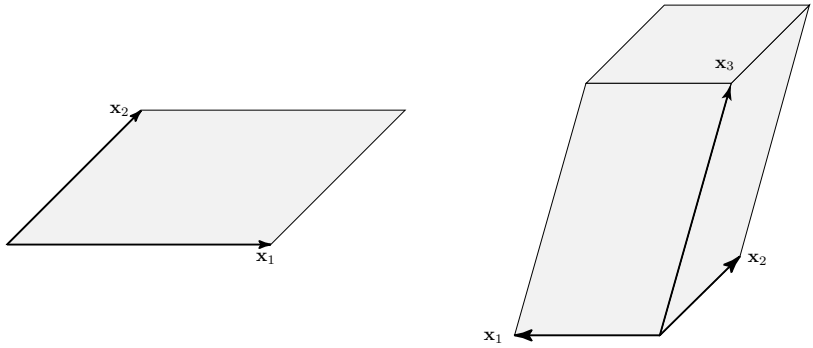


FIGURE 6.1.2

Problem: When $\mathbf{A} \in \Re^{m\times n}$ has linearly independent columns, explain why the volume of the n-dimensional parallelepiped generated by the columns of $\mathbf{A}$ is $V_n = \left[\det\left(\mathbf{A}^T\mathbf{A}\right)\right]^{1/2}$. In particular, if $\mathbf{A}$ is square, then $V_n = |\det(\mathbf{A})|$.

Solution: Recall from Example 5.13.2 on p. 431 that if $\mathbf{A}_{m\times n} = \mathbf{Q}_{m\times n}\mathbf{R}_{n\times n}$ is the (rectangular) QR factorization of $\mathbf{A}$, then the volume of the n-dimensional parallelepiped generated by the columns of $\mathbf{A}$ is $V_n = \nu_1\nu_2\cdots\nu_n = \det(\mathbf{R})$, where the ν_k's are the diagonal elements of the upper-triangular matrix $\mathbf{R}$. Use

$\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ together with the product rule (6.1.15) and the fact that transposition doesn't affect determinants (6.1.4) to write

$$\begin{aligned}\det\left(\mathbf{A}^T\mathbf{A}\right) &= \det\left(\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R}\right) = \det\left(\mathbf{R}^T\mathbf{R}\right) = \det\left(\mathbf{R}^T\right)\det\left(\mathbf{R}\right)\\ &= \left(\det\left(\mathbf{R}\right)\right)^2 = \left(\nu_1\nu_2\cdots\nu_n\right)^2 = V_n^2.\end{aligned} \tag{6.1.17}$$

If $\mathbf{A}$ is square, $\det\left(\mathbf{A}^T\mathbf{A}\right) = \det\left(\mathbf{A}^T\right)\det\left(\mathbf{A}\right) = \left(\det\left(\mathbf{A}\right)\right)^2$, so $V_n = \left|\det\left(\mathbf{A}\right)\right|$.

Hadamard's Inequality: Recall from (5.13.7) that if

$$\mathbf{A} = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n\right]_{n\times n} \quad \text{and} \quad \mathbf{A}_j = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_j\right]_{n\times j},$$

then $\nu_1 = \|\mathbf{x}_1\|_2$ and $\nu_k = \|(\mathbf{I} - \mathbf{P}_k)\mathbf{x}_k\|_2$ (the projected height of $\mathbf{x}_k$) for $k > 1$, where $\mathbf{P}_k$ is the orthogonal projector onto $R\left(\mathbf{A}_{k-1}\right)$. But

$$\nu_k^2 = \|(\mathbf{I} - \mathbf{P}_k)\mathbf{x}_k\|_2^2 \le \|(\mathbf{I} - \mathbf{P}_k)\|_2^2\,\|\mathbf{x}_k\|_2^2 = \|\mathbf{x}_k\|_2^2 \quad \text{(recall (5.13.10))},$$

so, by (6.1.17), $\det\left(\mathbf{A}^T\mathbf{A}\right) \le \|\mathbf{x}_1\|_2^2\,\|\mathbf{x}_2\|_2^2\cdots\|\mathbf{x}_n\|_2^2$ or, equivalently,

$$\left|\det\left(\mathbf{A}\right)\right| \le \prod_{k=1}^{n}\|\mathbf{x}_k\|_2 = \prod_{j=1}^{n}\left(\sum_{i=1}^{n}|a_{ij}|^2\right)^{1/2}, \tag{6.1.18}$$

with equality holding if and only if the $\mathbf{x}_k$'s are mutually orthogonal. This is ***Hadamard's inequality***.[64] In light of the preceding discussion, it simply asserts that the volume of the parallelepiped $\mathcal{P}$ generated by the columns of $\mathbf{A}$ can't exceed the volume of a rectangular box whose sides have length $\|\mathbf{x}_k\|_2$, a fact that is geometrically evident because $\mathcal{P}$ is a *skewed* rectangular box with sides of length $\|\mathbf{x}_k\|_2$.

The product rule (6.1.15) provides a practical way to compute determinants. Recall from §3.10 that for every nonsingular matrix $\mathbf{A}$, there is a permutation matrix $\mathbf{P}$ (which is a product of elementary interchange matrices) such that $\mathbf{PA} = \mathbf{LU}$ in which $\mathbf{L}$ is lower triangular with 1's on its diagonal, and $\mathbf{U}$ is upper triangular with the pivots on its diagonal. The product rule guarantees

[64] Jacques Hadamard (1865–1963), a leading French mathematician of the first half of the twentieth century, discovered this inequality in 1893. Influenced in part by the tragic death of his sons in World War I, Hadamard became a peace activist whose politics drifted far left to the extent that the United States was reluctant to allow him to enter the country to attend the International Congress of Mathematicians held in Cambridge, Massachusetts, in 1950. Due to support from influential mathematicians, Hadamard was made honorary president of the congress, and the resulting visibility together with pressure from important U.S. scientists forced officials to allow him to attend.

that $\det(\mathbf{P})\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U})$, and we know from (6.1.9) that if $\mathbf{E}$ is an elementary interchange matrix, then $\det(\mathbf{E}) = -1$, so

$$\det(\mathbf{P}) = \begin{cases} +1 & \text{if } \mathbf{P} \text{ is the product of an } \textit{even} \text{ number of interchanges,} \\ -1 & \text{if } \mathbf{P} \text{ is the product of an } \textit{odd} \text{ number of interchanges.} \end{cases}$$

The result concerning triangular determinants (6.1.3) shows that $\det(\mathbf{L}) = 1$ and $\det(\mathbf{U}) = u_{11}u_{22}\cdots u_{nn}$, where the u_{ii}'s are the pivots, so, putting these observations together yields $\det(\mathbf{A}) = \pm u_{11}u_{22}\cdots u_{nn}$, where the sign depends on the number of row interchanges used. Below is a summary.

Computing a Determinant

If $\mathbf{PA}_{n\times n} = \mathbf{LU}$ is an LU factorization obtained with row interchanges (use partial pivoting for numerical stability), then

$$\det(\mathbf{A}) = \sigma u_{11}u_{22}\cdots u_{nn}.$$

The u_{ii}'s are the pivots, and σ is the sign of the permutation. That is,

$$\sigma = \begin{cases} +1 & \text{if an } \textit{even} \text{ number of row interchanges are used,} \\ -1 & \text{if an } \textit{odd} \text{ number of row interchanges are used.} \end{cases}$$

If a zero pivot emerges that cannot be removed (because all entries below the pivot are zero), then $\mathbf{A}$ is singular and $\det(\mathbf{A}) = 0$. Exercise 6.2.18 discusses orthogonal reduction to compute $\det(\mathbf{A})$.

Example 6.1.5

Problem: Use partial pivoting to determine an LU decomposition $\mathbf{PA} = \mathbf{LU}$, and then evaluate the determinant of $\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}$.

Solution: The LU factors of $\mathbf{A}$ were computed in Example 3.10.4 as follows.

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The only modification needed is to keep track of how many row interchanges are used. Reviewing Example 3.10.4 reveals that the pivoting process required three interchanges, so $\sigma = -1$, and hence $\det(\mathbf{A}) = (-1)(4)(5)(-6)(1) = 120$.

It's sometimes necessary to compute the derivative of a determinant whose entries are differentiable functions. The following formula shows how this is done.

Derivative of a Determinant

If the entries in $\mathbf{A}_{n\times n} = [a_{ij}(t)]$ are differentiable functions of t, then

$$\frac{d\left(\det\left(\mathbf{A}\right)\right)}{dt} = \det\left(\mathbf{D}_1\right) + \det\left(\mathbf{D}_2\right) + \cdots + \det\left(\mathbf{D}_n\right), \tag{6.1.19}$$

where $\mathbf{D}_i$ is identical to $\mathbf{A}$ except that the entries in the i^{th} row are replaced by their derivatives—i.e., $\left[\mathbf{D}_i\right]_{k*} = \begin{cases} \mathbf{A}_{k*} & \text{if } i \neq k, \\ d\,\mathbf{A}_{k*}/dt & \text{if } i = k. \end{cases}$

Proof. This follows directly from the definition of a determinant by writing

$$\begin{aligned}
\frac{d\big(\det\left(\mathbf{A}\right)\big)}{dt} &= \frac{d}{dt}\sum_p \sigma(p) a_{1p_1}a_{2p_2}\cdots a_{np_n} = \sum_p \sigma(p)\frac{d\big(a_{1p_1}a_{2p_2}\cdots a_{np_n}\big)}{dt} \\
&= \sum_p \sigma(p)\Big(a'_{1p_1}a_{2p_2}\cdots a_{np_n} + a_{1p_1}a'_{2p_2}\cdots a_{np_n} + \cdots + a_{1p_1}a_{2p_2}\cdots a'_{np_n}\Big) \\
&= \sum_p \sigma(p)a'_{1p_1}a_{2p_2}\cdots a_{np_n} + \sum_p \sigma(p)a_{1p_1}a'_{2p_2}\cdots a_{np_n} \\
&\qquad\qquad + \cdots + \sum_p \sigma(p)a_{1p_1}a_{2p_2}\cdots a'_{np_n} \\
&= \det\left(\mathbf{D}_1\right) + \det\left(\mathbf{D}_2\right) + \cdots + \det\left(\mathbf{D}_n\right). \quad \blacksquare
\end{aligned}$$

Example 6.1.6

Problem: Evaluate the derivative $d\left(\det\left(\mathbf{A}\right)\right)/dt$ for $\mathbf{A} = \begin{pmatrix} \mathrm{e}^{t} & \mathrm{e}^{-t} \\ \cos t & \sin t \end{pmatrix}$.

Solution: Applying formula (6.1.19) yields

$$\frac{d\left(\det\left(\mathbf{A}\right)\right)}{dt} = \begin{vmatrix} \mathrm{e}^{t} & -\mathrm{e}^{-t} \\ \cos t & \sin t \end{vmatrix} + \begin{vmatrix} \mathrm{e}^{t} & \mathrm{e}^{-t} \\ -\sin t & \cos t \end{vmatrix} = \left(\mathrm{e}^{t} + \mathrm{e}^{-t}\right)\left(\cos t + \sin t\right).$$

Check this by first expanding $\det\left(\mathbf{A}\right)$ and then computing the derivative.

Exercises for section 6.1

6.1.1. Use the definition to evaluate $\det(\mathbf{A})$ for each of the following matrices.

(a) $\mathbf{A} = \begin{pmatrix} 3 & -2 & 1 \\ -5 & 4 & 0 \\ 2 & 1 & 6 \end{pmatrix}$. (b) $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 6 & 2 & 1 \\ -2 & 2 & 1 \end{pmatrix}$.

(c) $\mathbf{A} = \begin{pmatrix} 0 & 0 & \alpha \\ 0 & \beta & 0 \\ \gamma & 0 & 0 \end{pmatrix}$. (d) $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$.

6.1.2. What is the volume of the parallelepiped generated by the three vectors $\mathbf{x}_1 = (3, 0, -4, 0)^T$, $\mathbf{x}_2 = (0, 2, 0, -2)^T$, and $\mathbf{x}_3 = (0, 1, 0, 1)^T$?

6.1.3. Using Gaussian elimination to reduce $\mathbf{A}$ to an upper-triangular matrix, evaluate $\det(\mathbf{A})$ for each of the following matrices.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 1 \\ 1 & 4 & 4 \end{pmatrix}$. (b) $\mathbf{A} = \begin{pmatrix} 1 & 3 & 5 \\ -1 & 4 & 2 \\ 3 & -2 & 4 \end{pmatrix}$.

(c) $\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}$. (d) $\mathbf{A} = \begin{pmatrix} 0 & 0 & -2 & 3 \\ 1 & 0 & 1 & 2 \\ -1 & 1 & 2 & 1 \\ 0 & 2 & -3 & 0 \end{pmatrix}$.

(e) $\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$. (f) $\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 3 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & n \end{pmatrix}$.

6.1.4. Use determinants to compute the rank of $\mathbf{A} = \begin{pmatrix} 1 & 3 & -2 \\ 0 & 1 & 2 \\ -1 & -1 & 6 \\ 2 & 5 & -6 \end{pmatrix}$.

6.1.5. Use determinants to find the values of α for which the following system possesses a unique solution.

$$\begin{pmatrix} 1 & \alpha & 0 \\ 0 & 1 & -1 \\ \alpha & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -3 \\ 4 \\ 7 \end{pmatrix}.$$

6.1.6. If $\mathbf{A}$ is nonsingular, explain why $\det\left(\mathbf{A}^{-1}\right) = 1/\det(\mathbf{A})$.

6.1.7. Explain why determinants are invariant under similarity transformations. That is, show $\det\left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right) = \det(\mathbf{A})$ for all nonsingular $\mathbf{P}$.

6.1.8. Explain why $\det(\mathbf{A}^*) = \overline{\det(\mathbf{A})}$.

6.1.9. (a) Explain why $|\det(\mathbf{Q})| = 1$ when $\mathbf{Q}$ is unitary. In particular, $\det(\mathbf{Q}) = \pm 1$ if $\mathbf{Q}$ is an orthogonal matrix.

(b) How are the singular values of $\mathbf{A} \in \mathcal{C}^{n\times n}$ related to $\det(\mathbf{A})$?

6.1.10. Prove that if $\mathbf{A}$ is $m \times n$, then $\det(\mathbf{A}^*\mathbf{A}) \geq 0$, and explain why $\det(\mathbf{A}^*\mathbf{A}) > 0$ if and only if $rank(\mathbf{A}) = n$.

6.1.11. If $\mathbf{A}$ is $n \times n$, explain why $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$ for all scalars α.

6.1.12. If $\mathbf{A}$ is an $n \times n$ skew-symmetric matrix, prove that $\mathbf{A}$ is singular whenever n is odd. **Hint:** Use Exercise 6.1.11.

6.1.13. How can you build random integer matrices with $\det(\mathbf{A}) = 1$?

6.1.14. If the k^{th} row of $\mathbf{A}_{n\times n}$ is written as a sum $\mathbf{A}_{k*} = \mathbf{x}^T + \mathbf{y}^T + \cdots + \mathbf{z}^T$, where $\mathbf{x^T}, \mathbf{y^T}, \ldots, \mathbf{z^T}$ are row vectors, explain why

$$\det(\mathbf{A}) = \det\begin{pmatrix} \mathbf{A}_{1*} \\ \vdots \\ \mathbf{x}^T \\ \vdots \\ \mathbf{A}_{n*} \end{pmatrix} + \det\begin{pmatrix} \mathbf{A}_{1*} \\ \vdots \\ \mathbf{y}^T \\ \vdots \\ \mathbf{A}_{n*} \end{pmatrix} + \cdots + \det\begin{pmatrix} \mathbf{A}_{1*} \\ \vdots \\ \mathbf{z}^T \\ \vdots \\ \mathbf{A}_{n*} \end{pmatrix}.$$

6.1.15. The CBS inequality (p. 272) says that $|\mathbf{x}^*\mathbf{y}| \leq \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$ for vectors $\mathbf{x}, \mathbf{y} \in \mathcal{C}^{n\times 1}$. Use Exercise 6.1.10 to give an alternate proof of the CBS inequality along with an alternate explanation of why equality holds if and only if $\mathbf{y}$ is a scalar multiple of $\mathbf{x}$.

6.1.16. Determinant Formula for Pivots. Let $\mathbf{A}_k$ be the $k \times k$ leading principal submatrix of $\mathbf{A}_{n\times n}$ (p. 148). Prove that if $\mathbf{A}$ has an LU factorization $\mathbf{A} = \mathbf{LU}$, then $\det(\mathbf{A}_k) = u_{11}u_{22}\cdots u_{kk}$, and deduce that the k^{th} pivot is $u_{kk} = \begin{cases} \det(\mathbf{A}_1) = a_{11} & \text{for } k = 1, \\ \det(\mathbf{A}_k)/\det(\mathbf{A}_{k-1}) & \text{for } k = 2, 3, \ldots, n. \end{cases}$

6.1.17. Prove that if $rank(\mathbf{A}_{m\times n}) = n$, then $\mathbf{A}^T\mathbf{A}$ has an LU factorization with positive pivots—i.e., $\mathbf{A}^T\mathbf{A}$ is *positive definite* (pp. 154 and 559).

6.1.18. Let $\mathbf{A}(x) = \begin{pmatrix} 2-x & 3 & 4 \\ 0 & 4-x & -5 \\ 1 & -1 & 3-x \end{pmatrix}$.

(a) First evaluate $\det(\mathbf{A})$, and then compute $d\big(\det(\mathbf{A})\big)/dx$.

(b) Use formula (6.1.19) to evaluate $d\big(\det(\mathbf{A})\big)/dx$.

6.1.19. When the entries of $\mathbf{A} = [a_{ij}(x)]$ are differentiable functions of x, we define $d\,\mathbf{A}/dx = [d\,a_{ij}/dx]$ (the matrix of derivatives). For square matrices, is it always the case that $d\big(\det(\mathbf{A})\big)/dx = \det(d\mathbf{A}/dx)$?

6.1.20. For a set of functions $\mathcal{S} = \{f_1(x), f_2(x), \ldots, f_n(x)\}$ that are $n-1$ times differentiable, the determinant

$$w(x) = \begin{vmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1'(x) & f_2'(x) & \cdots & f_n'(x) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{vmatrix}$$

is called the ***Wronskian*** of $\mathcal{S}$. If $\mathcal{S}$ is a linearly dependent set, explain why $w(x) = 0$ for every value of x. **Hint:** Recall Example 4.3.6 (p. 189).

6.1.21. Consider evaluating an $n \times n$ determinant from the definition (6.1.1).

(a) How many multiplications are required?

(b) Assuming a computer will do 1,000,000 multiplications per second, and neglecting all other operations, what is the largest order determinant that can be evaluated in one hour?

(c) Under the same conditions of part (b), how long will it take to evaluate the determinant of a 100×100 matrix?
Hint: $100! \approx 9.33 \times 10^{157}$.

(d) If all other operations are neglected, how many multiplications per second must a computer perform if the task of evaluating the determinant of a 100×100 matrix is to be completed in 100 years?

6.2 ADDITIONAL PROPERTIES OF DETERMINANTS

The purpose of this section is to present some additional properties of determinants that will be helpful in later developments.

Block Determinants

If $\mathbf{A}$ and $\mathbf{D}$ are square matrices, then

$$\det\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix} = \begin{cases}\det(\mathbf{A})\det\left(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right) & \text{when } \mathbf{A}^{-1} \text{ exists,}\\ \det(\mathbf{D})\det\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right) & \text{when } \mathbf{D}^{-1} \text{ exists.}\end{cases} \tag{6.2.1}$$

The matrices $\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ are called the ***Schur complements*** of $\mathbf{A}$ and $\mathbf{D}$, respectively—see Exercise 3.7.11 on p. 123.

Proof. If $\mathbf{A}^{-1}$ exists, then $\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix} = \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{C}\mathbf{A}^{-1} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{0} & \mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\end{pmatrix}$, and the product rules (p. 467) produce the first formula in (6.2.1). The second formula follows by using a similar trick. ∎

Since the determinant of a product is equal to the product of the determinants, it's only natural to inquire if a similar result holds for sums. In other words, is $\det(\mathbf{A}+\mathbf{B}) = \det(\mathbf{A})+\det(\mathbf{B})$? *Almost never*! Try a couple of examples to convince yourself. Nevertheless, there are still some statements that can be made regarding the determinant of certain types of sums. In a loose sense, the result of Exercise 6.1.14 was a statement concerning determinants and sums, but the following result is a little more satisfying.

Rank-One Updates

If $\mathbf{A}_{n\times n}$ is nonsingular, and if $\mathbf{c}$ and $\mathbf{d}$ are $n\times 1$ columns, then

- $\det\left(\mathbf{I}+\mathbf{c}\mathbf{d}^T\right) = 1+\mathbf{d}^T\mathbf{c}$, (6.2.2)
- $\det\left(\mathbf{A}+\mathbf{c}\mathbf{d}^T\right) = \det(\mathbf{A})\left(1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}\right)$. (6.2.3)

Exercise 6.2.7 presents a generalized version of these formulas.

Proof. The proof of (6.2.2) follows by applying the product rules (p. 467) to

$$\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{d}^T & 1\end{pmatrix}\begin{pmatrix}\mathbf{I}+\mathbf{c}\mathbf{d}^T & \mathbf{c}\\ \mathbf{0} & 1\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{d}^T & 1\end{pmatrix} = \begin{pmatrix}\mathbf{I} & \mathbf{c}\\ \mathbf{0} & 1+\mathbf{d}^T\mathbf{c}\end{pmatrix}.$$

To prove (6.2.3), write $\mathbf{A}+\mathbf{c}\mathbf{d}^T = \mathbf{A}\left(\mathbf{I}+\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T\right)$, and apply the product rule (6.1.15) along with (6.2.2). ∎

Example 6.2.1

Problem: For $\mathbf{A} = \begin{pmatrix} 1+\lambda_1 & 1 & \cdots & 1 \\ 1 & 1+\lambda_2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1+\lambda_n \end{pmatrix}$, $\lambda_i \neq 0$, find $\det(\mathbf{A})$.

Solution: Express $\mathbf{A}$ as a rank-one updated matrix $\mathbf{A} = \mathbf{D} + \mathbf{e}\mathbf{e}^T$, where $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\mathbf{e}^T = (\,1 \quad 1 \quad \cdots \quad 1\,)$. Apply (6.2.3) to produce

$$\det(\mathbf{D} + \mathbf{e}\mathbf{e}^T) = \det(\mathbf{D})\left(1 + \mathbf{e}^T\mathbf{D}^{-1}\mathbf{e}\right) = \left(\prod_{i=1}^{n} \lambda_i\right)\left(1 + \sum_{i=1}^{n} \frac{1}{\lambda_i}\right).$$

The classical result known as Cramer's rule[65] is a corollary of the rank-one update formula (6.2.3).

Cramer's Rule

In a nonsingular system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$, the i^{th} unknown is

$$x_i = \frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})},$$

where $\mathbf{A}_i = \left[\mathbf{A}_{*1} \,\middle|\, \cdots \,\middle|\, \mathbf{A}_{*i-1} \,\middle|\, \mathbf{b} \,\middle|\, \mathbf{A}_{*i+1} \,\middle|\, \cdots \,\middle|\, \mathbf{A}_{*n}\right]$. That is, $\mathbf{A}_i$ is identical to $\mathbf{A}$ except that column $\mathbf{A}_{*i}$ has been replaced by $\mathbf{b}$.

Proof. Since $\mathbf{A}_i = \mathbf{A} + (\mathbf{b} - \mathbf{A}_{*i})\,\mathbf{e}_i^T$, where $\mathbf{e}_i$ is the i^{th} unit vector, (6.2.3) may be applied to yield

$$\begin{aligned} \det(\mathbf{A}_i) &= \det(\mathbf{A})\Big(1 + \mathbf{e}_i^T\mathbf{A}^{-1}(\mathbf{b} - \mathbf{A}_{*i})\Big) = \det(\mathbf{A})\Big(1 + \mathbf{e}_i^T(\mathbf{x} - \mathbf{e}_i)\Big) \\ &= \det(\mathbf{A})\,(1 + x_i - 1) = \det(\mathbf{A})\,x_i. \end{aligned}$$

Thus $x_i = \det(\mathbf{A}_i)/\det(\mathbf{A})$ because $\mathbf{A}$ being nonsingular insures $\det(\mathbf{A}) \neq 0$ by (6.1.13). ∎

[65] Gabriel Cramer (1704–1752) was a mathematician from Geneva, Switzerland. As mentioned in §6.1, Cramer's rule was apparently known to others long before Cramer rediscovered and published it in 1750. Nevertheless, Cramer's recognition is not undeserved because his work was responsible for a revived interest in determinants and systems of linear equations. After Cramer's publication, Cramer's rule met with instant success, and it quickly found its way into the textbooks and classrooms of Europe. It is reported that there was a time when students passed or failed the exams in the schools of public service in France according to their understanding of Cramer's rule.

Example 6.2.2

Problem: Determine the value of t for which $x_3(t)$ is minimized in

$$\begin{pmatrix} t & 0 & 1/t \\ 0 & t & t^2 \\ 1 & t^2 & t^3 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} = \begin{pmatrix} 1 \\ 1/t \\ 1/t^2 \end{pmatrix}.$$

Solution: Only one component of the solution is required, so it's wasted effort to solve the entire system. Use Cramer's rule to obtain

$$x_3(t) = \frac{\begin{vmatrix} t & 0 & 1 \\ 0 & t & 1/t \\ 1 & t^2 & 1/t^2 \end{vmatrix}}{\begin{vmatrix} t & 0 & 1/t \\ 0 & t & t^2 \\ 1 & t^2 & t^3 \end{vmatrix}} = \frac{1-t-t^2}{-1} = t^2+t-1, \quad \text{and set} \quad \frac{d\,x_3(t)}{dt} = 0$$

to conclude that $x_3(t)$ is minimized at $t = -1/2$.

Recall that minor determinants of $\mathbf{A}$ are simply determinants of submatrices of $\mathbf{A}$. We are now in a position to see that in an $n \times n$ matrix the $n-1 \times n-1$ minor determinants have a special significance.

Cofactors

The ***cofactor*** of $\mathbf{A}_{n\times n}$ associated with the (i,j)-position is defined as

$$\mathring{\mathrm{A}}_{ij} = (-1)^{i+j} M_{ij},$$

where M_{ij} is the $n-1 \times n-1$ minor obtained by deleting the i^{th} row and j^{th} column of $\mathbf{A}$. The matrix of cofactors is denoted by $\mathring{\mathbf{A}}$.

Example 6.2.3

Problem: For $\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & 6 \\ -3 & 9 & 1 \end{pmatrix}$, determine the cofactors $\mathring{\mathrm{A}}_{21}$ and $\mathring{\mathrm{A}}_{13}$.

Solution:

$\mathring{\mathrm{A}}_{21} = (-1)^{2+1} M_{21} = (-1)(-19) = 19$ and $\mathring{\mathrm{A}}_{13} = (-1)^{1+3} M_{13} = (+1)(18) = 18.$

The entire matrix of cofactors is $\mathring{\mathbf{A}} = \begin{pmatrix} -54 & -20 & 18 \\ 19 & 7 & -6 \\ -6 & -2 & 2 \end{pmatrix}$.

The cofactors of a square matrix $\mathbf{A}$ appear naturally in the expansion of $\det(\mathbf{A})$. For example,

$$\begin{aligned}\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\qquad - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{12}(a_{23}a_{31} - a_{21}a_{33}) \\ &\qquad + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \\ &= a_{11}\mathring{A}_{11} + a_{12}\mathring{A}_{12} + a_{13}\mathring{A}_{13}.\end{aligned} \tag{6.2.4}$$

Because this expansion is in terms of the entries of the first row and the corresponding cofactors, (6.2.4) is called *the cofactor expansion of* $\det(\mathbf{A})$ *in terms of the first row.* It should be clear that there is nothing special about the first row of $\mathbf{A}$. That is, it's just as easy to write an expression similar to (6.2.4) in which entries from any other row or column appear. For example, the terms in (6.2.4) can be rearranged to produce

$$\begin{aligned}\det(\mathbf{A}) &= a_{12}(a_{23}a_{31} - a_{21}a_{33}) + a_{22}(a_{11}a_{33} - a_{13}a_{31}) + a_{32}(a_{13}a_{21} - a_{11}a_{23}) \\ &= a_{12}\mathring{A}_{12} + a_{22}\mathring{A}_{22} + a_{32}\mathring{A}_{32}.\end{aligned}$$

This is called *the cofactor expansion for* $\det(\mathbf{A})$ *in terms of the second column.* The 3×3 case is typical, and exactly the same reasoning can be applied to a more general $n \times n$ matrix in order to obtain the following statements.

Cofactor Expansions

- $\det(\mathbf{A}) = a_{i1}\mathring{A}_{i1} + a_{i2}\mathring{A}_{i2} + \cdots + a_{in}\mathring{A}_{in}$ (about row i). (6.2.5)
- $\det(\mathbf{A}) = a_{1j}\mathring{A}_{1j} + a_{2j}\mathring{A}_{2j} + \cdots + a_{nj}\mathring{A}_{nj}$ (about column j). (6.2.6)

Example 6.2.4

Problem: Use cofactor expansions to evaluate $\det(\mathbf{A})$ for

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 2 \\ 7 & 1 & 6 & 5 \\ 3 & 7 & 2 & 0 \\ 0 & 3 & -1 & 4 \end{pmatrix}.$$

Solution: To minimize the effort, expand $\det(\mathbf{A})$ in terms of the row or column that contains a maximal number of zeros. For this example, the expansion in terms of the first row is most efficient because

$$\det(\mathbf{A}) = a_{11}\mathring{A}_{11} + a_{12}\mathring{A}_{12} + a_{13}\mathring{A}_{13} + a_{14}\mathring{A}_{14} = a_{14}\mathring{A}_{14} = (2)(-1)\begin{vmatrix} 7 & 1 & 6 \\ 3 & 7 & 2 \\ 0 & 3 & -1 \end{vmatrix}.$$

Now expand this remaining 3×3 determinant either in terms of the first column or the third row. Using the first column produces

$$\begin{vmatrix} 7 & 1 & 6 \\ 3 & 7 & 2 \\ 0 & 3 & -1 \end{vmatrix} = (7)(+1)\begin{vmatrix} 7 & 2 \\ 3 & -1 \end{vmatrix} + (3)(-1)\begin{vmatrix} 1 & 6 \\ 3 & -1 \end{vmatrix} = -91 + 57 = -34,$$

so $\det(\mathbf{A}) = (2)(-1)(-34) = 68$. You may wish to try an expansion using different rows or columns, and verify that the final result is the same.

In the previous example, we were able to take advantage of the fact that there were zeros in convenient positions. However, for a general matrix $\mathbf{A}_{n\times n}$ with no zero entries, it's not difficult to verify that successive application of cofactor expansions requires $n!\left(1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{(n-1)!}\right)$ multiplications to evaluate $\det(\mathbf{A})$. Even for moderate values of n, this number is too large for the cofactor expansion to be practical for computational purposes. Nevertheless, cofactors can be useful for theoretical developments such as the following determinant formula for $\mathbf{A}^{-1}$.

Determinant Formula for $\mathbf{A}^{-1}$

The ***adjugate*** of $\mathbf{A}_{n\times n}$ is defined to be $\operatorname{adj}(\mathbf{A}) = \mathring{\mathbf{A}}^T$, the transpose of the matrix of cofactors—some older texts call this the ***adjoint*** matrix. If $\mathbf{A}$ is nonsingular, then

$$\mathbf{A}^{-1} = \frac{\mathring{\mathbf{A}}^T}{\det(\mathbf{A})} = \frac{\operatorname{adj}(\mathbf{A})}{\det(\mathbf{A})}. \tag{6.2.7}$$

Proof. $\left[\mathbf{A}^{-1}\right]_{ij}$ is the i^{th} component in the solution to $\mathbf{Ax} = \mathbf{e}_j$, where $\mathbf{e}_j$ is the j^{th} unit vector. By Cramer's rule, this is

$$\left[\mathbf{A}^{-1}\right]_{ij} = x_i = \frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})},$$

where $\mathbf{A}_i$ is identical to $\mathbf{A}$ except that the i^{th} column has been replaced by $\mathbf{e}_j$, and the cofactor expansion in terms of the i^{th} column implies that

$$\det(\mathbf{A}_i) = \begin{vmatrix} a_{11} & \cdots & \overset{i^{th}}{\overset{\downarrow}{0}} & \cdots & a_{1n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{j1} & \cdots & 1 & \cdots & a_{jn} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{n1} & \cdots & 0 & \cdots & a_{nn} \end{vmatrix} = \mathring{A}_{ji}. \quad \blacksquare$$

Example 6.2.5

Problem: Use determinants to compute $\left[\mathbf{A}^{-1}\right]_{12}$ and $\left[\mathbf{A}^{-1}\right]_{31}$ for the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & 6 \\ -3 & 9 & 1 \end{pmatrix}.$$

Solution: The cofactors $\mathring{A}_{21}$ and $\mathring{A}_{13}$ were determined in Example 6.2.3 to be $\mathring{A}_{21} = 19$ and $\mathring{A}_{13} = 18$, and it's straightforward to compute $\det(\mathbf{A}) = 2$, so

$$\left[\mathbf{A}^{-1}\right]_{12} = \frac{\mathring{A}_{21}}{\det(\mathbf{A})} = \frac{19}{2} \quad \text{and} \quad \left[\mathbf{A}^{-1}\right]_{31} = \frac{\mathring{A}_{13}}{\det(\mathbf{A})} = \frac{18}{2} = 9.$$

Using the matrix of cofactors $\mathring{\mathbf{A}}$ computed in Example 6.2.3, we have that

$$\mathbf{A}^{-1} = \frac{\text{adj}(\mathbf{A})}{\det(\mathbf{A})} = \frac{\mathring{\mathbf{A}}^T}{\det(\mathbf{A})} = \frac{1}{2}\begin{pmatrix} -54 & 19 & -6 \\ -20 & 7 & -2 \\ 18 & -6 & 2 \end{pmatrix}.$$

Example 6.2.6

Problem: For $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, determine a general formula for $\mathbf{A}^{-1}$.

Solution: $\text{adj}(\mathbf{A}) = \mathbf{A}^T = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, and $\det(\mathbf{A}) = ad - bc$, so

$$\mathbf{A}^{-1} = \frac{\text{adj}(\mathbf{A})}{\det(\mathbf{A})} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Example 6.2.7

Problem: Explain why the entries in $\mathbf{A}^{-1}$ vary continuously with the entries in $\mathbf{A}$ when $\mathbf{A}$ is nonsingular. This is in direct contrast with the lack of continuity exhibited by pseudoinverses (p. 423).

Solution: Recall from elementary calculus that the sum, the product, and the quotient of continuous functions are each continuous functions. In particular, the sum and the product of any set of numbers varies continuously as the numbers vary, so $\det(\mathbf{A})$ is a continuous function of the a_{ij}'s. Since each entry in $\text{adj}(\mathbf{A})$ is a determinant, each quotient $[\mathbf{A}^{-1}]_{ij} = [\text{adj}(\mathbf{A})]_{ij}/\det(\mathbf{A})$ must be a continuous function of the a_{ij}'s.

The Moral: The formula $\mathbf{A}^{-1} = \text{adj}(\mathbf{A})/\det(\mathbf{A})$ is nearly worthless for actually computing the value of $\mathbf{A}^{-1}$, but, as this example demonstrates, the formula is nevertheless a useful mathematical tool. It's not uncommon for applied oriented students to fall into the trap of believing that the worth of a formula or an idea is tied to its utility for computing something. This example makes the point that things can have significant mathematical value without being computationally important. In fact, most of this chapter is in this category.

Example 6.2.8

Problem: Explain why the inner product of one row (or column) in $\mathbf{A}_{n\times n}$ with the cofactors of a different row (or column) in $\mathbf{A}$ must always be zero.

Solution: Let $\tilde{\mathbf{A}}$ be the result of replacing the j^{th} column in $\mathbf{A}$ by the k^{th} column of $\mathbf{A}$. Since $\tilde{\mathbf{A}}$ has two identical columns, $\det(\tilde{\mathbf{A}}) = 0$. Furthermore, the cofactor associated with the (i,j)-position in $\tilde{\mathbf{A}}$ is $\mathring{A}_{ij}$, the cofactor associated with the (i,j) in $\mathbf{A}$, so expansion of $\det(\tilde{\mathbf{A}})$ in terms of the j^{th} column yields

$$0 = \det(\tilde{\mathbf{A}}) = \begin{vmatrix} & & \overset{j^{th}}{\downarrow} & & \overset{k^{th}}{\downarrow} & & \\ a_{11} & \cdots & a_{1k} & \cdots & a_{1k} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots & & \\ a_{i1} & \cdots & a_{ik} & \cdots & a_{ik} & \cdots & a_{in} \\ \vdots & & \vdots & & \vdots & & \\ a_{n1} & \cdots & a_{nk} & \cdots & a_{nk} & \cdots & a_{nn} \end{vmatrix} = \sum_{i=1}^{n} a_{ik}\mathring{A}_{ij}.$$

Thus the inner product of the k^{th} column of $\mathbf{A}_{n\times n}$ with the cofactors of the j^{th} column of $\mathbf{A}$ is zero. A similar result holds for rows. Combining these observations with (6.2.5) and (6.2.6) produces

$$\sum_{j=1}^{n} a_{kj}\mathring{A}_{ij} = \begin{cases} \det(\mathbf{A}) & \text{if } k = i, \\ 0 & \text{if } k \neq i, \end{cases} \quad \text{and} \quad \sum_{i=1}^{n} a_{ik}\mathring{A}_{ij} = \begin{cases} \det(\mathbf{A}) & \text{if } k = j, \\ 0 & \text{if } k \neq j, \end{cases}$$

which is equivalent to saying that $\mathbf{A}[\text{adj}(\mathbf{A})] = [\text{adj}(\mathbf{A})]\mathbf{A} = \det(\mathbf{A})\,\mathbf{I}$.

Example 6.2.9

Differential Equations and Determinants. A system of n homogeneous first-order linear differential equations

$$\frac{dx_i(t)}{dt} = a_{i1}(t)x_1(t) + a_{i2}(t)x_2(t) + \cdots + a_{in}(t)x_n(t), \quad i = 1, 2, \ldots, n$$

can be expressed in matrix notation by writing

$$\begin{pmatrix} x_1'(t) \\ x_2'(t) \\ \vdots \\ x_n'(t) \end{pmatrix} = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}$$

or, equivalently, $\mathbf{x}' = \mathbf{A}\mathbf{x}$. Let $\mathcal{S} = \{\mathbf{w}_1(t), \mathbf{w}_2(t), \ldots, \mathbf{w}_n(t)\}$ be a set of $n \times 1$ vectors that are solutions to $\mathbf{x}' = \mathbf{A}\mathbf{x}$, and place these solutions as columns in a matrix $\mathbf{W}(t)_{n\times n} = [\mathbf{w}_1(t) \,|\, \mathbf{w}_2(t) \,|\, \cdots \,|\, \mathbf{w}_n(t)]$ so that $\mathbf{W}' = \mathbf{A}\mathbf{W}$.

Problem: Prove that if $w(t) = \det(\mathbf{W})$, (called the *Wronskian* (p. 474)), then

$$w(t) = w(\xi_0)\, e^{\int_{\xi_0}^{t} \text{trace}\,\mathbf{A}(\xi)\, d\xi}, \quad \text{where } \xi_0 \text{ is an arbitrary constant.} \qquad (6.2.8)$$

Solution: By (6.1.19), $d\,w(t)/dt = \sum_{i=1}^{n} \det(\mathbf{D}_i)$, where

$$\mathbf{D}_i = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ w'_{i1} & w'_{i2} & \cdots & w'_{in} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{pmatrix} = \mathbf{W} + \mathbf{e}_i\mathbf{e}_i^T\mathbf{W}' - \mathbf{e}_i\mathbf{e}_i^T\mathbf{W}.$$

Notice that $\left(-\mathbf{e}_i\mathbf{e}_i^T\mathbf{W}\right)$ subtracts $\mathbf{W}_{i*}$ from the i^{th} row while $\left(+\mathbf{e}_i\mathbf{e}_i^T\mathbf{W}'\right)$ adds $\mathbf{W}'_{i*}$ to the i^{th} row. Use the fact that $\mathbf{W}' = \mathbf{AW}$ to write

$$\mathbf{D}_i = \mathbf{W}+\mathbf{e}_i\mathbf{e}_i^T\mathbf{W}'-\mathbf{e}_i\mathbf{e}_i^T\mathbf{W} = \mathbf{W}+\mathbf{e}_i\mathbf{e}_i^T\mathbf{AW}-\mathbf{e}_i\mathbf{e}_i^T\mathbf{W} = \left(\mathbf{I}+\mathbf{e}_i\left(\mathbf{e}_i^T\mathbf{A}-\mathbf{e}_i^T\right)\right)\mathbf{W},$$

and apply formula (6.2.2) for the determinant of a rank-one updated matrix together with the product rule (6.1.15) to produce

$$\det(\mathbf{D}_i) = \left(1 + \mathbf{e}_i^T\mathbf{A}\mathbf{e}_i - \mathbf{e}_i^T\mathbf{e}_i\right)\det(\mathbf{W}) = a_{ii}(t)\,w(t),$$

so

$$\frac{d\,w(t)}{dt} = \sum_{i=1}^{n} \det(\mathbf{D}_i) = \left(\sum_{i=1}^{n} a_{ii}(t)\right) w(t) = trace\,\mathbf{A}(t)\,w(t).$$

In other words, $w(t)$ satisfies the first-order differential equation $w' = \tau\, w$, where $\tau = trace\,\mathbf{A}(t)$, and the solution of this equation is $w(t) = w(\xi_0)\,e^{\int_{\xi_0}^{t} \tau(\xi)\,d\xi}$.

Consequences: In addition to its aesthetic elegance, (6.2.8) is a useful result because it is the basis for the following theorems.

- If $\mathbf{x}' = \mathbf{Ax}$ has a set of solutions $\mathcal{S} = \{\mathbf{w}_1(t), \mathbf{w}_2(t), \ldots, \mathbf{w}_n(t)\}$ that is linearly independent at *some* point $\xi_0 \in (a,b)$, and if $\int_{\xi_0}^{t} \tau(\xi)\,d\xi$ is finite for $t \in (a,b)$, then $\mathcal{S}$ must be linearly independent at *every* point $t \in (a,b)$.
- If $\mathbf{A}$ is a constant matrix, and if $\mathcal{S}$ is a set of n solutions that is linearly independent at *some* value $t = \xi_0$, then $\mathcal{S}$ must be linearly independent for *all* values of t.

Proof. If $\mathcal{S}$ is linearly independent at ξ_0, then $\mathbf{W}(\xi_0)$ is nonsingular, so $w(\xi_0) \neq 0$. If $\int_{\xi_0}^{t} \tau(\xi)\,d\xi$ is finite when $t \in (a,b)$, then $e^{\int_{\xi_0}^{t} \tau(\xi)\,d\xi}$ is finite and nonzero on (a,b), so, by (6.2.8), $w(t) \neq 0$ on (a,b). Therefore, $\mathbf{W}(t)$ is nonsingular for $t \in (a,b)$, and thus $\mathcal{S}$ is linearly independent at each $t \in (a,b)$.

Exercises for section 6.2

6.2.1. Use a cofactor expansion to evaluate each of the following determinants.

(a) $\begin{vmatrix} 2 & 1 & 1 \\ 6 & 2 & 1 \\ -2 & 2 & 1 \end{vmatrix}$, (b) $\begin{vmatrix} 0 & 0 & -2 & 3 \\ 1 & 0 & 1 & 2 \\ -1 & 1 & 2 & 1 \\ 0 & 2 & -3 & 0 \end{vmatrix}$, (c) $\begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{vmatrix}$.

6.2.2. Use determinants to compute the following inverses.

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

6.2.3. (a) Use Cramer's rule to solve

$$\begin{aligned} x_1 + x_2 + x_3 &= 1, \\ x_1 + x_2 &= \alpha, \\ x_2 + x_3 &= \beta. \end{aligned}$$

(b) Evaluate $\lim_{t\to\infty} x_2(t)$, where $x_2(t)$ is defined by the system

$$\begin{aligned} x_1 + tx_2 + t^2x_3 &= t^4, \\ t^2x_1 + x_2 + tx_3 &= t^3, \\ tx_1 + t^2x_2 + x_3 &= 0. \end{aligned}$$

6.2.4. Is the following equation a valid derivation of Cramer's rule for solving a nonsingular system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}_i$ is as described on p. 476?

$$\frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})} = \det\left(\mathbf{A}^{-1}\mathbf{A}_i\right) = \det\left[\mathbf{e}_1 \cdots \mathbf{e}_{i-1}\ \mathbf{x}\ \mathbf{e}_{i+1} \cdots \mathbf{e}_n\right] = x_i.$$

6.2.5. (a) By example, show that $\det(\mathbf{A}+\mathbf{B}) \neq \det(\mathbf{A}) + \det(\mathbf{B})$.

(b) Using square matrices, construct an example that shows that

$$\det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \neq \det(\mathbf{A})\det(\mathbf{D}) - \det(\mathbf{B})\det(\mathbf{C}).$$

6.2.6. Suppose $rank(\mathbf{B}_{m\times n}) = n$, and let $\mathbf{Q}$ be the orthogonal projector onto $N\left(\mathbf{B}^T\right)$. For $\mathbf{A} = [\mathbf{B} \,|\, \mathbf{c}_{n\times 1}]$, prove $\mathbf{c}^T\mathbf{Qc} = \det\left(\mathbf{A}^T\mathbf{A}\right)/\det\left(\mathbf{B}^T\mathbf{B}\right)$.

6.2.7. If $\mathbf{A}_{n\times n}$ is a nonsingular matrix, and if $\mathbf{D}$ and $\mathbf{C}$ are $n \times k$ matrices, explain how to use (6.2.1) to derive the formula

$$\det\left(\mathbf{A} + \mathbf{CD}^T\right) = \det(\mathbf{A})\det\left(\mathbf{I}_k + \mathbf{D}^T\mathbf{A}^{-1}\mathbf{C}\right).$$

Note: This is a generalization of (6.2.3) because if $\mathbf{c}_i$ and $\mathbf{d}_i$ are the i^{th} columns of $\mathbf{C}$ and $\mathbf{D}$, respectively, then

$$\mathbf{A} + \mathbf{CD}^T = \mathbf{A} + \mathbf{c}_1\mathbf{d}_1^T + \mathbf{c}_2\mathbf{d}_2^T + \cdots + \mathbf{c}_k\mathbf{d}_k^T.$$

6.2.8. Explain why $\mathbf{A}$ is singular if and only if $\mathbf{A}[\text{adj}\,(\mathbf{A})] = \mathbf{0}$.

6.2.9. For a nonsingular linear system $\mathbf{Ax} = \mathbf{b}$, explain why each component of the solution must vary continuously with the entries of $\mathbf{A}$.

6.2.10. For scalars α, explain why $\text{adj}\,(\alpha\mathbf{A}) = \alpha^{n-1}\text{adj}\,(\mathbf{A})$. **Hint:** Recall Exercise 6.1.11.

6.2.11. For an $n \times n$ matrix $\mathbf{A}$, prove that the following statements are true.

(a) If $rank\,(\mathbf{A}) < n-1$, then $\text{adj}\,(\mathbf{A}) = \mathbf{0}$.

(b) If $rank\,(\mathbf{A}) = n-1$, then $rank\,(\text{adj}\,(\mathbf{A})) = 1$.

(c) If $rank\,(\mathbf{A}) = n$, then $rank\,(\text{adj}\,(\mathbf{A})) = n$.

6.2.12. In 1812, Cauchy discovered the formula that says that if $\mathbf{A}$ is $n \times n$, then $\det(\text{adj}\,(\mathbf{A})) = [\det(\mathbf{A})]^{n-1}$. Establish Cauchy's formula.

6.2.13. For the following tridiagonal matrix, $\mathbf{A}_n$, let $D_n = \det(\mathbf{A}_n)$, and derive the formula $D_n = 2D_{n-1} - D_{n-2}$ to deduce that $D_n = n+1$.

$$\mathbf{A}_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}_{n\times n}.$$

6.2.14. By considering rank-one updated matrices, derive the following formulas.

(a) $$\begin{vmatrix} \frac{1+\alpha_1}{\alpha_1} & 1 & \cdots & 1 \\ 1 & \frac{1+\alpha_2}{\alpha_2} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \frac{1+\alpha_n}{\alpha_n} \end{vmatrix} = \frac{1+\sum \alpha_i}{\prod \alpha_i}.$$

(b) $$\begin{vmatrix} \alpha & \beta & \beta & \cdots & \beta \\ \beta & \alpha & \beta & \cdots & \beta \\ \beta & \beta & \alpha & \cdots & \beta \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta & \beta & \beta & \cdots & \alpha \end{vmatrix}_{n\times n} = \begin{cases} (\alpha-\beta)^n\left(1+\frac{n\beta}{\alpha-\beta}\right) & \text{if } \alpha \neq \beta, \\ 0 & \text{if } \alpha = \beta. \end{cases}$$

(c) $$\begin{vmatrix} 1+\alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_1 & 1+\alpha_2 & \cdots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & 1+\alpha_n \end{vmatrix} = 1+\alpha_1+\alpha_2+\cdots+\alpha_n.$$

6.2.15. A ***bordered matrix*** has the form $\mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{y}^T & \alpha \end{pmatrix}$ in which $\mathbf{A}_{n\times n}$ is nonsingular, $\mathbf{x}$ is a column, $\mathbf{y}^T$ is a row, and α is a scalar. Explain why the following statements must be true.

$$\text{(a)} \quad \begin{vmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{y}^T & -1 \end{vmatrix} = -\det\left(\mathbf{A} + \mathbf{x}\mathbf{y}^T\right). \quad \text{(b)} \quad \begin{vmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{y}^T & 0 \end{vmatrix} = -\mathbf{y}^T \operatorname{adj}(\mathbf{A})\,\mathbf{x}.$$

6.2.16. If $\mathbf{B}$ is $m \times n$ and $\mathbf{C}$ is $n \times m$, explain why (6.2.1) guarantees that $\lambda^m \det(\lambda\mathbf{I}_n - \mathbf{CB}) = \lambda^n \det(\lambda\mathbf{I}_m - \mathbf{BC})$ is true for all scalars λ.

6.2.17. For a square matrix $\mathbf{A}$ and column vectors $\mathbf{c}$ and $\mathbf{d}$, derive the following two extensions of formula (6.2.3).

(a) If $\mathbf{Ax} = \mathbf{c}$, then $\det\left(\mathbf{A} + \mathbf{cd}^T\right) = \det(\mathbf{A})\left(1 + \mathbf{d}^T\mathbf{x}\right)$.

(b) If $\mathbf{y}^T\mathbf{A} = \mathbf{d}^T$, then $\det\left(\mathbf{A} + \mathbf{cd}^T\right) = \det(\mathbf{A})\left(1 + \mathbf{y}^T\mathbf{c}\right)$.

6.2.18. Describe the determinant of an elementary reflector (p. 324) and a plane rotation (p. 333), and then explain how to find $\det(\mathbf{A})$ using Householder reduction (p. 341) and Givens reduction (Example 5.7.2).

6.2.19. Suppose that $\mathbf{A}$ is a nonsingular matrix whose entries are integers. Prove that the entries in $\mathbf{A}^{-1}$ are integers if and only if $\det(\mathbf{A}) = \pm 1$.

6.2.20. Let $\mathbf{A} = \mathbf{I} - 2\mathbf{uv}^T$ be a matrix in which $\mathbf{u}$ and $\mathbf{v}$ are column vectors with integer entries.

(a) Prove that $\mathbf{A}^{-1}$ has integer entries if and only if $\mathbf{v}^T\mathbf{u} = 0$ or 1.

(b) A matrix is said to be ***involutory*** whenever $\mathbf{A}^{-1} = \mathbf{A}$. Explain why $\mathbf{A} = \mathbf{I} - 2\mathbf{uv}^T$ is involutory when $\mathbf{v}^T\mathbf{u} = 1$.

6.2.21. Use induction to argue that a cofactor expansion of $\det(\mathbf{A}_{n\times n})$ requires

$$c(n) = n!\left(1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{(n-1)!}\right)$$

multiplications for $n \geq 2$. Assume a computer will do 1,000,000 multiplications per second, and neglect all other operations to estimate how long it will take to evaluate the determinant of a 100×100 matrix using cofactor expansions. **Hint:** Recall the series expansion for e^x, and use $100! \approx 9.33 \times 10^{157}$.

6.2.22. Determine all values of λ for which the matrix $\mathbf{A}-\lambda\mathbf{I}$ is singular, where

$$\mathbf{A} = \begin{pmatrix} 0 & -3 & -2 \\ 2 & 5 & 2 \\ -2 & -3 & 0 \end{pmatrix}.$$

Hint: If $p(\lambda) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0$ is a monic polynomial with integer coefficients, then the integer roots of $p(\lambda)$ are a subset of the factors of α_0.

6.2.23. Suppose that $f_1(t),\ f_2(t), \ldots, f_n(t)$ are solutions of n^{th}-order linear differential equation $y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_{n-1}(t)y' + p_n(t)y = 0$, and let $w(t)$ be the Wronskian

$$w(t) = \begin{vmatrix} f_1(t) & f_2(t) & \cdots & f_n(t) \\ f_1'(t) & f_2'(t) & \cdots & f_n'(t) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(t) & f_2^{(n-1)}(t) & \cdots & f_n^{(n-1)}(t) \end{vmatrix}.$$

By converting the n^{th}-order equation into a system of n first-order equations with the substitutions $x_1 = y,\ x_2 = y', \ldots, x_n = y^{(n-1)}$, show that $w(t) = w(\xi_0)\, \mathrm{e}^{-\int_{\xi_0}^{t} p_1(\xi)\, d\xi}$ for an arbitrary constant ξ_0.

6.2.24. Evaluate the ***Vandermonde determinant*** by showing

$$\begin{vmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{j>i} (x_j - x_i).$$

When is this nonzero (compare with Example 4.3.4)? **Hint:** For the polynomial $p(\lambda) = \begin{vmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{k-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{k-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_k & x_k^2 & \cdots & x_k^{k-1} \end{vmatrix}_{k\times k}$, use induction to find the degree of $p(\lambda)$, the roots of $p(\lambda)$, and the coefficient of λ^{k-1} in $p(\lambda)$.

6.2.25. Suppose that each entry in $\mathbf{A}_{n\times n} = [a_{ij}(x)]$ is a differentiable function of a real variable x. Use formula (6.1.19) to derive the formula

$$\frac{d\big(\det(\mathbf{A})\big)}{dx} = \sum_{j=1}^{n}\sum_{i=1}^{n} \frac{d\, a_{ij}}{dx}\, \mathring{\mathrm{A}}_{ij}.$$

6.2.26. Consider the entries of $\mathbf{A}$ to be independent variables, and use formula (6.1.19) to derive the formula

$$\frac{\partial \det(\mathbf{A})}{\partial a_{ij}} = \mathring{\mathrm{A}}_{ij}.$$

6.2.27. Laplace's Expansion. In 1772, the French mathematician Pierre-Simon Laplace (1749–1827) presented the following generalized version of the cofactor expansion. For an $n \times n$ matrix $\mathbf{A}$, let

$$\mathbf{A}(i_1 i_2 \cdots i_k \,|\, j_1 j_2 \cdots j_k) = \text{the } k \times k \text{ submatrix of } \mathbf{A} \text{ that lies on the intersection of rows } i_1, i_2, \ldots, i_k \text{ with columns } j_1, j_2, \ldots, j_k,$$

and let

$$M(i_1 i_2 \cdots i_k \,|\, j_1 j_2 \cdots j_k) = \text{the } n-k \times n-k \text{ minor determinant obtained by deleting rows } i_1, i_2, \ldots, i_k \text{ and columns } j_1, j_2, \ldots, j_k \text{ from } \mathbf{A}.$$

The ***cofactor*** of $\mathbf{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k)$ is defined to be the signed minor

$$\mathring{\mathrm{A}}(i_1 \cdots i_k \,|\, j_1 \cdots j_k) = (-1)^{i_1 + \cdots + i_k + j_1 + \cdots + j_k} M(i_1 \cdots i_k \,|\, j_1 \cdots j_k).$$

This is consistent with the definition of cofactor given earlier because if $\mathbf{A}(i \,|\, j) = a_{ij}$, then $\mathring{\mathrm{A}}(i \,|\, j) = (-1)^{i+j} M(i \,|\, j) = (-1)^{i+j} M_{ij} = \mathring{\mathrm{A}}_{ij}$. For each fixed set of row indices $1 \le i_1 < \cdots < i_k \le n$,

$$\det(\mathbf{A}) = \sum_{1 \le j_1 < \cdots < j_k \le n} \det \mathbf{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k) \mathring{\mathrm{A}}(i_1 \cdots i_k \,|\, j_1 \cdots j_k).$$

Similarly, for each fixed set of column indices $1 \le j_1 < \cdots < j_k \le n$,

$$\det(\mathbf{A}) = \sum_{1 \le i_1 < \cdots < i_k \le n} \det \mathbf{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k) \mathring{\mathrm{A}}(i_1 \cdots i_k \,|\, j_1 \cdots j_k).$$

Each of these sums contains $\binom{n}{k}$ terms. Use Laplace's expansion to evaluate the determinant of

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & -2 & 3 \\ 1 & 0 & 1 & 2 \\ -1 & 1 & 2 & 1 \\ 0 & 2 & -3 & 0 \end{pmatrix}$$

in terms of the first and third rows.

You know that I write slowly. This is chiefly because I am never satisfied until I have said as much as possible in a few words, and writing briefly takes far more time than writing at length.
— *Carl Friedrich Gauss* (1777–1855)

CHAPTER 7

Eigenvalues and Eigenvectors

7.1 ELEMENTARY PROPERTIES OF EIGENSYSTEMS

Up to this point, almost everything was either motivated by or evolved from the consideration of systems of linear *algebraic* equations. But we have come to a turning point, and from now on the emphasis will be different. Rather than being concerned with systems of *algebraic* equations, many topics will be motivated or driven by applications involving systems of linear *differential* equations and their discrete counterparts, difference equations.

For example, consider the problem of solving the system of two first-order linear differential equations, $du_1/dt = 7u_1 - 4u_2$ and $du_2/dt = 5u_1 - 2u_2$. In matrix notation, this system is

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \text{or, equivalently,} \quad \mathbf{u}' = \mathbf{A}\mathbf{u}, \tag{7.1.1}$$

where $\mathbf{u}' = \begin{pmatrix} u_1' \\ u_2' \end{pmatrix}$, $\mathbf{A} = \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix}$, and $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$. Because solutions of a single equation $u' = \lambda u$ have the form $u = \alpha e^{\lambda t}$, we are motivated to seek solutions of (7.1.1) that also have the form

$$u_1 = \alpha_1 e^{\lambda t} \quad \text{and} \quad u_2 = \alpha_2 e^{\lambda t}. \tag{7.1.2}$$

Differentiating these two expressions and substituting the results in (7.1.1) yields

$$\begin{aligned} \alpha_1 \lambda e^{\lambda t} &= 7\alpha_1 e^{\lambda t} - 4\alpha_2 e^{\lambda t} \\ \alpha_2 \lambda e^{\lambda t} &= 5\alpha_1 e^{\lambda t} - 2\alpha_2 e^{\lambda t} \end{aligned} \Rightarrow \begin{aligned} \alpha_1 \lambda &= 7\alpha_1 - 4\alpha_2 \\ \alpha_2 \lambda &= 5\alpha_1 - 2\alpha_2 \end{aligned} \Rightarrow \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \lambda \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}.$$

In other words, solutions of (7.1.1) having the form (7.1.2) can be constructed provided solutions for λ and $\mathbf{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ in the matrix equation $\mathbf{Ax} = \lambda\mathbf{x}$ can be found. Clearly, $\mathbf{x} = \mathbf{0}$ trivially satisfies $\mathbf{Ax} = \lambda\mathbf{x}$, but $\mathbf{x} = \mathbf{0}$ provides no useful information concerning the solution of (7.1.1). What we really need are scalars λ and *nonzero* vectors $\mathbf{x}$ that satisfy $\mathbf{Ax} = \lambda\mathbf{x}$. Writing $\mathbf{Ax} = \lambda\mathbf{x}$ as $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}$ shows that the vectors of interest are the *nonzero* vectors in $N(\mathbf{A} - \lambda\mathbf{I})$. But $N(\mathbf{A} - \lambda\mathbf{I})$ contains nonzero vectors if and only if $\mathbf{A} - \lambda\mathbf{I}$ is singular. Therefore, the scalars of interest are precisely the values of λ that make $\mathbf{A} - \lambda\mathbf{I}$ singular or, equivalently, the λ's for which $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. These observations motivate the definition of eigenvalues and eigenvectors.[66]

Eigenvalues and Eigenvectors

For an $n \times n$ matrix $\mathbf{A}$, scalars λ and vectors $\mathbf{x}_{n\times 1} \neq \mathbf{0}$ satisfying $\mathbf{Ax} = \lambda\mathbf{x}$ are called ***eigenvalues*** and ***eigenvectors*** of $\mathbf{A}$, respectively, and any such pair, $(\lambda, \mathbf{x})$, is called an ***eigenpair*** for $\mathbf{A}$. The set of *distinct* eigenvalues, denoted by $\sigma(\mathbf{A})$, is called the ***spectrum*** of $\mathbf{A}$.

- $\lambda \in \sigma(\mathbf{A}) \iff \mathbf{A} - \lambda\mathbf{I}$ is singular $\iff \det(\mathbf{A} - \lambda\mathbf{I}) = 0.$ (7.1.3)
- $\{\mathbf{x} \neq \mathbf{0} \mid \mathbf{x} \in N(\mathbf{A} - \lambda\mathbf{I})\}$ is the set of all eigenvectors associated with λ. From now on, $N(\mathbf{A} - \lambda\mathbf{I})$ is called an ***eigenspace*** for $\mathbf{A}$.
- Nonzero row vectors $\mathbf{y}^*$ such that $\mathbf{y}^*(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$ are called ***left-hand eigenvectors*** for $\mathbf{A}$ (see Exercise 7.1.18 on p. 503).

Geometrically, $\mathbf{Ax} = \lambda\mathbf{x}$ says that under transformation by $\mathbf{A}$, eigenvectors experience only changes in magnitude or sign—the orientation of $\mathbf{Ax}$ in $\Re^n$ is the same as that of $\mathbf{x}$. The eigenvalue λ is simply the amount of "stretch" or "shrink" to which the eigenvector $\mathbf{x}$ is subjected when transformed by $\mathbf{A}$. Figure 7.1.1 depicts the situation in $\Re^2$.

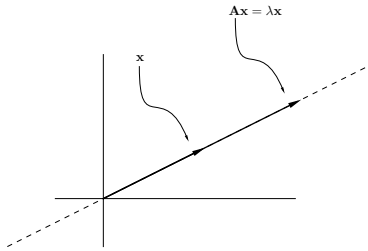


FIGURE 7.1.1

[66] The words *eigenvalue* and *eigenvector* are derived from the German word *eigen*, which means *owned by* or *peculiar to*. Eigenvalues and eigenvectors are sometimes called *characteristic values* and *characteristic vectors, proper values* and *proper vectors,* or *latent values* and *latent vectors.*

Let's now face the problem of finding the eigenvalues and eigenvectors of the matrix $\mathbf{A} = \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix}$ appearing in (7.1.1). As noted in (7.1.3), the eigenvalues are the scalars λ for which $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Expansion of $\det(\mathbf{A} - \lambda\mathbf{I})$ produces the second-degree polynomial

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 7-\lambda & -4 \\ 5 & -2-\lambda \end{vmatrix} = \lambda^2 - 5\lambda + 6 = (\lambda - 2)(\lambda - 3),$$

which is called the ***characteristic polynomial*** for $\mathbf{A}$. Consequently, the eigenvalues for $\mathbf{A}$ are the solutions of the ***characteristic equation*** $p(\lambda) = 0$ (i.e., the roots of the characteristic polynomial), and they are $\lambda = 2$ and $\lambda = 3$.

The eigenvectors associated with $\lambda = 2$ and $\lambda = 3$ are simply the nonzero vectors in the eigenspaces $N(\mathbf{A} - 2\mathbf{I})$ and $N(\mathbf{A} - 3\mathbf{I})$, respectively. But determining these eigenspaces amounts to nothing more than solving the two homogeneous systems, $(\mathbf{A} - 2\mathbf{I})\,\mathbf{x} = \mathbf{0}$ and $(\mathbf{A} - 3\mathbf{I})\,\mathbf{x} = \mathbf{0}$.

For $\lambda = 2$,

$$\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 5 & -4 \\ 5 & -4 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -4/5 \\ 0 & 0 \end{pmatrix} \implies \begin{array}{l} x_1 = (4/5)x_2 \\ x_2 \text{ is free} \end{array}$$

$$\implies N(\mathbf{A} - 2\mathbf{I}) = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \alpha \begin{pmatrix} 4/5 \\ 1 \end{pmatrix} \right\}.$$

For $\lambda = 3$,

$$\mathbf{A} - 3\mathbf{I} = \begin{pmatrix} 4 & -4 \\ 5 & -5 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} \implies \begin{array}{l} x_1 = x_2 \\ x_2 \text{ is free} \end{array}$$

$$\implies N(\mathbf{A} - 3\mathbf{I}) = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

In other words, the eigenvectors of $\mathbf{A}$ associated with $\lambda = 2$ are all nonzero multiples of $\mathbf{x} = \begin{pmatrix} 4/5 & 1 \end{pmatrix}^T$, and the eigenvectors associated with $\lambda = 3$ are all nonzero multiples of $\mathbf{y} = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$. Although there are an infinite number of eigenvectors associated with each eigenvalue, each eigenspace is one dimensional, so, for this example, there is only one *independent* eigenvector associated with each eigenvalue.

Let's complete the discussion concerning the system of differential equations $\mathbf{u}' = \mathbf{A}\mathbf{u}$ in (7.1.1). Coupling (7.1.2) with the eigenpairs $(\lambda_1, \mathbf{x})$ and $(\lambda_2, \mathbf{y})$ of $\mathbf{A}$ computed above produces two solutions of $\mathbf{u}' = \mathbf{A}\mathbf{u}$, namely,

$$\mathbf{u}_1 = e^{\lambda_1 t}\mathbf{x} = e^{2t} \begin{pmatrix} 4/5 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 = e^{\lambda_2 t}\mathbf{y} = e^{3t} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

It turns out that all other solutions are linear combinations of these two particular solutions—more is said in §7.4 on p. 541.

Below is a summary of some general statements concerning features of the characteristic polynomial and the characteristic equation.

Characteristic Polynomial and Equation

- The ***characteristic polynomial*** of $\mathbf{A}_{n\times n}$ is $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$. The degree of $p(\lambda)$ is n, and the leading term in $p(\lambda)$ is $(-1)^n\lambda^n$.
- The ***characteristic equation*** for $\mathbf{A}$ is $p(\lambda) = 0$.
- The eigenvalues of $\mathbf{A}$ are the solutions of the characteristic equation or, equivalently, the roots of the characteristic polynomial.
- Altogether, $\mathbf{A}$ has n eigenvalues, but some may be complex numbers (even if the entries of $\mathbf{A}$ are real numbers), and some eigenvalues may be repeated.
- If $\mathbf{A}$ contains only real numbers, then its complex eigenvalues must occur in conjugate pairs—i.e., if $\lambda \in \sigma(\mathbf{A})$, then $\overline{\lambda} \in \sigma(\mathbf{A})$.

Proof. The fact that $\det(\mathbf{A} - \lambda\mathbf{I})$ is a polynomial of degree n whose leading term is $(-1)^n\lambda^n$ follows from the definition of determinant given in (6.1.1). If

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

then

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \sum_p \sigma(p)(a_{1p_1} - \delta_{1p_1}\lambda)(a_{2p_2} - \delta_{2p_2}\lambda)\cdots(a_{np_n} - \delta_{np_n}\lambda)$$

is a polynomial in λ. The highest power of λ is produced by the term

$$(a_{11} - \lambda)(a_{22} - \lambda)\cdots(a_{nn} - \lambda),$$

so the degree is n, and the leading term is $(-1)^n\lambda^n$. The discussion given earlier contained the proof that the eigenvalues are precisely the solutions of the characteristic equation, but, for the sake of completeness, it's repeated below:

$$\lambda \in \sigma(\mathbf{A}) \iff \mathbf{A}\mathbf{x} = \lambda\mathbf{x} \text{ for some } \mathbf{x} \neq \mathbf{0} \iff (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \text{ for some } \mathbf{x} \neq \mathbf{0}$$
$$\iff \mathbf{A} - \lambda\mathbf{I} \text{ is singular} \iff \det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

The fundamental theorem of algebra is a deep result that insures every polynomial of degree n with real or complex coefficients has n roots, but some roots may be complex numbers (even if all the coefficients are real), and some roots may be repeated. Consequently, $\mathbf{A}$ has n eigenvalues, but some may be complex, and some may be repeated. The fact that complex eigenvalues of real matrices must occur in conjugate pairs is a consequence of the fact that the roots of a polynomial with real coefficients occur in conjugate pairs. ∎

Example 7.1.1

Problem: Determine the eigenvalues and eigenvectors of $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$.

Solution: The characteristic polynomial is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & -1 \\ 1 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 + 1 = \lambda^2 - 2\lambda + 2,$$

so the characteristic equation is $\lambda^2 - 2\lambda + 2 = 0$. Application of the quadratic formula yields

$$\lambda = \frac{2 \pm \sqrt{-4}}{2} = \frac{2 \pm 2\sqrt{-1}}{2} = 1 \pm \mathrm{i},$$

so the spectrum of $\mathbf{A}$ is $\sigma(\mathbf{A}) = \{1+\mathrm{i}, 1-\mathrm{i}\}$. Notice that the eigenvalues are complex conjugates of each other—as they must be because complex eigenvalues of real matrices must occur in conjugate pairs. Now find the eigenspaces.

For $\lambda = 1 + \mathrm{i}$,

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} -\mathrm{i} & -1 \\ 1 & -\mathrm{i} \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -\mathrm{i} \\ 0 & 0 \end{pmatrix} \implies N(\mathbf{A} - \lambda\mathbf{I}) = span\left\{ \begin{pmatrix} \mathrm{i} \\ 1 \end{pmatrix} \right\}.$$

For $\lambda = 1 - \mathrm{i}$,

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} \mathrm{i} & -1 \\ 1 & \mathrm{i} \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & \mathrm{i} \\ 0 & 0 \end{pmatrix} \implies N(\mathbf{A} - \lambda\mathbf{I}) = span\left\{ \begin{pmatrix} -\mathrm{i} \\ 1 \end{pmatrix} \right\}.$$

In other words, the eigenvectors associated with $\lambda_1 = 1 + \mathrm{i}$ are all nonzero multiples of $\mathbf{x}_1 = (\,\mathrm{i} \quad 1\,)^T$, and the eigenvectors associated with $\lambda_2 = 1 - \mathrm{i}$ are all nonzero multiples of $\mathbf{x}_2 = (\,-\mathrm{i} \quad 1\,)^T$. In previous sections, you could be successful by thinking only in terms of real numbers and by dancing around those statements and issues involving complex numbers. But this example makes it clear that avoiding complex numbers, even when dealing with real matrices, is no longer possible—very innocent looking matrices, such as the one in this example, can possess complex eigenvalues and eigenvectors.

As we have seen, computing eigenvalues boils down to solving a polynomial equation. But determining solutions to polynomial equations can be a formidable task. It was proven in the nineteenth century that it's impossible to express the roots of a general polynomial of degree five or higher using radicals of the coefficients. This means that there does not exist a generalized version of the quadratic formula for polynomials of degree greater than four, and general polynomial equations cannot be solved by a finite number of arithmetic operations involving $+, -, \times, \div, \sqrt[n]{}$. Unlike solving $\mathbf{Ax} = \mathbf{b}$, the eigenvalue problem generally requires an infinite algorithm, so all practical eigenvalue computations are accomplished by iterative methods—some are discussed later.

For theoretical work, and for textbook-type problems, it's helpful to express the characteristic equation in terms of the principal minors. Recall that an $r \times r$ ***principal submatrix*** of $\mathbf{A}_{n\times n}$ is a submatrix that lies on the same set of r rows and columns, and an $r \times r$ ***principal minor*** is the determinant of an $r \times r$ principal submatrix. In other words, $r \times r$ principal minors are obtained by deleting the same set of $n-r$ rows and columns, and there are $\binom{n}{r} = n!/r!(n-r)!$ such minors. For example, the 1×1 principal minors of

$$\mathbf{A} = \begin{pmatrix} -3 & 1 & -3 \\ 20 & 3 & 10 \\ 2 & -2 & 4 \end{pmatrix} \tag{7.1.4}$$

are the diagonal entries -3, 3, and 4. The 2×2 principal minors are

$$\begin{vmatrix} -3 & 1 \\ 20 & 3 \end{vmatrix} = -29, \quad \begin{vmatrix} -3 & -3 \\ 2 & 4 \end{vmatrix} = -6, \quad \text{and} \quad \begin{vmatrix} 3 & 10 \\ -2 & 4 \end{vmatrix} = 32,$$

and the only 3×3 principal minor is $\det(\mathbf{A}) = -18$.

Related to the principal minors are the symmetric functions of the eigenvalues. The k^{th} ***symmetric function*** of $\lambda_1, \lambda_2, \ldots, \lambda_n$ is defined to be the sum of the product of the eigenvalues taken k at a time. That is,

$$s_k = \sum_{1\le i_1<\cdots<i_k\le n} \lambda_{i_1} \cdots \lambda_{i_k}.$$

For example, when $n = 4$,

$$\begin{aligned} s_1 &= \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4, \\ s_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_4 + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \lambda_3\lambda_4, \\ s_3 &= \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_4 + \lambda_1\lambda_3\lambda_4 + \lambda_2\lambda_3\lambda_4, \\ s_4 &= \lambda_1\lambda_2\lambda_3\lambda_4. \end{aligned}$$

The connection between symmetric functions, principal minors, and the coefficients in the characteristic polynomial is given in the following theorem.

Coefficients in the Characteristic Equation

If $\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_{n-1}\lambda + c_n = 0$ is the characteristic equation for $\mathbf{A}_{n\times n}$, and if s_k is the k^{th} symmetric function of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\mathbf{A}$, then

- $c_k = (-1)^k \sum(\text{all } k \times k \text{ principal minors}),$ (7.1.5)
- $s_k = \sum(\text{all } k \times k \text{ principal minors}),$ (7.1.6)
- $\mathit{trace}\,(\mathbf{A}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n = -c_1,$ (7.1.7)
- $\det(\mathbf{A}) = \lambda_1\lambda_2\cdots\lambda_n = (-1)^n c_n.$ (7.1.8)

Proof. At least two proofs of (7.1.5) are possible, and although they are conceptually straightforward, each is somewhat tedious. One approach is to successively use the result of Exercise 6.1.14 to expand $\det(\mathbf{A}-\lambda\mathbf{I})$. Another proof rests on the observation that if

$$p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})=(-1)^n\lambda^n+a_1\lambda^{n-1}+a_2\lambda^{n-2}+\cdots+a_{n-1}\lambda+a_n$$

is the characteristic *polynomial* for $\mathbf{A}$, then the characteristic *equation* is

$$\lambda^n+c_1\lambda^{n-1}+c_2\lambda^{n-2}+\cdots+c_{n-1}\lambda+c_n=0, \quad \text{where} \quad c_i=(-1)^n a_i.$$

Taking the r^{th} derivative of $p(\lambda)$ yields $p^{(r)}(0)=r!a_{n-r}$, and hence

$$c_{n-r}=\frac{(-1)^n}{r!}p^{(r)}(0). \tag{7.1.9}$$

It's now a matter of repeatedly applying the formula (6.1.19) for differentiating a determinant to $p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})$. After r applications of (6.1.19),

$$p^{(r)}(\lambda)=\sum_{i_j\neq i_k} D_{i_1\cdots i_r}(\lambda),$$

where $D_{i_1\cdots i_r}(\lambda)$ is the determinant of the matrix identical to $\mathbf{A}-\lambda\mathbf{I}$ except that rows $i_1, i_2, \ldots, i_r$ have been replaced by $-\mathbf{e}_{i_1}^T, -\mathbf{e}_{i_2}^T, \ldots, -\mathbf{e}_{i_r}^T$, respectively. It follows that $D_{i_1\cdots i_r}(0)=(-1)^r\det(\mathbf{A}_{i_1\cdots i_r})$, where $\mathbf{A}_{i_1 i_2\cdots i_r}$ is identical to $\mathbf{A}$ except that rows $i_1, i_2, \ldots, i_r$ have been replaced by $\mathbf{e}_{i_1}^T, \mathbf{e}_{i_2}^T, \ldots, \mathbf{e}_{i_r}^T$, respectively, and $\det(\mathbf{A}_{i_1\cdots i_r})$ is the $n-r\times n-r$ principal minor obtained by deleting rows and columns $i_1, i_2, \ldots, i_r$ from $\mathbf{A}$. Consequently,

$$\begin{aligned} p^{(r)}(0)&=\sum_{i_j\neq i_k} D_{i_1\cdots i_r}(0)=(-1)^r\sum_{i_j\neq i_k}\det(\mathbf{A}_{i_1\cdots i_r}) \\ &=r!\times(-1)^r\sum(\text{all } n-r\times n-r \text{ principal minors}). \end{aligned}$$

The factor $r!$ appears because each of the $r!$ permutations of the subscripts on $\mathbf{A}_{i_1\cdots i_r}$ describes the same matrix. Therefore, (7.1.9) says

$$c_{n-r}=\frac{(-1)^n}{r!}p^{(r)}(0)=(-1)^{n-r}\sum(\text{all } n-r\times n-r \text{ principal minors}).$$

To prove (7.1.6), write the characteristic equation for $\mathbf{A}$ as

$$(\lambda-\lambda_1)(\lambda-\lambda_2)\cdots(\lambda-\lambda_n)=0, \tag{7.1.10}$$

and expand the left-hand side to produce

$$\lambda^n-s_1\lambda^{n-1}+\cdots+(-1)^k s_k\lambda^{n-k}+\cdots+(-1)^n s_n=0. \tag{7.1.11}$$

(Using $n=3$ or $n=4$ in (7.1.10) makes this clear.) Comparing (7.1.11) with (7.1.5) produces the desired conclusion. Statements (7.1.7) and (7.1.8) are obtained from (7.1.5) and (7.1.6) by setting $k=1$ and $k=n$. ∎

Example 7.1.2

Problem: Determine the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} -3 & 1 & -3 \\ 20 & 3 & 10 \\ 2 & -2 & 4 \end{pmatrix}.$$

Solution: Use the principal minors computed in (7.1.4) along with (7.1.5) to obtain the characteristic equation

$$\lambda^3 - 4\lambda^2 - 3\lambda + 18 = 0.$$

A result from elementary algebra states that if the coefficients α_i in

$$\lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0 = 0$$

are integers, then every integer solution is a factor of α_0. For our problem, this means that if there exist integer eigenvalues, then they must be contained in the set $\mathcal{S} = \{\pm 1, \pm 2, \pm 3, \pm 6, \pm 9, \pm 18\}$. Evaluating $p(\lambda)$ for each $\lambda \in \mathcal{S}$ reveals that $p(3) = 0$ and $p(-2) = 0$, so $\lambda = 3$ and $\lambda = -2$ are eigenvalues for $\mathbf{A}$. To determine the other eigenvalue, deflate the problem by dividing

$$\frac{\lambda^3 - 4\lambda^2 - 3\lambda + 18}{\lambda - 3} = \lambda^2 - \lambda - 6 = (\lambda - 3)(\lambda + 2).$$

Thus the characteristic equation can be written in factored form as

$$(\lambda - 3)^2(\lambda + 2) = 0,$$

so the spectrum of $\mathbf{A}$ is $\sigma(\mathbf{A}) = \{3, -2\}$ in which $\lambda = 3$ is repeated—we say that the ***algebraic multiplicity*** of $\lambda = 3$ is two. The eigenspaces are obtained as follows.

For $\lambda = 3$,

$$\mathbf{A} - 3\mathbf{I} \longrightarrow \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \implies N(\mathbf{A} - 3\mathbf{I}) = span\left\{ \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} \right\}.$$

For $\lambda = -2$,

$$\mathbf{A} + 2\mathbf{I} \longrightarrow \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{pmatrix} \implies N(\mathbf{A} + 2\mathbf{I}) = span\left\{ \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} \right\}.$$

Notice that although the algebraic multiplicity of $\lambda = 3$ is two, the dimension of the associated eigenspace is only one—we say that $\mathbf{A}$ is ***deficient*** in eigenvectors. As we will see later, deficient matrices pose significant difficulties.

Example 7.1.3

Continuity of Eigenvalues. A classical result (requiring complex analysis) states that the roots of a polynomial vary continuously with the coefficients. Since the coefficients of the characteristic polynomial $p(\lambda)$ of $\mathbf{A}$ can be expressed in terms of sums of principal minors, it follows that the coefficients of $p(\lambda)$ vary continuously with the entries of $\mathbf{A}$. Consequently, the eigenvalues of $\mathbf{A}$ must vary continuously with the entries of $\mathbf{A}$. **Caution!** Components of an eigenvector need not vary continuously with the entries of $\mathbf{A}$—e.g., consider $\mathbf{x} = (\epsilon^{-1}, 1)$ as an eigenvector for $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & \epsilon \end{pmatrix}$, and let $\epsilon \to 0$.

Example 7.1.4

Spectral Radius. For square matrices $\mathbf{A}$, the number

$$\rho(\mathbf{A}) = \max_{\lambda \in \sigma(\mathbf{A})} |\lambda|$$

is called the ***spectral radius*** of $\mathbf{A}$. It's not uncommon for applications to require only a bound on the eigenvalues of $\mathbf{A}$. That is, precise knowledge of each eigenvalue may not called for, but rather just an upper bound on $\rho(\mathbf{A})$ is all that's often needed. A rather crude (but cheap) upper bound on $\rho(\mathbf{A})$ is obtained by observing that $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$ for every matrix norm. This is true because if $(\lambda, \mathbf{x})$ is any eigenpair, then $\mathbf{X} = \big[\mathbf{x} \,|\, \mathbf{0} \,|\, \cdots \,|\, \mathbf{0}\big]_{n \times n} \neq \mathbf{0}$, and $\lambda \mathbf{X} = \mathbf{A}\mathbf{X}$ implies $|\lambda| \, \|\mathbf{X}\| = \|\lambda \mathbf{X}\| = \|\mathbf{A}\mathbf{X}\| \leq \|\mathbf{A}\| \, \|\mathbf{X}\|$, so

$$|\lambda| \leq \|\mathbf{A}\| \quad \text{for all } \lambda \in \sigma(\mathbf{A}). \tag{7.1.12}$$

This result is a precursor to a stronger relationship between spectral radius and norm that is hinted at in Exercise 7.3.12 and developed in Example 7.10.1 (p. 619).

The eigenvalue bound (7.1.12) given in Example 7.1.4 is cheap to compute, especially if the 1-norm or ∞-norm is used, but you often get what you pay for. You get one big circle whose radius is usually much larger than the spectral radius $\rho(\mathbf{A})$. It's possible to do better by using a set of Gerschgorin[67] circles as described below.

[67] S. A. Gerschgorin illustrated the use of Gerschgorin circles for estimating eigenvalues in 1931, but the concept appears earlier in work by L. Lévy in 1881, by H. Minkowski (p. 278) in 1900, and by J. Hadamard (p. 469) in 1903. However, each time the idea surfaced, it gained little attention and was quickly forgotten until Olga Taussky (1906–1995), the premier woman of linear algebra, and her fellow German emigrè Alfred Brauer (1894–1985) became captivated by the result. Taussky (who became Olga Taussky-Todd after marrying the numerical analyst John Todd) and Brauer devoted significant effort to strengthening, promoting, and popularizing Gerschgorin-type eigenvalue bounds. Their work during the 1940s and 1950s ended the periodic rediscoveries, and they made Gerschgorin (who might otherwise have been forgotten) famous.

Gerschgorin Circles

- The eigenvalues of $\mathbf{A} \in \mathcal{C}^{n\times n}$ are contained the union $\mathcal{G}_r$ of the n ***Gerschgorin circles*** defined by

$$|z - a_{ii}| \le r_i, \quad \text{where} \quad r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \text{ for } i = 1, 2, \ldots, n. \tag{7.1.13}$$

  In other words, the eigenvalues are trapped in the collection of circles centered at a_{ii} with radii given by the sum of absolute values in $\mathbf{A}_{i*}$ with a_{ii} deleted.

- Furthermore, if a union $\mathcal{U}$ of k Gerschgorin circles does not touch any of the other $n-k$ circles, then there are exactly k eigenvalues (counting multiplicities) in the circles in $\mathcal{U}$. (7.1.14)

- Since $\sigma(\mathbf{A}^T) = \sigma(\mathbf{A})$, the deleted absolute row sums in (7.1.13) can be replaced by deleted absolute column sums, so the eigenvalues of $\mathbf{A}$ are also contained in the union $\mathcal{G}_c$ of the circles defined by

$$|z - a_{jj}| \le c_j, \quad \text{where} \quad c_j = \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \text{ for } j = 1, 2, \ldots, n. \tag{7.1.15}$$

- Combining (7.1.13) and (7.1.15) means that the eigenvalues of $\mathbf{A}$ are contained in the intersection $\mathcal{G}_r \cap \mathcal{G}_c$. (7.1.16)

Proof. Let $(\lambda, \mathbf{x})$ be an eigenpair for $\mathbf{A}$, and assume $\mathbf{x}$ has been normalized so that $\|\mathbf{x}\|_\infty = 1$. If x_i is a component of $\mathbf{x}$ such that $|x_i| = 1$, then

$$\lambda x_i = [\lambda \mathbf{x}]_i = [\mathbf{A}\mathbf{x}]_i = \sum_{j=1}^{n} a_{ij}x_j \quad \Longrightarrow \quad (\lambda - a_{ii})x_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j,$$

and hence

$$|\lambda - a_{ii}| = |\lambda - a_{ii}|\,|x_i| = \left|\sum_{j\neq i} a_{ij}x_j\right| \le \sum_{j\neq i} |a_{ij}|\,|x_j| \le \sum_{j\neq i} |a_{ij}| = r_i.$$

Thus λ is in one of the Gerschgorin circles, so the union of all such circles contains $\sigma(\mathbf{A})$. To establish (7.1.14), let $\mathbf{D} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ and $\mathbf{B} = \mathbf{A} - \mathbf{D}$, and set $\mathbf{C}(t) = \mathbf{D} + t\mathbf{B}$ for $t \in [0, 1]$. The first part shows that the eigenvalues of $\lambda_i(t)$ of $\mathbf{C}(t)$ are contained in the union of the Gerschgorin circles $\mathcal{C}_i(t)$ defined by $|z - a_{ii}| \le t\, r_i$. The circles $\mathcal{C}_i(t)$ grow continuously with t from individual points a_{ii} when $t = 0$ to the Gerschgorin circles of $\mathbf{A}$ when $t = 1$,

so, if the circles in the isolated union $\mathcal{U}$ are centered at $a_{i_1i_1}, a_{i_2i_2}, \ldots, a_{i_ki_k}$, then for every $t \in [0,1]$ the union $\mathcal{U}(t) = \mathcal{C}_{i_1}(t) \cup \mathcal{C}_{i_2}(t) \cup \cdots \cup \mathcal{C}_{i_k}(t)$ is disjoint from the union $\overline{\mathcal{U}(t)}$ of the other $n-k$ Gerschgorin circles of $\mathbf{C}(t)$. Since (as mentioned in Example 7.1.3) each eigenvalue $\lambda_i(t)$ of $\mathbf{C}(t)$ also varies continuously with t, each $\lambda_i(t)$ is on a continuous curve Γ_i having one end at $\lambda_i(0) = a_{ii}$ and the other end at $\lambda_i(1) \in \sigma(\mathbf{A})$. But since $\mathcal{U}(t) \cap \overline{\mathcal{U}(t)} = \phi$ for all $t \in [0,1]$, the curves $\Gamma_{i_1}, \Gamma_{i_2}, \ldots, \Gamma_{i_k}$ are entirely contained in $\mathcal{U}$, and hence the end points $\lambda_{i_1}(1), \lambda_{i_2}(1), \ldots, \lambda_{i_k}(1)$ are in $\mathcal{U}$. Similarly, the other $n-k$ eigenvalues of $\mathbf{A}$ are in the union of the complementary set of circles. ∎

Example 7.1.5

Problem: Estimate the eigenvalues of $\mathbf{A} = \begin{pmatrix} 5 & 1 & 1 \\ 0 & 6 & 1 \\ 1 & 0 & -5 \end{pmatrix}$.

- A crude estimate is derived from the bound given in Example 7.1.4 on p. 497. Using the ∞-norm, (7.1.12) says that $|\lambda| \leq \|\mathbf{A}\|_\infty = 7$ for all $\lambda \in \sigma(\mathbf{A})$.
- Better estimates are produced by the Gerschgorin circles in Figure 7.1.2 that are derived from row sums. Statements (7.1.13) and (7.1.14) guarantee that one eigenvalue is in (or on) the circle centered at -5, while the remaining two eigenvalues are in (or on) the larger circle centered at $+5$.

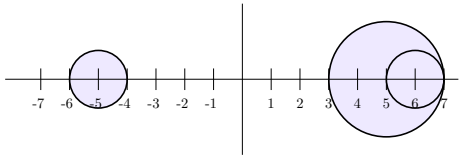


Figure 7.1.2. Gerschgorin circles derived from row sums.

- The best estimate is obtained from (7.1.16) by considering $\mathcal{G}_r \cap \mathcal{G}_c$.

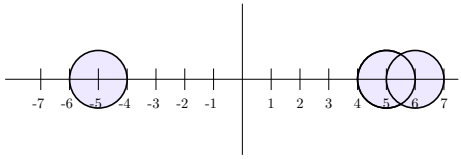


Figure 7.1.3. Gerschgorin circles derived from $\mathcal{G}_r \cap \mathcal{G}_c$.

In other words, one eigenvalue is in the circle centered at -5, while the other two eigenvalues are in the union of the other two circles in Figure 7.1.3. This is corroborated by computing $\sigma(\mathbf{A}) = \{5,\ (1 \pm 5\sqrt{5})/2\} \approx \{5,\ 6.0902,\ -5.0902\}$.

Example 7.1.6

Diagonally Dominant Matrices Revisited. Recall from Example 4.3.3 on p. 184 that $\mathbf{A}_{n\times n}$ is said to be ***diagonally dominant*** (some authors say *strictly* diagonally dominant) whenever

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \text{for each } i = 1, 2, \ldots, n.$$

Gerschgorin's theorem (7.1.13) guarantees that diagonally dominant matrices cannot possess a zero eigenvalue. But $0 \notin \sigma(\mathbf{A})$ if and only if $\mathbf{A}$ is nonsingular (Exercise 7.1.6), so Gerschgorin's theorem provides an alternative to the argument used in Example 4.3.3 to prove that *all diagonally dominant matrices are nonsingular*.[68] For example, the 3×3 matrix $\mathbf{A}$ in Example 7.1.5 is diagonally dominant, and thus $\mathbf{A}$ is nonsingular. Even when a matrix is not diagonally dominant, Gerschgorin estimates still may be useful in determining whether or not the matrix is nonsingular simply by observing if zero is excluded from $\sigma(\mathbf{A})$ based on the configuration of the Gerschgorin circles given in (7.1.16).

Exercises for section 7.1

7.1.1. Determine the eigenvalues and eigenvectors for the following matrices.

$$\mathbf{A} = \begin{pmatrix} -10 & -7 \\ 14 & 11 \end{pmatrix}. \quad \mathbf{B} = \begin{pmatrix} 2 & 16 & 8 \\ 4 & 14 & 8 \\ -8 & -32 & -18 \end{pmatrix}. \quad \mathbf{C} = \begin{pmatrix} 3 & -2 & 5 \\ 0 & 1 & 4 \\ 0 & -1 & 5 \end{pmatrix}.$$

$$\mathbf{D} = \begin{pmatrix} 0 & 6 & 3 \\ -1 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix}. \quad \mathbf{E} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Which, if any, are deficient in eigenvectors in the sense that there fails to exist a complete linearly independent set?

7.1.2. Without doing an eigenvalue–eigenvector computation, determine which of the following are eigenvectors for

$$\mathbf{A} = \begin{pmatrix} -9 & -6 & -2 & -4 \\ -8 & -6 & -3 & -1 \\ 20 & 15 & 8 & 5 \\ 32 & 21 & 7 & 12 \end{pmatrix},$$

and for those which are eigenvectors, identify the associated eigenvalue.

(a) $\begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$. (b) $\begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}$. (c) $\begin{pmatrix} -1 \\ 0 \\ 2 \\ 2 \end{pmatrix}$. (d) $\begin{pmatrix} 0 \\ 1 \\ -3 \\ 0 \end{pmatrix}$.

[68] In fact, this result was the motivation behind the original development of Gerschgorin's circles.

7.1.3. Explain why the eigenvalues of triangular and diagonal matrices

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

are simply the diagonal entries—the t_{ii}'s and λ_i's.

7.1.4. For $\mathbf{T} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$, prove $\det(\mathbf{T} - \lambda\mathbf{I}) = \det(\mathbf{A} - \lambda\mathbf{I})\det(\mathbf{C} - \lambda\mathbf{I})$ to conclude that $\sigma\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} = \sigma(\mathbf{A}) \cup \sigma(\mathbf{C})$ for square $\mathbf{A}$ and $\mathbf{C}$.

7.1.5. Determine the eigenvectors of $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. In particular, what is the eigenspace associated with λ_i?

7.1.6. Prove that $0 \in \sigma(\mathbf{A})$ if and only if $\mathbf{A}$ is a singular matrix.

7.1.7. Explain why it's apparent that $\mathbf{A}_{n\times n} = \begin{pmatrix} n & 1 & 1 & \cdots & 1 \\ 1 & n & 1 & \cdots & 1 \\ 1 & 1 & n & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & n \end{pmatrix}$ doesn't have a zero eigenvalue, and hence why $\mathbf{A}$ is nonsingular.

7.1.8. Explain why the eigenvalues of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ are real and nonnegative for every $\mathbf{A} \in \mathcal{C}^{m\times n}$. **Hint:** Consider $\|\mathbf{A}\mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2$. When are the eigenvalues of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ strictly positive?

7.1.9. (a) If $\mathbf{A}$ is nonsingular, and if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, show that $(\lambda^{-1}, \mathbf{x})$ is an eigenpair for $\mathbf{A}^{-1}$.

(b) For all $\alpha \notin \sigma(\mathbf{A})$, prove that $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ if and only if $\mathbf{x}$ is an eigenvector of $(\mathbf{A} - \alpha\mathbf{I})^{-1}$.

7.1.10. (a) Show that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(\lambda^k, \mathbf{x})$ is an eigenpair for $\mathbf{A}^k$ for each positive integer k.

(b) If $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k$ is any polynomial, then we define $p(\mathbf{A})$ to be the matrix

$$p(\mathbf{A}) = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_k\mathbf{A}^k.$$

Show that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(p(\lambda), \mathbf{x})$ is an eigenpair for $p(\mathbf{A})$.

7.1.11. Explain why (7.1.14) in Gerschgorin's theorem on p. 498 implies that $\mathbf{A} = \begin{pmatrix} 1 & 0 & -2 & 0 \\ 0 & 12 & 0 & -4 \\ 1 & 0 & -1 & 0 \\ 0 & 5 & 0 & 0 \end{pmatrix}$ must have at least two real eigenvalues. Corroborate this fact by computing the eigenvalues of $\mathbf{A}$.

7.1.12. If $\mathbf{A}$ is ***nilpotent*** ($\mathbf{A}^k = \mathbf{0}$ for some k), explain why $trace\,(\mathbf{A}) = 0$. **Hint:** What is $\sigma(\mathbf{A})$?

7.1.13. If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are eigenvectors of $\mathbf{A}$ associated with the same eigenvalue λ, explain why every nonzero linear combination

$$\mathbf{v} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_n\mathbf{x}_n$$

is also an eigenvector for $\mathbf{A}$ associated with the eigenvalue λ.

7.1.14. Explain why an eigenvector for a square matrix $\mathbf{A}$ cannot be associated with two distinct eigenvalues for $\mathbf{A}$.

7.1.15. Suppose $\sigma(\mathbf{A}_{n\times n}) = \sigma(\mathbf{B}_{n\times n})$. Does this guarantee that $\mathbf{A}$ and $\mathbf{B}$ have the same characteristic polynomial?

7.1.16. Construct 2×2 examples to prove the following statements.

(a) $\lambda \in \sigma(\mathbf{A})$ and $\mu \in \sigma(\mathbf{B}) \implies \lambda + \mu \in \sigma(\mathbf{A} + \mathbf{B})$.

(b) $\lambda \in \sigma(\mathbf{A})$ and $\mu \in \sigma(\mathbf{B}) \implies \lambda\mu \in \sigma(\mathbf{AB})$.

7.1.17. Suppose that $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are the eigenvalues for $\mathbf{A}_{n\times n}$, and let $(\lambda_k, \mathbf{c})$ be a particular eigenpair.

(a) For $\lambda \notin \sigma(\mathbf{A})$, explain why $(\mathbf{A} - \lambda\mathbf{I})^{-1}\mathbf{c} = \mathbf{c}/(\lambda_k - \lambda)$.

(b) For an arbitrary vector $\mathbf{d}_{n\times 1}$, prove that the eigenvalues of $\mathbf{A} + \mathbf{cd}^T$ agree with those of $\mathbf{A}$ except that λ_k is replaced by $\lambda_k + \mathbf{d}^T\mathbf{c}$.

(c) How can $\mathbf{d}$ be selected to guarantee that the eigenvalues of $\mathbf{A} + \mathbf{cd}^T$ and $\mathbf{A}$ agree except that λ_k is replaced by a specified number μ?

7.1.18. Suppose that $\mathbf{A}$ is a square matrix.

(a) Explain why $\mathbf{A}$ and $\mathbf{A}^T$ have the same eigenvalues.

(b) Explain why $\lambda \in \sigma(\mathbf{A}) \iff \overline{\lambda} \in \sigma(\mathbf{A}^*)$.
Hint: Recall Exercise 6.1.8.

(c) Do these results imply that $\lambda \in \sigma(\mathbf{A}) \iff \overline{\lambda} \in \sigma(\mathbf{A})$ when $\mathbf{A}$ is a square matrix of *real* numbers?

(d) A nonzero row vector $\mathbf{y}^*$ is called a ***left-hand*** eigenvector for $\mathbf{A}$ whenever there is a scalar $\mu \in \mathcal{C}$ such that $\mathbf{y}^*(\mathbf{A} - \mu\mathbf{I}) = \mathbf{0}$. Explain why μ must be an eigenvalue for $\mathbf{A}$ in the "right-hand" sense of the term when $\mathbf{A}$ is a square matrix of *real* numbers.

7.1.19. Consider matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times m}$.

(a) Explain why $\mathbf{AB}$ and $\mathbf{BA}$ have the same characteristic polynomial if $m = n$. **Hint:** Recall Exercise 6.2.16.

(b) Explain why the characteristic polynomials for $\mathbf{AB}$ and $\mathbf{BA}$ can't be the same when $m \neq n$, and then explain why $\sigma(\mathbf{AB})$ and $\sigma(\mathbf{BA})$ agree, with the possible exception of a zero eigenvalue.

7.1.20. If $\mathbf{AB} = \mathbf{BA}$, prove that $\mathbf{A}$ and $\mathbf{B}$ have a common eigenvector.

Hint: For $\lambda \in \sigma(\mathbf{A})$, let the columns of $\mathbf{X}$ be a basis for $N(\mathbf{A} - \lambda\mathbf{I})$ so that $(\mathbf{A} - \lambda\mathbf{I})\mathbf{BX} = \mathbf{0}$. Explain why there exists a matrix $\mathbf{P}$ such that $\mathbf{BX} = \mathbf{XP}$, and then consider any eigenpair for $\mathbf{P}$.

7.1.21. For fixed matrices $\mathbf{P}_{m\times m}$ and $\mathbf{Q}_{n\times n}$, let $\mathbf{T}$ be the linear operator on $\mathcal{C}^{m\times n}$ defined by $\mathbf{T}(\mathbf{A}) = \mathbf{PAQ}$.

(a) Show that if $\mathbf{x}$ is a right-hand eigenvector for $\mathbf{P}$ and $\mathbf{y}^*$ is a left-hand eigenvector for $\mathbf{Q}$, then $\mathbf{xy}^*$ is an eigenvector for $\mathbf{T}$.

(b) Explain why $trace(\mathbf{T}) = trace(\mathbf{P})\, trace(\mathbf{Q})$.

7.1.22. Let $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ be a diagonal real matrix such that $\lambda_1 < \lambda_2 < \cdots < \lambda_n$, and let $\mathbf{v}_{n\times 1}$ be a column of real nonzero numbers.

(a) Prove that if α is real and nonzero, then λ_i is not an eigenvalue for $\mathbf{D} + \alpha\mathbf{vv}^T$. Show that the eigenvalues of $\mathbf{D} + \alpha\mathbf{vv}^T$ are in fact given by the solutions of the *secular equation* $f(\xi) = 0$ defined by

$$f(\xi) = 1 + \alpha \sum_{i=1}^{n} \frac{v_i^2}{\lambda_i - \xi}.$$

For $n = 4$ and $\alpha > 0$, verify that the graph of $f(\xi)$ is as depicted in Figure 7.1.4, and thereby conclude that the eigenvalues of $\mathbf{D} + \alpha\mathbf{vv}^T$ interlace with those of $\mathbf{D}$.

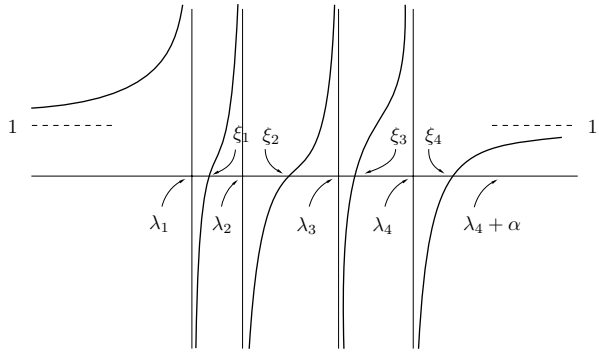


FIGURE 7.1.4

(b) Verify that $(\mathbf{D} - \xi_i \mathbf{I})^{-1}\mathbf{v}$ is an eigenvector for $\mathbf{D} + \alpha \mathbf{v}\mathbf{v}^T$ that is associated with the eigenvalue ξ_i.

7.1.23. Newton's Identities. Let $\lambda_1, \ldots, \lambda_n$ be the roots of the polynomial $p(\lambda) = \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_n$, and let $\tau_k = \lambda_1^k + \lambda_2^k + \cdots + \lambda_n^k$. Newton's identities say $c_k = -(\tau_1 c_{k-1} + \tau_2 c_{k-2} + \cdots + \tau_{k-1}c_1 + \tau_k)/k$. Derive these identities by executing the following steps:

(a) Show $p'(\lambda) = p(\lambda)\sum_{i=1}^{n}(\lambda - \lambda_i)^{-1}$ (logarithmic differentiation).

(b) Use the geometric series expansion for $(\lambda - \lambda_i)^{-1}$ to show that for $|\lambda| > \max_i |\lambda_i|$,

$$\sum_{i=1}^{n} \frac{1}{(\lambda - \lambda_i)} = \frac{n}{\lambda} + \frac{\tau_1}{\lambda^2} + \frac{\tau_2}{\lambda^3} + \cdots.$$

(c) Combine these two results, and equate like powers of λ.

7.1.24. Leverrier–Souriau–Frame Algorithm.[69] Let the characteristic equation for $\mathbf{A}$ be given by $\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_n = 0$, and define a sequence by taking $\mathbf{B}_0 = \mathbf{I}$ and

$$\mathbf{B}_k = -\frac{trace\,(\mathbf{A}\mathbf{B}_{k-1})}{k}\mathbf{I} + \mathbf{A}\mathbf{B}_{k-1} \quad \text{for} \quad k = 1, 2, \ldots, n.$$

Prove that for each k,

$$c_k = -\frac{trace\,(\mathbf{A}\mathbf{B}_{k-1})}{k}.$$

Hint: Use Newton's identities, and recall Exercise 7.1.10(a).

[69] This algorithm has been rediscovered and modified several times. In 1840, the Frenchman U. J. J. Leverrier provided the basic connection with Newton's identities. J. M. Souriau, also from France, and J. S. Frame, from Michigan State University, independently modified the algorithm to its present form—Souriau's formulation was published in France in 1948, and Frame's method appeared in the United States in 1949. Paul Horst (USA, 1935) along with Faddeev and Sominskii (USSR, 1949) are also credited with rediscovering the technique. Although the algorithm is intriguingly beautiful, it is not practical for floating-point computations.

7.2 DIAGONALIZATION BY SIMILARITY TRANSFORMATIONS

The correct choice of a coordinate system (or basis) often can simplify the form of an equation or the analysis of a particular problem. For example, consider the obliquely oriented ellipse in Figure 7.2.1 whose equation in the xy-coordinate system is

$$13x^2 + 10xy + 13y^2 = 72.$$

By rotating the xy-coordinate system counterclockwise through an angle of $45°$

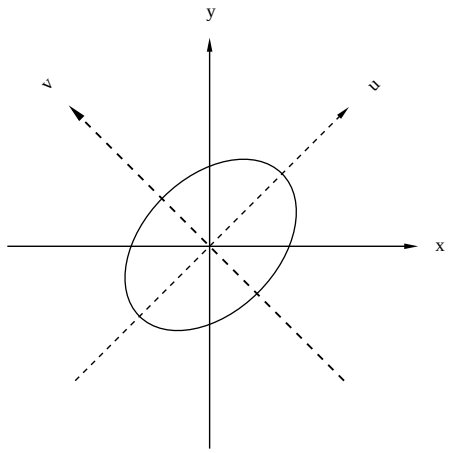


FIGURE 7.2.1

into a uv-coordinate system by means of (5.6.13) on p. 326, the cross-product term is eliminated, and the equation of the ellipse simplifies to become

$$\frac{u^2}{9} + \frac{v^2}{4} = 1.$$

It's shown in Example 7.6.3 on p. 567 that we can do a similar thing for quadratic equations in $\Re^n$.

Choosing or changing to the most appropriate coordinate system (or basis) is always desirable, but in linear algebra it is fundamental. For a linear operator $\mathbf{L}$ on a finite-dimensional space $\mathcal{V}$, the goal is to find a basis $\mathcal{B}$ for $\mathcal{V}$ such that the matrix representation of $\mathbf{L}$ with respect to $\mathcal{B}$ is as simple as possible. Since different matrix representations $\mathbf{A}$ and $\mathbf{B}$ of $\mathbf{L}$ are related by a similarity transformation $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{B}$ (recall §4.8),[70] the fundamental problem for linear operators is strictly a matrix issue—i.e., find a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{AP}$ is as simple as possible. The concept of similarity was first introduced on p. 255, but in the interest of continuity it is reviewed below.

[70] While it is helpful to have covered the topics in §§4.7–4.9, much of the subsequent development is accessible without an understanding of this material.

Similarity

- Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be ***similar*** whenever there exists a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{B}$. The product $\mathbf{P}^{-1}\mathbf{AP}$ is called a ***similarity transformation*** on $\mathbf{A}$.
- **A Fundamental Problem.** Given a square matrix $\mathbf{A}$, reduce it to the simplest possible form by means of a similarity transformation.

Diagonal matrices have the simplest form, so we first ask, "Is every square matrix similar to a diagonal matrix?" Linear algebra and matrix theory would be simpler subjects if this were true, but it's not. For example, consider

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{7.2.1}$$

and observe that $\mathbf{A}^2 = \mathbf{0}$ ($\mathbf{A}$ is nilpotent). If there exists a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$, where $\mathbf{D}$ is diagonal, then

$$\mathbf{D}^2 = \mathbf{P}^{-1}\mathbf{APP}^{-1}\mathbf{AP} = \mathbf{P}^{-1}\mathbf{A}^2\mathbf{P} = \mathbf{0} \implies \mathbf{D} = \mathbf{0} \implies \mathbf{A} = \mathbf{0},$$

which is false. Thus $\mathbf{A}$, as well as any other nonzero nilpotent matrix, is not similar to a diagonal matrix. Nonzero nilpotent matrices are not the only ones that can't be diagonalized, but, as we will see, nilpotent matrices play a particularly important role in nondiagonalizability.

So, if not all square matrices can be diagonalized by a similarity transformation, what are the characteristics of those that can? An answer is easily derived by examining the equation

$$\mathbf{P}^{-1}\mathbf{A}_{n\times n}\mathbf{P} = \mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

which implies $\mathbf{A}\,[\mathbf{P}_{*1} \,|\, \cdots \,|\, \mathbf{P}_{*n}] = [\mathbf{P}_{*1} \,|\, \cdots \,|\, \mathbf{P}_{*n}] \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$ or, equivalently, $[\mathbf{AP}_{*1} \,|\, \cdots \,|\, \mathbf{AP}_{*n}] = [\lambda_1\mathbf{P}_{*1} \,|\, \cdots \,|\, \lambda_n\mathbf{P}_{*n}]$. Consequently, $\mathbf{AP}_{*j} = \lambda_j\mathbf{P}_{*j}$ for each j, so each $(\lambda_j, \mathbf{P}_{*j})$ is an eigenpair for $\mathbf{A}$. In other words, $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$ implies that $\mathbf{P}$ must be a matrix whose columns constitute n linearly independent eigenvectors, and $\mathbf{D}$ is a diagonal matrix whose diagonal entries are the corresponding eigenvalues. It's straightforward to reverse the above argument to prove the converse—i.e., if there exists a linearly independent set of n eigenvectors that are used as columns to build a nonsingular matrix $\mathbf{P}$, and if $\mathbf{D}$ is the diagonal matrix whose diagonal entries are the corresponding eigenvalues, then $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$. Below is a summary.

Diagonalizability

- A square matrix $\mathbf{A}$ is said to be ***diagonalizable*** whenever $\mathbf{A}$ is similar to a diagonal matrix.
- A ***complete set of eigenvectors*** for $\mathbf{A}_{n\times n}$ is any set of n linearly independent eigenvectors for $\mathbf{A}$. Not all matrices have complete sets of eigenvectors—e.g., consider (7.2.1) or Example 7.1.2. Matrices that fail to possess complete sets of eigenvectors are sometimes called ***deficient*** or ***defective*** matrices.
- $\mathbf{A}_{n\times n}$ is diagonalizable if and only if $\mathbf{A}$ possesses a complete set of eigenvectors. Moreover, $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ if and only if the columns of $\mathbf{P}$ constitute a complete set of eigenvectors and the λ_j's are the associated eigenvalues—i.e., each $(\lambda_j, \mathbf{P}_{*j})$ is an eigenpair for $\mathbf{A}$.

Example 7.2.1

Problem: If possible, diagonalize the following matrix with a similarity transformation:

$$\mathbf{A} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}.$$

Solution: Determine whether or not $\mathbf{A}$ has a complete set of three linearly independent eigenvectors. The characteristic equation—perhaps computed by using (7.1.5)—is

$$\lambda^3 + 5\lambda^2 + 3\lambda - 9 = (\lambda - 1)(\lambda + 3)^2 = 0.$$

Therefore, $\lambda = 1$ is a simple eigenvalue, and $\lambda = -3$ is repeated twice (we say its algebraic multiplicity is 2). Bases for the eigenspaces $N(\mathbf{A} - 1\mathbf{I})$ and $N(\mathbf{A} + 3\mathbf{I})$ are determined in the usual way to be

$$N(\mathbf{A} - 1\mathbf{I}) = span\left\{\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}\right\} \quad \text{and} \quad N(\mathbf{A} + 3\mathbf{I}) = span\left\{\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\right\},$$

and it's easy to check that when combined these three eigenvectors constitute a linearly independent set. Consequently, $\mathbf{A}$ must be diagonalizable. To explicitly exhibit the similarity transformation that diagonalizes $\mathbf{A}$, set

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}, \quad \text{and verify} \quad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -3 \end{pmatrix} = \mathbf{D}.$$

Since not all square matrices are diagonalizable, it's natural to inquire about the next best thing—i.e., can every square matrix be *triangularized* by similarity? This time the answer is *yes,* but before explaining why, we need to make the following observation.

Similarity Preserves Eigenvalues

Row reductions don't preserve eigenvalues (try a simple example). However, similar matrices have the same characteristic polynomial, so they have the same eigenvalues with the same multiplicities. **Caution!** Similar matrices need not have the same eigenvectors—see Exercise 7.2.3.

Proof. Use the product rule for determinants in conjunction with the fact that $\det\left(\mathbf{P}^{-1}\right) = 1/\det\left(\mathbf{P}\right)$ (Exercise 6.1.6) to write

$$\begin{aligned}\det\left(\mathbf{A}-\lambda\mathbf{I}\right) &= \det\left(\mathbf{P}^{-1}\mathbf{B}\mathbf{P}-\lambda\mathbf{I}\right) = \det\left(\mathbf{P}^{-1}(\mathbf{B}-\lambda\mathbf{I})\mathbf{P}\right)\\ &= \det\left(\mathbf{P}^{-1}\right)\det\left(\mathbf{B}-\lambda\mathbf{I}\right)\det\left(\mathbf{P}\right) = \det\left(\mathbf{B}-\lambda\mathbf{I}\right). \quad \blacksquare\end{aligned}$$

In the context of linear operators, this means that the eigenvalues of a matrix representation of an operator $\mathbf{L}$ are invariant under a change of basis. In other words, the eigenvalues are intrinsic to $\mathbf{L}$ in the sense that they are independent of any coordinate representation.

Now we can establish the fact that every square matrix can be triangularized by a similarity transformation. In fact, as Issai Schur (p. 123) realized in 1909, the similarity transformation always can be made to be unitary.

Schur's Triangularization Theorem

Every square matrix is unitarily similar to an upper-triangular matrix. That is, for each $\mathbf{A}_{n\times n}$, there exists a unitary matrix $\mathbf{U}$ (not unique) and an upper-triangular matrix $\mathbf{T}$ (not unique) such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$, and the diagonal entries of $\mathbf{T}$ are the eigenvalues of $\mathbf{A}$.

Proof. Use induction on n, the size of the matrix. For $n = 1$, there is nothing to prove. For $n > 1$, assume that all $n-1 \times n-1$ matrices are unitarily similar to an upper-triangular matrix, and consider an $n \times n$ matrix $\mathbf{A}$. Suppose that $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, and suppose that $\mathbf{x}$ has been normalized so that $\|\mathbf{x}\|_2 = 1$. As discussed on p. 325, we can construct an elementary reflector $\mathbf{R} = \mathbf{R}^* = \mathbf{R}^{-1}$ with the property that $\mathbf{R}\mathbf{x} = \mathbf{e}_1$ or, equivalently, $\mathbf{x} = \mathbf{R}\mathbf{e}_1$ (set $\mathbf{R} = \mathbf{I}$ if $\mathbf{x} = \mathbf{e}_1$). Thus $\mathbf{x}$ is the first column in $\mathbf{R}$, so $\mathbf{R} = \left(\mathbf{x} \,|\, \mathbf{V}\right)$, and

$$\mathbf{R}\mathbf{A}\mathbf{R} = \mathbf{R}\mathbf{A}\left(\mathbf{x} \,|\, \mathbf{V}\right) = \mathbf{R}\left(\lambda\mathbf{x} \,|\, \mathbf{A}\mathbf{V}\right) = \left(\lambda\mathbf{e}_1 \,|\, \mathbf{R}\mathbf{A}\mathbf{V}\right) = \begin{pmatrix}\lambda & \mathbf{x}^*\mathbf{A}\mathbf{V}\\ \mathbf{0} & \mathbf{V}^*\mathbf{A}\mathbf{V}\end{pmatrix}.$$

Since $\mathbf{V}^*\mathbf{AV}$ is $n-1 \times n-1$, the induction hypothesis insures that there exists a unitary matrix $\mathbf{Q}$ such that $\mathbf{Q}^*(\mathbf{V}^*\mathbf{AV})\mathbf{Q} = \tilde{\mathbf{T}}$ is upper triangular. If $\mathbf{U} = \mathbf{R}\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix}$, then $\mathbf{U}$ is unitary (because $\mathbf{U}^* = \mathbf{U}^{-1}$), and

$$\mathbf{U}^*\mathbf{AU} = \begin{pmatrix} \lambda & \mathbf{x}^*\mathbf{AVQ} \\ \mathbf{0} & \mathbf{Q}^*\mathbf{V}^*\mathbf{AVQ} \end{pmatrix} = \begin{pmatrix} \lambda & \mathbf{x}^*\mathbf{AVQ} \\ \mathbf{0} & \tilde{\mathbf{T}} \end{pmatrix} = \mathbf{T}$$

is upper triangular. Since similar matrices have the same eigenvalues, and since the eigenvalues of a triangular matrix are its diagonal entries (Exercise 7.1.3), the diagonal entries of $\mathbf{T}$ must be the eigenvalues of $\mathbf{A}$. ∎

Example 7.2.2

The Cayley–Hamilton[71] **theorem** asserts that every square matrix satisfies its own characteristic equation $p(\lambda) = 0$. That is, $p(\mathbf{A}) = \mathbf{0}$.

Problem: Show how the Cayley–Hamilton theorem follows from Schur's triangularization theorem.

Solution: Schur's theorem insures the existence of a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{AU} = \mathbf{T}$ is triangular, and the development allows for the eigenvalues $\mathbf{A}$ to appear in any given order on the diagonal of $\mathbf{T}$. So, if $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ with λ_i repeated a_i times, then there is a unitary $\mathbf{U}$ such that

$$\mathbf{U}^*\mathbf{AU} = \mathbf{T} = \begin{pmatrix} \mathbf{T}_1 & \star & \cdots & \star \\ & \mathbf{T}_2 & \cdots & \star \\ & & \ddots & \vdots \\ & & & \mathbf{T}_k \end{pmatrix}, \quad \text{where } \mathbf{T}_i = \underbrace{\begin{pmatrix} \lambda_i & \star & \cdots & \star \\ & \lambda_i & \cdots & \star \\ & & \ddots & \vdots \\ & & & \lambda_i \end{pmatrix}}_{a_i \times a_i}.$$

Consequently, $(\mathbf{T}_i - \lambda_i\mathbf{I})^{a_i} = \mathbf{0}$, so $(\mathbf{T} - \lambda_i\mathbf{I})^{a_i}$ has the form

$$(\mathbf{T} - \lambda_i\mathbf{I})^{a_i} = \begin{pmatrix} \star & \cdots & \star & \cdots & \star \\ & \ddots & \vdots & & \vdots \\ & & \mathbf{0} & \cdots & \star \\ & & & \ddots & \vdots \\ & & & & \star \end{pmatrix} \longleftarrow i^{th} \text{ row of blocks.}$$

[71] William Rowan Hamilton (1805–1865), an Irish mathematical astronomer, established this result in 1853 for his quaternions, matrices of the form $\begin{pmatrix} a+b\mathrm{i} & c+d\mathrm{i} \\ -c+d\mathrm{i} & a-b\mathrm{i} \end{pmatrix}$ that resulted from his attempt to generalize complex numbers. In 1858 Arthur Cayley (p. 80) enunciated the general result, but his argument was simply to make direct computations for 2×2 and 3×3 matrices. Cayley apparently didn't appreciate the subtleties of the result because he stated that a formal proof "was not necessary." Hamilton's quaternions took shape in his mind while walking with his wife along the Royal Canal in Dublin, and he was so inspired that he stopped to carve his idea in the stone of the Brougham Bridge. He believed quaternions would revolutionize mathematical physics, and he spent the rest of his life working on them. But the world did not agree. Hamilton became an unhappy man addicted to alcohol who is reported to have died from a severe attack of gout.

This form insures that $(\mathbf{T}-\lambda_1\mathbf{I})^{a_1}(\mathbf{T}-\lambda_2\mathbf{I})^{a_2}\cdots(\mathbf{T}-\lambda_k\mathbf{I})^{a_k}=\mathbf{0}$. The characteristic equation for $\mathbf{A}$ is $p(\lambda)=(\lambda-\lambda_1)^{a_1}(\lambda-\lambda_2)^{a_2}\cdots(\lambda-\lambda_k)^{a_k}=0$, so

$$\begin{aligned}\mathbf{U}^*p(\mathbf{A})\mathbf{U}&=\mathbf{U}^*(\mathbf{A}-\lambda_1\mathbf{I})^{a_1}(\mathbf{A}-\lambda_2\mathbf{I})^{a_2}\cdots(\mathbf{A}-\lambda_k\mathbf{I})^{a_k}\mathbf{U}\\&=(\mathbf{T}-\lambda_1\mathbf{I})^{a_1}(\mathbf{T}-\lambda_2\mathbf{I})^{a_2}\cdots(\mathbf{T}-\lambda_k\mathbf{I})^{a_k}=\mathbf{0},\end{aligned}$$

and thus $p(\mathbf{A})=\mathbf{0}$. **Note:** A completely different approach to the Cayley–Hamilton theorem is discussed on p. 532.

Schur's theorem is not the complete story on triangularizing by similarity. By allowing nonunitary similarity transformations, the structure of the upper-triangular matrix $\mathbf{T}$ can be simplified to contain zeros everywhere except on the diagonal and the superdiagonal (the diagonal immediately above the main diagonal). This is the Jordan form developed on p. 590, but some of the seeds are sown here.

Multiplicities

For $\lambda\in\sigma(\mathbf{A})=\{\lambda_1,\lambda_2,\ldots,\lambda_s\}$, we adopt the following definitions.

- The ***algebraic multiplicity*** of λ is the number of times it is repeated as a root of the characteristic polynomial. In other words, $alg\ mult_{\mathbf{A}}(\lambda_i)=a_i$ if and only if $(x-\lambda_1)^{a_1}\cdots(x-\lambda_s)^{a_s}=0$ is the characteristic equation for $\mathbf{A}$.
- When $alg\ mult_{\mathbf{A}}(\lambda)=1$, λ is called a ***simple eigenvalue.***
- The ***geometric multiplicity*** of λ is $\dim N(\mathbf{A}-\lambda\mathbf{I})$. In other words, $geo\ mult_{\mathbf{A}}(\lambda)$ is the maximal number of linearly independent eigenvectors associated with λ.
- Eigenvalues such that $alg\ mult_{\mathbf{A}}(\lambda)=geo\ mult_{\mathbf{A}}(\lambda)$ are called ***semisimple eigenvalues*** of $\mathbf{A}$. It follows from (7.2.2) on p. 511 that a simple eigenvalue is always semisimple, but not conversely.

Example 7.2.3

The algebraic and geometric multiplicity need not agree. For example, the nilpotent matrix $\mathbf{A}=\begin{pmatrix}0&1\\0&0\end{pmatrix}$ in (7.2.1) has only one distinct eigenvalue, $\lambda=0$, that is repeated twice, so $alg\ mult_{\mathbf{A}}(0)=2$. But

$$\dim N(\mathbf{A}-0\mathbf{I})=\dim N(\mathbf{A})=1\implies geo\ mult_{\mathbf{A}}(0)=1.$$

In other words, there is only one linearly independent eigenvector associated with $\lambda=0$ even though $\lambda=0$ is repeated twice as an eigenvalue.

Example 7.2.3 shows that $geo\ mult_{\mathbf{A}}(\lambda)<alg\ mult_{\mathbf{A}}(\lambda)$ is possible. However, the inequality can never go in the reverse direction.

Multiplicity Inequality

For every $\mathbf{A} \in \mathcal{C}^{n\times n}$, and for each $\lambda \in \sigma(\mathbf{A})$,

$$geo\ mult_{\mathbf{A}}(\lambda) \leq alg\ mult_{\mathbf{A}}(\lambda). \tag{7.2.2}$$

Proof. Suppose $alg\ mult_{\mathbf{A}}(\lambda) = k$. Schur's triangularization theorem (p. 508) insures the existence of a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}_{n\times n}\mathbf{U} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix}$, where $\mathbf{T}_{11}$ is a $k \times k$ upper-triangular matrix whose diagonal entries are equal to λ, and $\mathbf{T}_{22}$ is an $n-k \times n-k$ upper-triangular matrix with $\lambda \notin \sigma(\mathbf{T}_{22})$. Consequently, $\mathbf{T}_{22} - \lambda\mathbf{I}$ is nonsingular, so

$$\begin{aligned} rank(\mathbf{A} - \lambda\mathbf{I}) = rank(\mathbf{U}^*(\mathbf{A} - \lambda\mathbf{I})\mathbf{U}) &= rank\begin{pmatrix} \mathbf{T}_{11} - \lambda\mathbf{I} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} - \lambda\mathbf{I} \end{pmatrix} \\ &\geq rank(\mathbf{T}_{22} - \lambda\mathbf{I}) = n - k. \end{aligned}$$

The inequality follows from the fact that the rank of a matrix is at least as great as the rank of any submatrix—recall the result on p. 215. Therefore,

$$alg\ mult_{\mathbf{A}}(\lambda) = k \geq n - rank(\mathbf{A} - \lambda\mathbf{I}) = \dim N(\mathbf{A} - \lambda\mathbf{I}) = geo\ mult_{\mathbf{A}}(\lambda). \quad \blacksquare$$

Determining whether or not $\mathbf{A}_{n\times n}$ is diagonalizable is equivalent to determining whether or not $\mathbf{A}$ has a complete linearly independent set of eigenvectors, and this can be done if you are willing and able to compute all of the eigenvalues and eigenvectors for $\mathbf{A}$. But this brute force approach can be a monumental task. Fortunately, there are some theoretical tools to help determine how many linearly independent eigenvectors a given matrix possesses.

Independent Eigenvectors

Let $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ be a set of distinct eigenvalues for $\mathbf{A}$.

- If $\{(\lambda_1, \mathbf{x}_1), (\lambda_2, \mathbf{x}_2), \ldots, (\lambda_k, \mathbf{x}_k)\}$ is a set of eigenpairs for $\mathbf{A}$, then $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is a linearly independent set. (7.2.3)
- If $\mathcal{B}_i$ is a basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$, then $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ is a linearly independent set. (7.2.4)

Proof of (7.2.3). Suppose $\mathcal{S}$ is a dependent set. If the vectors in $\mathcal{S}$ are arranged so that $\mathcal{M} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ is a maximal linearly independent subset, then

$$\mathbf{x}_{r+1} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_i,$$

and multiplication on the left by $\mathbf{A} - \lambda_{r+1}\mathbf{I}$ produces

$$\mathbf{0} = \sum_{i=1}^{r} \alpha_i \left(\mathbf{A}\mathbf{x}_i - \lambda_{r+1}\mathbf{x}_i\right) = \sum_{i=1}^{r} \alpha_i \left(\lambda_i - \lambda_{r+1}\right)\mathbf{x}_i.$$

Because $\mathcal{M}$ is linearly independent, $\alpha_i\left(\lambda_i - \lambda_{r+1}\right) = 0$ for each i. Consequently, $\alpha_i = 0$ for each i (because the eigenvalues are distinct), and hence $\mathbf{x}_{r+1} = \mathbf{0}$. But this is impossible because eigenvectors are nonzero. Therefore, the supposition that $\mathcal{S}$ is a dependent set must be false. ∎

Proof of (7.2.4). The result of Exercise 5.9.14 guarantees that $\mathcal{B}$ is linearly independent if and only if

$$\mathcal{M}_j = N\left(\mathbf{A} - \lambda_j\mathbf{I}\right) \cap \Big[N\left(\mathbf{A} - \lambda_1\mathbf{I}\right) + N\left(\mathbf{A} - \lambda_2\mathbf{I}\right) + \cdots + N\left(\mathbf{A} - \lambda_{j-1}\mathbf{I}\right)\Big] = \mathbf{0}$$

for each $j = 1, 2, \ldots, k$. Suppose we have $\mathbf{0} \neq \mathbf{x} \in \mathcal{M}_j$ for some j. Then $\mathbf{A}\mathbf{x} = \lambda_j\mathbf{x}$ and $\mathbf{x} = \mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_{j-1}$ for $\mathbf{v}_i \in N\left(\mathbf{A} - \lambda_i\mathbf{I}\right)$, which implies

$$\sum_{i=1}^{j-1}(\lambda_i - \lambda_j)\mathbf{v}_i = \sum_{i=1}^{j-1}\lambda_i\mathbf{v}_i - \lambda_j\sum_{i=1}^{j-1}\mathbf{v}_i = \mathbf{A}\mathbf{x} - \lambda_j\mathbf{x} = \mathbf{0}.$$

By (7.2.3), the $\mathbf{v}_i$'s are linearly independent, and hence $\lambda_i - \lambda_j = 0$ for each $i = 1, 2, \ldots, j-1$. But this is impossible because the eigenvalues are distinct. Therefore, $\mathcal{M}_j = \mathbf{0}$ for each j, and thus $\mathcal{B}$ is linearly independent. ∎

These results lead to the following characterization of diagonalizability.

Diagonalizability and Multiplicities

A matrix $\mathbf{A}_{n\times n}$ is diagonalizable if and only if

$$\textit{geo mult}_{\mathbf{A}}(\lambda) = \textit{alg mult}_{\mathbf{A}}(\lambda) \tag{7.2.5}$$

for each $\lambda \in \sigma(\mathbf{A})$—i.e., if and only if every eigenvalue is semisimple.

Proof. Suppose $\textit{geo mult}_{\mathbf{A}}(\lambda_i) = \textit{alg mult}_{\mathbf{A}}(\lambda_i) = a_i$ for each eigenvalue λ_i. If there are k distinct eigenvalues, and if $\mathcal{B}_i$ is a basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$, then $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ contains $\sum_{i=1}^{k} a_i = n$ vectors. We just proved in (7.2.4) that $\mathcal{B}$ is a linearly independent set, so $\mathcal{B}$ represents a complete set of linearly independent eigenvectors of $\mathbf{A}$, and we know this insures that $\mathbf{A}$ must be diagonalizable. Conversely, if $\mathbf{A}$ is diagonalizable, and if λ is an eigenvalue for $\mathbf{A}$ with $\textit{alg mult}_{\mathbf{A}}(\lambda) = a$, then there is a nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D} = \begin{pmatrix} \lambda\mathbf{I}_{a\times a} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix},$$

where $\lambda \notin \sigma(\mathbf{B})$. Consequently,

$$\textit{rank}\,(\mathbf{A} - \lambda\mathbf{I}) = \textit{rank}\,\mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} - \lambda\mathbf{I} \end{pmatrix}\mathbf{P}^{-1} = \textit{rank}\,(\mathbf{B} - \lambda\mathbf{I}) = n - a,$$

and thus

$$\textit{geo mult}_{\mathbf{A}}(\lambda) = \dim N(\mathbf{A} - \lambda\mathbf{I}) = n - \textit{rank}\,(\mathbf{A} - \lambda\mathbf{I}) = a = \textit{alg mult}_{\mathbf{A}}(\lambda). \quad \blacksquare$$

Example 7.2.4

Problem: Determine if either of the following matrices is diagonalizable:

$$\mathbf{A} = \begin{pmatrix} -1 & -1 & -2 \\ 8 & -11 & -8 \\ -10 & 11 & 7 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}.$$

Solution: Each matrix has exactly the same characteristic equation

$$\lambda^3 + 5\lambda^2 + 3\lambda - 9 = (\lambda - 1)(\lambda + 3)^2 = 0,$$

so $\sigma(\mathbf{A}) = \{1, -3\} = \sigma(\mathbf{B})$, where $\lambda = 1$ has algebraic multiplicity 1 and $\lambda = -3$ has algebraic multiplicity 2. Since

$$\textit{geo mult}_{\mathbf{A}}(-3) = \dim N(\mathbf{A} + 3\mathbf{I}) = 1 < \textit{alg mult}_{\mathbf{A}}(-3),$$

$\mathbf{A}$ is *not* diagonalizable. On the other hand,

$$\textit{geo mult}_{\mathbf{B}}(-3) = \dim N(\mathbf{B} + 3\mathbf{I}) = 2 = \textit{alg mult}_{\mathbf{B}}(-3),$$

and $\textit{geo mult}_{\mathbf{B}}(1) = 1 = \textit{alg mult}_{\mathbf{B}}(1)$, so $\mathbf{B}$ *is* diagonalizable.

If $\mathbf{A}_{n\times n}$ happens to have n distinct eigenvalues, then each eigenvalue is simple. This means that $\textit{geo mult}_{\mathbf{A}}(\lambda) = \textit{alg mult}_{\mathbf{A}}(\lambda) = 1$ for each λ, so (7.2.5) produces the following corollary guaranteeing diagonalizability.

Distinct Eigenvalues

If no eigenvalue of $\mathbf{A}$ is repeated, then $\mathbf{A}$ is diagonalizable. (7.2.6)
Caution! The converse is not true—see Example 7.2.4.

Example 7.2.5

Toeplitz[72] **matrices** have constant entries on each diagonal parallel to the main diagonal. For example, a 4×4 Toeplitz matrix $\mathbf{T}$ along with a *tridiagonal* Toeplitz matrix $\mathbf{A}$ are shown below:

$$\mathbf{T} = \begin{pmatrix} t_0 & t_1 & t_2 & t_3 \\ t_{-1} & t_0 & t_1 & t_2 \\ t_{-2} & t_{-1} & t_0 & t_1 \\ t_{-3} & t_{-2} & t_{-1} & t_0 \end{pmatrix}, \qquad \mathbf{A} = \begin{pmatrix} t_0 & t_1 & 0 & 0 \\ t_{-1} & t_0 & t_1 & 0 \\ 0 & t_{-1} & t_0 & t_1 \\ 0 & 0 & t_{-1} & t_0 \end{pmatrix}.$$

Toeplitz structures occur naturally in a variety of applications, and tridiagonal Toeplitz matrices are commonly the result of discretizing differential equation problems—e.g., see §1.4 (p. 18) and Example 7.6.1 (p. 559). The Toeplitz structure is rich in special properties, but tridiagonal Toeplitz matrices are particularly nice because they are among the few nontrivial structures that admit formulas for their eigenvalues and eigenvectors.

Problem: Show that the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} b & a & & & \\ c & b & a & & \\ & \ddots & \ddots & \ddots & \\ & & c & b & a \\ & & & c & b \end{pmatrix}_{n \times n} \qquad \text{with} \qquad a \neq 0 \neq c$$

are given by

$$\lambda_j = b + 2a\sqrt{c/a}\,\cos\left(\frac{j\pi}{n+1}\right) \qquad \text{and} \qquad \mathbf{x}_j = \begin{pmatrix} (c/a)^{1/2}\sin\left(1j\pi/(n+1)\right) \\ (c/a)^{2/2}\sin\left(2j\pi/(n+1)\right) \\ (c/a)^{3/2}\sin\left(3j\pi/(n+1)\right) \\ \vdots \\ (c/a)^{n/2}\sin\left(nj\pi/(n+1)\right) \end{pmatrix}$$

[72] Otto Toeplitz (1881–1940) was a professor in Bonn, Germany, but because of his Jewish background he was dismissed from his chair by the Nazis in 1933. In addition to the matrix that bears his name, Toeplitz is known for his general theory of infinite-dimensional spaces developed in the 1930s.

for $j = 1, 2, \ldots, n,$ and conclude that $\mathbf{A}$ is diagonalizable.

Solution: For an eigenpair $(\lambda, \mathbf{x}),$ the components in $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}$ are $cx_{k-1}+(b-\lambda)x_k+ax_{k+1} = 0,\ k = 1, \ldots, n$ with $x_0 = x_{n+1} = 0$ or, equivalently,

$$x_{k+2}+\left(\frac{b-\lambda}{a}\right)x_{k+1}+\left(\frac{c}{a}\right)x_k = 0 \quad \text{for } k = 0, \ldots, n-1 \text{ with } x_0 = x_{n+1} = 0.$$

These are second-order homogeneous difference equations, and solving them is similar to solving analogous differential equations. The technique is to seek solutions of the form $x_k = \xi r^k$ for constants ξ and r. This produces the quadratic equation $r^2 + (b-\lambda)r/a + c/a = 0$ with roots r_1 and $r_2,$ and it can be argued that the general solution of $x_{k+2} + ((b-\lambda)/a)x_{k+1} + (c/a)x_k = 0$ is

$$x_k = \begin{cases} \alpha r_1^k + \beta r_2^k & \text{if } r_1 \neq r_2, \\ \alpha\rho^k + \beta k\rho^k & \text{if } r_1 = r_2 = \rho, \end{cases} \qquad \text{where } \alpha \text{ and } \beta \text{ are arbitrary constants.}$$

For the eigenvalue problem at hand, r_1 and r_2 must be distinct—otherwise $x_k = \alpha\rho^k + \beta k\rho^k,$ and $x_0 = x_{n+1} = 0$ implies each $x_k = 0,$ which is impossible because $\mathbf{x}$ is an eigenvector. Hence $x_k = \alpha r_1^k + \beta r_2^k,$ and $x_0 = x_{n+1} = 0$ yields

$$\left\{\begin{array}{l} 0 = \alpha + \beta \\ 0 = \alpha r_1^{n+1} + \beta r_2^{n+1} \end{array}\right\} \implies \left(\frac{r_1}{r_2}\right)^{n+1} = \frac{-\beta}{\alpha} = 1 \implies \frac{r_1}{r_2} = \mathrm{e}^{\mathrm{i}2\pi j/(n+1)},$$

so $r_1 = r_2\mathrm{e}^{\mathrm{i}2\pi j/(n+1)}$ for some $1 \leq j \leq n.$ Couple this with

$$r^2 + \frac{(b-\lambda)r}{a} + \frac{c}{a} = (r-r_1)(r-r_2) \implies \left\{\begin{array}{l} r_1 r_2 = c/a \\ r_1 + r_2 = -(b-\lambda)/a \end{array}\right.$$

to conclude that $r_1 = \sqrt{c/a}\,\mathrm{e}^{\mathrm{i}\pi j/(n+1)},$ $r_2 = \sqrt{c/a}\,\mathrm{e}^{-\mathrm{i}\pi j/(n+1)},$ and

$$\lambda = b + a\sqrt{c/a}\left(\mathrm{e}^{\mathrm{i}\pi j/(n+1)} + \mathrm{e}^{-\mathrm{i}\pi j/(n+1)}\right) = b + 2a\sqrt{c/a}\,\cos\left(\frac{j\pi}{n+1}\right).$$

Therefore, the eigenvalues of $\mathbf{A}$ must be given by

$$\lambda_j = b + 2a\sqrt{c/a}\,\cos\left(\frac{j\pi}{n+1}\right), \quad j = 1, 2, \ldots, n.$$

Since these λ_j's are all distinct ($\cos\theta$ is a strictly decreasing function of θ on $(0, \pi),$ and $a \neq 0 \neq c$), $\mathbf{A}$ must be diagonalizable—recall (7.2.6). Finally, the k^{th} component of any eigenvector associated with λ_j satisfies $x_k = \alpha r_1^k + \beta r_2^k$ with $\alpha + \beta = 0,$ so

$$x_k = \alpha\left(\frac{c}{a}\right)^{k/2}\left(\mathrm{e}^{\mathrm{i}\pi jk/(n+1)} - \mathrm{e}^{-\mathrm{i}\pi jk/(n+1)}\right) = 2\mathrm{i}\alpha\left(\frac{c}{a}\right)^{k/2}\sin\left(\frac{jk\pi}{n+1}\right).$$

Setting $\alpha = 1/2\mathrm{i}$ yields a particular eigenvector associated with λ_j as

$$\mathbf{x}_j = \begin{pmatrix} (c/a)^{1/2} \sin(1j\pi/(n+1)) \\ (c/a)^{2/2} \sin(2j\pi/(n+1)) \\ (c/a)^{3/2} \sin(3j\pi/(n+1)) \\ \vdots \\ (c/a)^{n/2} \sin(nj\pi/(n+1)) \end{pmatrix}.$$

Because the λ_j's are distinct, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a complete linearly independent set—recall (7.2.3)—so $\mathbf{P} = \begin{pmatrix} \mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n \end{pmatrix}$ diagonalizes $\mathbf{A}$.

It's often the case that a right-hand and left-hand eigenvector for some eigenvalue is known. Rather than starting from scratch to find additional eigenpairs, the known information can be used to reduce or "deflate" the problem to a smaller one as described in the following example.

Example 7.2.6

Deflation. Suppose that right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$ for an eigenvalue λ of $\mathbf{A} \in \Re^{n\times n}$ are already known, so $\mathbf{Ax} = \lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*$. Furthermore, suppose $\mathbf{y}^*\mathbf{x} \neq 0$—such eigenvectors are guaranteed to exist if λ is simple or if $\mathbf{A}$ is diagonalizable (Exercises 7.2.23 and 7.2.22).

Problem: Use $\mathbf{x}$ and $\mathbf{y}^*$ to deflate the size of the remaining eigenvalue problem.

Solution: Scale $\mathbf{x}$ and $\mathbf{y}^*$ so that $\mathbf{y}^*\mathbf{x} = 1$, and construct $\mathbf{X}_{n\times n-1}$ so that its columns are an orthonormal basis for $\mathbf{y}^{\perp}$. An easy way of doing this is to build a reflector $\mathbf{R} = \begin{bmatrix} \tilde{\mathbf{y}} \,|\, \mathbf{X} \end{bmatrix}$ having $\tilde{\mathbf{y}} = \mathbf{y}/\|\mathbf{y}\|_2$ as its first column as described on p. 325. If $\mathbf{P} = \begin{bmatrix} \mathbf{x} \,|\, \mathbf{X} \end{bmatrix}$, then straightforward multiplication shows that

$$\mathbf{P}^{-1} = \begin{pmatrix} \mathbf{y}^* \\ \mathbf{X}^*(\mathbf{I} - \mathbf{x}\mathbf{y}^*) \end{pmatrix} \quad \text{and} \quad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix},$$

where $\mathbf{B} = \mathbf{X}^*\mathbf{A}\mathbf{X}$ is $n-1 \times n-1$. The eigenvalues of $\mathbf{B}$ constitute the remaining eigenvalues of $\mathbf{A}$ (Exercise 7.1.4), and thus an $n \times n$ eigenvalue problem is deflated to become one of size $n-1 \times n-1$.

Note: When $\mathbf{A}$ is symmetric, we can take $\mathbf{x} = \mathbf{y}$ to be an eigenvector with $\|\mathbf{x}\|_2 = 1$, so $\mathbf{P} = \mathbf{R} = \mathbf{R}^{-1}$, and $\mathbf{RAR} = \begin{pmatrix} \lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$ in which $\mathbf{B} = \mathbf{B}^T$.

An elegant and more geometrical way of expressing diagonalizability is now presented to help simplify subsequent analyses and pave the way for extensions.

Spectral Theorem for Diagonalizable Matrices

A matrix $\mathbf{A}_{n\times n}$ with spectrum $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ is diagonalizable if and only if there exist matrices $\{\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_k\}$ such that

$$\mathbf{A} = \lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \cdots + \lambda_k\mathbf{G}_k, \tag{7.2.7}$$

where the $\mathbf{G}_i$'s have the following properties.

- $\mathbf{G}_i$ is the projector onto $N(\mathbf{A} - \lambda_i\mathbf{I})$ along $R(\mathbf{A} - \lambda_i\mathbf{I})$. (7.2.8)
- $\mathbf{G}_i\mathbf{G}_j = \mathbf{0}$ whenever $i \neq j$. (7.2.9)
- $\mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_k = \mathbf{I}$. (7.2.10)

The expansion (7.2.7) is known as the ***spectral decomposition*** of $\mathbf{A}$, and the $\mathbf{G}_i$'s are called the ***spectral projectors*** associated with $\mathbf{A}$.

Proof. If $\mathbf{A}$ is diagonalizable, and if $\mathbf{X}_i$ is a matrix whose columns form a basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$, then $\mathbf{P} = \left(\mathbf{X}_1 \,|\, \mathbf{X}_2 \,|\, \cdots \,|\, \mathbf{X}_k\right)$ is nonsingular. If $\mathbf{P}^{-1}$ is partitioned in a conformable manner, then we must have

$$\mathbf{A} = \mathbf{PDP}^{-1} = \left(\mathbf{X}_1 \,|\, \mathbf{X}_2 \,|\, \cdots \,|\, \mathbf{X}_k\right) \begin{pmatrix} \lambda_1\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_k\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1^T \\ \hline \mathbf{Y}_2^T \\ \hline \vdots \\ \hline \mathbf{Y}_k^T \end{pmatrix} \tag{7.2.11}$$

$$= \lambda_1\mathbf{X}_1\mathbf{Y}_1^T + \lambda_2\mathbf{X}_2\mathbf{Y}_2^T + \cdots + \lambda_k\mathbf{X}_k\mathbf{Y}_k^T$$
$$= \lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \cdots + \lambda_k\mathbf{G}_k.$$

For $\mathbf{G}_i = \mathbf{X}_i\mathbf{Y}_i^T$, the statement $\mathbf{PP}^{-1} = \mathbf{I}$ translates to $\sum_{i=1}^{k}\mathbf{G}_i = \mathbf{I}$, and

$$\mathbf{P}^{-1}\mathbf{P} = \mathbf{I} \implies \mathbf{Y}_i^T\mathbf{X}_j = \begin{cases} \mathbf{I} & \text{when } i = j, \\ \mathbf{0} & \text{when } i \neq j, \end{cases} \implies \begin{cases} \mathbf{G}_i^2 = \mathbf{G}_i, \\ \mathbf{G}_i\mathbf{G}_j = \mathbf{0} & \text{when } i \neq j. \end{cases}$$

To establish that $R(\mathbf{G}_i) = N(\mathbf{A} - \lambda_i\mathbf{I})$, use $R(\mathbf{AB}) \subseteq R(\mathbf{A})$ (Exercise 4.2.12) and $\mathbf{Y}_i^T\mathbf{X}_i = \mathbf{I}$ to write

$$R(\mathbf{G}_i) = R(\mathbf{X}_i\mathbf{Y}_i^T) \subseteq R(\mathbf{X}_i) = R(\mathbf{X}_i\mathbf{Y}_i^T\mathbf{X}_i) = R(\mathbf{G}_i\mathbf{X}_i) \subseteq R(\mathbf{G}_i).$$

Thus $R(\mathbf{G}_i) = R(\mathbf{X}_i) = N(\mathbf{A} - \lambda_i\mathbf{I})$. To show $N(\mathbf{G}_i) = R(\mathbf{A} - \lambda_i\mathbf{I})$, use $\mathbf{A} = \sum_{j=1}^{k}\lambda_j\mathbf{G}_j$ with the already established properties of the $\mathbf{G}_i$'s to conclude

$$\mathbf{G}_i(\mathbf{A} - \lambda_i\mathbf{I}) = \mathbf{G}_i\left(\sum_{j=1}^{k}\lambda_j\mathbf{G}_j - \lambda_i\sum_{j=1}^{k}\mathbf{G}_j\right) = \mathbf{0} \implies R(\mathbf{A} - \lambda_i\mathbf{I}) \subseteq N(\mathbf{G}_i).$$

But we already know that $N(\mathbf{A}-\lambda_i\mathbf{I}) = R(\mathbf{G}_i)$, so

$$\dim R(\mathbf{A}-\lambda_i\mathbf{I}) = n - \dim N(\mathbf{A}-\lambda_i\mathbf{I}) = n - \dim R(\mathbf{G}_i) = \dim N(\mathbf{G}_i),$$

and therefore, by (4.4.6), $R(\mathbf{A}-\lambda_i\mathbf{I}) = N(\mathbf{G}_i)$. Conversely, if there exist matrices $\mathbf{G}_i$ satisfying (7.2.8)–(7.2.10), then $\mathbf{A}$ must be diagonalizable. To see this, note that (7.2.8) insures $\dim R(\mathbf{G}_i) = \dim N(\mathbf{A}-\lambda_i\mathbf{I}) = \textit{geo mult}_{\mathbf{A}}(\lambda_i)$, while (7.2.9) implies $R(\mathbf{G}_i) \cap R(\mathbf{G}_j) = \mathbf{0}$ and $R\big(\sum_{i=1}^{k}\mathbf{G}_i\big) = \sum_{i=1}^{k} R(\mathbf{G}_i)$ (Exercise 5.9.17). Use these with (7.2.10) in the formula for the dimension of a sum (4.4.19) to write

$$\begin{aligned} n = \dim R(\mathbf{I}) &= \dim R(\mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_k) \\ &= \dim\left[R(\mathbf{G}_1) + R(\mathbf{G}_2) + \cdots + R(\mathbf{G}_k)\right] \\ &= \dim R(\mathbf{G}_1) + \dim R(\mathbf{G}_2) + \cdots + \dim R(\mathbf{G}_k) \\ &= \textit{geo mult}_{\mathbf{A}}(\lambda_1) + \textit{geo mult}_{\mathbf{A}}(\lambda_2) + \cdots + \textit{geo mult}_{\mathbf{A}}(\lambda_k). \end{aligned}$$

Since $\textit{geo mult}_{\mathbf{A}}(\lambda_i) \le \textit{alg mult}_{\mathbf{A}}(\lambda_i)$ and $\sum_{i=1}^{k} \textit{alg mult}_{\mathbf{A}}(\lambda_i) = n$, the above equation insures that $\textit{geo mult}_{\mathbf{A}}(\lambda_i) = \textit{alg mult}_{\mathbf{A}}(\lambda_i)$ for each i, and, by (7.2.5), this means $\mathbf{A}$ is diagonalizable. ∎

Simple Eigenvalues and Projectors

If $\mathbf{x}$ and $\mathbf{y}^*$ are respective right-hand and left-hand eigenvectors associated with a *simple* eigenvalue $\lambda \in \sigma(\mathbf{A})$, then

$$\mathbf{G} = \mathbf{x}\mathbf{y}^*/\mathbf{y}^*\mathbf{x} \tag{7.2.12}$$

is the projector onto $N(\mathbf{A}-\lambda\mathbf{I})$ along $R(\mathbf{A}-\lambda\mathbf{I})$. In the context of the spectral theorem (p. 517), this means that $\mathbf{G}$ is the spectral projector associated with λ.

Proof. It's not difficult to prove $\mathbf{y}^*\mathbf{x} \neq 0$ (Exercise 7.2.23), and it's clear that $\mathbf{G}$ is a projector because $\mathbf{G}^2 = \mathbf{x}(\mathbf{y}^*\mathbf{x})\mathbf{y}^*/(\mathbf{y}^*\mathbf{x})^2 = \mathbf{G}$. Now determine $R(\mathbf{G})$. The image of any $\mathbf{z}$ is $\mathbf{G}\mathbf{z} = \alpha\mathbf{x}$ with $\alpha = \mathbf{y}^*\mathbf{z}/\mathbf{y}^*\mathbf{x}$, so

$$R(\mathbf{G}) \subseteq \textit{span}\{\mathbf{x}\} = N(\mathbf{A}-\lambda\mathbf{I}) \quad \text{and} \quad \dim R(\mathbf{G}) = 1 = \dim N(\mathbf{A}-\lambda\mathbf{I}).$$

Thus $R(\mathbf{G}) = N(\mathbf{A}-\lambda\mathbf{I})$. To find $N(\mathbf{G})$, recall $N(\mathbf{G}) = R(\mathbf{I}-\mathbf{G})$ (see (5.9.11), p. 386), and observe that $\mathbf{y}^*(\mathbf{A}-\lambda\mathbf{I}) = \mathbf{0} \implies \mathbf{y}^*(\mathbf{I}-\mathbf{G}) = \mathbf{0}$, so

$$R(\mathbf{A}-\lambda\mathbf{I})^{\perp} \subseteq R(\mathbf{I}-\mathbf{G})^{\perp} = N(\mathbf{G})^{\perp} \Longrightarrow N(\mathbf{G}) \subseteq R(\mathbf{A}-\lambda\mathbf{I}) \text{ (Exercise 5.11.5).}$$

But $\dim N(\mathbf{G}) = n - \dim R(\mathbf{G}) = n-1 = n - \dim N(\mathbf{A}-\lambda\mathbf{I}) = \dim R(\mathbf{A}-\lambda\mathbf{I})$, so $N(\mathbf{G}) = R(\mathbf{A}-\lambda\mathbf{I})$. ∎

Example 7.2.7

Problem: Determine the spectral projectors for $\mathbf{A} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}$.

Solution: This is the diagonalizable matrix from Example 7.2.1 (p. 507). Since there are two distinct eigenvalues, $\lambda_1 = 1$ and $\lambda_2 = -3$, there are two spectral projectors,

$$\begin{aligned} \mathbf{G}_1 &= \text{ the projector onto } N(\mathbf{A} - 1\mathbf{I}) \text{ along } R(\mathbf{A} - 1\mathbf{I}), \\ \mathbf{G}_2 &= \text{ the projector onto } N(\mathbf{A} + 3\mathbf{I}) \text{ along } R(\mathbf{A} + 3\mathbf{I}). \end{aligned}$$

There are several different ways to find these projectors.

1. Compute bases for the necessary nullspaces and ranges, and use (5.9.12).

2. Compute $\mathbf{G}_i = \mathbf{X}_i\mathbf{Y}_i^T$ as described in (7.2.11). The required computations are essentially the same as those needed above. Since much of the work has already been done in Example 7.2.1, let's complete the arithmetic. We have

$$\mathbf{P} = \left(\begin{array}{c|cc} 1 & 1 & 1 \\ 2 & 1 & 0 \\ -2 & 0 & 1 \end{array}\right) = \Big(\mathbf{X}_1 \,|\, \mathbf{X}_2\Big), \quad \mathbf{P}^{-1} = \left(\begin{array}{ccc} 1 & -1 & -1 \\ \hline -2 & 3 & 2 \\ 2 & -2 & -1 \end{array}\right) = \begin{pmatrix} \mathbf{Y}_1^T \\ \hline \mathbf{Y}_2^T \end{pmatrix},$$

so

$$\mathbf{G}_1 = \mathbf{X}_1\mathbf{Y}_1^T = \begin{pmatrix} 1 & -1 & -1 \\ 2 & -2 & -2 \\ -2 & 2 & 2 \end{pmatrix}, \quad \mathbf{G}_2 = \mathbf{X}_2\mathbf{Y}_2^T = \begin{pmatrix} 0 & 1 & 1 \\ -2 & 3 & 2 \\ 2 & -2 & -1 \end{pmatrix}.$$

Check that these are correct by confirming the validity of (7.2.7)–(7.2.10).

3. Since $\lambda_1 = 1$ is a simple eigenvalue, (7.2.12) may be used to compute $\mathbf{G}_1$ from any pair of associated right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^T$. Of course, $\mathbf{P}$ and $\mathbf{P}^{-1}$ are not needed to determine such a pair, but since $\mathbf{P}$ and $\mathbf{P}^{-1}$ have been computed above, we can use $\mathbf{X}_1$ and $\mathbf{Y}_1^T$ to make the point that *any* right-hand and left-hand eigenvectors associated with $\lambda_1 = 1$ will do the job because they are all of the form $\mathbf{x} = \alpha\mathbf{X}_1$ and $\mathbf{y}^T = \beta\mathbf{Y}_1^T$ for $\alpha \neq 0 \neq \beta$. Consequently,

$$\mathbf{G}_1 = \frac{\mathbf{x}\mathbf{y}^T}{\mathbf{y}^T\mathbf{x}} = \frac{\alpha \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \beta \begin{pmatrix} 1 & -1 & -1 \end{pmatrix}}{\alpha\beta} = \begin{pmatrix} 1 & -1 & -1 \\ 2 & -2 & -2 \\ -2 & 2 & 2 \end{pmatrix}.$$

Invoking (7.2.10) yields the other spectral projector as $\mathbf{G}_2 = \mathbf{I} - \mathbf{G}_1$.

4. An even easier solution is obtained from the spectral theorem by writing

$$\begin{aligned} \mathbf{A} - \mathbf{I} &= (1\mathbf{G}_1 - 3\mathbf{G}_2) - (\mathbf{G}_1 + \mathbf{G}_2) = -4\mathbf{G}_2, \\ \mathbf{A} + 3\mathbf{I} &= (1\mathbf{G}_1 - 3\mathbf{G}_2) + 3(\mathbf{G}_1 + \mathbf{G}_2) = 4\mathbf{G}_1, \end{aligned}$$

so that

$$\mathbf{G}_1 = \frac{(\mathbf{A}+3\mathbf{I})}{4} \quad \text{and} \quad \mathbf{G}_2 = \frac{-(\mathbf{A}-\mathbf{I})}{4}.$$

Can you see how to make this rather ad hoc technique work in more general situations?

5. In fact, the technique above is really a special case of a completely general formula giving each $\mathbf{G}_i$ as a function $\mathbf{A}$ and λ_i as

$$\mathbf{G}_i = \frac{\prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\mathbf{A}-\lambda_j\mathbf{I})}{\prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\lambda_i-\lambda_j)}.$$

This "interpolation formula" is developed on p. 529.

Below is a summary of the facts concerning diagonalizability.

Summary of Diagonalizability

For an $n \times n$ matrix $\mathbf{A}$ with spectrum $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, the following statements are equivalent.

- $\mathbf{A}$ is similar to a diagonal matrix—i.e., $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$.
- $\mathbf{A}$ has a complete linearly independent set of eigenvectors.
- Every λ_i is semisimple—i.e., $\textit{geo mult}_{\mathbf{A}}(\lambda_i) = \textit{alg mult}_{\mathbf{A}}(\lambda_i)$.
- $\mathbf{A} = \lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \cdots + \lambda_k\mathbf{G}_k$, where
  - $\mathbf{G}_i$ is the projector onto $N(\mathbf{A}-\lambda_i\mathbf{I})$ along $R(\mathbf{A}-\lambda_i\mathbf{I})$,
  - $\mathbf{G}_i\mathbf{G}_j = \mathbf{0}$ whenever $i \neq j$,
  - $\mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_k = \mathbf{I}$,
  - $\mathbf{G}_i = \prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\mathbf{A}-\lambda_j\mathbf{I}) \Big/ \prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\lambda_i-\lambda_j)$ (see (7.3.11) on p. 529).
  - If λ_i is a simple eigenvalue associated with right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$, respectively, then $\mathbf{G}_i = \mathbf{x}\mathbf{y}^*/\mathbf{y}^*\mathbf{x}$.

Exercises for section 7.2

7.2.1. Diagonalize $\mathbf{A} = \begin{pmatrix} -8 & -6 \\ 12 & 10 \end{pmatrix}$ with a similarity transformation, or else explain why $\mathbf{A}$ can't be diagonalized.

7.2.2. (a) Verify that $alg\ mult_{\mathbf{A}}(\lambda) = geo\ mult_{\mathbf{A}}(\lambda)$ for each eigenvalue of

$$\mathbf{A} = \begin{pmatrix} -4 & -3 & -3 \\ 0 & -1 & 0 \\ 6 & 6 & 5 \end{pmatrix}.$$

(b) Find a nonsingular $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is a diagonal matrix.

7.2.3. Show that similar matrices need not have the same eigenvectors by giving an example of two matrices that are similar but have different eigenspaces.

7.2.4. $\lambda = 2$ is an eigenvalue for $\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 \\ 0 & 2 & 0 \\ -2 & -3 & 0 \end{pmatrix}$. Find $alg\ mult_{\mathbf{A}}(\lambda)$ as well as $geo\ mult_{\mathbf{A}}(\lambda)$. Can you conclude anything about the diagonalizability of $\mathbf{A}$ from these results?

7.2.5. If $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, explain why $\mathbf{B}^k = \mathbf{P}^{-1}\mathbf{A}^k\mathbf{P}$.

7.2.6. Compute $\lim_{n\to\infty} \mathbf{A}^n$ for $\mathbf{A} = \begin{pmatrix} 7/5 & 1/5 \\ -1 & 1/2 \end{pmatrix}$.

7.2.7. Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$ be a set of linearly independent eigenvectors for $\mathbf{A}_{n\times n}$ associated with respective eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_t\}$, and let $\mathbf{X}$ be any $n \times (n-t)$ matrix such that $\mathbf{P}_{n\times n} = \big(\mathbf{x}_1 \,|\, \cdots \,|\, \mathbf{x}_t \,|\, \mathbf{X}\big)$ is nonsingular. Prove that if $\mathbf{P}^{-1} = \begin{pmatrix} \mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_t^* \\ \mathbf{Y}^* \end{pmatrix}$, where the $\mathbf{y}_i^*$'s are rows and $\mathbf{Y}^*$ is $(n-t) \times n$, then $\{\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_t^*\}$ is a set of linearly independent left-hand eigenvectors associated with $\{\lambda_1, \lambda_2, \ldots, \lambda_t\}$, respectively (i.e., $\mathbf{y}_i^*\mathbf{A} = \lambda_i\mathbf{y}_i^*$).

7.2.8. Let $\mathbf{A}$ be a diagonalizable matrix, and let $\rho(\star)$ denote the spectral radius (recall Example 7.1.4 on p. 497). Prove that $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$ if and only if $\rho(\mathbf{A}) < 1$. **Note:** It is demonstrated on p. 617 that this result holds for nondiagonalizable matrices as well.

7.2.9. Apply the technique used to prove Schur's triangularization theorem (p. 508) to construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P}$ is upper triangular for $\mathbf{A} = \begin{pmatrix} 13 & -9 \\ 16 & -11 \end{pmatrix}$.

7.2.10. Verify the Cayley–Hamilton theorem for $\mathbf{A} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}$.
Hint: This is the matrix from Example 7.2.1 on p. 507.

7.2.11. Since each row sum in the following symmetric matrix $\mathbf{A}$ is 4, it's clear that $\mathbf{x} = (1, 1, 1, 1)^T$ is both a right-hand and left-hand eigenvector associated with $\lambda = 4 \in \sigma(\mathbf{A})$. Use the deflation technique of Example 7.2.6 (p. 516) to determine the remaining eigenvalues of

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 2 & 1 & 1 \\ 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}.$$

7.2.12. Explain why $\mathbf{A}\mathbf{G}_i = \mathbf{G}_i\mathbf{A} = \lambda_i\mathbf{G}_i$ for the spectral projector $\mathbf{G}_i$ associated with the eigenvalue λ_i of a diagonalizable matrix $\mathbf{A}$.

7.2.13. Prove that $\mathbf{A} = \mathbf{c}_{n\times 1}\mathbf{d}^T_{1\times n}$ is diagonalizable if and only if $\mathbf{d}^T\mathbf{c} \neq 0$.

7.2.14. Prove that $\mathbf{A} = \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$ is diagonalizable if and only if $\mathbf{W}_{s\times s}$ and $\mathbf{Z}_{t\times t}$ are each diagonalizable.

7.2.15. Prove that if $\mathbf{AB} = \mathbf{BA}$, then $\mathbf{A}$ and $\mathbf{B}$ can be *simultaneously* triangularized by a unitary similarity transformation—i.e., $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}_1$ and $\mathbf{U}^*\mathbf{B}\mathbf{U} = \mathbf{T}_2$ for some unitary matrix $\mathbf{U}$. **Hint:** Recall Exercise 7.1.20 (p. 503) along with the development of Schur's triangularization theorem (p. 508).

7.2.16. For diagonalizable matrices, prove that $\mathbf{AB} = \mathbf{BA}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ can be *simultaneously* diagonalized—i.e., $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}_1$ and $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \mathbf{D}_2$ for some $\mathbf{P}$. **Hint:** If $\mathbf{A}$ and $\mathbf{B}$ commute, then so do $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \lambda_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}$ and $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \begin{pmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}$.

7.2.17. Explain why the following "proof" of the Cayley–Hamilton theorem is not valid. $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) \implies p(\mathbf{A}) = \det(\mathbf{A} - \mathbf{A}\mathbf{I}) = \det(\mathbf{0}) = 0$.

7.2.18. Show that the eigenvalues of the finite difference matrix (p. 19)

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}_{n\times n} \text{ are } \lambda_j = 4\sin^2\frac{j\pi}{2(n+1)},\ 1 \leq j \leq n.$$

7.2.19. Let $\mathbf{N} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}_{n\times n}$.

(a) Show that $\lambda \in \sigma\left(\mathbf{N}+\mathbf{N}^T\right)$ if and only if $\mathrm{i}\lambda \in \sigma\left(\mathbf{N}-\mathbf{N}^T\right)$.
(b) Explain why $\mathbf{N}+\mathbf{N}^T$ is nonsingular if and only if n is even.
(c) Evaluate $\det\left(\mathbf{N}-\mathbf{N}^T\right)/\det\left(\mathbf{N}+\mathbf{N}^T\right)$ when n is even.

7.2.20. A Toeplitz matrix having the form

$$\mathbf{C} = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ c_2 & c_1 & c_0 & \cdots & c_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_0 \end{pmatrix}_{n\times n}$$

is called a ***circulant matrix***. If $p(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1}$, and if $\{1, \xi, \xi^2, \ldots, \xi^{n-1}\}$ are the n^{th} roots of unity, then the results of Exercise 5.8.12 (p. 379) insure that

$$\mathbf{F}_n\mathbf{C}\mathbf{F}_n^{-1} = \begin{pmatrix} p(1) & 0 & \cdots & 0 \\ 0 & p(\xi) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\xi^{n-1}) \end{pmatrix}$$

in which $\mathbf{F}_n$ is the Fourier matrix of order n. Verify these facts for the circulant below by computing its eigenvalues and eigenvectors directly:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

7.2.21. Suppose that $(\lambda, \mathbf{x})$ and $(\mu, \mathbf{y}^*)$ are right-hand and left-hand eigenpairs for $\mathbf{A} \in \Re^{n\times n}$—i.e., $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{A} = \mu\mathbf{y}^*$. Explain why $\mathbf{y}^*\mathbf{x} = 0$ whenever $\lambda \neq \mu$.

7.2.22. Consider $\mathbf{A} \in \Re^{n\times n}$.

(a) Show that if $\mathbf{A}$ is diagonalizable, then there are right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$ associated with $\lambda \in \sigma(\mathbf{A})$ such that $\mathbf{y}^*\mathbf{x} \neq 0$ so that we can make $\mathbf{y}^*\mathbf{x} = 1$.
(b) Show that not every right-hand and left-hand eigenvector $\mathbf{x}$ and $\mathbf{y}^*$ associated with $\lambda \in \sigma(\mathbf{A})$ must satisfy $\mathbf{y}^*\mathbf{x} \neq 0$.
(c) Show that (a) need not be true when $\mathbf{A}$ is not diagonalizable.

7.2.23. Consider $\mathbf{A} \in \Re^{n\times n}$ with $\lambda \in \sigma(\mathbf{A})$.

(a) Prove that if λ is simple, then $\mathbf{y}^*\mathbf{x} \neq 0$ for every pair of respective right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$ associated with λ regardless of whether or not $\mathbf{A}$ is diagonalizable. **Hint:** Use the core-nilpotent decomposition on p. 397.

(b) Show that $\mathbf{y}^*\mathbf{x} = 0$ is possible when λ is not simple.

7.2.24. For $\mathbf{A} \in \Re^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, show $\mathbf{A}$ is diagonalizable if and only if $\Re^n = N(\mathbf{A} - \lambda_1\mathbf{I}) \oplus N(\mathbf{A} - \lambda_2\mathbf{I}) \oplus \cdots \oplus N(\mathbf{A} - \lambda_k\mathbf{I})$. **Hint:** Recall Exercise 5.9.14.

7.2.25. The Real Schur Form. Schur's triangularization theorem (p. 508) insures that every square matrix $\mathbf{A}$ is unitarily similar to an upper-triangular matrix—say, $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$. But even when $\mathbf{A}$ is real, $\mathbf{U}$ and $\mathbf{T}$ may have to be complex if $\mathbf{A}$ has some complex eigenvalues. However, the matrices (and the arithmetic) can be constrained to be real by settling for a block-triangular result with 2×2 or scalar entries on the diagonal. Prove that for each $\mathbf{A} \in \Re^{n\times n}$ there exists an orthogonal matrix $\mathbf{P} \in \Re^{n\times n}$ and real matrices $\mathbf{B}_{ij}$ such that

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1k} \\ \mathbf{0} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_{kk} \end{pmatrix}, \quad \text{where } \mathbf{B}_{jj} \text{ is } 1\times 1 \text{ or } 2\times 2.$$

If $\mathbf{B}_{jj} = [\lambda_j]$ is 1×1, then $\lambda_j \in \sigma(\mathbf{A})$, and if $\mathbf{B}_{jj}$ is 2×2, then $\sigma(\mathbf{B}_{jj}) = \{\lambda_j, \overline{\lambda}_j\} \subseteq \sigma(\mathbf{A})$.

7.2.26. When $\mathbf{A} \in \Re^{n\times n}$ is diagonalizable by a similarity transformation $\mathbf{S}$, then $\mathbf{S}$ may have to be complex if $\mathbf{A}$ has some complex eigenvalues. Analogous to Exercise 7.2.25, we can stay in the realm of real numbers by settling for a block-diagonal result with 1×1 or 2×2 entries on the diagonal. Prove that if $\mathbf{A} \in \Re^{n\times n}$ is diagonalizable with real eigenvalues $\{\rho_1, \ldots, \rho_r\}$ and complex eigenvalues $\{\lambda_1, \overline{\lambda}_1, \lambda_2, \overline{\lambda}_2, \ldots, \lambda_t, \overline{\lambda}_t\}$ with $2t + r = n$, then there exists a nonsingular $\mathbf{P} \in \Re^{n\times n}$ and $\mathbf{B}_j$'s $\in \Re^{2\times 2}$ such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_t \end{pmatrix}, \quad \text{where} \quad \mathbf{D} = \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_r \end{pmatrix},$$

and where $\mathbf{B}_j$ has eigenvalues λ_j and $\overline{\lambda}_j$.

7.3 FUNCTIONS OF DIAGONALIZABLE MATRICES

For square matrices $\mathbf{A}$, what should it mean to write $\sin\mathbf{A}$, $e^{\mathbf{A}}$, $\ln\mathbf{A}$, etc.? A naive approach might be to simply apply the given function to each entry of $\mathbf{A}$ such as

$$\sin\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \stackrel{?}{=} \begin{pmatrix} \sin a_{11} & \sin a_{12} \\ \sin a_{21} & \sin a_{22} \end{pmatrix}. \tag{7.3.1}$$

But doing so results in matrix functions that fail to have the same properties as their scalar counterparts. For example, since $\sin^2 x + \cos^2 x = 1$ for all scalars x, we would like our definitions of $\sin\mathbf{A}$ and $\cos\mathbf{A}$ to result in the analogous matrix identity $\sin^2\mathbf{A} + \cos^2\mathbf{A} = \mathbf{I}$ for all square matrices $\mathbf{A}$. The entrywise approach (7.3.1) clearly fails in this regard.

One way to define matrix functions possessing properties consistent with their scalar counterparts is to use infinite series expansions. For example, consider the exponential function

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!}\cdots. \tag{7.3.2}$$

Formally replacing the scalar argument z by a square matrix $\mathbf{A}$ ($z^0 = 1$ is replaced with $\mathbf{A}^0 = \mathbf{I}$) results in the infinite series of matrices

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2!} + \frac{\mathbf{A}^3}{3!}\cdots, \tag{7.3.3}$$

called the *matrix exponential.* While this results in a matrix that has properties analogous to its scalar counterpart, it suffers from the fact that convergence must be dealt with, and then there is the problem of describing the entries in the limit. These issues are handled by deriving a closed form expression for (7.3.3).

If $\mathbf{A}$ is diagonalizable, then $\mathbf{A} = \mathbf{PDP}^{-1} = \mathbf{P}\,\mathrm{diag}\,(\lambda_1,\ldots,\lambda_n)\,\mathbf{P}^{-1}$, and $\mathbf{A}^k = \mathbf{PD}^k\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\left(\lambda_1^k,\ldots,\lambda_n^k\right)\mathbf{P}^{-1}$, so

$$e^{\mathbf{A}} = \sum_{k=0}^{\infty}\frac{\mathbf{A}^k}{k!} = \sum_{k=0}^{\infty}\frac{\mathbf{PD}^k\mathbf{P}^{-1}}{k!} = \mathbf{P}\left(\sum_{k=0}^{\infty}\frac{\mathbf{D}^k}{k!}\right)\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\left(e^{\lambda_1},\ldots,e^{\lambda_n}\right)\mathbf{P}^{-1}.$$

In other words, we don't have to use the infinite series (7.3.3) to define $e^{\mathbf{A}}$. Instead, define $e^{\mathbf{D}} = \mathrm{diag}\,(e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n})$, and set

$$e^{\mathbf{A}} = \mathbf{P}e^{\mathbf{D}}\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\,(e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n})\,\mathbf{P}^{-1}.$$

This idea can be generalized to any function $f(z)$ that is defined on the eigenvalues λ_i of a diagonalizable matrix $\mathbf{A} = \mathbf{PDP}^{-1}$ by defining $f(\mathbf{D})$ to be $f(\mathbf{D}) = \mathrm{diag}\,(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n))$ and by setting

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{D})\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\,(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n))\,\mathbf{P}^{-1}. \tag{7.3.4}$$

At first glance this definition seems to have an edge over the infinite series approach because there are no convergence issues to deal with. But convergence worries have been traded for uniqueness worries. Because $\mathbf{P}$ is not unique, it's not apparent that (7.3.4) is well defined. The eigenvector matrix $\mathbf{P}$ you compute for a given $\mathbf{A}$ need not be the same as the eigenvector matrix I compute, so what insures that your $f(\mathbf{A})$ will be the same as mine? The spectral theorem (p. 517) does. Suppose there are k distinct eigenvalues that are grouped according to repetition, and expand (7.3.4) just as (7.2.11) is expanded to produce

$$f(\mathbf{A}) = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} = \Big(\mathbf{X}_1|\mathbf{X}_2|\cdots|\mathbf{X}_k\Big)\begin{pmatrix} f(\lambda_1)\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & f(\lambda_2)\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & f(\lambda_k)\mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{Y}_1^T \\ \hline \mathbf{Y}_2^T \\ \hline \vdots \\ \hline \mathbf{Y}_k^T \end{pmatrix}$$

$$= \sum_{i=1}^{k} f(\lambda_i)\mathbf{X}_i\mathbf{Y}_i^T = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i.$$

Since $\mathbf{G}_i$ is the projector onto $N(\mathbf{A}-\lambda_i\mathbf{I})$ along $R(\mathbf{A}-\lambda_i\mathbf{I})$, $\mathbf{G}_i$ is uniquely determined by $\mathbf{A}$. Therefore, (7.3.4) uniquely defines $f(\mathbf{A})$ regardless of the choice of $\mathbf{P}$. We can now make a formal definition.

Functions of Diagonalizable Matrices

Let $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ be a diagonalizable matrix where the eigenvalues in $\mathbf{D} = \operatorname{diag}(\lambda_1\mathbf{I}, \lambda_2\mathbf{I}, \ldots, \lambda_k\mathbf{I})$ are grouped by repetition. For a function $f(z)$ that is defined at each $\lambda_i \in \sigma(\mathbf{A})$, define

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{D})\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} f(\lambda_1)\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & f(\lambda_2)\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & f(\lambda_k)\mathbf{I} \end{pmatrix}\mathbf{P}^{-1} \quad (7.3.5)$$

$$= f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 + \cdots + f(\lambda_k)\mathbf{G}_k, \quad (7.3.6)$$

where $\mathbf{G}_i$ is the i^{th} spectral projector as described on pp. 517, 529. The generalization to nondiagonalizable matrices is on p. 603.

The discussion of matrix functions was initiated by considering infinite series, so, to complete the circle, a formal statement connecting infinite series with (7.3.5) and (7.3.6) is needed. By replacing $\mathbf{A}$ by $\mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ in $\sum_{n=0}^{\infty} c_n(\mathbf{A}-z_0\mathbf{I})^n$ and expanding the result, the following result is established.

Infinite Series

If $f(z) = \sum_{n=0}^{\infty} c_n(z - z_0)^n$ converges when $|z - z_0| < r$, and if $|\lambda_i - z_0| < r$ for each eigenvalue λ_i of a diagonalizable matrix $\mathbf{A}$, then

$$f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n(\mathbf{A} - z_0\mathbf{I})^n. \tag{7.3.7}$$

It can be argued that the matrix series on the right-hand side of (7.3.7) converges if and only if $|\lambda_i - z_0| < r$ for each λ_i, regardless of whether or not $\mathbf{A}$ is diagonalizable. So (7.3.7) serves to define $f(\mathbf{A})$ for functions with series expansions regardless of whether or not $\mathbf{A}$ is diagonalizable. More is said in Example 7.9.3 (p. 605).

Example 7.3.1

Neumann Series Revisited. The function $f(z) = (1-z)^{-1}$ has the geometric series expansion $(1-z)^{-1} = \sum_{k=1}^{\infty} z^k$ that converges if and only if $|z| < 1$. This means that the associated matrix function $f(\mathbf{A}) = (\mathbf{I} - \mathbf{A})^{-1}$ is given by

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k \quad \text{if and only if } |\lambda| < 1 \text{ for all } \lambda \in \sigma(\mathbf{A}). \tag{7.3.8}$$

This is the ***Neumann series*** discussed on p. 126, where it was argued that if $\lim_{n\to\infty} \mathbf{A}^n = \mathbf{0}$, then $(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k$. The two approaches are the same because it turns out that $\lim_{n\to\infty} \mathbf{A}^n = \mathbf{0} \iff |\lambda| < 1$ for all $\lambda \in \sigma(\mathbf{A})$. This is immediate for diagonalizable matrices, but the nondiagonalizable case is a bit more involved—the complete statement is developed on p. 618. Because $\max_i |\lambda_i| \leq \|\mathbf{A}\|$ for all matrix norms (Example 7.1.4, p. 497), a corollary of (7.3.8) is that $(\mathbf{I} - \mathbf{A})^{-1}$ exists and

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k \quad \text{when } \|\mathbf{A}\| < 1 \text{ for any matrix norm.} \tag{7.3.9}$$

Caution! $(\mathbf{I} - \mathbf{A})^{-1}$ can exist without the Neumann series expansion being valid because all that's needed for $\mathbf{I} - \mathbf{A}$ to be nonsingular is $1 \notin \sigma(\mathbf{A})$, while convergence of the Neumann series requires each $|\lambda| < 1$.

Example 7.3.2

Eigenvalue Perturbations. It's often important to understand how the eigenvalues of a matrix are affected by perturbations. In general, this is a complicated issue, but for diagonalizable matrices the problem is more tractable.

Problem: Suppose $\mathbf{B} = \mathbf{A} + \mathbf{E}$, where $\mathbf{A}$ is diagonalizable, and let $\beta \in \sigma(\mathbf{B})$. If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$, explain why

$$\min_{\lambda_i \in \sigma(\mathbf{A})} |\beta - \lambda_i| \leq \kappa(\mathbf{P})\,\|\mathbf{E}\|, \quad \text{where} \quad \kappa(\mathbf{P}) = \|\mathbf{P}\|\,\|\mathbf{P}^{-1}\| \tag{7.3.10}$$

for matrix norms satisfying $\|\mathbf{D}\| = \max_i |\lambda_i|$ (e.g., any standard induced norm).

Solution: Assume $\beta \notin \sigma(\mathbf{A})$—(7.3.10) is trivial if $\beta \in \sigma(\mathbf{A})$—and observe that

$$(\beta\mathbf{I} - \mathbf{A})^{-1}(\beta\mathbf{I} - \mathbf{B}) = (\beta\mathbf{I} - \mathbf{A})^{-1}(\beta\mathbf{I} - \mathbf{A} - \mathbf{E}) = \mathbf{I} - (\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}$$

implies that $1 \leq \|(\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}\|$—otherwise $\mathbf{I} - (\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}$ is nonsingular by (7.3.9), which is impossible because $(\beta\mathbf{I} - \mathbf{B})$ (and hence $(\beta\mathbf{I} - \mathbf{A})^{-1}(\beta\mathbf{I} - \mathbf{B})$ is singular). Consequently,

$$\begin{aligned} 1 &\leq \|(\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}\| = \|\mathbf{P}(\beta\mathbf{I} - \mathbf{D})^{-1}\mathbf{P}^{-1}\mathbf{E}\| \leq \|\mathbf{P}\|\,\|(\beta\mathbf{I} - \mathbf{D})^{-1}\|\,\|\mathbf{P}^{-1}\|\,\|\mathbf{E}\| \\ &= \kappa(\mathbf{P})\,\|\mathbf{E}\| \max_i |\beta - \lambda_i|^{-1} = \kappa(\mathbf{P})\,\|\mathbf{E}\| \frac{1}{\min_i |\beta - \lambda_i|}, \end{aligned}$$

and this produces (7.3.10). Similar to the case of linear systems (Example 5.12.1, p. 414), the expression $\kappa(\mathbf{P})$ is a *condition number* in the sense that if $\kappa(\mathbf{P})$ is relatively small, then the λ_i's are relatively insensitive, but if $\kappa(\mathbf{P})$ is relatively large, we must be suspicious. **Note:** Because it's a corollary of their 1960 results, the bound (7.3.10) is often referred to as the ***Bauer–Fike bound***.

Infinite series representations can always be avoided because *every function of* $\mathbf{A}_{n\times n}$ *can be expressed as a polynomial in* $\mathbf{A}$. In other words, when $f(\mathbf{A})$ exists, there is a polynomial $p(z)$ such that $p(\mathbf{A}) = f(\mathbf{A})$. This is true for all matrices, but the development here is limited to diagonalizable matrices—nondiagonalizable matrices are treated in Exercise 7.3.7. In the diagonalizable case, $f(\mathbf{A})$ exists if and only if $f(\lambda_i)$ exists for each $\lambda_i \in \sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, and, by (7.3.6), $f(\mathbf{A}) = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i$, where $\mathbf{G}_i$ is the i^{th} spectral projector. Any polynomial $p(z)$ agreeing with $f(z)$ on $\sigma(\mathbf{A})$ does the job because if $p(\lambda_i) = f(\lambda_i)$ for each $\lambda_i \in \sigma(\mathbf{A})$, then

$$p(\mathbf{A}) = \sum_{i=1}^{k} p(\lambda_i)\mathbf{G}_i = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i = f(\mathbf{A}).$$

But is there always a polynomial satisfying $p(\lambda_i) = f(\lambda_i)$ for each $\lambda_i \in \sigma(\mathbf{A})$? Sure—that's what the *Lagrange interpolating polynomial* from Example 4.3.5 (p. 186) does. It's given by

$$p(z) = \sum_{i=1}^{k} \left(f(\lambda_i) \frac{\prod_{\substack{j=1 \\ j \neq i}}^{k} (z - \lambda_j)}{\prod_{\substack{j=1 \\ j \neq i}}^{k} (\lambda_i - \lambda_j)} \right), \text{ so } f(\mathbf{A}) = p(\mathbf{A}) = \sum_{i=1}^{k} \left(f(\lambda_i) \frac{\prod_{\substack{j=1 \\ j \neq i}}^{k} (\mathbf{A} - \lambda_j \mathbf{I})}{\prod_{\substack{j=1 \\ j \neq i}}^{k} (\lambda_i - \lambda_j)} \right).$$

Using the function $g_i(z) = \begin{cases} 1 & \text{if } z = \lambda_i, \\ 0 & \text{if } z \neq \lambda_i, \end{cases}$ with this representation as well as that in (7.3.6) yields $\prod_{\substack{j=1 \\ j \neq i}}^{k} (\mathbf{A} - \lambda_j \mathbf{I}) \Big/ \prod_{\substack{j=1 \\ j \neq i}}^{k} (\lambda_i - \lambda_j) = g_i(\mathbf{A}) = \mathbf{G}_i$. For example, if $\sigma(\mathbf{A}_{n \times n}) = \{\lambda_1, \lambda_2, \lambda_3\}$, then $f(\mathbf{A}) = f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 + f(\lambda_3)\mathbf{G}_3$ with

$$\mathbf{G}_1 = \frac{(\mathbf{A}-\lambda_2\mathbf{I})(\mathbf{A}-\lambda_3\mathbf{I})}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)}, \quad \mathbf{G}_2 = \frac{(\mathbf{A}-\lambda_1\mathbf{I})(\mathbf{A}-\lambda_3\mathbf{I})}{(\lambda_2-\lambda_1)(\lambda_2-\lambda_3)}, \quad \mathbf{G}_3 = \frac{(\mathbf{A}-\lambda_1\mathbf{I})(\mathbf{A}-\lambda_2\mathbf{I})}{(\lambda_3-\lambda_1)(\lambda_3-\lambda_2)}.$$

Below is a summary of these observations.

Spectral Projectors

If $\mathbf{A}$ is diagonalizable with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then the spectral projector onto $N(\mathbf{A} - \lambda_i \mathbf{I})$ along $R(\mathbf{A} - \lambda_i \mathbf{I})$ is given by

$$\mathbf{G}_i = \prod_{\substack{j=1 \\ j \neq i}}^{k} (\mathbf{A} - \lambda_j \mathbf{I}) \Big/ \prod_{\substack{j=1 \\ j \neq i}}^{k} (\lambda_i - \lambda_j) \quad \text{for } i = 1, 2, \ldots, k. \tag{7.3.11}$$

Consequently, if $f(z)$ is defined on $\sigma(\mathbf{A})$, then $f(\mathbf{A}) = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i$ is a polynomial in $\mathbf{A}$ of degree at most $k - 1$.

Example 7.3.3

Problem: For a scalar t, determine the matrix exponential $e^{\mathbf{A}t}$, where

$$\mathbf{A} = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix} \quad \text{with } \alpha + \beta \neq 0.$$

Solution 1: The characteristic equation for $\mathbf{A}$ is $\lambda^2 + (\alpha + \beta)\lambda = 0$, so the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 0$ and $\lambda_2 = -(\alpha+\beta)$. Note that $\mathbf{A}$ is diagonalizable

because no eigenvalue is repeated—recall (7.2.6). Using the function $f(z) = e^{zt}$, the spectral representation (7.3.6) says that

$$e^{\mathbf{A}t} = f(\mathbf{A}) = f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 = e^{\lambda_1 t}\mathbf{G}_1 + e^{\lambda_2 t}\mathbf{G}_2.$$

The spectral projectors $\mathbf{G}_1$ and $\mathbf{G}_2$ are determined from (7.3.11) to be

$$\mathbf{G}_1 = \frac{\mathbf{A} - \lambda_2\mathbf{I}}{-\lambda_2} = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix} \quad \text{and} \quad \mathbf{G}_2 = \frac{\mathbf{A}}{\lambda_2} = \frac{1}{\alpha+\beta}\begin{pmatrix} \alpha & -\beta \\ -\alpha & \beta \end{pmatrix},$$

so

$$e^{\mathbf{A}t} = \mathbf{G}_1 + e^{-(\alpha+\beta)t}\mathbf{G}_2 = \frac{1}{\alpha+\beta}\left[\begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix} + e^{-(\alpha+\beta)t}\begin{pmatrix} \alpha & -\beta \\ -\alpha & \beta \end{pmatrix}\right].$$

Solution 2: Compute eigenpairs $(\lambda_1, \mathbf{x}_1)$ and $(\lambda_2, \mathbf{x}_2)$, construct $\mathbf{P} = \begin{bmatrix}\mathbf{x}_1 \,|\, \mathbf{x}_2\end{bmatrix}$, and compute

$$e^{\mathbf{A}t} = \mathbf{P}\begin{pmatrix} f(\lambda_1) & 0 \\ 0 & f(\lambda_2) \end{pmatrix}\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix}\mathbf{P}^{-1}.$$

The computational details are called for in Exercise 7.3.2.

Example 7.3.4

Problem: For $\mathbf{T} = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$, evaluate $\lim_{k\to\infty} \mathbf{T}^k$.

Solution 1: Compute two eigenpairs, $\lambda_1 = 1$, $\mathbf{x}_1 = (1,\ 1)^T$, and $\lambda_2 = 1/4$, $\mathbf{x}_2 = (-2,\ 1)^T$. If $\mathbf{P} = [\mathbf{x}_1 \,|\, \mathbf{x}_2]$, then $\mathbf{T} = \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & 1/4 \end{pmatrix}\mathbf{P}^{-1}$, so

$$\mathbf{T}^k = \mathbf{P}\begin{pmatrix} 1^k & 0 \\ 0 & 1/4^k \end{pmatrix}\mathbf{P}^{-1} \to \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathbf{P}^{-1} = \frac{1}{3}\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}. \tag{7.3.12}$$

Solution 2: We know from (7.3.6) that $\mathbf{T}^k = 1^k\mathbf{G}_1 + (1/4)^k\mathbf{G}_2 \to \mathbf{G}_1$. Since $\lambda_1 = 1$ is a simple eigenvalue, formula (7.2.12) on p. 518 can be used to compute $\mathbf{G}_1 = \mathbf{x}_1\mathbf{y}_1^T/\mathbf{y}_1^T\mathbf{x}_1$, where $\mathbf{x}_1$ and $\mathbf{y}_1^T$ are any right- and left-hand eigenvectors associated with $\lambda_1 = 1$. A right-hand eigenvector $\mathbf{x}_1$ was computed above. Computing a left-hand eigenvector $\mathbf{y}_1^T = (1,\ 2)$ yields

$$\mathbf{T}^k \to \mathbf{G}_1 = \frac{\mathbf{x}_1\mathbf{y}_1^T}{\mathbf{y}_1^T\mathbf{x}_1} = \frac{1}{3}\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}. \tag{7.3.13}$$

Example 7.3.5

Population Migration. Suppose that the population migration between two geographical regions—say, the North and the South—is as follows. Each year, 50% of the population in the North migrates to the South, while only 25% of the population in the South moves to the North. This situation is depicted by drawing a transition diagram as shown below in Figure 7.3.1.

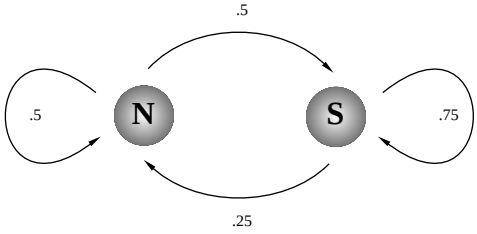


FIGURE 7.3.1

Problem: If this migration pattern continues, will the population in the North continually shrink until the entire population is eventually in the South, or will the population distribution somehow stabilize before the North is completely deserted?

Solution: Let n_k and s_k denote the respective proportions of the total population living in the North and South at the end of year k, and assume $n_k + s_k = 1$. The migration pattern dictates that the fractions of the population in each region at the end of year $k+1$ are

$$\left\{ \begin{array}{l} n_{k+1} = n_k(.5) + s_k(.25) \\ s_{k+1} = n_k(.5) + s_k(.75) \end{array} \right\} \quad \text{or, equivalently,} \quad \mathbf{p}_{k+1}^T = \mathbf{p}_k^T \mathbf{T}, \tag{7.3.14}$$

where $\mathbf{p}_k^T = (\, n_k \quad s_k \,)$ and $\mathbf{p}_{k+1}^T = (\, n_{k+1} \quad s_{k+1} \,)$ are the respective population distributions at the end of years k and $k+1$, and where

$$\mathbf{T} = \begin{array}{c} \\ \text{N} \\ \text{S} \end{array} \begin{array}{c} \begin{array}{cc} \text{N} & \text{S} \end{array} \\ \begin{pmatrix} .5 & .5 \\ .25 & .75 \end{pmatrix} \end{array}$$

is the associated ***transition matrix*** (recall Example 3.6.3). Inducting on

$$\mathbf{p}_1^T = \mathbf{p}_0^T \mathbf{T}, \quad \mathbf{p}_2^T = \mathbf{p}_1^T \mathbf{T} = \mathbf{p}_0^T \mathbf{T}^2, \quad \mathbf{p}_3^T = \mathbf{p}_2^T \mathbf{T} = \mathbf{p}_0^T \mathbf{T}^3, \quad \cdots$$

leads to $\mathbf{p}_k^T = \mathbf{p}_0^T \mathbf{T}^k$, which indicates that the powers of $\mathbf{T}$ determine how the process evolves. Determining the long-run population distribution[73] is therefore

[73] The long-run distribution goes by a lot of different names. It's also called the *limiting* distribution, the *steady-state* distribution, and the *stationary* distribution.

accomplished by analyzing $\lim_{k\to\infty}\mathbf{T}^k$. The results of Example 7.3.4 together with $n_0+s_0=1$ yield the long-run (or limiting) population distribution as

$$\mathbf{p}_\infty^T=\lim_{k\to\infty}\mathbf{p}_k^T=\lim_{k\to\infty}\mathbf{p}_0^T\mathbf{T}^k=\mathbf{p}_0^T\lim_{k\to\infty}\mathbf{T}^k=\begin{pmatrix}n_0 & s_0\end{pmatrix}\begin{pmatrix}1/3 & 2/3\\ 1/3 & 2/3\end{pmatrix}$$
$$=\begin{pmatrix}\dfrac{n_0+s_0}{3} & \dfrac{2(n_0+s_0)}{3}\end{pmatrix}=\begin{pmatrix}\dfrac{1}{3} & \dfrac{2}{3}\end{pmatrix}.$$

So if the migration pattern continues to hold, then the population distribution will eventually stabilize with 1/3 of the population being in the North and 2/3 of the population in the South. And this is independent of the initial distribution!

Observations: This is an example of a broader class of evolutionary processes known as ***Markov chains*** (p. 687), and the following observations are typical.

- It's clear from (7.3.12) or (7.3.13) that the rate at which the population distribution stabilizes is governed by how fast $(1/4)^k\to 0$. In other words, the magnitude of the largest subdominant eigenvalue of $\mathbf{T}$ determines the rate of evolution.
- For the dominant eigenvalue $\lambda_1=1$, the column, $\mathbf{x}_1$, of 1's is a right-hand eigenvector (because $\mathbf{T}$ has unit row sums). This forces the limiting distribution $\mathbf{p}_\infty^T$ to be a particular left-hand eigenvector associated with $\lambda_1=1$ because for an arbitrary left-hand eigenvector $\mathbf{y}_1^T$ associated with $\lambda_1=1$, equation (7.3.13) in Example 7.3.4 insures that

$$\mathbf{p}_\infty^T=\lim_{k\to\infty}\mathbf{p}_0^T\mathbf{T}^k=\mathbf{p}_0^T\lim_{k\to\infty}\mathbf{T}^k=\mathbf{p}_0^T\mathbf{G}_1=\frac{(\mathbf{p}_0^T\mathbf{x}_1)\mathbf{y}_1^T}{\mathbf{y}_1^T\mathbf{x}_1}=\frac{\mathbf{y}_1^T}{\mathbf{y}_1^T\mathbf{x}_1}.\quad(7.3.15)$$

  The fact that $\mathbf{p}_0^T\mathbf{T}^k$ converges to an eigenvector is a special case of the power method discussed in Example 7.3.7.
- Equation (7.3.15) shows why the initial distribution $\mathbf{p}_0^T$ always drops away in the limit. But $\mathbf{p}_0^T$ is not completely irrelevant because it always affects the transient behavior—i.e., the behavior of $\mathbf{p}_k^T=\mathbf{p}_0^T\mathbf{T}^k$ for smaller k's.

Example 7.3.6

Cayley–Hamilton Revisited. The Cayley–Hamilton theorem (p. 509) says that if $p(\lambda)=0$ is the characteristic equation for $\mathbf{A}$, then $p(\mathbf{A})=\mathbf{0}$. This is evident for diagonalizable $\mathbf{A}$ because $p(\lambda_i)=0$ for each $\lambda_i\in\sigma(\mathbf{A})$, so, by (7.3.6), $p(\mathbf{A})=p(\lambda_1)\mathbf{G}_1+p(\lambda_2)\mathbf{G}_2+\cdots+p(\lambda_k)\mathbf{G}_k=\mathbf{0}$.

Problem: Establish the Cayley–Hamilton theorem for nondiagonalizable matrices by using the diagonalizable result together with a continuity argument.

Solution: Schur's triangularization theorem (p. 508) insures $\mathbf{A}_{n\times n}=\mathbf{UTU}^*$ for a unitary $\mathbf{U}$ and an upper triangular $\mathbf{T}$ having the eigenvalues of $\mathbf{A}$ on the

diagonal. For each $\epsilon \neq 0$, it's possible to find numbers ϵ_i such that $(\lambda_1 + \epsilon_1)$, $(\lambda_2 + \epsilon_2)$, $\ldots$, $(\lambda_n + \epsilon_n)$ are distinct and $\sum \epsilon_i^2 = |\epsilon|$. Set

$$\mathbf{D}(\epsilon) = \text{diag}\,(\epsilon_1, \epsilon_2, \ldots, \epsilon_n) \quad \text{and} \quad \mathbf{B}(\epsilon) = \mathbf{U}\big(\mathbf{T} + \mathbf{D}(\epsilon)\big)\mathbf{U}^* = \mathbf{A} + \mathbf{E}(\epsilon),$$

where $\mathbf{E}(\epsilon) = \mathbf{U}\mathbf{D}(\epsilon)\mathbf{U}^*$. The $(\lambda_i + \epsilon_i)$'s are the eigenvalues of $\mathbf{B}(\epsilon)$ and they are distinct, so $\mathbf{B}(\epsilon)$ is diagonalizable—by (7.2.6). Consequently, $\mathbf{B}(\epsilon)$ satisfies its own characteristic equation $0 = p_\epsilon(\lambda) = \det(\mathbf{A} + \mathbf{E}(\epsilon) - \lambda\mathbf{I})$ for each $\epsilon \neq 0$. The coefficients of $p_\epsilon(\lambda)$ are continuous functions of the entries in $\mathbf{E}(\epsilon)$ (recall (7.1.6)) and hence are continuous functions of the ϵ_i's. Combine this with $\lim_{\epsilon\to 0} \mathbf{E}(\epsilon) = \mathbf{0}$ to obtain $\mathbf{0} = \lim_{\epsilon\to 0} p_\epsilon(\mathbf{B}(\epsilon)) = p(\mathbf{A})$.

Note: Embedded in the above development is the fact that every square complex matrix is arbitrarily close to some diagonalizable matrix because for each $\epsilon \neq 0$, we have $\|\mathbf{A} - \mathbf{B}(\epsilon)\|_F = \|\mathbf{E}(\epsilon)\|_F = \epsilon$ (recall Exercise 5.6.9).

Example 7.3.7

Power method[74] is an iterative technique for computing a dominant eigenpair $(\lambda_1, \mathbf{x})$ of a diagonalizable $\mathbf{A} \in \Re^{m\times m}$ with eigenvalues

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_k|.$$

Note that this implies λ_1 is real—otherwise $\overline{\lambda}_1$ is another eigenvalue with the same magnitude as λ_1. Consider $f(z) = (z/\lambda_1)^n$, and use the spectral representation (7.3.6) along with $|\lambda_i/\lambda_1| < 1$ for $i = 2, 3, \ldots, k$ to conclude that

$$\begin{aligned}\left(\frac{\mathbf{A}}{\lambda_1}\right)^n &= f(\mathbf{A}) = f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 + \cdots + f(\lambda_k)\mathbf{G}_k \\ &= \mathbf{G}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^n \mathbf{G}_2 + \cdots + \left(\frac{\lambda_k}{\lambda_1}\right)^n \mathbf{G}_k \to \mathbf{G}_1\end{aligned} \tag{7.3.16}$$

as $n \to \infty$. Consequently, $(\mathbf{A}^n\mathbf{x}_0/\lambda_1^n) \to \mathbf{G}_1\mathbf{x}_0 \in N(\mathbf{A} - \lambda_1\mathbf{I})$ for all $\mathbf{x}_0$. So if $\mathbf{G}_1\mathbf{x}_0 \neq \mathbf{0}$ or, equivalently, $\mathbf{x}_0 \notin R(\mathbf{A} - \lambda_1\mathbf{I})$, then $\mathbf{A}^n\mathbf{x}_0/\lambda_1^n$ converges to an eigenvector associated with λ_1. This means that the direction of $\mathbf{A}^n\mathbf{x}_0$ tends toward the direction of an eigenvector because λ_1^n acts only as a scaling factor to keep the length of $\mathbf{A}^n\mathbf{x}_0$ under control. Rather than using λ_1^n, we can scale $\mathbf{A}^n\mathbf{x}_0$ with something more convenient. For example, $\|\mathbf{A}^n\mathbf{x}_0\|$ (for any vector norm) is a reasonable scaling factor, but there are even better choices. For vectors $\mathbf{v}$, let $m(\mathbf{v})$ denote the component of maximal magnitude, and if there is more

[74] While the development of the power method was considered to be a great achievement when R. von Mises introduced it in 1929, later algorithms relegated its computational role to that of a special purpose technique. Nevertheless, it's still an important idea because, in some way or another, most practical algorithms for eigencomputations implicitly rely on the mathematical essence of the power method.

than one maximal component, let $m(\mathbf{v})$ be the *first* maximal component—e.g., $m(1,3,-2)=3$, and $m(-3,3,-2)=-3$. It's clear that $m(\alpha\mathbf{v})=\alpha m(\mathbf{v})$ for all scalars α. Suppose $m(\mathbf{A}^n\mathbf{x}_0/\lambda_1^n)\to\gamma$. Since $(\mathbf{A}^n/\lambda_1^n)\to\mathbf{G}_1$, we see that

$$\lim_{n\to\infty}\frac{\mathbf{A}^n\mathbf{x}_0}{m(\mathbf{A}^n\mathbf{x}_0)}=\lim_{n\to\infty}\frac{(\mathbf{A}^n/\lambda_1^n)\mathbf{x}_0}{m(\mathbf{A}^n\mathbf{x}_0/\lambda_1^n)}=\frac{\mathbf{G}_1\mathbf{x}_0}{\gamma}=\mathbf{x}$$

is an eigenvector associated with λ_1. But rather than successively powering $\mathbf{A}$, the sequence $\mathbf{A}^n\mathbf{x}_0/m(\mathbf{A}^n\mathbf{x}_0)$ is more efficiently generated by starting with $\mathbf{x}_0\notin R(\mathbf{A}-\lambda_1\mathbf{I})$ and setting

$$\mathbf{y}_n=\mathbf{A}\mathbf{x}_n,\quad \nu_n=m(\mathbf{y}_n),\quad \mathbf{x}_{n+1}=\frac{\mathbf{y}_n}{\nu_n},\quad \text{for } n=0,1,2,\ldots. \tag{7.3.17}$$

Not only does $\mathbf{x}_n\to\mathbf{x}$, but as a bonus we get $\nu_n\to\lambda_1$ because for all n, $\mathbf{A}\mathbf{x}_{n+1}=\mathbf{A}^2\mathbf{x}_n/\nu_n$, so if $\nu_n\to\nu$ as $n\to\infty$, the limit on the left-hand side is $\mathbf{A}\mathbf{x}=\lambda_1\mathbf{x}$, while the limit on the right-hand side is $\mathbf{A}^2\mathbf{x}/\nu=\lambda_1^2\mathbf{x}/\nu$. Since these two limits must agree, $\lambda_1\mathbf{x}=(\lambda_1^2/\nu)\mathbf{x}$, and this implies $\nu=\lambda_1$.

Summary. The sequence $(\nu_n,\mathbf{x}_n)$ defined by (7.3.17) converges to an eigenpair $(\lambda_1,\mathbf{x})$ for $\mathbf{A}$ provided that $\mathbf{G}_1\mathbf{x}_0\neq\mathbf{0}$ or, equivalently, $\mathbf{x}_0\notin R(\mathbf{A}-\lambda_1\mathbf{I})$.

- ▷ **Advantages.** Each iteration requires only one matrix–vector product, and this can be exploited to reduce the computational effort when $\mathbf{A}$ is large and sparse—assuming that a dominant eigenpair is the only one of interest.
- ▷ **Disadvantages.** Only a dominant eigenpair is determined—something else must be done if others are desired. Furthermore, it's clear from (7.3.16) that the rate at which (7.3.17) converges depends on how fast $(\lambda_2/\lambda_1)^n\to 0$, so convergence is slow when $|\lambda_1|$ is close to $|\lambda_2|$.

Example 7.3.8

Inverse Power Method. Given a real approximation $\alpha\notin\sigma(\mathbf{A})$ to any real $\lambda\in\sigma(\mathbf{A})$, this algorithm (also called the ***inverse iteration***) determines an eigenpair $(\lambda,\mathbf{x})$ for a diagonalizable matrix $\mathbf{A}\in\Re^{m\times m}$ by applying the power method[75] to $\mathbf{B}=(\mathbf{A}-\alpha\mathbf{I})^{-1}$. Recall from Exercise 7.1.9 that

$$\begin{gathered}\mathbf{x}\text{ is an eigenvector for }\mathbf{A}\Longleftrightarrow\mathbf{x}\text{ is an eigenvector for }\mathbf{B},\\ \lambda\in\sigma(\mathbf{A})\Longleftrightarrow(\lambda-\alpha)^{-1}\in\sigma(\mathbf{B}).\end{gathered} \tag{7.3.18}$$

If $|\lambda-\alpha|<|\lambda_i-\alpha|$ for all other $\lambda_i\in\sigma(\mathbf{A})$, then $(\lambda-\alpha)^{-1}$ is the dominant eigenvalue of $\mathbf{B}$ because $|\lambda-\alpha|^{-1}>|\lambda_i-\alpha|^{-1}$. Therefore, applying the power

[75] The relation between the power method and inverse iteration is clear to us now, but it originally took 15 years to make the connection. Inverse iteration was not introduced until 1944 by the German mathematician Helmut Wielandt (1910–).

method to $\mathbf{B}$ produces an eigenpair $\left((\lambda-\alpha)^{-1}, \mathbf{x}\right)$ for $\mathbf{B}$ from which the eigenpair $(\lambda, \mathbf{x})$ for $\mathbf{A}$ is determined. That is, if $\mathbf{x}_0 \notin R(\mathbf{B}-\lambda\mathbf{I})$, and if

$$\mathbf{y}_n = \mathbf{B}\mathbf{x}_n = (\mathbf{A}-\alpha\mathbf{I})^{-1}\mathbf{x}_n, \quad \nu_n = m(\mathbf{y}_n), \quad \mathbf{x}_{n+1} = \frac{\mathbf{y}_n}{\nu_n} \quad \text{for } n = 0, 1, 2, \ldots,$$

then $(\nu_n, \mathbf{x}_n) \to \left((\lambda-\alpha)^{-1}, \mathbf{x}\right)$, an eigenpair for $\mathbf{B}$, so (7.3.18) guarantees that $(\nu_n^{-1}+\alpha,\ \mathbf{x}_n) \to (\lambda, \mathbf{x})$, an eigenpair for $\mathbf{A}$. Rather than using matrix inversion to compute $\mathbf{y}_n = (\mathbf{A}-\alpha\mathbf{I})^{-1}\mathbf{x}_n$, it's more efficient to solve the linear system $(\mathbf{A}-\alpha\mathbf{I})\mathbf{y}_n = \mathbf{x}_n$ for $\mathbf{y}_n$. Because this is a system in which the coefficient matrix remains the same from step to step, the efficiency is further enhanced by computing an LU factorization of $(\mathbf{A}-\alpha\mathbf{I})$ at the outset so that at each step only one forward solve and one back solve (as described on pp. 146 and 153) are needed to determine $\mathbf{y}_n$.

- ▷ **Advantages.** Striking results are often obtained (particularly in the case of symmetric matrices) with only one or two iterations, even when $\mathbf{x}_0$ is nearly in $R(\mathbf{B}-\lambda\mathbf{I}) = R(\mathbf{A}-\lambda\mathbf{I})$. For α close to λ, computing an accurate floating-point solution of $(\mathbf{A}-\alpha\mathbf{I})\mathbf{y}_n = \mathbf{x}_n$ is difficult because $\mathbf{A}-\alpha\mathbf{I}$ is nearly singular, and this almost surely guarantees that $(\mathbf{A}-\alpha\mathbf{I})\mathbf{y}_n = \mathbf{x}_n$ is an ill-conditioned system. But only the direction of the solution is important, and the direction of a computed solution is usually reasonable in spite of conditioning problems. Finally, the algorithm can be adapted to compute approximations of eigenvectors associated with complex eigenvalues.
- ▷ **Disadvantages.** Only one eigenpair at a time is computed, and an approximate eigenvalue must be known in advance. Furthermore, the rate of convergence depends on how fast $[(\lambda-\alpha)/(\lambda_i-\alpha)]^n \to 0$, and this can be slow when there is another eigenvalue λ_i close to the desired λ. If λ_i is too close to λ, roundoff error can divert inverse iteration toward an eigenvector associated with λ_i instead of λ in spite of a theoretically correct α.

Note: In the standard version of inverse iteration a constant value of α is used at each step to approximate an eigenvalue λ, but there is variation called ***Rayleigh quotient iteration*** that uses the current iterate $\mathbf{x}_n$ to improve the value of α at each step by setting $\alpha = \mathbf{x}_n^T\mathbf{A}\mathbf{x}_n/\mathbf{x}_n^T\mathbf{x}_n$. The function $R(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}/\mathbf{x}^T\mathbf{x}$ is called the *Rayleigh quotient*. It can be shown that if $\mathbf{x}$ is a good approximation to an eigenvector, then $R(\mathbf{x})$ is a good approximation of the associated eigenvalue. More is said about this in Example 7.5.1 (p. 549).

Example 7.3.9

The QR Iteration algorithm for computing the eigenvalues of a general matrix came from an elegantly simple idea that was proposed by Heinz Rutishauser in 1958 and refined by J. F. G. Francis in 1961-1962. The underlying concept is to alternate between computing QR factors (Rutishauser used LU factors) and

reversing their order as shown below. Starting with $\mathbf{A}_1 = \mathbf{A} \in \Re^{n\times n}$,

$$\begin{aligned} \text{Factor:} \quad & \mathbf{A}_1 = \mathbf{Q}_1\mathbf{R}_1, \\ \text{Set:} \quad & \mathbf{A}_2 = \mathbf{R}_1\mathbf{Q}_1, \\ \text{Factor:} \quad & \mathbf{A}_2 = \mathbf{Q}_2\mathbf{R}_2, \\ \text{Set:} \quad & \mathbf{A}_3 = \mathbf{R}_2\mathbf{Q}_2, \\ & \vdots \end{aligned}$$

In general, $\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k$, where $\mathbf{Q}_k$ and $\mathbf{R}_k$ are the QR factors of $\mathbf{A}_k$. Notice that if $\mathbf{P}_k = \mathbf{Q}_1\mathbf{Q}_2\cdots\mathbf{Q}_k$, then each $\mathbf{P}_k$ is an orthogonal matrix such that

$$\begin{aligned} \mathbf{P}_1^T\mathbf{A}\mathbf{P}_1 &= \mathbf{Q}_1^T\mathbf{Q}_1\mathbf{R}_1\mathbf{Q}_1 = \mathbf{A}_2, \\ \mathbf{P}_2^T\mathbf{A}\mathbf{P}_2 &= \mathbf{Q}_2^T\mathbf{Q}_1^T\mathbf{A}\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2^T\mathbf{A}_2\mathbf{Q}_2 = \mathbf{A}_3, \\ &\vdots \\ \mathbf{P}_k^T\mathbf{A}\mathbf{P}_k &= \mathbf{A}_{k+1}. \end{aligned}$$

In other words, $\mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4, \ldots$ are each orthogonally similar to $\mathbf{A}$, and hence $\sigma(\mathbf{A}_k) = \sigma(\mathbf{A})$ for each k. But the process does more than just create a matrix that is similar to $\mathbf{A}$ at each step. The magic lies in the fact that if the process converges, then $\lim_{k\to\infty}\mathbf{A}_k = \mathbf{R}$ is an upper-triangular matrix in which the diagonal entries are the eigenvalues of $\mathbf{A}$. Indeed, if $\mathbf{P}_k \to \mathbf{P}$, then

$$\mathbf{Q}_k = \mathbf{P}_{k-1}^T\mathbf{P}_k \to \mathbf{P}^T\mathbf{P} = \mathbf{I} \quad \text{and} \quad \mathbf{R}_k = \mathbf{A}_{k+1}\mathbf{Q}_k^T \to \mathbf{R}\,\mathbf{I} = \mathbf{R},$$

so

$$\lim_{k\to\infty}\mathbf{A}_k = \lim_{k\to\infty}\mathbf{Q}_k\mathbf{R}_k = \mathbf{R},$$

which is necessarily upper triangular having diagonal entries equal to the eigenvalues of $\mathbf{A}$. However, as is often the case, there is a big gap between theory and practice, and turning this clever idea into a practical algorithm requires significant effort. For example, one obvious hurdle that needs to be overcome is the fact that the R factor in a QR factorization has positive diagonal entries, so, unless modifications are made, the "vanilla" version of the QR iteration can't converge for matrices with complex or nonpositive eigenvalues. Laying out all of the details and analyzing the rigors that constitute the practical implementation of the QR iteration is tedious and would take us too far astray, but the basic principals are within our reach.

- **Hessenberg Matrices.** A big step in turning the QR iteration into a practical method is to realize that everything can be done with upper-Hessenberg matrices. As discussed in Example 5.7.4 (p. 350), Householder reduction can be used to produce an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{H}_1$, and Example 5.7.5 (p. 352) shows that Givens reduction easily produces

the QR factors of any Hessenberg matrix. Givens reduction on $\mathbf{H}_1$ produces the Q factor of $\mathbf{H}_1$ as the transposed product of plane rotations $\mathbf{Q}_1 = \mathbf{P}_{12}^T\mathbf{P}_{23}^T\cdots\mathbf{P}_{(n-1)n}^T$, and this is also upper Hessenberg (constructing a 4×4 example will convince you). Since multiplication by an upper-triangular matrix can't alter the upper-Hessenberg structure, the matrix $\mathbf{R}_1\mathbf{Q}_1 = \mathbf{H}_2$ at the second step of the QR iteration is again upper Hessenberg, and so on for each successive step. Being able to iterate with Hessenberg matrices results in a significant reduction of arithmetic. Note that if $\mathbf{A} = \mathbf{A}^T$, then $\mathbf{H}_k = \mathbf{H}_k^T$ for each k, which means that each $\mathbf{H}_k$ is tridiagonal in structure.

- **Convergence.** When the $\mathbf{H}_k$'s converge, the entries at the bottom of the first subdiagonal tend to die first—i.e., a typical pattern might be

$$\mathbf{H}_k = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & \epsilon & \star \end{pmatrix}.$$

When ϵ is satisfactorily small, take $\star$ (the (n,n)-entry) to be an eigenvalue, and deflate the problem. An even nicer state of affairs is to have a zero (or a satisfactorily small) entry in row $n-1$ and column 2 (illustrated below for $n=4$)

$$\mathbf{H}_k = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & \epsilon & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix} \tag{7.3.19}$$

because the trailing 2×2 block $\begin{pmatrix} \star & \star \\ \star & \star \end{pmatrix}$ will yield two eigenvalues by the quadratic formula, and thus complex eigenvalues can be revealed.

- **Shifts.** Instead of factoring $\mathbf{H}_k$ at the k^{th} step, factor a shifted matrix $\mathbf{H}_k - \alpha_k\mathbf{I} = \mathbf{Q}_k\mathbf{R}_k$, and set $\mathbf{H}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + \alpha_k\mathbf{I}$, where α_k is an approximate *real* eigenvalue—a good candidate is $\alpha_k = [\mathbf{H}_k]_{nn}$. Notice that $\sigma(\mathbf{H}_{k+1}) = \sigma(\mathbf{H}_k)$ because $\mathbf{H}_{k+1} = \mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k$. The inverse power method is now at work. To see how, drop the subscripts, and write $\mathbf{H} - \alpha\mathbf{I} = \mathbf{QR}$ as $\mathbf{Q}^T = \mathbf{R}(\mathbf{H} - \alpha\mathbf{I})^{-1}$. If $\alpha \approx \lambda \in \sigma(\mathbf{H}) = \sigma(\mathbf{A})$ (say, $|\lambda - \alpha| = \epsilon$ with $\alpha, \lambda \in \Re$), then the discussion concerning the inverse power method in Example 7.3.8 insures that the rows in $\mathbf{Q}^T$ are close to being left-hand eigenvectors of $\mathbf{H}$ associated with λ. In particular, if $\mathbf{q}_n^T$ is the last row in $\mathbf{Q}^T$, then

$$r_{nn}\mathbf{e}_n^T = \mathbf{e}_n^T\mathbf{R} = \mathbf{q}_n^T\mathbf{QR} = \mathbf{q}_n^T(\mathbf{H} - \alpha\mathbf{I}) = \mathbf{q}_n^T\mathbf{H} - \alpha\mathbf{q}_n^T \approx (\lambda - \alpha)\mathbf{q}_n^T,$$

so $r_{nn} = |r_{nn}| \approx \left\|(\lambda - \alpha)\mathbf{q}_n^T\right\|_2 = \epsilon$ and $\mathbf{q}_n^T \approx \pm\mathbf{e}_n^T$. The significance of this

is revealed by looking at a generic 4×4 pattern for

$$\begin{aligned}\mathbf{H}_{k+1} &= \mathbf{RQ} + \alpha\mathbf{I} \\ &\approx \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & \epsilon \end{pmatrix} \begin{pmatrix} * & * & * & 0 \\ * & * & * & 0 \\ 0 & * & * & 0 \\ 0 & 0 & \star & \pm 1 \end{pmatrix} + \begin{pmatrix} \alpha & & & \\ & \alpha & & \\ & & \alpha & \\ & & & \alpha \end{pmatrix} \\ &= \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & \epsilon\star & \alpha \pm \epsilon \end{pmatrix} \approx \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & 0 & \alpha \pm \epsilon \end{pmatrix}.\end{aligned}$$

The strength of the last approximation rests not only on the size of ϵ, but it is also reinforced by the fact that $\star \approx 0$ because the 2-norm of the last row of $\mathbf{Q}$ must be 1. This indicates why this technique (called the *single shifted* QR iteration) can provide rapid convergence to a real eigenvalue. To extract complex eigenvalues, a *double shift* strategy is employed in which the eigenvalues α_k and β_k of the lower 2×2 block of $\mathbf{H}_k$ are used as shifts as indicated below:

$$\begin{aligned} &\text{Factor:} && \mathbf{H}_k - \alpha_k\mathbf{I} = \mathbf{Q}_k\mathbf{R}_k, \\ &\text{Set:} && \mathbf{H}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + \alpha_k\mathbf{I} && (\text{so } \mathbf{H}_{k+1} = \mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k), \\ &\text{Factor:} && \mathbf{H}_{k+1} - \beta_k\mathbf{I} = \mathbf{Q}_{k+1}\mathbf{R}_{k+1}, \\ &\text{Set:} && \mathbf{H}_{k+2} = \mathbf{R}_{k+1}\mathbf{Q}_{k+1} + \beta_k\mathbf{I} && (\text{so } \mathbf{H}_{k+2} = \mathbf{Q}_{k+1}^T\mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k\mathbf{Q}_{k+1}), \\ &&& \quad\vdots \end{aligned}$$

The nice thing about the double shift strategy is that even when α_k is complex (so that $\beta_k = \overline{\alpha}_k$) the matrix $\mathbf{Q}_k\mathbf{Q}_{k+1}$ (and hence $\mathbf{H}_{k+2}$) is real, and there are efficient ways to form $\mathbf{Q}_k\mathbf{Q}_{k+1}$ by computing only the first column of the product. The double shift method typically requires very few iterations (using only real arithmetic) to produce a small entry in the $(n-2, 2)$-position as depicted in (7.3.19) for a generic 4×4 pattern.

Exercises for section 7.3

7.3.1. Determine $\cos \mathbf{A}$ for $\mathbf{A} = \begin{pmatrix} -\pi/2 & \pi/2 \\ \pi/2 & -\pi/2 \end{pmatrix}$.

7.3.2. For the matrix $\mathbf{A}$ in Example 7.3.3, verify with direct computation that $e^{\lambda_1 t}\mathbf{G}_1 + e^{\lambda_2 t}\mathbf{G}_2 = \mathbf{P}\begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix}\mathbf{P}^{-1} = e^{\mathbf{A}t}$.

7.3.3. Explain why $\sin^2 \mathbf{A} + \cos^2 \mathbf{A} = \mathbf{I}$ for a diagonalizable matrix $\mathbf{A}$.

7.3.4. Explain $e^{\mathbf{0}} = \mathbf{I}$ for every square zero matrix.

7.3.5. The ***spectral mapping property*** for diagonalizable matrices says that if $f(\mathbf{A})$ exists, and if $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are the eigenvalues of $\mathbf{A}_{n\times n}$ (including multiplicities), then $\{f(\lambda_1), \ldots, f(\lambda_n)\}$ are the eigenvalues of $f(\mathbf{A})$.

(a) Establish this for diagonalizable matrices.

(b) Establish this when an infinite series $f(z) = \sum_{n=0}^{\infty} c_n(z - z_0)^n$ defines $f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n(\mathbf{A} - z_0\mathbf{I})^n$ as discussed in (7.3.7).

7.3.6. Explain why $\det\left(e^{\mathbf{A}}\right) = e^{trace(\mathbf{A})}$.

7.3.7. Suppose that for nondiagonalizable matrices $\mathbf{A}_{m\times m}$ an infinite series $f(z) = \sum_{n=0}^{\infty} c_n(z - z_0)^n$ is used to define $f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n(\mathbf{A} - z_0\mathbf{I})^n$ as suggested in (7.3.7). Neglecting convergence issues, explain why there is a polynomial $p(z)$ of at most degree $m-1$ such that $f(\mathbf{A}) = p(\mathbf{A})$.

7.3.8. If $f(\mathbf{A})$ exists for a diagonalizable $\mathbf{A}$, explain why $\mathbf{A}f(\mathbf{A}) = f(\mathbf{A})\mathbf{A}$. What can you say when $\mathbf{A}$ is not diagonalizable?

7.3.9. Explain why $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$ whenever $\mathbf{AB} = \mathbf{BA}$. Give an example to show that $e^{\mathbf{A}+\mathbf{B}}$, $e^{\mathbf{A}}e^{\mathbf{B}}$, and $e^{\mathbf{B}}e^{\mathbf{A}}$ all can differ when $\mathbf{AB} \neq \mathbf{BA}$. **Hint:** Exercise 7.2.16 can be used for the diagonalizable case. For the general case, consider $\mathbf{F}(t) = e^{(\mathbf{A}+\mathbf{B})t} - e^{\mathbf{A}t}e^{\mathbf{B}t}$ and $\mathbf{F}'(t)$.

7.3.10. Show that $e^{\mathbf{A}}$ is an orthogonal matrix whenever $\mathbf{A}$ is skew symmetric.

7.3.11. A particular electronic device consists of a collection of switching circuits that can be either in an ON state or an OFF state. These electronic switches are allowed to change state at regular time intervals called *clock cycles.* Suppose that at the end of each clock cycle, 30% of the switches currently in the OFF state change to ON, while 90% of those in the ON state revert to the OFF state.

(a) Show that the device approaches an equilibrium in the sense that the proportion of switches in each state eventually becomes constant, and determine these equilibrium proportions.

(b) Independent of the initial proportions, about how many clock cycles does it take for the device to become essentially stable?

7.3.12. The ***spectral radius*** of $\mathbf{A}$ is $\rho(\mathbf{A}) = \max_{\lambda_i \in \sigma(\mathbf{A})} |\lambda_i|$ (p. 497). Prove that if $\mathbf{A}$ is diagonalizable, then

$$\rho(\mathbf{A}) = \lim_{n\to\infty} \|\mathbf{A}^n\|^{1/n} \quad \text{for every matrix norm.}$$

This result is true for nondiagonalizable matrices as well, but the proof at this point in the game is more involved. The full development is given in Example 7.10.1 (p. 619).

7.3.13. Find a dominant eigenpair for $\mathbf{A} = \begin{pmatrix} 7 & 2 & 3 \\ 0 & 2 & 0 \\ -6 & -2 & -2 \end{pmatrix}$ by the power method.

7.3.14. Apply the inverse power method (Example 7.3.8, p. 534) to find an eigenvector for each of the eigenvalues of the matrix $\mathbf{A}$ in Exercise 7.3.13.

7.3.15. Explain why the function $m(\mathbf{v})$ used in the development of the power method in Example 7.3.7 is not a continuous function, so statements like $m(\mathbf{x}_n) \to m(\mathbf{x})$ when $\mathbf{x}_n \to \mathbf{x}$ are not valid. Nevertheless, if $\lim_{n\to\infty} \mathbf{x}_n \neq \mathbf{0}$, then $\lim_{n\to\infty} m(\mathbf{x}_n) \neq 0$.

7.3.16. Let $\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & -2 & -1 \\ 0 & 2 & 1 \end{pmatrix}$.

(a) Apply the "vanilla" QR iteration to $\mathbf{H}$.
(b) Apply the the single shift QR iteration on $\mathbf{H}$.

7.3.17. Show that the QR iteration can fail to converge using $\mathbf{H} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$.

(a) First use the "vanilla" QR iteration on $\mathbf{H}$ to see what happens.
(b) Now try the single shift QR iteration on $\mathbf{H}$.
(c) Finally, execute the double shift QR iteration on $\mathbf{H}$.

7.4 SYSTEMS OF DIFFERENTIAL EQUATIONS

Systems of first-order linear differential equations with constant coefficients were used in §7.1 to motivate the introduction of eigenvalues and eigenvectors, but now we can delve a little deeper. For constants a_{ij}, the goal is to solve the following system for the unknown functions $u_i(t)$.

$$\begin{array}{ll} u_1' = a_{11}u_1 + a_{12}u_2 + \cdots + a_{1n}u_n, & u_1(0) = c_1, \\ u_2' = a_{21}u_1 + a_{22}u_2 + \cdots + a_{2n}u_n, & u_2(0) = c_2, \\ \quad\vdots \qquad\qquad \text{with} & \quad\vdots \\ u_n' = a_{n1}u_1 + a_{n2}u_2 + \cdots + a_{nn}u_n, & u_n(0) = c_n. \end{array} \tag{7.4.1}$$

Since the scalar exponential provides the unique solution to a single differential equation $u'(t) = \alpha u(t)$ with $u(0) = c$ as $u(t) = \mathrm{e}^{\alpha t}c$, it's only natural to try to use the matrix exponential in an analogous way to solve a system of differential equations. Begin by writing (7.4.1) in matrix form as $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where

$$\mathbf{u} = \begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_n(t) \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}.$$

If $\mathbf{A}$ is diagonalizable with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then (7.3.6) guarantees

$$\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\lambda_1 t}\mathbf{G}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{G}_2 + \cdots + \mathrm{e}^{\lambda_k t}\mathbf{G}_k. \tag{7.4.2}$$

The following identities are derived from properties of the $\mathbf{G}_i$'s given on p. 517.

- $d\mathrm{e}^{\mathbf{A}t}/dt = \sum_{i=1}^{k} \lambda_i \mathrm{e}^{\lambda_i t}\mathbf{G}_i = \left(\sum_{i=1}^{k} \lambda_i \mathbf{G}_i\right)\left(\sum_{i=1}^{k} \mathrm{e}^{\lambda_i t}\mathbf{G}_i\right) = \mathbf{A}\mathrm{e}^{\mathbf{A}t}.$ (7.4.3)
- $\mathbf{A}\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\mathbf{A}t}\mathbf{A}$ (by a similar argument). (7.4.4)
- $\mathrm{e}^{-\mathbf{A}t}\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\mathbf{A}t}\mathrm{e}^{-\mathbf{A}t} = \mathbf{I} = \mathrm{e}^{\mathbf{0}}$ (by a similar argument). (7.4.5)

Equation (7.4.3) insures that $\mathbf{u} = \mathrm{e}^{\mathbf{A}t}\mathbf{c}$ is *one* solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$. To see that $\mathbf{u} = \mathrm{e}^{\mathbf{A}t}\mathbf{c}$ is the *only* solution, suppose $\mathbf{v}(t)$ is another solution so that $\mathbf{v}' = \mathbf{A}\mathbf{v}$ with $\mathbf{v}(0) = \mathbf{c}$. Differentiating $\mathrm{e}^{-\mathbf{A}t}\mathbf{v}$ produces

$$\frac{d\left[\mathrm{e}^{-\mathbf{A}t}\mathbf{v}\right]}{dt} = \mathrm{e}^{-\mathbf{A}t}\mathbf{v}' - \mathrm{e}^{-\mathbf{A}t}\mathbf{A}\mathbf{v} = \mathbf{0}, \quad \text{so} \quad \mathrm{e}^{-\mathbf{A}t}\mathbf{v} \text{ is constant for all } t.$$

At $t = 0$ we have $e^{-\mathbf{A}t}\mathbf{v}\big|_{t=0} = e^{\mathbf{0}}\mathbf{v}(0) = \mathbf{Ic} = \mathbf{c}$, and hence $e^{-\mathbf{A}t}\mathbf{v} = \mathbf{c}$ for all t. Multiply both sides of this equation by $e^{\mathbf{A}t}$ and use (7.4.5) to conclude $\mathbf{v} = e^{\mathbf{A}t}\mathbf{c}$. Thus $\mathbf{u} = e^{\mathbf{A}t}\mathbf{c}$ is the unique solution to $\mathbf{u}' = \mathbf{Au}$ with $\mathbf{u}(0) = \mathbf{c}$.

Finally, notice that $\mathbf{v}_i = \mathbf{G}_i\mathbf{c} \in N(\mathbf{A} - \lambda_i\mathbf{I})$ is an eigenvector associated with λ_i, so that the solution to $\mathbf{u}' = \mathbf{Au}$, $\mathbf{u}(0) = \mathbf{c}$, is

$$\mathbf{u} = e^{\lambda_1 t}\mathbf{v}_1 + e^{\lambda_2 t}\mathbf{v}_2 + \cdots + e^{\lambda_k t}\mathbf{v}_k, \tag{7.4.6}$$

and this solution is completely determined by the eigenpairs $(\lambda_i, \mathbf{v}_i)$. It turns out that $\mathbf{u}$ also can be expanded in terms of *any* complete set of independent eigenvectors—see Exercise 7.4.1. Let's summarize what's been said so far.

Differential Equations

If $\mathbf{A}_{n\times n}$ is diagonalizable with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then the unique solution of $\mathbf{u}' = \mathbf{Au}$, $\mathbf{u}(0) = \mathbf{c}$, is given by

$$\mathbf{u} = e^{\mathbf{A}t}\mathbf{c} = e^{\lambda_1 t}\mathbf{v}_1 + e^{\lambda_2 t}\mathbf{v}_2 + \cdots + e^{\lambda_k t}\mathbf{v}_k \tag{7.4.7}$$

in which $\mathbf{v}_i$ is the eigenvector $\mathbf{v}_i = \mathbf{G}_i\mathbf{c}$, where $\mathbf{G}_i$ is the i^{th} spectral projector. (See Exercise 7.4.1 for an alternate eigenexpansion.) Nonhomogeneous systems as well as the nondiagonalizable case are treated in Example 7.9.6 (p. 608).

Example 7.4.1

An Application to Diffusion. Important issues in medicine and biology involve the question of how drugs or chemical compounds move from one cell to another by means of diffusion through cell walls. Consider two cells, as depicted in Figure 7.4.1, which are both devoid of a particular compound. A unit amount of the compound is injected into the first cell at time $t = 0$, and as time proceeds the compound diffuses according to the following assumption.

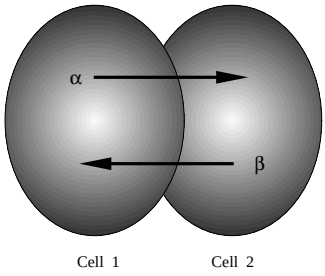


FIGURE 7.4.1

At each point in time the rate (amount per second) of diffusion from one cell to the other is proportional to the concentration (amount per unit volume) of the compound in the cell giving up the compound—say the rate of diffusion from cell 1 to cell 2 is α times the concentration in cell 1, and the rate of diffusion from cell 2 to cell 1 is β times the concentration in cell 2. Assume $\alpha,\ \beta > 0$.

Problem: Determine the concentration of the compound in each cell at any given time t, and, in the long run, determine the steady-state concentrations.

Solution: If $u_k = u_k(t)$ denotes the concentration of the compound in cell k at time t, then the statements in the above assumption are translated as follows:

$$\begin{aligned}
\frac{du_1}{dt} &= \text{ rate in } - \text{ rate out } = \beta u_2 - \alpha u_1, \quad \text{where} \quad u_1(0) = 1, \\
\frac{du_2}{dt} &= \text{ rate in } - \text{ rate out } = \alpha u_1 - \beta u_2, \quad \text{where} \quad u_2(0) = 0.
\end{aligned}$$

In matrix notation this system is $\mathbf{u}' = \mathbf{A}\mathbf{u},\ \ \mathbf{u}(0) = \mathbf{c},$ where

$$\mathbf{A} = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Since $\mathbf{A}$ is the matrix of Example 7.3.3 we can use the results from Example 7.3.3 to write the solution as

$$\mathbf{u}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{c} = \frac{1}{\alpha+\beta}\left[\begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix} + \mathrm{e}^{-(\alpha+\beta)t}\begin{pmatrix} \alpha & -\beta \\ -\alpha & \beta \end{pmatrix}\right]\begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

so that

$$u_1(t) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\mathrm{e}^{-(\alpha+\beta)t} \quad \text{and} \quad u_2(t) = \frac{\alpha}{\alpha+\beta}\left(1 - \mathrm{e}^{-(\alpha+\beta)t}\right).$$

In the long run, the concentrations in each cell stabilize in the sense that

$$\lim_{t\to\infty} u_1(t) = \frac{\beta}{\alpha+\beta} \quad \text{and} \quad \lim_{t\to\infty} u_2(t) = \frac{\alpha}{\alpha+\beta}.$$

An innumerable variety of physical situations can be modeled by $\mathbf{u}' = \mathbf{A}\mathbf{u}$, and the form of the solution (7.4.6) makes it clear that the eigenvalues and eigenvectors of $\mathbf{A}$ are intrinsic to the underlying physical phenomenon being investigated. We might say that the eigenvalues and eigenvectors of $\mathbf{A}$ act as its genes and chromosomes because they are the basic components that either dictate or govern all other characteristics of $\mathbf{A}$ along with the physics of associated phenomena.

For example, consider the long-run behavior of a physical system that can be modeled by $\mathbf{u}' = \mathbf{A}\mathbf{u}$. We usually want to know whether the system will eventually blow up or will settle down to some sort of stable state. Might it neither blow up nor settle down but rather oscillate indefinitely? These are questions concerning the nature of the limit

$$\lim_{t\to\infty} \mathbf{u}(t) = \lim_{t\to\infty} \mathrm{e}^{\mathbf{A}t}\mathbf{c} = \lim_{t\to\infty} \left(\mathrm{e}^{\lambda_1 t}\mathbf{G}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{G}_2 + \cdots + \mathrm{e}^{\lambda_k t}\mathbf{G}_k\right)\mathbf{c},$$

and the answers depend only on the eigenvalues. To see how, recall that for a complex number $\lambda = x + \mathrm{i}y$ and a real parameter $t > 0$,

$$\mathrm{e}^{\lambda t} = \mathrm{e}^{(x+\mathrm{i}y)t} = \mathrm{e}^{xt}\mathrm{e}^{\mathrm{i}yt} = \mathrm{e}^{xt}\left(\cos yt + \mathrm{i}\sin yt\right). \tag{7.4.8}$$

The term $\mathrm{e}^{\mathrm{i}yt} = (\cos yt + \mathrm{i}\sin yt)$ is a point on the unit circle that oscillates as a function of t, so $|\mathrm{e}^{\mathrm{i}yt}| = |\cos yt + \mathrm{i}\sin yt| = 1$ and $\left|\mathrm{e}^{\lambda t}\right| = |\mathrm{e}^{xt}\mathrm{e}^{\mathrm{i}yt}| = |\mathrm{e}^{xt}| = \mathrm{e}^{xt}$. This makes it clear that if $\mathrm{Re}\,(\lambda_i) < 0$ for each i, then, as $t \to \infty$, $\mathrm{e}^{\mathbf{A}t} \to \mathbf{0}$, and $\mathbf{u}(t) \to \mathbf{0}$ for every initial vector $\mathbf{c}$. Thus the system eventually settles down to zero, and we say the system is *stable.* On the other hand, if $\mathrm{Re}\,(\lambda_i) > 0$ for some i, then components of $\mathbf{u}(t)$ may become unbounded as $t \to \infty$, and we say the system is *unstable.* Finally, if $\mathrm{Re}\,(\lambda_i) \leq 0$ for each i, then the components of $\mathbf{u}(t)$ remain finite for all t, but some may oscillate indefinitely, and this is called a *semistable* situation. Below is a summary of stability.

Stability

Let $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where $\mathbf{A}$ is diagonalizable with eigenvalues λ_i.

- If $\mathrm{Re}\,(\lambda_i) < 0$ for each i, then $\lim_{t\to\infty} \mathrm{e}^{\mathbf{A}t} = \mathbf{0}$, and $\lim_{t\to\infty} \mathbf{u}(t) = \mathbf{0}$ for every initial vector $\mathbf{c}$. In this case $\mathbf{u}' = \mathbf{A}\mathbf{u}$ is said to be a ***stable system,*** and $\mathbf{A}$ is called a ***stable matrix.***
- If $\mathrm{Re}\,(\lambda_i) > 0$ for some i, then components of $\mathbf{u}(t)$ can become unbounded as $t \to \infty$, in which case the system $\mathbf{u}' = \mathbf{A}\mathbf{u}$ as well as the underlying matrix $\mathbf{A}$ are said to be ***unstable.***
- If $\mathrm{Re}\,(\lambda_i) \leq 0$ for each i, then the components of $\mathbf{u}(t)$ remain finite for all t, but some can oscillate indefinitely. This is called a ***semistable*** situation.

Example 7.4.2

Predator–Prey Application. Consider two species of which one is the predator and the other is the prey, and assume there are initially 100 in each population. Let $u_1(t)$ and $u_2(t)$ denote the respective population of the predator and

prey species at time t, and suppose their growth rates are given by

$$u_1' = u_1 + u_2,$$
$$u_2' = -u_1 + u_2.$$

Problem: Determine the size of each population at all future times, and decide if (and when) either population will eventually become extinct.

Solution: Write the system as $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} 100 \\ 100 \end{pmatrix}.$$

The characteristic equation for $\mathbf{A}$ is $p(\lambda) = \lambda^2 - 2\lambda + 2 = 0$, so the eigenvalues for $\mathbf{A}$ are $\lambda_1 = 1 + \mathrm{i}$ and $\lambda_2 = 1 - \mathrm{i}$. We know from (7.4.7) that

$$\mathbf{u}(t) = \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{v}_2 \quad (\text{where } \mathbf{v}_i = \mathbf{G}_i\mathbf{c}\,) \tag{7.4.9}$$

is the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$. The spectral theorem on p. 517 implies $\mathbf{A} - \lambda_2\mathbf{I} = (\lambda_1 - \lambda_2)\mathbf{G}_1$ and $\mathbf{I} = \mathbf{G}_1 + \mathbf{G}_2$, so $(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c} = (\lambda_1 - \lambda_2)\mathbf{v}_1$ and $\mathbf{c} = \mathbf{v}_1 + \mathbf{v}_2$, and consequently

$$\mathbf{v}_1 = \frac{(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c}}{(\lambda_1 - \lambda_2)} = 50\begin{pmatrix} \lambda_2 \\ \lambda_1 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_2 = \mathbf{c} - \mathbf{v}_1 = 50\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}.$$

With the aid of (7.4.8) we obtain the solution components from (7.4.9) as

$$u_1(t) = 50\left(\lambda_2\mathrm{e}^{\lambda_1 t} + \lambda_1\mathrm{e}^{\lambda_2 t}\right) = 100\mathrm{e}^t(\cos t + \sin t)$$

and

$$u_2(t) = 50\left(\lambda_1\mathrm{e}^{\lambda_1 t} + \lambda_2\mathrm{e}^{\lambda_2 t}\right) = 100\mathrm{e}^t(\cos t - \sin t).$$

The system is unstable because $\mathrm{Re}\,(\lambda_i) > 0$ for each eigenvalue. Indeed, $u_1(t)$ and $u_2(t)$ both become unbounded as $t \to \infty$. However, a population cannot become negative–once it's zero, it's extinct. Figure 7.4.2 shows that the graph of $u_2(t)$ will cross the horizontal axis before that of $u_1(t)$.

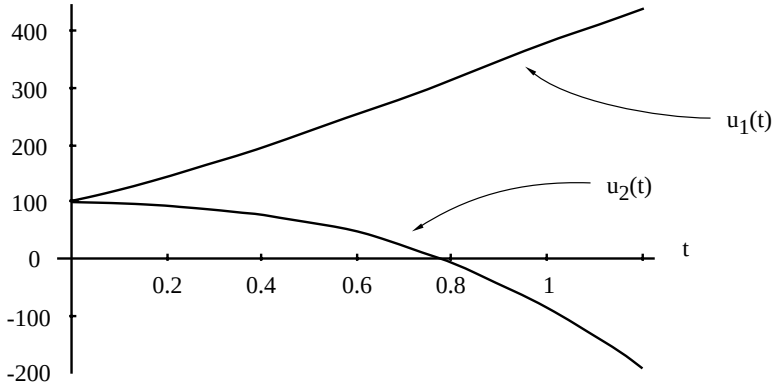


FIGURE 7.4.2

Therefore, the prey species will become extinct at the value of t for which $u_2(t) = 0$—i.e., when

$$100e^t(\cos t - \sin t) = 0 \implies \cos t = \sin t \implies t = \frac{\pi}{4}.$$

Exercises for section 7.4

7.4.1. Suppose that $\mathbf{A}_{n\times n}$ is diagonalizable, and let $\mathbf{P} = [\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n]$ be a matrix whose columns are a complete set of linearly independent eigenvectors corresponding to eigenvalues λ_i. Show that the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, can be written as

$$\mathbf{u}(t) = \xi_1 e^{\lambda_1 t}\mathbf{x}_1 + \xi_2 e^{\lambda_2 t}\mathbf{x}_2 + \cdots + \xi_n e^{\lambda_n t}\mathbf{x}_n$$

in which the coefficients ξ_i satisfy the algebraic system $\mathbf{P}\boldsymbol{\xi} = \mathbf{c}$.

7.4.2. Using only the eigenvalues, determine the long-run behavior of the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$ for each of the following matrices.

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

7.4.3. Competing Species. Consider two species that coexist in the same environment but compete for the same resources. Suppose that the population of each species increases proportionally to the number of its own kind but decreases proportionally to the number in the competing species—say that the population of each species increases at a rate equal to twice its existing number but decreases at a rate equal to the number in the other population. Suppose that there are initially 100 of species I and 200 of species II.

(a) Determine the number of each species at all future times.
(b) Determine which species is destined to become extinct, and compute the time to extinction.

7.4.4. Cooperating Species. Consider two species that survive in a symbiotic relationship in the sense that the population of each species decreases at a rate equal to its existing number but increases at a rate equal to the existing number in the other population.

(a) If there are initially 200 of species I and 400 of species II, determine the number of each species at all future times.
(b) Discuss the long-run behavior of each species.

7.5 NORMAL MATRICES

A matrix $\mathbf{A}$ is diagonalizable if and only if $\mathbf{A}$ possesses a complete independent set of eigenvectors, and if such a complete set is used for columns of $\mathbf{P}$, then $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$ is diagonal (p. 507). But even when $\mathbf{A}$ possesses a complete independent set of eigenvectors, there's no guarantee that a complete *orthonormal* set of eigenvectors can be found. In other words, there's no assurance that $\mathbf{P}$ can be taken to be unitary (or orthogonal). And the Gram–Schmidt procedure (p. 309) doesn't help—Gram–Schmidt can turn a basis of eigenvectors into an orthonormal basis but not into an orthonormal basis of eigenvectors. So when (or how) are complete orthonormal sets of eigenvectors produced? In other words, when is $\mathbf{A}$ *unitarily* similar to a diagonal matrix?

Unitary Diagonalization

$\mathbf{A} \in \mathcal{C}^{n\times n}$ is unitarily similar to a diagonal matrix (i.e., $\mathbf{A}$ has a complete orthonormal set of eigenvectors) if and only if $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$, in which case $\mathbf{A}$ is said to be a ***normal matrix.***

- Whenever $\mathbf{U}^*\mathbf{AU} = \mathbf{D}$ with $\mathbf{U}$ unitary and $\mathbf{D}$ diagonal, the columns of $\mathbf{U}$ must be a complete orthonormal set of eigenvectors for $\mathbf{A}$, and the diagonal entries of $\mathbf{D}$ are the associated eigenvalues.

Proof. If $\mathbf{A}$ is normal with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then $\mathbf{A} - \lambda_k\mathbf{I}$ is also normal. All normal matrices are RPN (range is perpendicular to nullspace, p. 409), so there is a unitary matrix $\mathbf{U}_k$ such that

$$\mathbf{U}_k^*(\mathbf{A} - \lambda_k\mathbf{I})\mathbf{U}_k = \begin{pmatrix} \mathbf{C}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{(by (5.11.15) on p. 408)}$$

or, equivalently

$$\mathbf{U}_k^*\mathbf{AU}_k = \begin{pmatrix} \mathbf{C}_k + \lambda_k\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \lambda_k\mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{k-1} & \mathbf{0} \\ \mathbf{0} & \lambda_k\mathbf{I} \end{pmatrix},$$

where $\mathbf{C}_k$ is nonsingular and $\mathbf{A}_{k-1} = \mathbf{C}_k + \lambda_k\mathbf{I}$. Note that $\lambda_k \notin \sigma(\mathbf{A}_{k-1})$ (otherwise $\mathbf{A}_{k-1} - \lambda_k\mathbf{I} = \mathbf{C}_k$ would be singular), so $\sigma(\mathbf{A}_{k-1}) = \{\lambda_1, \lambda_2, \ldots, \lambda_{k-1}\}$ (Exercise 7.1.4). Because $\mathbf{A}_{k-1}$ is also normal, the same argument can be repeated with $\mathbf{A}_{k-1}$ and λ_{k-1} in place $\mathbf{A}$ and λ_k to insure the existence of a unitary matrix $\mathbf{U}_{k-1}$ such that

$$\mathbf{U}_{k-1}^*\mathbf{A}_{k-1}\mathbf{U}_{k-1} = \begin{pmatrix} \mathbf{A}_{k-2} & \mathbf{0} \\ \mathbf{0} & \lambda_{k-1}\mathbf{I} \end{pmatrix},$$

where $\mathbf{A}_{k-2}$ is normal and $\sigma(\mathbf{A}_{k-2}) = \{\lambda_1, \lambda_2, \ldots, \lambda_{k-2}\}$. After k such repetitions, $\mathbf{U}_k \begin{pmatrix} \mathbf{U}_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \cdots \begin{pmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \mathbf{U}$ is a unitary matrix such that

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \begin{pmatrix} \lambda_1\mathbf{I}_{a_1} & 0 & \cdots & 0 \\ 0 & \lambda_2\mathbf{I}_{a_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k\mathbf{I}_{a_k} \end{pmatrix} = \mathbf{D}, \quad a_i = alg\ mult_{\mathbf{A}}(\lambda_i). \tag{7.5.1}$$

Conversely, if there is a unitary matrix $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D}$ is diagonal, then $\mathbf{A}^*\mathbf{A} = \mathbf{U}\mathbf{D}^*\mathbf{D}\mathbf{U}^* = \mathbf{U} = \mathbf{U}\mathbf{D}\mathbf{D}^*\mathbf{U}^* = \mathbf{A}\mathbf{A}^*$, so $\mathbf{A}$ is normal. ∎

Caution! While it's true that normal matrices possess a complete orthonormal set of eigenvectors, not all complete independent sets of eigenvectors of a normal $\mathbf{A}$ are orthonormal (or even orthogonal)—see Exercise 7.5.6. Below are some things that are true.

Properties of Normal Matrices

If $\mathbf{A}$ is a normal matrix with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then

- $\mathbf{A}$ is RPN—i.e., $R(\mathbf{A}) \perp N(\mathbf{A})$ (see p. 408).
- Eigenvectors corresponding to distinct eigenvalues are orthogonal. In other words,

$$N(\mathbf{A} - \lambda_i\mathbf{I}) \perp N(\mathbf{A} - \lambda_j\mathbf{I}) \quad \text{for } \lambda_i \neq \lambda_j. \tag{7.5.2}$$

- The spectral theorems (7.2.7) and (7.3.6) on pp. 517 and 526 hold, but the spectral projectors $\mathbf{G}_i$ on p. 529 specialize to become *orthogonal* projectors because $R(\mathbf{A} - \lambda_i\mathbf{I}) \perp N(\mathbf{A} - \lambda_i\mathbf{I})$ for each λ_i.

Proof of (7.5.2). If $\mathbf{A}$ is normal, so is $\mathbf{A} - \lambda_j\mathbf{I}$, and hence $\mathbf{A} - \lambda_j\mathbf{I}$ is RPN. Consequently, $N(\mathbf{A} - \lambda_j\mathbf{I})^* = N(\mathbf{A} - \lambda_j\mathbf{I})$—recall (5.11.14) from p. 408. If $(\lambda_i, \mathbf{x}_i)$ and $(\lambda_j, \mathbf{x}_j)$ are distinct eigenpairs, then $(\mathbf{A} - \lambda_j\mathbf{I})^*\mathbf{x}_j = \mathbf{0}$, and $0 = \mathbf{x}_j^*(\mathbf{A} - \lambda_j\mathbf{I})\mathbf{x}_i = \mathbf{x}_j^*\mathbf{A}\mathbf{x}_i - \mathbf{x}_j^*\lambda_j\mathbf{x}_i = (\lambda_i - \lambda_j)\mathbf{x}_j^*\mathbf{x}_i$ implies $0 = \mathbf{x}_j^*\mathbf{x}_i$. ∎

Several common types of matrices are normal. For example, real-symmetric and hermitian matrices are normal, real skew-symmetric and skew-hermitian matrices are normal, and orthogonal and unitary matrices are normal. By virtue of being normal, these kinds of matrices inherit all of the above properties, but it's worth looking a bit closer at the real-symmetric and hermitian matrices because they have some special eigenvalue properties.

If $\mathbf{A}$ is real symmetric or hermitian, and if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $\mathbf{x}^*\mathbf{x} \neq 0$, and $\lambda\mathbf{x} = \mathbf{A}\mathbf{x}$ implies $\overline{\lambda}\mathbf{x}^* = \mathbf{x}^*\mathbf{A}^*$, so

$$\mathbf{x}^*\mathbf{x}(\lambda - \overline{\lambda}) = \mathbf{x}^*(\lambda - \overline{\lambda})\mathbf{x} = \mathbf{x}^*\mathbf{A}\mathbf{x} - \mathbf{x}^*\mathbf{A}^*\mathbf{x} = 0 \implies \lambda = \overline{\lambda}.$$

In other words, eigenvalues of real-symmetric and hermitian matrices are real. A similar argument (Exercise 7.5.4) shows that the eigenvalues of a real skew-symmetric or skew-hermitian matrix are pure imaginary numbers.

Eigenvectors for a hermitian $\mathbf{A} \in \mathcal{C}^{n\times n}$ may have to involve complex numbers, but a real-symmetric matrix possesses a complete orthonormal set of *real* eigenvectors. Consequently, the real-symmetric case can be distinguished by observing that $\mathbf{A}$ is real symmetric if and only if $\mathbf{A}$ is *orthogonally* similar to a real-diagonal matrix $\mathbf{D}$. Below is a summary of these observations.

Symmetric and Hermitian Matrices

In addition to the properties inherent to all normal matrices,

- Real-symmetric and hermitian matrices have real eigenvalues. (7.5.3)
- $\mathbf{A}$ is real symmetric if and only if $\mathbf{A}$ is *orthogonally* similar to a real-diagonal matrix $\mathbf{D}$—i.e., $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D}$ for some orthogonal $\mathbf{P}$.
- Real skew-symmetric and skew-hermitian matrices have pure imaginary eigenvalues.

Example 7.5.1

Largest and Smallest Eigenvalues. Since the eigenvalues of a hermitian matrix $\mathbf{A}_{n\times n}$ are real, they can be ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

Problem: Explain why the largest and smallest eigenvalues can be described as

$$\lambda_1 = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x} \quad \text{and} \quad \lambda_n = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x}. \tag{7.5.4}$$

Solution: There is a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$ or, equivalently, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$. Since $\|\mathbf{x}\|_2 = 1 \Longleftrightarrow \|\mathbf{y}\|_2 = 1$ for $\mathbf{y} = \mathbf{U}^*\mathbf{x}$,

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x} = \max_{\|\mathbf{y}\|_2=1} \mathbf{y}^*\mathbf{D}\mathbf{y} = \max_{\|\mathbf{y}\|_2=1} \sum_{i=1}^{n} \lambda_i |y_i|^2 \leq \max_{\|\mathbf{y}\|_2=1} \lambda_1 \sum_{i=1}^{n} |y_i|^2 = \lambda_1$$

with equality being attained when $\mathbf{x}$ is an eigenvector of unit norm associated with λ_1. The expression for the smallest eigenvalue λ_n is obtained by writing

$$\min_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x} = \min_{\|\mathbf{y}\|_2=1} \mathbf{y}^*\mathbf{D}\mathbf{y} = \min_{\|\mathbf{y}\|_2=1} \sum_{i=1}^{n} \lambda_i |y_i|^2 \geq \min_{\|\mathbf{y}\|_2=1} \lambda_n \sum_{i=1}^{n} |y_i|^2 = \lambda_n,$$

where equality is attained at an eigenvector of unit norm associated with λ_n.
Note: The characterizations in (7.5.4) often appear in the equivalent forms

$$\lambda_1 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \quad \text{and} \quad \lambda_n = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}}.$$

Consequently, $\lambda_1 \geq (\mathbf{x}^*\mathbf{A}\mathbf{x}/\mathbf{x}^*\mathbf{x}) \geq \lambda_n$ for all $\mathbf{x} \neq \mathbf{0}$. The term $\mathbf{x}^*\mathbf{A}\mathbf{x}/\mathbf{x}^*\mathbf{x}$ is referred to as a ***Rayleigh quotient*** in honor of the famous English physicist John William Strutt (1842–1919) who became Baron Rayleigh in 1873.

It's only natural to wonder if the intermediate eigenvalues of a hermitian matrix have representations similar to those for the extreme eigenvalues as described in (7.5.4). Ernst Fischer (1875–1954) gave the answer for matrices in 1905, and Richard Courant (1888–1972) provided extensions for infinite-dimensional operators in 1920.

Courant–Fischer Theorem

The eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of a hermitian matrix $\mathbf{A}_{n\times n}$ are

$$\lambda_i = \max_{\dim \mathcal{V}=i} \min_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{A}\mathbf{x} \quad \text{and} \quad \lambda_i = \min_{\dim \mathcal{V}=n-i+1} \max_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{A}\mathbf{x}. \tag{7.5.5}$$

When $i = 1$ in the min-max formula and when $i = n$ in the max-min formula, $\mathcal{V} = \mathcal{C}^n$, so these cases reduce to the equations in (7.5.4). Alternate max-min and min-max formulas are given in Exercise 7.5.12.

Proof. Only the min-max characterization is proven—the max-min proof is analogous (Exercise 7.5.11). As shown in Example 7.5.1, a change of coordinates $\mathbf{y} = \mathbf{U}^*\mathbf{x}$ with a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$ has the effect of replacing $\mathbf{A}$ by $\mathbf{D}$, so we need only establish that

$$\lambda_i = \min_{\dim \mathcal{V}=n-i+1} \max_{\substack{\mathbf{y}\in\mathcal{V} \\ \|\mathbf{y}\|_2=1}} \mathbf{y}^*\mathbf{D}\mathbf{y}.$$

For a subspace $\mathcal{V}$ of dimension $n-i+1$, let $\mathcal{S}_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V},\, \|\mathbf{y}\|_2 = 1\}$, and let

$$\mathcal{S}'_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V} \cap \mathcal{F},\, \|\mathbf{y}\|_2 = 1\}, \quad \text{where} \quad \mathcal{F} = \textit{span}\,\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i\}.$$

Note that $\mathcal{V} \cap \mathcal{F} \neq \mathbf{0}$, for otherwise $\dim(\mathcal{V} + \mathcal{F}) = \dim \mathcal{V} + \dim \mathcal{F} = n + 1$, which is impossible. In other words, $\mathcal{S}'_{\mathcal{V}}$ contains those vectors of $\mathcal{S}_{\mathcal{V}}$ of the form $\mathbf{y} = (y_1, \ldots, y_i, 0, \ldots, 0)^T$ with $\sum_{j=1}^{i} |y_j|^2 = 1$. So for each subspace $\mathcal{V}$ with $\dim \mathcal{V} = n - i + 1$,

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=1}^{i} \lambda_j |y_j|^2 \geq \lambda_i \sum_{j=1}^{i} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}'_{\mathcal{V}}.$$

Since $\mathcal{S}'_{\mathcal{V}} \subseteq \mathcal{S}_{\mathcal{V}}$, it follows that $\max_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \max_{\mathcal{S}'_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \lambda_i$, and hence

$$\min_{\mathcal{V}} \max_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \lambda_i.$$

But this inequality is reversible because if $\tilde{\mathcal{V}} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}\}^{\perp}$, then every $\mathbf{y} \in \tilde{\mathcal{V}}$ has the form $\mathbf{y} = (0, \ldots, 0, y_i, \ldots, y_n)^T$, and hence

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=i}^{n} \lambda_j |y_j|^2 \leq \lambda_i \sum_{j=i}^{n} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}_{\tilde{\mathcal{V}}}.$$

So $\min\limits_{\mathcal{V}} \max\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \max\limits_{\mathcal{S}_{\tilde{\mathcal{V}}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \lambda_i$, and thus $\min\limits_{\mathcal{V}} \max\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} = \lambda_i$. ∎

The value of the Courant–Fischer theorem is its ability to produce inequalities concerning eigenvalues of hermitian matrices without involving the associated eigenvectors. This is illustrated in the following two important examples.

Example 7.5.2

Eigenvalue Perturbations. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of a hermitian $\mathbf{A} \in \mathcal{C}^{n\times n}$, and suppose $\mathbf{A}$ is perturbed by a hermitian $\mathbf{E}$ with eigenvalues $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_n$ to produce $\mathbf{B} = \mathbf{A} + \mathbf{E}$, which is also hermitian.

Problem: If $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ are the eigenvalues of $\mathbf{B}$, explain why

$$\lambda_i + \epsilon_1 \geq \beta_i \geq \lambda_i + \epsilon_n \quad \text{for each } i. \tag{7.5.6}$$

Solution: If $\mathbf{U}$ is a unitary matrix such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, then $\tilde{\mathbf{B}} = \mathbf{U}^*\mathbf{B}\mathbf{U}$ and $\tilde{\mathbf{E}} = \mathbf{U}^*\mathbf{E}\mathbf{U}$ have the same eigenvalues as $\mathbf{B}$ and $\mathbf{E}$, respectively, and $\tilde{\mathbf{B}} = \mathbf{D} + \tilde{\mathbf{E}}$. For $\mathbf{x} \in \mathcal{F} = \mathit{span}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i\}$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathbf{x} = (x_1, \ldots, x_i, 0, \ldots, 0)^T \quad \text{and} \quad \mathbf{x}^*\mathbf{D}\mathbf{x} = \sum_{j=1}^{i} \lambda_j |x_j|^2 \geq \lambda_i \sum_{j=1}^{i} |x_j|^2 = \lambda_i,$$

so applying the max-min part of the Courant–Fischer theorem to $\tilde{\mathbf{B}}$ yields

$$\begin{aligned}\beta_i &= \max_{\dim \mathcal{V} = i} \min_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \geq \min_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \min_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \left(\mathbf{x}^*\mathbf{D}\mathbf{x} + \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x}\right) \\ &\geq \min_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\mathbf{D}\mathbf{x} + \min_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} \geq \lambda_i + \min_{\substack{\mathbf{x}\in\mathcal{C}^n \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} = \lambda_i + \epsilon_n,\end{aligned}$$

where the last equality is the result of the "min" part of (7.5.4). Similarly, for $\mathbf{x} \in \mathcal{T} = \mathit{span}\{\mathbf{e}_i, \ldots, \mathbf{e}_n\}$ with $\|\mathbf{x}\|_2 = 1$, we have $\mathbf{x}^*\mathbf{D}\mathbf{x} \leq \lambda_i$, and

$$\begin{aligned}\beta_i &= \min_{\dim \mathcal{V} = n-i+1} \max_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \leq \max_{\substack{\mathbf{x}\in\mathcal{T} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \max_{\substack{\mathbf{x}\in\mathcal{T} \\ \|\mathbf{x}\|_2 = 1}} \left(\mathbf{x}^*\mathbf{D}\mathbf{x} + \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x}\right) \\ &\leq \max_{\substack{\mathbf{x}\in\mathcal{T} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\mathbf{D}\mathbf{x} + \max_{\substack{\mathbf{x}\in\mathcal{T} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} \leq \lambda_i + \max_{\substack{\mathbf{x}\in\mathcal{C}^n \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} = \lambda_i + \epsilon_1.\end{aligned}$$

Note: Because $\mathbf{E}$ often represents an error, only $\|\mathbf{E}\|$ (or an estimate thereof) is known. But for every matrix norm, $|\epsilon_j| \leq \|\mathbf{E}\|$ for each j (Example 7.1.4, p. 497). Since the ϵ_j's are real, $-\|\mathbf{E}\| \leq \epsilon_j \leq \|\mathbf{E}\|$, so (7.5.6) guarantees that

$$\lambda_i - \|\mathbf{E}\| \leq \beta_i \leq \lambda_i + \|\mathbf{E}\|. \tag{7.5.7}$$

In other words,

- the eigenvalues of a hermitian matrix $\mathbf{A}$ are perfectly conditioned because a hermitian perturbation $\mathbf{E}$ changes no eigenvalue of $\mathbf{A}$ by more than $\|\mathbf{E}\|$.

It's interesting to compare (7.5.7) with the Bauer–Fike bound of Example 7.3.2 (p. 528). When $\mathbf{A}$ is hermitian, (7.3.10) reduces to $\min_{\lambda_i \in \sigma(\mathbf{A})} |\beta - \lambda_i| \leq \|\mathbf{E}\|$ because $\mathbf{P}$ can be made unitary, so, for induced matrix norms, $\kappa(\mathbf{P}) = 1$. The two results differ in that Bauer–Fike does not assume $\mathbf{E}$ and $\mathbf{B}$ are hermitian.

Example 7.5.3

Interlaced Eigenvalues. For a hermitian matrix $\mathbf{A} \in \mathcal{C}^{n\times n}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and for $\mathbf{c} \in \mathcal{C}^{n\times 1}$, let $\mathbf{B}$ be the bordered matrix

$$\mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{c} \\ \mathbf{c}^* & \alpha \end{pmatrix}_{n+1\times n+1} \quad \text{with eigenvalues} \quad \beta_1 \geq \beta_2 \geq \cdots \geq \beta_n \geq \beta_{n+1}.$$

Problem: Explain why the eigenvalues of $\mathbf{A}$ interlace with those of $\mathbf{B}$ in that

$$\beta_1 \geq \lambda_1 \geq \beta_2 \geq \lambda_2 \geq \cdots \geq \beta_n \geq \lambda_n \geq \beta_{n+1}. \tag{7.5.8}$$

Solution: To see that $\beta_i \geq \lambda_i \geq \beta_{i+1}$ for $1 \leq i \leq n$, let $\mathbf{U}$ be a unitary matrix such that $\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Since $\mathbf{V} = \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}$ is also unitary, the eigenvalues of $\mathbf{B}$ agree with those of

$$\tilde{\mathbf{B}} = \mathbf{V}^*\mathbf{B}\mathbf{V} = \begin{pmatrix} \mathbf{D} & \mathbf{y} \\ \mathbf{y}^* & \alpha \end{pmatrix}, \quad \text{where} \quad \mathbf{y} = \mathbf{U}^*\mathbf{c}.$$

For $\mathbf{x} \in \mathcal{F} = span\,\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i\} \subset \mathcal{C}^{n+1\times 1}$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathbf{x} = (x_1, \ldots, x_i, 0, \ldots, 0)^T \quad \text{and} \quad \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \sum_{j=1}^{n} \lambda_j |x_j|^2 \geq \lambda_i \sum_{j=1}^{n} |x_j|^2 = \lambda_i,$$

so applying the max-min part of the Courant–Fisher theorem to $\tilde{\mathbf{B}}$ yields

$$\beta_i = \max_{\dim \mathcal{V} = i} \min_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \geq \min_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \geq \lambda_i.$$

For $\mathbf{x} \in \mathcal{T} = span\{\mathbf{e}_{i-1}, \mathbf{e}_i, \ldots, \mathbf{e}_n\} \subset \mathcal{C}^{n+1\times 1}$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathbf{x} = (0, \ldots, 0, x_{i-1}, \ldots, x_n, 0)^T \text{ and } \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \sum_{j=i-1}^{n} \lambda_j |x_j|^2 \leq \lambda_{i-1} \sum_{j=i}^{n} |x_j|^2 = \lambda_{i-1},$$

so the min-max part of the Courant–Fisher theorem produces

$$\beta_i = \min_{\dim \mathcal{V} = n-i+2} \max_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \leq \max_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \leq \lambda_{i-1}.$$

Note: If $\mathbf{A}$ is any $n \times n$ principal submatrix of $\mathbf{B}$, then (7.5.8) still holds because each principal submatrix can be brought to the upper-left-hand corner by a similarity transformation $\mathbf{P}^T\mathbf{B}\mathbf{P}$, where $\mathbf{P}$ is a permutation matrix. In other words,

- the eigenvalues of an $n+1 \times n+1$ hermitian matrix are interlaced with the eigenvalues of each of its $n \times n$ principal submatrices.

For $\mathbf{A} \in \mathcal{C}^{m\times n}$ (or $\Re^{m\times n}$), the products $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ (or $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$) are hermitian (or real symmetric), so they are diagonalizable by a unitary (or orthogonal) similarity transformation, and their eigenvalues are necessarily real. But in addition to being real, the eigenvalues of these matrices are always nonnegative. For example, if $(\lambda, \mathbf{x})$ is an eigenpair of $\mathbf{A}^*\mathbf{A}$, then $\lambda = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}/\mathbf{x}^*\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2 \geq 0$, and similarly for the other products. In fact, these λ's are the squares of the singular values for $\mathbf{A}$ developed in §5.12 (p. 411) because if

$$\mathbf{A} = \mathbf{U} \begin{pmatrix} \mathbf{D}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n} \mathbf{V}^*$$

is a singular value decomposition, where $\mathbf{D} = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_r)$ contains the nonzero singular values of $\mathbf{A}$, then

$$\mathbf{V}^*\mathbf{A}^*\mathbf{A}\mathbf{V} = \begin{pmatrix} \mathbf{D}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{7.5.9}$$

and this means that $(\sigma_i^2, \mathbf{v}_i)$ for $i = 1, 2, \ldots, r$ is an eigenpair for $\mathbf{A}^*\mathbf{A}$. In other words, *the nonzero singular values of* $\mathbf{A}$ *are precisely the positive square roots of the nonzero eigenvalues of* $\mathbf{A}^*\mathbf{A}$, *and right-hand singular vectors* $\mathbf{v}_i$ *of* $\mathbf{A}$ *are particular eigenvectors of* $\mathbf{A}^*\mathbf{A}$. Note that this establishes the uniqueness of the σ_i's (but not the $\mathbf{v}_i$'s), and pay attention to the fact that the number of zero singular values of $\mathbf{A}$ need not agree with the number of zero eigenvalues of $\mathbf{A}^*\mathbf{A}$—e.g., $\mathbf{A}_{1\times 2} = (1, 1)$ has no zero singular values, but $\mathbf{A}^*\mathbf{A}$ has one zero eigenvalue. The same game can be played with $\mathbf{A}\mathbf{A}^*$ in place of $\mathbf{A}^*\mathbf{A}$ to argue that the nonzero singular values of $\mathbf{A}$ are the positive square roots of

the nonzero eigenvalues of $\mathbf{AA}^*$, and left-hand singular vectors $\mathbf{u}_i$ of $\mathbf{A}$ are particular eigenvectors of $\mathbf{AA}^*$.

Caution! The statement that right-hand singular vectors $\mathbf{v}_i$ of $\mathbf{A}$ are eigenvectors of $\mathbf{A}^*\mathbf{A}$ and left-hand singular vectors $\mathbf{u}_i$ of $\mathbf{A}$ are eigenvectors of $\mathbf{AA}^*$ is a one-way street—it doesn't mean that just any orthonormal sets of eigenvectors for $\mathbf{A}^*\mathbf{A}$ and $\mathbf{AA}^*$ can be used as respective right-hand and left-hand singular vectors for $\mathbf{A}$. The columns $\mathbf{v}_i$ of any unitary matrix $\mathbf{V}$ that diagonalizes $\mathbf{A}^*\mathbf{A}$ as in (7.5.9) can serve as right-hand singular vectors for $\mathbf{A}$, but corresponding left-hand singular vectors $\mathbf{u}_i$ are constrained by the relationships

$$\begin{aligned} &\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad i = 1, 2, \ldots, r \quad \Longrightarrow \quad \mathbf{u}_i = \frac{\mathbf{A}\mathbf{v}_i}{\sigma_i} = \frac{\mathbf{A}\mathbf{v}_i}{\|\mathbf{A}\mathbf{v}_i\|_2}, \quad i = 1, 2, \ldots, r, \\ &\mathbf{u}_i^*\mathbf{A} = \mathbf{0}, \quad i = r+1, \ldots, m \quad \Longrightarrow \quad span\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\} = N(\mathbf{A}^*). \end{aligned}$$

In other words, the first r left-hand singular vectors for $\mathbf{A}$ are uniquely determined by the first r right-hand singular vectors, while the last $m-r$ left-hand singular vectors can be any orthonormal basis for $N(\mathbf{A}^*)$. If $\mathbf{U}$ is constructed from $\mathbf{V}$ as described above, then $\mathbf{U}$ is guaranteed to be unitary because for

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 \cdots \mathbf{u}_r | \mathbf{u}_{r+1} \cdots \mathbf{u}_m \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 | \mathbf{U}_2 \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} \mathbf{v}_1 \cdots \mathbf{v}_r | \mathbf{v}_{r+1} \cdots \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{V}_1 | \mathbf{V}_2 \end{bmatrix},$$

$\mathbf{U}_1$ and $\mathbf{U}_2$ each contain orthonormal columns, and, by using (7.5.9),

$$\begin{aligned} R(\mathbf{U}_1) &= R\left(\mathbf{A}\mathbf{V}_1\mathbf{D}^{-1}\right) = R(\mathbf{A}\mathbf{V}_1) = R(\mathbf{A}\mathbf{V}_1\mathbf{D}) = R\left([\mathbf{A}\mathbf{V}_1\mathbf{D}][\mathbf{A}\mathbf{V}_1\mathbf{D}]^*\right) \\ &= R(\mathbf{A}\mathbf{A}^*\mathbf{A}\mathbf{A}^*) = R(\mathbf{A}\mathbf{A}^*) = R(\mathbf{A}) = N(\mathbf{A}^*)^{\perp} = R(\mathbf{U}_2)^{\perp}. \end{aligned}$$

The matrix $\mathbf{V}$ is unitary to start with, but, in addition,

$$\begin{aligned} &R(\mathbf{V}_1) = R(\mathbf{V}_1\mathbf{D}) = R([\mathbf{V}_1\mathbf{D}][\mathbf{V}_1\mathbf{D}]^*) = R(\mathbf{A}^*\mathbf{A}) = R(\mathbf{A}^*) \text{ and} \\ &R(\mathbf{V}_2) = R(\mathbf{A}^*)^{\perp} = N(\mathbf{A}). \end{aligned}$$

These observations are consistent with those established on p. 407 for any URV factorization. Otherwise something would be terribly wrong because an SVD is just a special kind of a URV factorization. Finally, notice that there is nothing special about starting with $\mathbf{V}$ to build a $\mathbf{U}$—we can also take the columns of any unitary $\mathbf{U}$ that diagonalizes $\mathbf{AA}^*$ as left-hand singular vectors for $\mathbf{A}$ and build corresponding right-hand singular vectors in a manner similar to that described above. Below is a summary of the preceding developments concerning singular values together with an additional observation connecting singular values with eigenvalues.

Singular Values and Eigenvalues

For $\mathbf{A} \in \mathcal{C}^{m\times n}$ with $rank\,(\mathbf{A}) = r$, the following statements are valid.

- The nonzero eigenvalues of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ are equal and positive.
- The nonzero singular values of $\mathbf{A}$ are the positive square roots of the nonzero eigenvalues of $\mathbf{A}^*\mathbf{A}$ (and $\mathbf{A}\mathbf{A}^*$).
- If $\mathbf{A}$ is normal with nonzero eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$, then the nonzero singular values of $\mathbf{A}$ are $\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_r|\}$.
- Right-hand and left-hand singular vectors for $\mathbf{A}$ are special eigenvectors for $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$, respectively.
- Any complete orthonormal set of eigenvectors $\mathbf{v}_i$ for $\mathbf{A}^*\mathbf{A}$ can serve as a complete set of right-hand singular vectors for $\mathbf{A}$, and a corresponding complete set of left-hand singular vectors is given by $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\left\|\mathbf{A}\mathbf{v}_i\right\|_2$, $i = 1, 2, \ldots, r$, together with any orthonormal basis $\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$ for $N(\mathbf{A}^*)$. Similarly, any complete orthonormal set of eigenvectors for $\mathbf{A}\mathbf{A}^*$ can be used as left-hand singular vectors for $\mathbf{A}$, and corresponding right-hand singular vectors can be built in an analogous way.
- The hermitian matrix $\mathbf{B} = \begin{pmatrix} \mathbf{0}_{m\times m} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0}_{n\times n} \end{pmatrix}$ of order $m+n$ has nonzero eigenvalues $\{\pm\sigma_1, \pm\sigma_2, \ldots, \pm\sigma_r\}$ in which $\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$ are the nonzero singular values of $\mathbf{A}$.

Proof. Only the last point requires proof, and this follows by observing that if λ is an eigenvalue of $\mathbf{B}$, then

$$\begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \lambda\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \implies \left\{\begin{array}{r} \mathbf{A}\mathbf{x}_2 = \lambda\mathbf{x}_1 \\ \mathbf{A}^*\mathbf{x}_1 = \lambda\mathbf{x}_2 \end{array}\right\} \implies \mathbf{A}^*\mathbf{A}\mathbf{x}_2 = \lambda^2\mathbf{x}_2,$$

so each eigenvalue of $\mathbf{B}$ is the square of a singular value of $\mathbf{A}$. But $\mathbf{B}$ is hermitian with $rank\,(\mathbf{B}) = 2r$, so there are exactly $2r$ nonzero eigenvalues of $\mathbf{B}$. Therefore, each pair $\pm\sigma_i$, $i = 1, 2, \ldots, r$, must be an eigenvalue for $\mathbf{B}$. ∎

Example 7.5.4

Min-Max Singular Values. Since the singular values of $\mathbf{A}$ are the positive square roots of the eigenvalues of $\mathbf{A}^*\mathbf{A}$, and since $\left\|\mathbf{A}\mathbf{x}\right\|_2 = (\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x})^{1/2}$, it's a corollary of the Courant–Fischer theorem (p. 550) that if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ are the singular values for $\mathbf{A}_{m\times n}$ $(n \leq m)$, then

$$\sigma_i = \max_{\dim \mathcal{V}=i} \min_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2=1}} \left\|\mathbf{A}\mathbf{x}\right\|_2 \quad \text{and} \quad \sigma_i = \min_{\dim \mathcal{V}=n-i+1} \max_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2=1}} \left\|\mathbf{A}\mathbf{x}\right\|_2.$$

These expressions provide intermediate values between the extremes

$$\sigma_1 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \quad \text{and} \quad \sigma_n = \min_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \qquad \text{(see p. 414)}.$$

Exercises for section 7.5

7.5.1. Is $\mathbf{A} = \begin{pmatrix} 5+\mathrm{i} & -2\,\mathrm{i} \\ 2 & 4+2\,\mathrm{i} \end{pmatrix}$ a normal matrix?

7.5.2. Give examples of two distinct classes of normal matrices that are real but not symmetric.

7.5.3. Show that $\mathbf{A} \in \Re^{n\times n}$ is normal and has real eigenvalues if and only if $\mathbf{A}$ is symmetric.

7.5.4. Prove that the eigenvalues of a real skew-symmetric or skew-hermitian matrix must be pure imaginary numbers (i.e., multiples of i).

7.5.5. When trying to decide what's true about matrices and what's not, it helps to think in terms of the following associations.

Hermitian matrices	$\longleftrightarrow$	Real numbers ($z = \overline{z}$).
Skew-hermitian matrices	$\longleftrightarrow$	Imaginary numbers ($z = -\overline{z}$).
Unitary matrices	$\longleftrightarrow$	Points on the unit circle ($z = \mathrm{e}^{\mathrm{i}\theta}$).

For example, the complex function $f(z) = (1-z)(1+z)^{-1}$ maps the imaginary axis in the complex plane to points on the unit circle because $|f(z)|^2 = 1$ whenever $\overline{z} = -z$. It's therefore reasonable to conjecture (as Cayley did in 1846) that if $\mathbf{A}$ is skew hermitian (or real skew symmetric), then

$$f(\mathbf{A}) = (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A}) \tag{7.5.10}$$

is unitary (or orthogonal). Prove this is indeed correct. **Note:** Expression (7.5.10) has come to be known as the ***Cayley transformation***.

7.5.6. Show by example that a normal matrix can have a complete independent set of eigenvectors that are not orthonormal, and then explain how every complete independent set of eigenvectors for a normal matrix can be transformed into a complete orthonormal set of eigenvectors.

7.5.7. Construct an example to show that the converse of (7.5.2) is false. In other words, show that it is possible for $N(\mathbf{A}-\lambda_i\mathbf{I}) \perp N(\mathbf{A}-\lambda_j\mathbf{I})$ whenever $i \neq j$ without $\mathbf{A}$ being normal.

7.5.8. Explain why a triangular matrix is normal if and only if it is diagonal.

7.5.9. Use the result of Exercise 7.5.8 to give an alternate proof of the unitary diagonalization theorem given on p. 547.

7.5.10. For a normal matrix $\mathbf{A}$, explain why $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$ if and only if $(\overline{\lambda}, \mathbf{x})$ is an eigenpair for $\mathbf{A}^*$.

7.5.11. To see if you understand the proof of the min-max part of the Courant–Fischer theorem (p. 550), construct an analogous proof for the max-min part of (7.5.5).

7.5.12. The Courant–Fischer theorem has the following alternate formulation.

$$\lambda_i = \max_{\mathbf{v}_1,\ldots,\mathbf{v}_{n-i}\in\mathcal{C}^n} \min_{\substack{\mathbf{x}\perp\mathbf{v}_1,\ldots,\mathbf{v}_{n-i}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{A}\mathbf{x} \quad \text{and} \quad \lambda_i = \min_{\mathbf{v}_1,\ldots,\mathbf{v}_{i-1}\in\mathcal{C}^n} \max_{\substack{\mathbf{x}\perp\mathbf{v}_1,\ldots,\mathbf{v}_{i-1}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{A}\mathbf{x}$$

for $1 < i < n$. To see if you *really* understand the proof of the min-max part of (7.5.5), adapt it to prove the alternate min-max formulation given above.

7.5.13. (a) Explain why every unitary matrix is unitarily similar to a diagonal matrix of the form

$$\mathbf{D} = \begin{pmatrix} e^{i\theta_1} & 0 & \cdots & 0 \\ 0 & e^{i\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{i\theta_n} \end{pmatrix}.$$

(b) Prove that every orthogonal matrix is orthogonally similar to a real block-diagonal matrix of the form

$$\mathbf{B} = \begin{pmatrix} \pm 1 & & & & & \\ & \ddots & & & & \\ & & \pm 1 & & & \\ & & & \cos\theta_1 & \sin\theta_1 & \\ & & & -\sin\theta_1 & \cos\theta_1 & \\ & & & & & \ddots \\ & & & & & \begin{matrix}\cos\theta_t & \sin\theta_t \\ -\sin\theta_t & \cos\theta_t\end{matrix} \end{pmatrix}.$$

7.6 POSITIVE DEFINITE MATRICES

Since the symmetric structure of a matrix forces its eigenvalues to be real, what additional property will force all eigenvalues to be *positive* (or perhaps just nonnegative)? To answer this, let's deal with real-symmetric matrices—the hermitian case follows along the same lines. If $\mathbf{A} \in \Re^{n\times n}$ is symmetric, then, as observed above, there is an orthogonal matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T$, where $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is real. If $\lambda_i \geq 0$ for each i, then $\mathbf{D}^{1/2}$ exists, so

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T = \mathbf{P}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{P}^T = \mathbf{B}^T\mathbf{B} \quad \text{for} \quad \mathbf{B} = \mathbf{D}^{1/2}\mathbf{P}^T,$$

and $\lambda_i > 0$ for each i if and only if $\mathbf{B}$ is nonsingular. Conversely, if $\mathbf{A}$ can be factored as $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, then all eigenvalues of $\mathbf{A}$ are nonnegative because for any eigenpair $(\lambda, \mathbf{x})$,

$$\lambda = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\|\mathbf{B}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \geq 0.$$

Moreover, if $\mathbf{B}$ is nonsingular, then $N(\mathbf{B}) = \mathbf{0} \Longrightarrow \mathbf{B}\mathbf{x} \neq \mathbf{0}$, so $\lambda > 0$. In other words, *a real-symmetric matrix* $\mathbf{A}$ *has nonnegative eigenvalues if and only if* $\mathbf{A}$ *can be factored as* $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, *and all eigenvalues are positive if and only if* $\mathbf{B}$ *is nonsingular*. A symmetric matrix $\mathbf{A}$ whose eigenvalues are positive is called ***positive definite***, and when the eigenvalues are just nonnegative, $\mathbf{A}$ is said to be ***positive semidefinite***.

The use of this terminology is consistent with that introduced in Example 3.10.7 (p. 154), where the term "positive definite" was used to designate symmetric matrices possessing an LU factorization with positive pivots. It was demonstrated in Example 3.10.7 that possessing positive pivots is equivalent to the existence of a *Cholesky factorization* $\mathbf{A} = \mathbf{R}^T\mathbf{R}$, where $\mathbf{R}$ is upper triangular with positive diagonal entries. By the result of the previous paragraph this means that *all eigenvalues of a symmetric matrix* $\mathbf{A}$ *are positive if and only if* $\mathbf{A}$ *has an LU factorization with positive pivots*.

But the pivots are intimately related to the leading principal minor determinants. Recall from Exercise 6.1.16 (p. 474) that if $\mathbf{A}_k$ is the k^{th} leading principal submatrix of $\mathbf{A}_{n\times n}$, then the k^{th} pivot is given by

$$u_{kk} = \begin{cases} \det(\mathbf{A}_1) = a_{11} & \text{for } k = 1, \\ \det(\mathbf{A}_k)/\det(\mathbf{A}_{k-1}) & \text{for } k = 2, 3, \ldots, n. \end{cases}$$

Consequently, *a symmetric matrix is positive definite if and only if each of its leading principal minors is positive*. However, if each leading principal minor is positive, then *all* principal minors must be positive because if $\mathbf{P}_k$ is any principal submatrix of $\mathbf{A}$, then there is a permutation matrix $\mathbf{Q}$ such that

$\mathbf{P}_k$ is a leading principal submatrix in $\mathbf{C} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{P}_k & \star \\ \star & \star \end{pmatrix}$, and, since $\sigma(\mathbf{A}) = \sigma(\mathbf{C})$, we have, with some obvious shorthand notation,

$$\mathbf{A}\text{'s leading pm's} > 0 \Rightarrow \mathbf{A} \text{ pd} \Rightarrow \mathbf{C} \text{ pd} \Rightarrow \det(\mathbf{P}_k) > 0 \Rightarrow \text{all of } \mathbf{A}\text{'s pm's} > 0.$$

Finally, observe that $\mathbf{A}$ *is positive definite if and only if* $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ *for every nonzero* $\mathbf{x} \in \Re^{n\times 1}$. If $\mathbf{A}$ is positive definite, then $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ for a nonsingular $\mathbf{B}$, so $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} = \|\mathbf{B}\mathbf{x}\|_2^2 \geq 0$ with equality if and only if $\mathbf{B}\mathbf{x} = \mathbf{0}$ or, equivalently, $\mathbf{x} = \mathbf{0}$. Conversely, if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, then for every eigenpair $(\lambda, \mathbf{x})$ we have $\lambda = (\mathbf{x}^T\mathbf{A}\mathbf{x}/\mathbf{x}^T\mathbf{x}) > 0$.

Below is a formal summary of the results for positive definite matrices.

Positive Definite Matrices

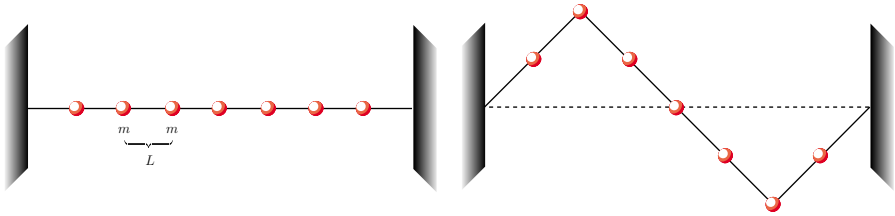
For real-symmetric matrices $\mathbf{A}$, the following statements are equivalent, and any one can serve as the definition of a ***positive definite*** matrix.

- $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for every nonzero $\mathbf{x} \in \Re^{n\times 1}$ (most commonly used as the definition).
- All eigenvalues of $\mathbf{A}$ are positive.
- $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ for some nonsingular $\mathbf{B}$.
  - ▷ While $\mathbf{B}$ is not unique, there is one and only one *upper-triangular* matrix $\mathbf{R}$ with positive diagonals such that $\mathbf{A} = \mathbf{R}^T\mathbf{R}$. This is the ***Cholesky factorization*** of $\mathbf{A}$ (Example 3.10.7, p. 154).
- $\mathbf{A}$ has an LU (or LDU) factorization with all pivots being positive.
  - ▷ The LDU factorization is of the form $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{R}^T\mathbf{R}$, where $\mathbf{R} = \mathbf{D}^{1/2}\mathbf{L}^T$ is the ***Cholesky factor*** of $\mathbf{A}$ (also see p. 154).
- The leading principal minors of $\mathbf{A}$ are positive.
- All principal minors of $\mathbf{A}$ are positive.

For hermitian matrices, replace $(\star)^T$ by $(\star)^*$ and $\Re$ by $\mathcal{C}$.

Example 7.6.1

Vibrating Beads on a String. Consider n small beads, each having mass m, spaced at equal intervals of length L on a very tightly stretched string or wire under a tension T as depicted in Figure 7.6.1. Each bead is initially displaced from its equilibrium position by a small vertical distance—say bead k is displaced by an amount c_k at $t = 0$. The beads are then released so that they can vibrate freely.



Equilibrium Position A Typical Initial Position

FIGURE 7.6.1

Problem: For small vibrations, determine the position of each bead at time $t > 0$ for any given initial configuration.

Solution: The small vibration hypothesis validates the following assumptions.

- The tension T remains constant for all time.
- There is only vertical motion (the horizontal forces cancel each other).
- Only small angles are involved, so the approximation $\sin\theta \approx \tan\theta$ is valid.

Let $y_k(t) = y_k$ be the vertical distance of the k^{th} bead from equilibrium at time t, and set $y_0 = 0 = y_{n+1}$.

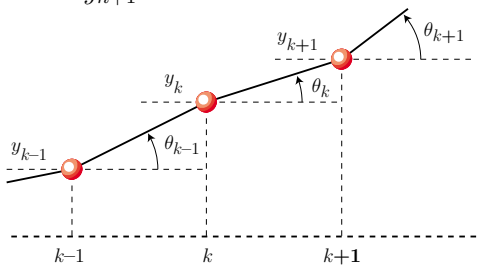


FIGURE 7.6.2

If θ_k is the angle depicted in Figure 7.6.2, the diagram above, then the upward force on the k^{th} bead at time t is $F_u = T\sin\theta_k$, while the downward force is $F_d = T\sin\theta_{k-1}$, so the total force on the k^{th} bead at time t is

$$F = F_u - F_d = T(\sin\theta_k - \sin\theta_{k-1}) \approx T(\tan\theta_k - \tan\theta_{k-1})$$
$$= T\left(\frac{y_{k+1} - y_k}{L} - \frac{y_k - y_{k-1}}{L}\right) = \frac{T}{L}(y_{k-1} - 2y_k + y_{k+1}).$$

Newton's second law says force = mass × acceleration, so we set

$$my_k'' = \frac{T}{L}(y_{k-1} - 2y_k + y_{k+1}) \implies y_k'' + \frac{T}{mL}(-y_{k-1} + 2y_k - y_{k+1}) = 0 \quad (7.6.1)$$

together with $y_k(0) = c_k$ and $y_k'(0) = 0$ to model the motion of the k^{th} bead. Altogether, equations (7.6.1) represent a system of n second-order linear differential equations, and each is coupled to its neighbors so that no single

equation can be solved in isolation. To extract solutions, the equations must somehow be uncoupled, and here's where matrix diagonalization works its magic. Write equations (7.6.1) in matrix form as

$$\begin{pmatrix} y_1'' \\ y_2'' \\ y_3'' \\ \vdots \\ y_n'' \end{pmatrix} + \frac{T}{mL} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{or} \quad \mathbf{y}'' + \mathbf{A}\mathbf{y} = \mathbf{0}, \tag{7.6.2}$$

with $\mathbf{y}(0) = \mathbf{c} = (c_1 c_2 \cdots c_n)^T$ and $\mathbf{y}'(0) = \mathbf{0}$. Since $\mathbf{A}$ is symmetric, there is an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$, where the λ_i's are the eigenvalues of $\mathbf{A}$. By making the substitution $\mathbf{y} = \mathbf{P}\mathbf{z}$ (or, equivalently, by changing the coordinate system), (7.6.2) is transformed into

$$\begin{array}{l} \mathbf{z}'' + \mathbf{D}\mathbf{z} = \mathbf{0}, \\ \mathbf{z}(0) = \mathbf{P}^T\mathbf{c} = \tilde{\mathbf{c}}, \\ \mathbf{z}'(0) = \mathbf{0}, \end{array} \quad \text{or} \quad \begin{pmatrix} z_1'' \\ z_2'' \\ \vdots \\ z_n'' \end{pmatrix} + \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

In other words, by changing to a coordinate system defined by a complete set of orthonormal eigenvectors for $\mathbf{A}$, the original system (7.6.2) is completely uncoupled so that each equation $z_k'' + \lambda_k z_k = 0$ with $z_k(0) = \tilde{c}_k$ and $z_k'(0) = 0$ can be solved independently. This helps reveal why diagonalizability is a fundamentally important concept. Recall from elementary differential equations that

$$z_k'' + \lambda_k z_k = 0 \implies z_k(t) = \begin{cases} \alpha_k \mathrm{e}^{t\sqrt{-\lambda_k}} + \beta_k \mathrm{e}^{-t\sqrt{-\lambda_k}} & \text{when } \lambda_k < 0, \\ \alpha_k \cos\left(t\sqrt{\lambda_k}\right) + \beta_k \sin\left(t\sqrt{\lambda_k}\right) & \text{when } \lambda_k \geq 0. \end{cases}$$

Vibrating beads suggest sinusoidal solutions, so we expect each $\lambda_k > 0$. In other words, the mathematical model would be grossly inconsistent with reality if the symmetric matrix $\mathbf{A}$ in (7.6.2) were not positive definite. It turns out that $\mathbf{A}$ *is* positive definite because there is a Cholesky factorization $\mathbf{A} = \mathbf{R}^T\mathbf{R}$ with

$$\mathbf{R} = \sqrt{\frac{T}{mL}} \begin{pmatrix} r_1 & -1/r_1 & & & \\ & r_2 & -1/r_2 & & \\ & & \ddots & \ddots & \\ & & r_{n-1} & -1/r_{n-1} & \\ & & & r_n & \end{pmatrix} \quad \text{with} \quad r_k = \sqrt{2 - \frac{k-1}{k}},$$

and thus we are insured that each $\lambda_k > 0$. In fact, since $\mathbf{A}$ is a tridiagonal Toeplitz matrix, the results of Example 7.2.5 (p. 514) can be used to show that

$$\lambda_k = \frac{2T}{mL}\left(1 - \cos\frac{k\pi}{n+1}\right) = \frac{4T}{mL}\sin^2\frac{k\pi}{2(n+1)} \qquad \text{(see Exercise 7.2.18).}$$

Therefore,

$$\left.\begin{array}{rcl} z_k & = & \alpha_k \cos\left(t\sqrt{\lambda_k}\right) + \beta_k \sin\left(t\sqrt{\lambda_k}\right) \\ z_k(0) & = & \tilde{c}_k \\ z_k'(0) & = & 0 \end{array}\right\} \implies z_k = \tilde{c}_k \cos\left(t\sqrt{\lambda_k}\right), \quad (7.6.3)$$

and for $\mathbf{P} = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n\right]$,

$$\mathbf{y} = \mathbf{P}\mathbf{z} = z_1\mathbf{x}_1 + z_2\mathbf{x}_2 + \cdots + z_n\mathbf{x}_n = \sum_{j=1}^{n} \left(\tilde{c}_j \cos\left(t\sqrt{\lambda_k}\right)\right)\mathbf{x}_j. \quad (7.6.4)$$

This means that every possible mode of vibration is a combination of modes determined by the eigenvectors $\mathbf{x}_j$. To understand this more clearly, suppose that the beads are initially positioned according to the components of $\mathbf{x}_j$ —i.e., $\mathbf{c} = \mathbf{y}(0) = \mathbf{x}_j$. Then $\tilde{\mathbf{c}} = \mathbf{P}^T\mathbf{c} = \mathbf{P}^T\mathbf{x}_j = \mathbf{e}_j$, so (7.6.3) and (7.6.4) reduce to

$$z_k = \begin{cases} \cos\left(t\sqrt{\lambda_k}\right) & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases} \implies \mathbf{y} = \left(\cos\left(t\sqrt{\lambda_k}\right)\right)\mathbf{x}_j. \quad (7.6.5)$$

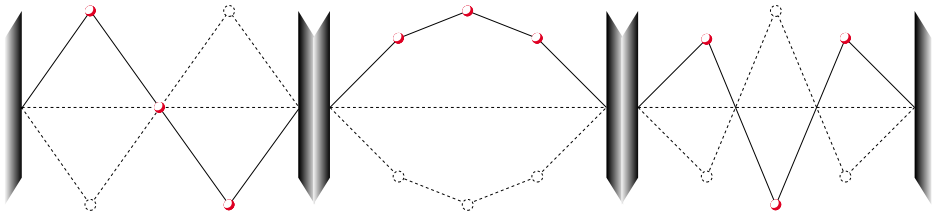
In other words, when $\mathbf{y}(0) = \mathbf{x}_j$, the j^{th} eigenpair $(\lambda_j, \mathbf{x}_j)$ completely determines the mode of vibration because the amplitudes are determined by $\mathbf{x}_j$, and each bead vibrates with a common frequency $f = \sqrt{\lambda_j}/2\pi$. This type of motion (7.6.5) is called a ***fundamental mode of vibration***. In these terms, equation (7.6.4) translates to say that *every possible mode of vibration is a combination of the fundamental modes*. For example, when $n = 3$, the matrix in (7.6.2) is

$$\mathbf{A} = \frac{T}{mL}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad \text{with} \quad \left\{\begin{array}{l} \lambda_1 = (T/mL)(2) \\ \lambda_2 = (T/mL)(2+\sqrt{2}) \\ \lambda_3 = (T/mL)(2-\sqrt{2}) \end{array}\right\},$$

and a complete orthonormal set of eigenvectors is

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

The three corresponding fundamental modes are shown in Figure 7.6.3.



Mode for $(\lambda_1, \mathbf{x}_1)$ Mode for $(\lambda_2, \mathbf{x}_2)$ Mode for $(\lambda_3, \mathbf{x}_3)$

FIGURE 7.6.3

Example 7.6.2

Discrete Laplacian. According to the laws of physics, the temperature at time t at a point (x, y, z) in a solid body is a function $u(x, y, z, t)$ satisfying the *diffusion equation*

$$\frac{\partial u}{\partial t} = K\nabla^2 u, \quad \text{where} \quad \nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

is the ***Laplacian*** of u and K is a constant of thermal diffusivity. At steady state the temperature at each point does not vary with time, so $\partial u/\partial t = 0$ and $u = u(x, y, z)$ satisfy *Laplace's equation* $\nabla^2 u = 0$. Solutions of this equation are often called *harmonic functions*. The nonhomogeneous equation $\nabla^2 u = f$ (*Poisson's equation*) is addressed in Exercise 7.6.9. To keep things simple, let's confine our attention to the following two-dimensional problem.

Problem: For a square plate as shown in Figure 7.6.4(a), explain how to numerically determine the steady-state temperature at interior grid points when the temperature around the boundary is prescribed to be $u(x, y) = g(x, y)$ for a given function g. In other words, explain how to extract a numerical solution to $\nabla^2 u = 0$ in the interior of the square when $u(x, y) = g(x, y)$ on the square's boundary. This is called a *Dirichlet problem*.[76]

Solution: Discretize the problem by overlaying the plate with a square mesh containing n^2 interior points at equally spaced intervals of length h. As illustrated in Figure 7.6.4(b) for $n = 4$, label the grid points using a rowwise ordering scheme—i.e., label them as you would label matrix entries.

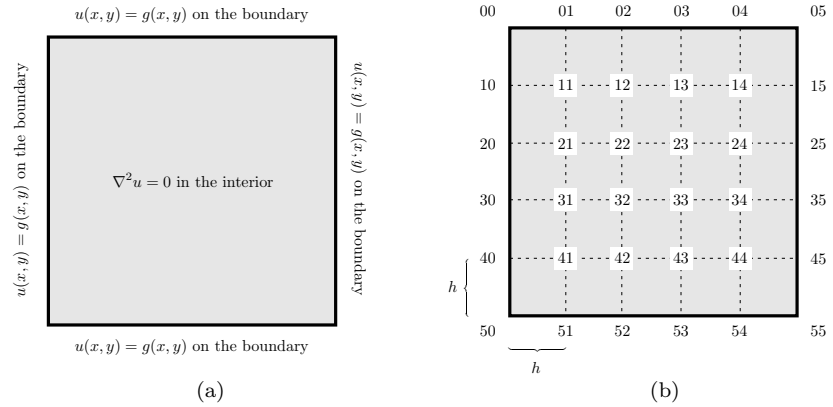


FIGURE 7.6.4

[76] Johann Peter Gustav Lejeune Dirichlet (1805–1859) held the chair at Göttingen previously occupied by Gauss. Because of his work on the convergence of trigonometric series, Dirichlet is generally considered to be the founder of the theory of Fourier series, but much of the groundwork was laid by S. D. Poisson (p. 572) who was Dirichlet's Ph.D. advisor.

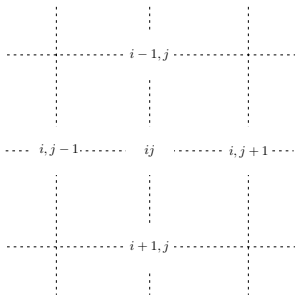
Approximate $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial y^2$ at the interior grid points (x_i, y_j) by using the second-order centered difference formula (1.4.3) developed on p. 19 to write

$$\begin{aligned}\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_i,y_j)} &= \frac{u(x_i-h,y_j)-2u(x_i,y_j)+u(x_i+h,y_j)}{h^2}+O(h^2),\\ \left.\frac{\partial^2 u}{\partial y^2}\right|_{(x_i,y_j)} &= \frac{u(x_i,y_j-h)-2u(x_i,y_j)+u(x_i,y_j+h)}{h^2}+O(h^2).\end{aligned} \tag{7.6.6}$$

Adopt the notation $u_{ij} = u(x_i, y_j)$, and add the expressions in (7.6.6) using $\nabla^2 u|_{(x_i,y_j)} = 0$ for interior points (x_i, y_j) to produce

$$4u_{ij} = (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) + O(h^4) \quad \text{for} \quad i,j = 1,2,\ldots,n.$$

In other words, the steady-state temperature at an interior grid point is approximately the average of the steady-state temperatures at the four neighboring grid points as illustrated in Figure 7.6.5.



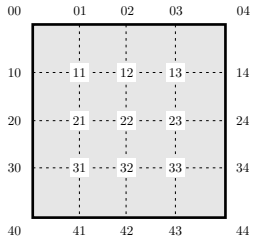
$$u_{ij} = \frac{u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}}{4} + O(h^4)$$

FIGURE 7.6.5

If the $O(h^4)$ terms are neglected, the resulting five-point difference equations,

$$4u_{ij} - (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) = 0 \quad \text{for} \quad i,j = 1,2,\ldots,n,$$

constitute an $n^2 \times n^2$ linear system $\mathbf{Lu} = \mathbf{g}$ in which the unknowns are the u_{ij}'s, and the right-hand side contains boundary values. For example, a mesh with nine interior points produces the 9×9 system in Figure 7.6.6.



$$\left(\begin{array}{ccc|ccc|ccc}
4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0\\
-1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0\\
0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0\\
\hline
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0\\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0\\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1\\
\hline
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0\\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1\\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4
\end{array}\right)
\left(\begin{array}{c} u_{11}\\ u_{12}\\ u_{13}\\ \hline u_{21}\\ u_{22}\\ u_{23}\\ \hline u_{31}\\ u_{32}\\ u_{33}\end{array}\right)
=
\left(\begin{array}{c} g_{01}+g_{10}\\ g_{02}\\ g_{03}+g_{14}\\ \hline g_{20}\\ 0\\ g_{24}\\ \hline g_{30}+g_{41}\\ g_{42}\\ g_{43}+g_{34}\end{array}\right)$$

FIGURE 7.6.6

The coefficient matrix of this system is the ***discrete Laplacian,*** and in general it has the symmetric block-tridiagonal form

$$\mathbf{L}=\begin{pmatrix} \mathbf{T} & -\mathbf{I} & & & \\ -\mathbf{I} & \mathbf{T} & -\mathbf{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I} & \mathbf{T} & -\mathbf{I} \\ & & & -\mathbf{I} & \mathbf{T} \end{pmatrix}_{n^2\times n^2} \text{ with } \mathbf{T}=\begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}_{n\times n}.$$

In addition, $\mathbf{L}$ is positive definite. In fact, the discrete Laplacian is a primary example of how positive definite matrices arise in practice. Note that $\mathbf{L}$ is the two-dimensional version of the one-dimensional finite-difference matrix in Example 1.4.1 (p. 19).

Problem: Show $\mathbf{L}$ is positive definite by explicitly exhibiting its eigenvalues.

Solution: Example 7.2.5 (p. 514) insures that the n eigenvalues of $\mathbf{T}$ are

$$\lambda_i = 4 - 2\cos\left(\frac{i\pi}{n+1}\right), \quad i = 1, 2, \ldots, n. \tag{7.6.7}$$

If $\mathbf{U}$ is an orthogonal matrix such that $\mathbf{U}^T\mathbf{T}\mathbf{U} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and if $\mathbf{B}$ is the $n^2 \times n^2$ block-diagonal orthogonal matrix

$$\mathbf{B} = \begin{pmatrix} \mathbf{U} & 0 & \cdots & 0 \\ 0 & \mathbf{U} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{U} \end{pmatrix}, \quad \text{then} \quad \mathbf{B}^T\mathbf{L}\mathbf{B} = \tilde{\mathbf{L}} = \begin{pmatrix} \mathbf{D} & -\mathbf{I} & & & \\ -\mathbf{I} & \mathbf{D} & -\mathbf{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I} & \mathbf{D} & -\mathbf{I} \\ & & & -\mathbf{I} & \mathbf{D} \end{pmatrix}.$$

Consider the permutation obtained by placing the numbers $1, 2, \ldots, n^2$ rowwise in a square matrix, and then reordering them by listing the entries columnwise. For example, when $n = 3$ this permutation is generated as follows:

$$\mathbf{v} = (1,2,3,4,5,6,7,8,9) \to \mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \to (1,4,7,2,5,8,3,6,9) = \tilde{\mathbf{v}}.$$

Equivalently, this can be described in terms of wrapping and unwrapping rows by writing $\mathbf{v} \xrightarrow{wrap} \mathbf{A} \longrightarrow \mathbf{A}^T \xrightarrow{unwrap} \tilde{\mathbf{v}}$. If $\mathbf{P}$ is the associated $n^2 \times n^2$ permutation matrix, then

$$\mathbf{P}^T\tilde{\mathbf{L}}\mathbf{P}=\begin{pmatrix} \mathbf{T}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{T}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{T}_n \end{pmatrix} \text{ with } \mathbf{T}_i=\begin{pmatrix} \lambda_i & -1 & & & \\ -1 & \lambda_i & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & \lambda_i & -1 \\ & & & -1 & \lambda_i \end{pmatrix}_{n\times n}.$$

If you try it on the 9×9 case, you will see why it works. Now, $\mathbf{T}_i$ is another tridiagonal Toeplitz matrix, so Example 7.2.5 (p. 514) again applies to yield $\sigma(\mathbf{T}_i) = \{\lambda_i - 2\cos(j\pi/n+1),\ j = 1, 2, \ldots, n\}$. This together with (7.6.7) produces the n^2 eigenvalues of $\mathbf{L}$ as

$$\lambda_{ij} = 4 - 2\left[\cos\left(\frac{i\pi}{n+1}\right) + \cos\left(\frac{j\pi}{n+1}\right)\right], \quad i, j = 1, 2, \ldots, n,$$

or, by using the identity $1 - \cos\theta = 2\sin^2(\theta/2)$,

$$\lambda_{ij} = 4\left[\sin^2\left(\frac{i\pi}{2(n+1)}\right) + \sin^2\left(\frac{j\pi}{2(n+1)}\right)\right], \quad i, j = 1, 2, \ldots, n. \tag{7.6.8}$$

Since each λ_{ij} is positive, $\mathbf{L}$ must be positive definite. As a corollary, $\mathbf{L}$ is nonsingular, and hence $\mathbf{Lu} = \mathbf{g}$ yields a unique solution for the steady-state temperatures on the square plate (otherwise something would be amiss).

At first glance it's tempting to think that statements about positive definite matrices translate to positive semidefinite matrices simply by replacing the word "positive" by "nonnegative," but this is not always true. When $\mathbf{A}$ has zero eigenvalues (i.e., when $\mathbf{A}$ is singular) there is no LU factorization, and, unlike the positive definite case, having nonnegative leading principal minors doesn't insure that $\mathbf{A}$ is positive semidefinite—e.g., consider $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$. The positive definite properties that have semidefinite analogues are listed below.

Positive Semidefinite Matrices

For real-symmetric matrices such that $rank(\mathbf{A}_{n\times n}) = r$, the following statements are equivalent, so any one of them can serve as the definition of a ***positive semidefinite*** matrix.

- $\mathbf{x}^T\mathbf{Ax} \geq 0$ for all $\mathbf{x} \in \Re^{n\times 1}$ (the most common definition). (7.6.9)
- All eigenvalues of $\mathbf{A}$ are nonnegative. (7.6.10)
- $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ for some $\mathbf{B}$ with $rank(\mathbf{B}) = r$. (7.6.11)
- All principal minors of $\mathbf{A}$ are nonnegative. (7.6.12)

For hermitian matrices, replace $(\star)^T$ by $(\star)^*$ and $\Re$ by $\mathcal{C}$.

Proof of (7.6.9) $\Longrightarrow$ (7.6.10). The hypothesis insures $\mathbf{x}^T\mathbf{Ax} \geq 0$ for eigenvectors of $\mathbf{A}$. If $(\lambda, \mathbf{x})$ is an eigenpair, then $\lambda = \mathbf{x}^T\mathbf{Ax}/\mathbf{x}^T\mathbf{x} = \|\mathbf{Bx}\|_2^2 / \|\mathbf{x}\|_2^2 \geq 0$.

Proof of (7.6.10) $\Longrightarrow$ (7.6.11). Similar to the positive definite case, if each $\lambda_i \geq 0$, write $\mathbf{A} = \mathbf{P}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{P}^T = \mathbf{B}^T\mathbf{B}$, where $\mathbf{B} = \mathbf{D}^{1/2}\mathbf{P}^T$ has rank r.

Proof of (7.6.11) $\Longrightarrow$ (7.6.12). If $\mathbf{P}_k$ is a principal submatrix of $\mathbf{A}$, then

$$\begin{pmatrix} \mathbf{P}_k & \star \\ \star & \star \end{pmatrix} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{Q}^T\mathbf{B}^T\mathbf{B}\mathbf{Q} = \begin{pmatrix} \mathbf{F}^T \\ \star \end{pmatrix} \begin{bmatrix} \mathbf{F} \mid \star \end{bmatrix} \quad \Longrightarrow \quad \mathbf{P}_k = \mathbf{F}^T\mathbf{F}$$

for a permutation matrix $\mathbf{Q}$. Thus $\det(\mathbf{P}_k) = \det\left(\mathbf{F}^T\mathbf{F}\right) \geq 0$ (Exercise 6.1.10).

Proof of (7.6.12) $\Longrightarrow$ (7.6.9). If $\mathbf{A}_k$ is the leading $k \times k$ principal submatrix of $\mathbf{A}$, and if $\{\mu_1, \mu_2, \ldots, \mu_k\}$ are the eigenvalues (including repetitions) of $\mathbf{A}_k$, then $\epsilon\mathbf{I} + \mathbf{A}_k$ has eigenvalues $\{\epsilon + \mu_1, \epsilon + \mu_2, \ldots, \epsilon + \mu_k\}$, so, for every $\epsilon > 0$,

$$\det(\epsilon\mathbf{I} + \mathbf{A}_k) = (\epsilon + \mu_1)(\epsilon + \mu_2)\cdots(\epsilon + \mu_k) = \epsilon^k + s_1\epsilon^{k-1} + \cdots + \epsilon s_{k-1} + s_k > 0$$

because s_j is the j^{th} symmetric function of the μ_i's (p. 494), and, by (7.1.6), s_j is the sum of the $j \times j$ principal minors of $\mathbf{A}_k$, which are principal minors of $\mathbf{A}$. In other words, each leading principal minor of $\epsilon\mathbf{I} + \mathbf{A}$ is positive, so $\epsilon\mathbf{I}+\mathbf{A}$ is positive definite by the results on p. 559. Consequently, for each nonzero $\mathbf{x} \in \Re^{n\times 1}$, we must have $\mathbf{x}^T(\epsilon\mathbf{I} + \mathbf{A})\mathbf{x} > 0$ for every $\epsilon > 0$. Let $\epsilon \to 0^+$ (i.e., through positive values) to conclude that $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for each $\mathbf{x} \in \Re^{n\times 1}$. ∎

Quadratic Forms

For a vector $\mathbf{x} \in \Re^{n\times 1}$ and a matrix $\mathbf{A} \in \Re^{n\times n}$, the scalar function defined by

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \tag{7.6.13}$$

is called a ***quadratic form***. A quadratic form is said to be *positive definite* whenever $\mathbf{A}$ is a positive definite matrix. In other words, (7.6.13) is a positive definite form if and only if $f(\mathbf{x}) > 0$ for all $\mathbf{0} \neq \mathbf{x} \in \Re^{n\times 1}$.

Because $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\left[(\mathbf{A} + \mathbf{A}^T)/2\right]\mathbf{x}$, and because $(\mathbf{A} + \mathbf{A}^T)/2$ is symmetric, the matrix of a quadratic form can always be forced to be symmetric. For this reason it is assumed that the matrix of *every* quadratic form is symmetric. When $\mathbf{x} \in \mathcal{C}^{n\times 1}$, $\mathbf{A} \in \mathcal{C}^{n\times n}$, and $\mathbf{A}$ is hermitian, the expression $\mathbf{x}^H\mathbf{A}\mathbf{x}$ is known as a *complex quadratic form*.

Example 7.6.3

Diagonalization of a Quadratic Form. A quadratic form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{D}\mathbf{x}$ is said to be a *diagonal form* whenever $\mathbf{D}_{n\times n}$ is a diagonal matrix, in which case $\mathbf{x}^T\mathbf{D}\mathbf{x} = \sum_{i=1}^{n} d_{ii}x_i^2$ (there are no cross-product terms). Every quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ can be diagonalized by making a change of variables (coordinates)

$\mathbf{y} = \mathbf{Q}^T\mathbf{x}$. This follows because $\mathbf{A}$ is symmetric, so there is an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$, where $\lambda_i \in \sigma(\mathbf{A})$, and setting $\mathbf{y} = \mathbf{Q}^T\mathbf{x}$ (or, equivalently, $\mathbf{x} = \mathbf{Q}\mathbf{y}$) gives

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{Q}^T\mathbf{A}\mathbf{Q}\mathbf{y} = \mathbf{y}^T\mathbf{D}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2. \tag{7.6.14}$$

This shows that the nature of the quadratic form is determined by the eigenvalues of $\mathbf{A}$ (which are necessarily real). The effect of diagonalizing a quadratic form in this way is to rotate the standard coordinate system so that in the new coordinate system the graph of $\mathbf{x}^T\mathbf{A}\mathbf{x} = \alpha$ is in "standard form." If $\mathbf{A}$ is positive definite, then all of its eigenvalues are positive (p. 559), so (7.6.14) makes it clear that the graph of $\mathbf{x}^T\mathbf{A}\mathbf{x} = \alpha$ for a constant $\alpha > 0$ is an ellipsoid centered at the origin. Go back and look at Figure 7.2.1 (p. 505), and see Exercise 7.6.4 (p. 571).

Example 7.6.4

Congruence. It's not necessary to solve an eigenvalue problem to diagonalize a quadratic form because a ***congruence transformation*** $\mathbf{C}^T\mathbf{A}\mathbf{C}$ in which $\mathbf{C}$ is nonsingular (but not necessarily orthogonal) can be found that will do the job. A particularly convenient congruence transformation is produced by the LDU factorization for $\mathbf{A}$, which is $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ because $\mathbf{A}$ is symmetric—see Exercise 3.10.9 (p. 157). This factorization is relatively cheap, and the diagonal entries in $\mathbf{D} = \text{diag}\,(p_1, p_2, \ldots, p_n)$ are the pivots that emerge during Gaussian elimination (p. 154). Setting $\mathbf{y} = \mathbf{L}^T\mathbf{x}$ (or, equivalently, $\mathbf{x} = (\mathbf{L}^T)^{-1}\mathbf{y}$) yields

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{D}\mathbf{y} = \sum_{i=1}^{n} p_i y_i^2.$$

The ***inertia*** of a real-symmetric matrix $\mathbf{A}$ is defined to be the triple (ρ, ν, ζ) in which ρ, ν, and ζ are the respective number of positive, negative, and zero eigenvalues, counting algebraic multiplicities. In 1852 J. J. Sylvester (p. 80) discovered that the inertia of $\mathbf{A}$ is invariant under congruence transformations.

Sylvester's Law of Inertia

Let $\mathbf{A} \cong \mathbf{B}$ denote the fact that real-symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ are congruent (i.e., $\mathbf{C}^T\mathbf{A}\mathbf{C} = \mathbf{B}$ for some nonsingular $\mathbf{C}$). Sylvester's law of inertia states that:

$\mathbf{A} \cong \mathbf{B}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ have the same inertia.

Proof.[77] Observe that if $\mathbf{A}_{n\times n}$ is real and symmetric with inertia (p,j,s), then

$$\mathbf{A} \cong \begin{pmatrix} \mathbf{I}_{p\times p} & & \\ & -\mathbf{I}_{j\times j} & \\ & & \mathbf{0}_{s\times s} \end{pmatrix} = \mathbf{E}, \tag{7.6.15}$$

because if $\{\lambda_1,\ldots,\lambda_p,-\lambda_{p+1},\ldots,-\lambda_{p+j},0,\ldots,0\}$ are the eigenvalues of $\mathbf{A}$ (counting multiplicities) with each $\lambda_i > 0$, there is an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \text{diag}\,(\lambda_1,\ldots,\lambda_p,-\lambda_{p+1},\ldots,-\lambda_{p+j},0,\ldots,0)$, so $\mathbf{C} = \mathbf{P}\mathbf{D}$, where $\mathbf{D} = \text{diag}\left(\lambda_1^{-1/2},\ldots,\lambda_{p+j}^{-1/2},1,\ldots,1\right)$, is nonsingular and $\mathbf{C}^T\mathbf{A}\mathbf{C} = \mathbf{E}$. Let $\mathbf{B}$ be a real-symmetric matrix with inertia (q,k,t) so that

$$\mathbf{B} \cong \begin{pmatrix} \mathbf{I}_{q\times q} & & \\ & -\mathbf{I}_{k\times k} & \\ & & \mathbf{0}_{t\times t} \end{pmatrix} = \mathbf{F}.$$

If $\mathbf{B} \cong \mathbf{A}$, then $\mathbf{F} \cong \mathbf{E}$ (congruence is transitive), so $rank\,(\mathbf{F}) = rank\,(\mathbf{E})$, and hence $s = t$. To show that $p = q$, assume to the contrary that $p > q$, and write $\mathbf{F} = \mathbf{K}^T\mathbf{E}\mathbf{K}$ for some nonsingular $\mathbf{K} = \left(\mathbf{X}_{n\times q} \,|\, \mathbf{Y}_{n\times n-q}\right)$. If $\mathcal{M} = R\,(\mathbf{Y}) \subseteq \Re^n$ and $\mathcal{N} = span\,\{\mathbf{e}_1,\ldots,\mathbf{e}_p\} \subseteq \Re^n$, then using the formula (4.4.19) for the dimension of a sum (p. 205) yields

$$\dim(\mathcal{M}\cap\mathcal{N}) = \dim\mathcal{M} + \dim\mathcal{N} - \dim(\mathcal{M}+\mathcal{N}) = (n-q)+p-\dim(\mathcal{M}+\mathcal{N}) > 0.$$

Consequently, there exists a nonzero vector $\mathbf{x} \in \mathcal{M}\cap\mathcal{N}$. For such a vector,

$$\mathbf{x}\in\mathcal{M} \implies \mathbf{x} = \mathbf{Y}\mathbf{y} = \mathbf{K}\begin{pmatrix}\mathbf{0}\\ \mathbf{y}\end{pmatrix} \implies \mathbf{x}^T\mathbf{E}\mathbf{x} = \left(\mathbf{0}^T \,|\, \mathbf{y}^T\right)\mathbf{F}\begin{pmatrix}\mathbf{0}\\ \mathbf{y}\end{pmatrix} \le 0,$$

and

$$\mathbf{x}\in\mathcal{N} \implies \mathbf{x} = (x_1,\ldots,x_p,0,\ldots,0)^T \implies \mathbf{x}^T\mathbf{E}\mathbf{x} > 0,$$

which is impossible. Therefore, we can't have $p > q$. A similar argument shows that it's also impossible to have $p < q$, so $p = q$. Thus it is proved that if $\mathbf{A} \cong \mathbf{B}$, then $\mathbf{A}$ and $\mathbf{B}$ have the same inertia. Conversely, if $\mathbf{A}$ and $\mathbf{B}$ have inertia (p,j,s), then the argument that produced (7.6.15) yields $\mathbf{A} \cong \mathbf{E} \cong \mathbf{B}$. ∎

[77] The fact that inertia is invariant under congruence is also a corollary of a deeper theorem stating that the eigenvalues of $\mathbf{A}$ vary continuously with the entries. The argument is as follows. Assume $\mathbf{A}$ is nonsingular (otherwise consider $\mathbf{A}+\epsilon\mathbf{I}$ for small ϵ), and set $\mathbf{X}(t) = t\mathbf{Q} + (1-t)\mathbf{Q}\mathbf{R}$ for $t \in [0,1]$, where $\mathbf{C} = \mathbf{Q}\mathbf{R}$ is the QR factorization. Both $\mathbf{X}(t)$ and $\mathbf{Y}(t) = \mathbf{X}^T(t)\mathbf{A}\mathbf{X}(t)$ are nonsingular on $[0,1]$, so continuity of eigenvalues insures that no eigenvalue $\mathbf{Y}(t)$ can cross the origin as t goes from 0 to 1. Hence $\mathbf{Y}(0) = \mathbf{C}^T\mathbf{A}\mathbf{C}$ has the same number of positive (and negative) eigenvalues as $\mathbf{Y}(1) = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$, which is similar to $\mathbf{A}$. Thus $\mathbf{C}^T\mathbf{A}\mathbf{C}$ and $\mathbf{A}$ have the same inertia.

Example 7.6.5

Taylor's theorem in $\Re^n$ says that if f is a smooth real-valued function defined on $\Re^n$, and if $\mathbf{x}_0 \in \Re^{n\times 1}$, then the value of f at $\mathbf{x} \in \Re^{n\times 1}$ is given by

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T\mathbf{g}(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T\mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + O(\|\mathbf{x}-\mathbf{x}_0\|^3),$$

where $\mathbf{g}(\mathbf{x}_0) = \nabla f(\mathbf{x}_0)$ (the gradient of f evaluated at $\mathbf{x}_0$) has components $g_i = \partial f/\partial x_i\Big|_{\mathbf{x}_0}$, and where $\mathbf{H}(\mathbf{x}_0)$ is the ***Hessian matrix*** whose entries are given by $h_{ij} = \partial^2 f/\partial x_i \partial x_j\Big|_{\mathbf{x}_0}$. Just as in the case of one variable, the vector $\mathbf{x}_0$ is called a *critical point* when $\mathbf{g}(\mathbf{x}_0) = \mathbf{0}$. If $\mathbf{x}_0$ is a critical point, then Taylor's theorem shows that $(\mathbf{x}-\mathbf{x}_0)^T\mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)$ governs the behavior of f at points $\mathbf{x}$ near to $\mathbf{x}_0$. This observation yields the following conclusions regarding local maxima or minima.

- If $\mathbf{x}_0$ is a critical point such that $\mathbf{H}(\mathbf{x}_0)$ is positive definite, then f has a local minimum at $\mathbf{x}_0$.
- If $\mathbf{x}_0$ is a critical point such that $\mathbf{H}(\mathbf{x}_0)$ is *negative definite* (i.e., $\mathbf{z}^T\mathbf{H}\mathbf{z} < 0$ for all $\mathbf{z} \neq \mathbf{0}$ or, equivalently, $-\mathbf{H}$ is positive definite), then f has a local maximum at $\mathbf{x}_0$.

Exercises for section 7.6

7.6.1. Which of the following matrices are positive definite?

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 5 & 1 \\ -1 & 1 & 5 \end{pmatrix}. \quad \mathbf{B} = \begin{pmatrix} 20 & 6 & 8 \\ 6 & 3 & 0 \\ 8 & 0 & 8 \end{pmatrix}. \quad \mathbf{C} = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 6 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

7.6.2. Spring-Mass Vibrations. Two masses m_1 and m_2 are suspended between three identical springs (with spring constant k) as shown in Figure 7.6.7. Each mass is initially displaced from its equilibrium position by a horizontal distance and released to vibrate freely (assume there is no vertical displacement).

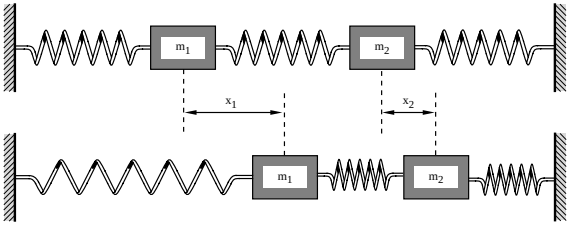


FIGURE 7.6.7

(a) If $x_i(t)$ denotes the horizontal displacement of m_i from equilibrium at time t, show that $\mathbf{Mx}'' = \mathbf{Kx}$, where

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \quad \text{and} \quad \mathbf{K} = k \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

(Consider a force directed to the left to be positive.) Notice that the *mass-stiffness equation* $\mathbf{Mx}'' = \mathbf{Kx}$ is the matrix version of Hooke's law $F = kx$, and $\mathbf{K}$ is positive definite.

(b) Look for a solution of the form $\mathbf{x} = \mathrm{e}^{i\theta t}\mathbf{v}$ for a constant vector $\mathbf{v}$, and show that this reduces the problem to solving an algebraic equation of the form $\mathbf{Kv} = \lambda\mathbf{Mv}$ (for $\lambda = -\theta^2$). This is called a ***generalized eigenvalue problem*** because when $\mathbf{M} = \mathbf{I}$ we are back to the ordinary eigenvalue problem. The *generalized eigenvalues* λ_1 and λ_2 are the roots of the equation $\det(\mathbf{K} - \lambda\mathbf{M}) = 0$—find them when $k = 1$, $m_1 = 1$, and $m_2 = 2$, and describe the two modes of vibration.

(c) Take $m_1 = m_2 = m$, and apply the technique used in the vibrating beads problem in Example 7.6.1 (p. 559) to determine the fundamental modes. Compare the results with those of part (b).

7.6.3. Three masses m_1, m_2, and m_3 are suspended on three identical springs (with spring constant k) as shown below. Each mass is initially displaced from its equilibrium position by a vertical distance and then released to vibrate freely.

(a) If $y_i(t)$ denotes the displacement of m_i from equilibrium at time t, show that the mass-stiffness equation is $\mathbf{My}'' = \mathbf{Ky}$, where

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix}, \ \mathbf{K} = k \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

($k_{33} = 1$ is not a mistake!).

(b) Show that $\mathbf{K}$ is positive definite.

(c) Find the fundamental modes when $m_1 = m_2 = m_3 = m$.

7.6.4. By diagonalizing the quadratic form $13x^2 + 10xy + 13y^2$, show that the rotated graph of $13x^2 + 10xy + 13y^2 = 72$ is an ellipse in standard form as shown in Figure 7.2.1 on p. 505.

7.6.5. Suppose that $\mathbf{A}$ is a real-symmetric matrix. Explain why the signs of the pivots in the LDU factorization for $\mathbf{A}$ reveal the inertia of $\mathbf{A}$.

7.6.6. Consider the quadratic form

$$f(\mathbf{x}) = \frac{1}{9}(-2x_1^2 + 7x_2^2 + 4x_3^2 + 4x_1x_2 + 16x_1x_3 + 20x_2x_3).$$

(a) Find a symmetric matrix $\mathbf{A}$ so that $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$.
(b) Diagonalize the quadratic form using the $\mathbf{LDL}^T$ factorization as described in Example 7.6.4, and determine the inertia of $\mathbf{A}$.
(c) Is this a positive definite form?
(d) Verify the inertia obtained above is correct by computing the eigenvalues of $\mathbf{A}$.
(e) Verify Sylvester's law of inertia by making up a congruence transformation $\mathbf{C}$ and then computing the inertia of $\mathbf{C}^T\mathbf{AC}$.

7.6.7. Polar Factorization. Explain why each nonsingular $\mathbf{A} \in \mathcal{C}^{n\times n}$ can be uniquely factored as $\mathbf{A} = \mathbf{RU}$, where $\mathbf{R}$ is hermitian positive definite and $\mathbf{U}$ is unitary. This is the matrix analog of the polar form of a complex number $z = re^{i\theta}$, $r > 0$, because 1×1 hermitian positive definite matrices are positive real numbers, and 1×1 unitary matrices are points on the unit circle. **Hint:** First explain why $\mathbf{R} = (\mathbf{AA}^*)^{1/2}$.

7.6.8. Explain why trying to produce better approximations to the solution of the Dirichlet problem in Example 7.6.2 by using finer meshes with more grid points results in an increasingly ill-conditioned linear system $\mathbf{Lu} = \mathbf{g}$.

7.6.9. For a given function f the equation $\nabla^2 u = f$ is called *Poisson's*[78] *equation.* Consider Poisson's equation on a square in two dimensions with Dirichlet boundary conditions. That is,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \quad \text{with} \quad u(x, y) = g(x, y) \quad \text{on the boundary.}$$

[78] Siméon Denis Poisson (1781–1840) was a prolific French scientist who was originally encouraged to study medicine but was seduced by mathematics. While he was still a teenager, his work attracted the attention of the reigning scientific elite of France such as Legendre, Laplace, and Lagrange. The latter two were originally his teachers (Lagrange was his thesis director) at the École Polytechnique, but they eventually became his friends and collaborators. It is estimated that Poisson published about 400 scientific articles, and his 1811 book *Traité de mécanique* was the standard reference for mechanics for many years. Poisson began his career as an astronomer, but he is primarily remembered for his impact on applied areas such as mechanics, probability, electricity and magnetism, and Fourier series. This seems ironic because he held the chair of "pure mathematics" in the Faculté des Sciences. The next time you find yourself on the streets of Paris, take a stroll on the Rue Denis Poisson, or you can check out Poisson's plaque, along with those of Lagrange, Laplace, and Legendre, on the first stage of the Eiffel Tower.

Discretize the problem by overlaying the square with a regular mesh containing n^2 interior points at equally spaced intervals of length h as explained in Example 7.6.2 (p. 563). Let $f_{ij} = f(x_i, y_j)$, and define $\mathbf{f}$ to be the vector $\mathbf{f} = (f_{11}, f_{12}, \ldots, f_{1n}|f_{21}, f_{22} \ldots, f_{2n}|\cdots|f_{n1}, f_{n2}, \ldots, f_{nn})^T$. Show that the discretization of Poisson's equation produces a system of linear equations of the form $\mathbf{Lu} = \mathbf{g} - h^2\mathbf{f}$, where $\mathbf{L}$ is the discrete Laplacian and where $\mathbf{u}$ and $\mathbf{g}$ are as described in Example 7.6.2.

7.6.10. As defined in Exercise 5.8.15 (p. 380) and discussed in Exercise 7.8.11 (p. 597) the *Kronecker product* (sometimes called *tensor product*, or *direct product*) of matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{p\times q}$ is the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}.$$

Verify that if $\mathbf{I}_n$ is the $n \times n$ identity matrix, and if

$$\mathbf{A}_n = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}_{n\times n}$$

is the n^{th}-order finite difference matrix of Example 1.4.1 (p. 19), then the discrete Laplacian is given by

$$\mathbf{L}_{n^2\times n^2} = (\mathbf{I}_n \otimes \mathbf{A}_n) + (\mathbf{A}_n \otimes \mathbf{I}_n).$$

Thus we have an elegant matrix connection between the finite difference approximations of the one-dimensional and two-dimensional Laplacians. This formula leads to a simple alternate derivation of (7.6.8)—see Exercise 7.8.12 (p. 598). As you might guess, the discrete three-dimensional Laplacian is

$$\mathbf{L}_{n^3\times n^3} = (\mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{A}_n) + (\mathbf{I}_n \otimes \mathbf{A}_n \otimes \mathbf{I}_n) + (\mathbf{A}_n \otimes \mathbf{I}_n \otimes \mathbf{I}_n).$$

7.7 NILPOTENT MATRICES AND JORDAN STRUCTURE

While it's not always possible to diagonalize a matrix $\mathbf{A} \in \mathcal{C}^{m\times m}$ with a similarity transformation, Schur's theorem (p. 508) guarantees that every $\mathbf{A} \in \mathcal{C}^{m\times m}$ is *unitarily* similar to an upper-triangular matrix—say $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$. But other than the fact that the diagonal entries of $\mathbf{T}$ are the eigenvalues of $\mathbf{A}$, there is no pattern to the nonzero part of $\mathbf{T}$. So to what extent can this be remedied by giving up the unitary nature of $\mathbf{U}$? In other words, is there a nonunitary $\mathbf{P}$ for which $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ has a simpler and more predictable pattern than that of $\mathbf{T}$? We have already made the first step in answering this question. The core-nilpotent decomposition (p. 397) says that for every singular matrix $\mathbf{A}$ of index k and rank r, there is a nonsingular matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{L} \end{pmatrix}, \text{ where } rank\,(\mathbf{C}) = r \text{ and } \mathbf{L} \text{ is nilpotent of index } k.$$

Consequently, any further simplification by means of similarity transformations can revolve around $\mathbf{C}$ and $\mathbf{L}$. Let's begin by examining the degree to which nilpotent matrices can be reduced by similarity transformations.

In what follows, let $\mathbf{L}_{n\times n}$ be a nilpotent matrix of index k so that $\mathbf{L}^k = \mathbf{0}$ but $\mathbf{L}^{k-1} \neq \mathbf{0}$. The first question is, "Can $\mathbf{L}$ be diagonalized by a similarity transformation?" To answer this, notice that $\lambda = 0$ is the only eigenvalue of $\mathbf{L}$ because

$$\mathbf{L}\mathbf{x} = \lambda\mathbf{x} \implies \mathbf{L}^k\mathbf{x} = \lambda^k\mathbf{x} \implies \mathbf{0} = \lambda^k\mathbf{x} \implies \lambda = 0 \quad (\text{since } \mathbf{x} \neq \mathbf{0}\,).$$

So if $\mathbf{L}$ is to be diagonalized by a similarity transformation, it must be the case that $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{D} = \mathbf{0}$ (diagonal entries of $\mathbf{D}$ must be eigenvalues of $\mathbf{L}$), and this forces $\mathbf{L} = \mathbf{0}$. In other words, the *only* nilpotent matrix that is similar to a diagonal matrix is the zero matrix.

Assume $\mathbf{L} \neq \mathbf{0}$ from now on so that $\mathbf{L}$ is not diagonalizable. Since $\mathbf{L}$ can always be triangularized (Schur's theorem again), our problem boils down to finding a nonsingular $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{L}\mathbf{P}$ is an upper-triangular matrix possessing a simple and predictable form. This turns out to be a fundamental problem, and the rest of this section is devoted to its solution. But before diving in, let's set the stage by thinking about some possibilities.

If $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{T}$ is upper triangular, then the diagonal entries of $\mathbf{T}$ must be the eigenvalues of $\mathbf{L}$, so $\mathbf{T}$ must have the form

$$\mathbf{T} = \begin{pmatrix} 0 & \star & \cdots & \star \\ & \ddots & \ddots & \vdots \\ & & \ddots & \star \\ & & & 0 \end{pmatrix}.$$

One way to simplify the form of $\mathbf{T}$ is to allow nonzero entries only on the superdiagonal (the diagonal immediately above the main diagonal) of $\mathbf{T}$, so we might try to construct a nonsingular $\mathbf{P}$ such that $\mathbf{T}$ has the form

$$\mathbf{T} = \begin{pmatrix} 0 & \star & & \\ & \ddots & \ddots & \\ & & \ddots & \star \\ & & & 0 \end{pmatrix}.$$

To gain some insight on how this might be accomplished, let $\mathbf{L}$ be a 3×3 nilpotent matrix for which $\mathbf{L}^3 = \mathbf{0}$ and $\mathbf{L}^2 \neq \mathbf{0}$, and search for a $\mathbf{P}$ such that

$$\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \Longleftrightarrow \mathbf{L}[\,\mathbf{P}_{*1}\,\mathbf{P}_{*2}\,\mathbf{P}_{*3}\,] = [\,\mathbf{P}_{*1}\,\mathbf{P}_{*2}\,\mathbf{P}_{*3}\,] \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$
$$\Longleftrightarrow \mathbf{L}\mathbf{P}_{*1} = \mathbf{0}, \quad \mathbf{L}\mathbf{P}_{*2} = \mathbf{P}_{*1}, \quad \mathbf{L}\mathbf{P}_{*3} = \mathbf{P}_{*2}.$$

Since $\mathbf{L}^3 = \mathbf{0}$, we can set $\mathbf{P}_{*1} = \mathbf{L}^2\mathbf{x}$ for any $\mathbf{x}_{3\times 1}$ such that $\mathbf{L}^2\mathbf{x} \neq \mathbf{0}$. This in turn allows us to set $\mathbf{P}_{*2} = \mathbf{L}\mathbf{x}$ and $\mathbf{P}_{*3} = \mathbf{x}$. Because $\mathcal{J} = \{\mathbf{L}^2\mathbf{x}, \mathbf{L}\mathbf{x}, \mathbf{x}\}$ is a linearly independent set (Exercise 5.10.8), $\mathbf{P} = [\,\mathbf{L}^2\mathbf{x}\,|\,\mathbf{L}\mathbf{x}\,|\,\mathbf{x}\,]$ will do the job. $\mathcal{J}$ is called a ***Jordan chain***, and it is characterized by the fact that its first vector is a somewhat special eigenvector for $\mathbf{L}$ while the other vectors are built (or "chained") on top of this eigenvector to form a special basis for $\mathcal{C}^3$. There are a few more wrinkles in the development of a general theory for $n \times n$ nilpotent matrices, but the features illustrated here illuminate the path.

For a general nilpotent matrix $\mathbf{L}_{n\times n} \neq \mathbf{0}$ of index k, we know that $\lambda = 0$ is the only eigenvalue, so the set of eigenvectors of $\mathbf{L}$ is $N(\mathbf{L})$ (excluding the zero vector of course). Realizing that $\mathbf{L}$ is not diagonalizable is equivalent to realizing that $\mathbf{L}$ does not possess a complete linearly independent set of eigenvectors or, equivalently, $\dim N(\mathbf{L}) < n$. As in the 3×3 example above, the strategy for building a similarity transformation $\mathbf{P}$ that reduces $\mathbf{L}$ to a simple triangular form is as follows.

(1) Construct a somewhat special basis $\mathcal{B}$ for $N(\mathbf{L})$.

(2) Extend $\mathcal{B}$ to a basis for $\mathcal{C}^n$ by building Jordan chains on top of the eigenvectors in $\mathcal{B}$.

To accomplish (1), consider the subspaces defined by

$$\mathcal{M}_i = R\left(\mathbf{L}^i\right) \cap N(\mathbf{L}) \quad \text{for } i = 0, 1, \ldots, k, \tag{7.7.1}$$

and notice (Exercise 7.7.4) that these subspaces are nested as

$$\mathbf{0} = \mathcal{M}_k \subseteq \mathcal{M}_{k-1} \subseteq \mathcal{M}_{k-2} \subseteq \cdots \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_0 = N(\mathbf{L}).$$

Use these nested spaces to construct a basis for $N(\mathbf{L}) = \mathcal{M}_0$ by starting with any basis $\mathcal{S}_{k-1}$ for $\mathcal{M}_{k-1}$ and by sequentially extending $\mathcal{S}_{k-1}$ with additional sets $\mathcal{S}_{k-2}, \mathcal{S}_{k-3}, \ldots, \mathcal{S}_0$ such that $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2}$ is a basis for $\mathcal{M}_{k-2}$, $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \mathcal{S}_{k-3}$ is a basis for $\mathcal{M}_{k-3}$, etc. In general, $\mathcal{S}_i$ is a set of vectors that extends $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_{i-1}$ to a basis for $\mathcal{M}_i$. Figure 7.7.1 is a heuristic diagram depicting an example of $k = 5$ nested subspaces $\mathcal{M}_i$ along with some typical extension sets $\mathcal{S}_i$ that combine to form a basis for $N(\mathbf{L})$.

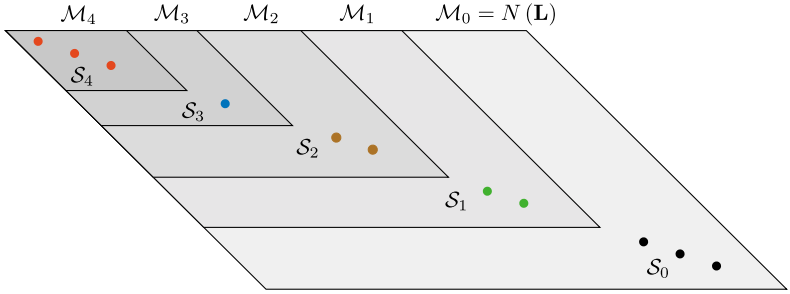


FIGURE 7.7.1

Now extend the basis $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$ for $N(\mathbf{L})$ to a basis for $\mathcal{C}^n$ by building Jordan chains on top of each $\mathbf{b} \in \mathcal{B}$. If $\mathbf{b} \in \mathcal{S}_i$, then there exists a vector $\mathbf{x}$ such that $\mathbf{L}^i\mathbf{x} = \mathbf{b}$ because each $\mathbf{b} \in \mathcal{S}_i$ belongs to $\mathcal{M}_i = R\left(\mathbf{L}^i\right) \cap N(\mathbf{L}) \subseteq R\left(\mathbf{L}^i\right)$. A ***Jordan chain*** is built on top of each $\mathbf{b} \in \mathcal{S}_i$ by solving the system $\mathbf{L}^i\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ and by setting

$$\mathcal{J}_{\mathbf{b}} = \{\mathbf{L}^i\mathbf{x}, \mathbf{L}^{i-1}\mathbf{x}, \ldots, \mathbf{L}\mathbf{x}, \mathbf{x}\}. \tag{7.7.2}$$

Notice that chains built on top of vectors from $\mathcal{S}_i$ each have length $i+1$. The heuristic diagram in Figure 7.7.2 depicts Jordan chains built on top of the basis vectors illustrated in Figure 7.7.1—the chain that is labeled is built on top of a vector $\mathbf{b} \in \mathcal{S}_3$.

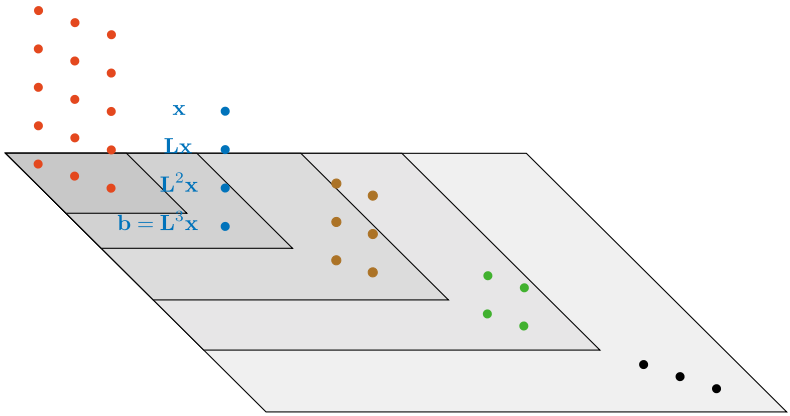


FIGURE 7.7.2

The collection of vectors in all of these Jordan chains is a basis for $\mathcal{C}^n$. To demonstrate this, first it must be argued that the total number of vectors in all Jordan chains is n, and then it must be proven that this collection is a linearly independent set. To count the number of vectors in all Jordan chains $\mathcal{J}_{\mathbf{b}}$, first recall from (4.5.1) that the rank of a product is given by the formula $rank\,(\mathbf{AB}) = rank\,(\mathbf{B}) - \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B})$, and apply this to conclude that $\dim \mathcal{M}_i = \dim R\left(\mathbf{L}^i\right) \cap N\,(\mathbf{L}) = rank\left(\mathbf{L}^i\right) - rank\left(\mathbf{L}\mathbf{L}^i\right)$. In other words, if we set $d_i = \dim \mathcal{M}_i$ and $r_i = rank\left(\mathbf{L}^i\right)$, then

$$d_i = \dim \mathcal{M}_i = rank\left(\mathbf{L}^i\right) - rank\left(\mathbf{L}^{i+1}\right) = r_i - r_{i+1}, \tag{7.7.3}$$

so the number of vectors in $\mathcal{S}_i$ is

$$\nu_i = d_i - d_{i+1} = r_i - 2r_{i+1} + r_{i+2}. \tag{7.7.4}$$

Since every chain emanating from a vector in $\mathcal{S}_i$ contains $i+1$ vectors, and since $d_k = 0 = r_k$, the total number of vectors in all Jordan chains is

$$\begin{aligned}
\text{total} &= \sum_{i=0}^{k-1}(i+1)\nu_i = \sum_{i=0}^{k-1}(i+1)(d_i - d_{i+1}) \\
&= d_0 - d_1 + 2(d_1 - d_2) + 3(d_2 - d_3) + \cdots + k(d_{k-1} - d_k) \\
&= d_0 + d_1 + \cdots + d_{k-1} \\
&= (r_0 - r_1) + (r_1 - r_2) + (r_2 - r_3) + \cdots + (r_{k-1} - r_k) \\
&= r_0 = n.
\end{aligned}$$

To prove that the set of all vectors from all Jordan chains is linearly independent, place these vectors as columns in a matrix $\mathbf{Q}_{n\times n}$ and show that $N\,(\mathbf{Q}) = \mathbf{0}$. The trick in doing so is to arrange the vectors from the $\mathcal{J}_{\mathbf{b}}$'s in just the right order. Begin by placing the vectors at the top level in chains emanating from $\mathcal{S}_i$ as columns in a matrix $\mathbf{X}_i$ as depicted in the heuristic diagram in Figure 7.7.3.

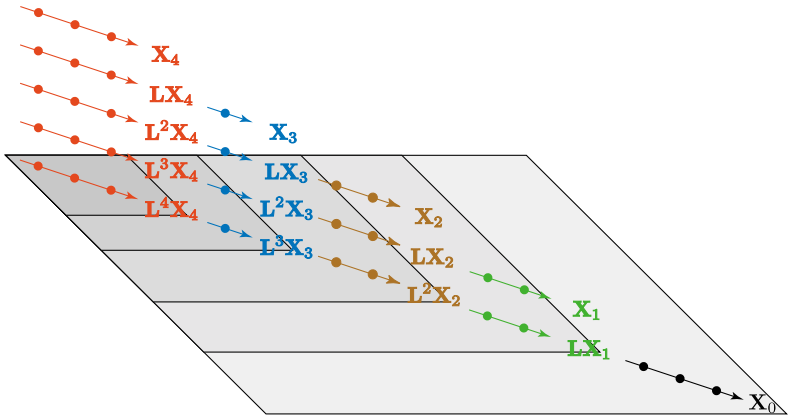


Figure 7.7.3

The matrix $\mathbf{LX}_i$ contains all vectors at the second highest level of those chains emanating from $\mathcal{S}_i$, while $\mathbf{L}^2\mathbf{X}_i$ contains all vectors at the third highest level of those chains emanating from $\mathcal{S}_i$, and so on. In general, $\mathbf{L}^j\mathbf{X}_i$ contains all vectors at the $(j+1)^{st}$ highest level of those chains emanating from $\mathcal{S}_i$. Proceed by filling in $\mathbf{Q} = [\,\mathbf{Q}_0\,|\,\mathbf{Q}_1\,|\cdots|\,\mathbf{Q}_{k-1}\,]$ from the bottom up by letting $\mathbf{Q}_j$ be the matrix whose columns are all vectors at the j^{th} level from the bottom in all chains. For the example illustrated in Figures 7.7.1–7.7.3 with $k = 5$,

$$\mathbf{Q}_0 = [\,\mathbf{X}_0\,|\,\mathbf{LX}_1\,|\,\mathbf{L}^2\mathbf{X}_2\,|\,\mathbf{L}^3\mathbf{X}_3\,|\,\mathbf{L}^4\mathbf{X}_4\,] = \text{vectors at level 0} = \text{basis } \mathcal{B} \text{ for } N(\mathbf{L}),$$

$$\mathbf{Q}_1 = [\,\mathbf{X}_1\,|\,\mathbf{LX}_2\,|\,\mathbf{L}^2\mathbf{X}_3\,|\,\mathbf{L}^3\mathbf{X}^4\,] = \text{ vectors at level 1 (from the bottom)},$$

$$\mathbf{Q}_2 = [\,\mathbf{X}_2\,|\,\mathbf{LX}_3\,|\,\mathbf{L}^2\mathbf{X}^4\,] = \text{ vectors at level 2 (from the bottom)},$$

$$\mathbf{Q}_3 = [\,\mathbf{X}_3\,|\,\mathbf{LX}^4\,] = \text{ vectors at level 3 (from the bottom)},$$

$$\mathbf{Q}_4 = [\,\mathbf{X}_4\,] = \text{ vectors at level 4 (from the bottom)}.$$

In general, $\mathbf{Q}_j = [\,\mathbf{X}_j\,|\,\mathbf{LX}_{j+1}\,|\,\mathbf{L}^2\mathbf{X}_{j+2}\,|\cdots|\,\mathbf{L}^{k-1-j}\mathbf{X}_{k-1}\,]$. Since the columns of $\mathbf{L}^j\mathbf{X}_j$ are all on the bottom level (level 0), they are part of the basis $\mathcal{B}$ for $N(\mathbf{L})$. This means that the columns of $\mathbf{L}^j\mathbf{Q}_j$ are also part of the basis $\mathcal{B}$ for $N(\mathbf{L})$, so they are linearly independent, and thus $N\left(\mathbf{L}^j\mathbf{Q}_j\right) = \mathbf{0}$. Furthermore, since the columns of $\mathbf{L}^j\mathbf{Q}_j$ are in $N(\mathbf{L})$, we have $\mathbf{L}\left(\mathbf{L}^j\mathbf{Q}_j\right) = \mathbf{0}$, and hence $\mathbf{L}^{j+h}\mathbf{Q}_j = \mathbf{0}$ for all $h \geq 1$. Now use these observations to prove $N(\mathbf{Q}) = \mathbf{0}$. If $\mathbf{Qz} = \mathbf{0}$, then multiplication by $\mathbf{L}^{k-1}$ yields

$$\begin{aligned}
\mathbf{0} = \mathbf{L}^{k-1}\mathbf{Qz} &= [\,\mathbf{L}^{k-1}\mathbf{Q}_0\,|\,\mathbf{L}^{k-1}\mathbf{Q}_1\,|\cdots|\,\mathbf{L}^{k-1}\mathbf{Q}_{k-1}\,]\,\mathbf{z} \\
&= [\,\mathbf{0}\,|\,\mathbf{0}\,|\cdots|\,\mathbf{L}^{k-1}\mathbf{Q}_{k-1}\,]\begin{pmatrix} \mathbf{z}_0 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_{k-1} \end{pmatrix} \implies \mathbf{z}_{k-1} \in N\left(\mathbf{L}^{k-1}\mathbf{Q}_{k-1}\right) \\
&\implies \mathbf{z}_{k-1} = \mathbf{0}.
\end{aligned}$$

This conclusion with the same argument applied to $\mathbf{0} = \mathbf{L}^{k-2}\mathbf{Qz}$ produces $\mathbf{z}_{k-2} = \mathbf{0}$. Similar repetitions show that $\mathbf{z}_i = \mathbf{0}$ for each i, and thus $N(\mathbf{Q}) = \mathbf{0}$.

It has now been proven that if $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$ is the basis for $N(\mathbf{L})$ derived from the nested subspaces $\mathcal{M}_i$, then the set of all Jordan chains $\mathcal{J} = \mathcal{J}_{\mathbf{b}_1} \cup \mathcal{J}_{\mathbf{b}_2} \cup \cdots \cup \mathcal{J}_{\mathbf{b}_t}$ is a basis for $\mathcal{C}^n$. If the vectors from $\mathcal{J}$ are placed as columns (in the order in which they appear in $\mathcal{J}$) in a matrix $\mathbf{P}_{n\times n} = [\,\mathbf{J}_1\,|\,\mathbf{J}_2\,|\cdots|\,\mathbf{J}_t\,]$, then $\mathbf{P}$ is nonsingular, and if $\mathbf{b}_j \in \mathcal{S}_i$, then $\mathbf{J}_j = [\,\mathbf{L}^i\mathbf{x}\,|\,\mathbf{L}^{i-1}\mathbf{x}\,|\cdots|\,\mathbf{Lx}\,|\,\mathbf{x}\,]$ for some $\mathbf{x}$ such that $\mathbf{L}^i\mathbf{x} = \mathbf{b}_j$ so that

$$\mathbf{LJ}_j = [\,\mathbf{0}\,|\,\mathbf{L}^i\mathbf{x}\,|\cdots|\,\mathbf{Lx}\,] = [\,\mathbf{L}^i\mathbf{x}\,|\cdots|\,\mathbf{Lx}\,|\,\mathbf{x}\,]\begin{pmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & 1 & \\ & & & 0 & \end{pmatrix} = \mathbf{J}_j\mathbf{N}_j,$$

where $\mathbf{N}_j$ is an $(i+1)\times(i+1)$ matrix whose entries are equal to 1 along the superdiagonal and zero elsewhere. Therefore,

$$\mathbf{LP} = [\,\mathbf{LJ}_1\,|\,\mathbf{LJ}_2\,|\cdots|\,\mathbf{LJ}_t\,] = [\,\mathbf{J}_1\,|\,\mathbf{J}_2\,|\cdots|\,\mathbf{J}_t\,]\begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix}$$

or, equivalently,

$$\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N} = \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix}, \text{ where } \mathbf{N}_j = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}. \quad (7.7.5)$$

Each $\mathbf{N}_j$ is a nilpotent matrix whose index is given by its size. The $\mathbf{N}_j$'s are called ***nilpotent Jordan blocks***, and the block-diagonal matrix $\mathbf{N}$ is called the ***Jordan form*** for $\mathbf{L}$. Below is a summary.

Jordan Form for a Nilpotent Matrix

Every nilpotent matrix $\mathbf{L}_{n\times n}$ of index k is similar to a block-diagonal matrix

$$\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N} = \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix} \quad (7.7.6)$$

in which each $\mathbf{N}_j$ is a nilpotent matrix having ones on the superdiagonal and zeros elsewhere—see (7.7.5).

- The number of blocks in $\mathbf{N}$ is given by $t = \dim N(\mathbf{L})$.
- The size of the largest block in $\mathbf{N}$ is $k \times k$.
- The number of $i \times i$ blocks in $\mathbf{N}$ is $\nu_i = r_{i-1} - 2r_i + r_{i+1}$, where $r_i = rank\left(\mathbf{L}^i\right)$—this follows from (7.7.4).
- If $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$ is a basis for $N(\mathbf{L})$ derived from the nested subspaces $\mathcal{M}_i = R\left(\mathbf{L}^i\right) \cap N(\mathbf{L})$, then
  - ▷ the set of vectors $\mathcal{J} = \mathcal{J}_{\mathbf{b}_1} \cup \mathcal{J}_{\mathbf{b}_2} \cup \cdots \cup \mathcal{J}_{\mathbf{b}_t}$ from all Jordan chains is a basis for $\mathcal{C}^n$;
  - ▷ $\mathbf{P}_{n\times n} = [\,\mathbf{J}_1\,|\,\mathbf{J}_2\,|\cdots|\,\mathbf{J}_t\,]$ is the nonsingular matrix containing these Jordan chains in the order in which they appear in $\mathcal{J}$.

The following theorem demonstrates that the ***Jordan structure*** (the number and the size of the blocks in $\mathbf{N}$) is uniquely determined by $\mathbf{L}$, but $\mathbf{P}$ is not. In other words, the Jordan form is unique up to the arrangement of the individual Jordan blocks.

Uniqueness of the Jordan Structure

The structure of the Jordan form for a nilpotent matrix $\mathbf{L}_{n\times n}$ of index k is uniquely determined by $\mathbf{L}$ in the sense that whenever $\mathbf{L}$ is similar to a block-diagonal matrix $\mathbf{B} = \operatorname{diag}(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_t)$ in which each $\mathbf{B}_i$ has the form

$$\mathbf{B}_i = \begin{pmatrix} 0 & \epsilon_i & 0 & \cdots & 0 \\ 0 & 0 & \epsilon_i & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \epsilon_i \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}_{n_i \times n_i} \quad \text{for} \quad \epsilon_i \neq 0,$$

then it must be the case that $t = \dim N(\mathbf{L})$, and the number of blocks having size $i \times i$ must be given by $r_{i-1} - 2r_i + r_{i+1}$, where $r_i = rank\left(\mathbf{L}^i\right)$.

Proof. Suppose that $\mathbf{L}$ is similar to both $\mathbf{B}$ and $\mathbf{N}$, where $\mathbf{B}$ is as described above and $\mathbf{N}$ is as described in (7.7.6). This implies that $\mathbf{B}$ and $\mathbf{N}$ are similar, and hence $rank\left(\mathbf{B}^i\right) = rank\left(\mathbf{L}^i\right) = r_i$ for every nonnegative integer i. In particular, $index(\mathbf{B}) = index(\mathbf{L})$. Each time a block $\mathbf{B}_i$ is powered, the line of ϵ_i's moves to the next higher diagonal level so that

$$rank\left(\mathbf{B}_i^p\right) = \begin{cases} n_i - p & \text{if } p < n_i, \\ 0 & \text{if } p \geq n_i. \end{cases}$$

Since $r_p = rank\left(\mathbf{B}^p\right) = \sum_{i=1}^{t} rank\left(\mathbf{B}_i^p\right)$, it follows that if ω_i is the number of $i \times i$ blocks in $\mathbf{B}$, then

$$\begin{aligned} r_{k-1} &= \omega_k, \\ r_{k-2} &= \omega_{k-1} + 2\omega_k, \\ r_{k-3} &= \omega_{k-2} + 2\omega_{k-1} + 3\omega_k, \\ &\vdots \end{aligned}$$

and, in general, $r_i = \omega_{i+1} + 2\omega_{i+2} + \cdots + (k-i)\omega_k$. It's now straightforward to verify that $r_{i-1} - 2r_i + r_{i+1} = \omega_i$. Finally, using this equation together with (7.7.4) guarantees that the number of blocks in $\mathbf{B}$ must be

$$t = \sum_{i=1}^{k} \omega_i = \sum_{i=1}^{k} (r_{i-1} - 2r_i + r_{i+1}) = \sum_{i=1}^{k} \nu_i = \dim N(\mathbf{L}). \quad \blacksquare$$

The manner in which we developed the Jordan theory spawned 1's on the superdiagonals of the Jordan blocks $\mathbf{N}_i$ in (7.7.5). But it was not necessary to do so—it was simply a matter of convenience. In fact, any nonzero value can be forced onto the superdiagonal of any $\mathbf{N}_i$—see Exercise 7.7.9. In other words, the fact that 1's appear on the superdiagonals of the $\mathbf{N}_i$'s is artificial and is not important to the structure of the Jordan form for $\mathbf{L}$. What's important, and what constitutes the "Jordan structure," is the number and sizes of the Jordan blocks (or chains) and not the values appearing on the superdiagonals of these blocks.

Example 7.7.1

Problem: Determine the Jordan forms for 3×3 nilpotent matrices $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$ that have respective indices $k = 1, 2, 3$.

Solution: The size of the largest block must be $k \times k$, so

$$\mathbf{N}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{N}_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{N}_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Example 7.7.2

For a nilpotent matrix $\mathbf{L}$, the theoretical development relies on a complicated basis for $N(\mathbf{L})$ to derive the structure of the Jordan form $\mathbf{N}$ as well as the Jordan chains that constitute a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{N}$. But, after the dust settled, we saw that a basis for $N(\mathbf{L})$ is not needed to construct $\mathbf{N}$ because $\mathbf{N}$ is completely determined simply by ranks of powers of $\mathbf{L}$. A basis for $N(\mathbf{L})$ is only required to construct the Jordan chains in $\mathbf{P}$.

Question: For the purpose of constructing Jordan chains in $\mathbf{P}$, can we use an arbitrary basis for $N(\mathbf{L})$ instead of the complicated basis built from the $\mathcal{M}_i$'s?

Answer: No! Consider the nilpotent matrix

$$\mathbf{L} = \begin{pmatrix} 2 & 0 & 1 \\ -4 & 0 & -2 \\ -4 & 0 & -2 \end{pmatrix} \quad \text{and its Jordan form} \quad \mathbf{N} = \left(\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right).$$

If $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{N}$, where $\mathbf{P} = [\,\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \mathbf{x}_3\,]$, then $\mathbf{L}\mathbf{P} = \mathbf{P}\mathbf{N}$ implies that $\mathbf{L}\mathbf{x}_1 = \mathbf{0}$, $\mathbf{L}\mathbf{x}_2 = \mathbf{x}_1$, and $\mathbf{L}\mathbf{x}_3 = \mathbf{0}$. In other words, $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_3\}$ must be a basis for $N(\mathbf{L})$, and $\mathcal{J}_{\mathbf{x}_1} = \{\mathbf{x}_1, \mathbf{x}_2\}$ must be a Jordan chain built on top of $\mathbf{x}_1$. If we try to construct such vectors by starting with the naive basis

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{7.7.7}$$

for $N(\mathbf{L})$ obtained by solving $\mathbf{Lx} = \mathbf{0}$ with straightforward Gaussian elimination, we immediately hit a brick wall because $\mathbf{x}_1 \notin R(\mathbf{L})$ means $\mathbf{Lx}_2 = \mathbf{x}_1$ is an inconsistent system, so $\mathbf{x}_2$ cannot be determined. Similarly, $\mathbf{x}_3 \notin R(\mathbf{L})$ insures that the same difficulty occurs if $\mathbf{x}_3$ is used in place of $\mathbf{x}_1$. In other words, even though the vectors in (7.7.7) constitute an otherwise perfectly good basis for $N(\mathbf{L})$, they can't be used to build $\mathbf{P}$.

Example 7.7.3

Problem: Let $\mathbf{L}_{n\times n}$ be a nilpotent matrix of index k. Provide an algorithm for constructing the Jordan chains that generate a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$ is in Jordan form.

Solution:

1. Start with the fact that $\mathcal{M}_{k-1} = R\left(\mathbf{L}^{k-1}\right)$ (Exercise 7.7.5), and determine a basis $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\}$ for $R\left(\mathbf{L}^{k-1}\right)$.
2. Extend $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\}$ to a basis for $\mathcal{M}_{k-2} = R\left(\mathbf{L}^{k-2}\right) \cap N(\mathbf{L})$ as follows.
   - ▷ Find a basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s\}$ for $N(\mathbf{LB})$, where $\mathbf{B}$ is a matrix containing a basis for $R\left(\mathbf{L}^{k-2}\right)$—e.g., the basic columns of $\mathbf{L}^{k-2}$. The set $\{\mathbf{Bv}_1, \mathbf{Bv}_2, \ldots, \mathbf{Bv}_s\}$ is a basis for $\mathcal{M}_{k-2}$ (see p. 211).
   - ▷ Find the basic columns in $[\,\mathbf{y}_1 \,|\, \mathbf{y}_2 \,|\, \cdots \,|\, \mathbf{y}_q \,|\, \mathbf{Bv}_1 \,|\, \mathbf{Bv}_2 \,|\, \cdots \,|\, \mathbf{Bv}_s\,]$. Say they are $\{\mathbf{y}_1, \ldots, \mathbf{y}_q, \mathbf{Bv}_{\beta_1}, \ldots, \mathbf{Bv}_{\beta_j}\}$ (all of the $\mathbf{y}_j$'s are basic because they are a leading linearly independent subset). This is a basis for $\mathcal{M}_{k-2}$ that contains a basis for $\mathcal{M}_{k-1}$. In other words,

$$\mathcal{S}_{k-1} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\} \quad \text{and} \quad \mathcal{S}_{k-2} = \{\mathbf{Bv}_{\beta_1}, \mathbf{Bv}_{\beta_2}, \ldots, \mathbf{Bv}_{\beta_j}\}.$$

3. Repeat the above procedure $k-1$ times to construct a basis for $N(\mathbf{L})$ that is of the form $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$, where $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_i$ is a basis for $\mathcal{M}_i$ for each $i = k-1, k-2, \ldots, 0$.
4. Build a Jordan chain on top of each $\mathbf{b}_j \in \mathcal{B}$. If $\mathbf{b}_j \in \mathcal{S}_i$, then we solve $\mathbf{L}^i\mathbf{x}_j = \mathbf{b}_j$ and set $\mathbf{J}_j = [\,\mathbf{L}^i\mathbf{x}_j \,|\, \mathbf{L}^{i-1}\mathbf{x}_j \,|\, \cdots \,|\, \mathbf{Lx}_j \,|\, \mathbf{x}_j\,]$. The desired similarity transformation is $\mathbf{P}_{n\times n} = [\,\mathbf{J}_1 \,|\, \mathbf{J}_2 \,|\, \cdots \,|\, \mathbf{J}_t\,]$.

Example 7.7.4

Problem: Find $\mathbf{P}$ and $\mathbf{N}$ such that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$ is in Jordan form, where

$$\mathbf{L} = \begin{pmatrix} 1 & 1 & -2 & 0 & 1 & -1 \\ 3 & 1 & 5 & 1 & -1 & 3 \\ -2 & -1 & 0 & 0 & -1 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ -5 & -3 & -1 & -1 & -1 & -1 \\ -3 & -2 & -1 & -1 & 0 & -1 \end{pmatrix}.$$

Solution: First determine the Jordan form for $\mathbf{L}$. Computing $r_i = rank\left(\mathbf{L}^i\right)$ reveals that $r_1 = 3,\ r_2 = 1,$ and $r_3 = 0,$ so the index of $\mathbf{L}$ is $k = 3,$ and

$$\begin{aligned}\text{the number of } 3\times 3 \text{ blocks } &= r_2 - 2r_3 + r_4 = 1,\\ \text{the number of } 2\times 2 \text{ blocks } &= r_1 - 2r_2 + r_3 = 1,\\ \text{the number of } 1\times 1 \text{ blocks } &= r_0 - 2r_1 + r_2 = 1.\end{aligned}$$

Consequently, the Jordan form of $\mathbf{L}$ is

$$\mathbf{N} = \left(\begin{array}{ccc|cc|c} 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ \hline 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ \hline 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Notice that three Jordan blocks were found, and this agrees with the fact that $\dim N(\mathbf{L}) = 6 - rank(\mathbf{L}) = 3$. Determine $\mathbf{P}$ by following the procedure described in Example 7.7.3.

1. Since $rank\left(\mathbf{L}^2\right) = 1,$ any nonzero column from $\mathbf{L}^2$ will be a basis for $\mathcal{M}_2 = R\left(\mathbf{L}^2\right),$ so set $\mathbf{y}_1 = [\mathbf{L}^2]_{*1} = (6,-6,0,0,-6,-6)^T$.
2. To extend $\mathbf{y}_1$ to a basis for $\mathcal{M}_1 = R(\mathbf{L}) \cap N(\mathbf{L}),$ use

$$\mathbf{B} = [\,\mathbf{L}_{*1}\,|\,\mathbf{L}_{*2}\,|\,\mathbf{L}_{*3}\,] = \begin{pmatrix} 1 & 1 & -2\\ 3 & 1 & 5\\ -2 & -1 & 0\\ 2 & 1 & 0\\ -5 & -3 & -1\\ -3 & -2 & -1 \end{pmatrix} \Longrightarrow \mathbf{LB} = \begin{pmatrix} 6 & 3 & 3\\ -6 & -3 & -3\\ 0 & 0 & 0\\ 0 & 0 & 0\\ -6 & -3 & -3\\ -6 & -3 & -3 \end{pmatrix},$$

and determine a basis for $N(\mathbf{LB})$ to be $\left\{\mathbf{v}_1 = \begin{pmatrix} -1\\ 2\\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -1\\ 0\\ 2 \end{pmatrix}\right\}$. Reducing $[\,\mathbf{y}_1\,|\,\mathbf{B}\mathbf{v}_1\,|\,\mathbf{B}\mathbf{v}_2\,]$ to echelon form shows that its basic columns are in the first and third positions, so $\{\mathbf{y}_1, \mathbf{B}\mathbf{v}_2\}$ is a basis for $\mathcal{M}_1$ with

$$\mathcal{S}_2 = \left\{\begin{pmatrix} 6\\ -6\\ 0\\ 0\\ -6\\ -6 \end{pmatrix} = \mathbf{b}_1\right\} \quad \text{and} \quad \mathcal{S}_1 = \left\{\begin{pmatrix} -5\\ 7\\ 2\\ -2\\ 3\\ 1 \end{pmatrix} = \mathbf{b}_2\right\}.$$

3. Now extend $\mathcal{S}_2 \cup \mathcal{S}_1 = \{\mathbf{b}_1, \mathbf{b}_2\}$ to a basis for $\mathcal{M}_0 = N(\mathbf{L})$. This time, $\mathbf{B} = \mathbf{I}$, and a basis for $N(\mathbf{LB}) = N(\mathbf{L})$ can be computed to be

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ -4 \\ -1 \\ 3 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} -4 \\ 5 \\ 2 \\ 0 \\ 3 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -2 \\ 0 \\ 0 \\ 3 \end{pmatrix},$$

and $\{\mathbf{Bv}_1, \mathbf{Bv}_2, \mathbf{Bv}_3\} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. Reducing $[\,\mathbf{b}_1 \,|\, \mathbf{b}_2 \,|\, \mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \mathbf{v}_3\,]$ to echelon form reveals that its basic columns are in positions one, two, and three, so $\mathbf{v}_1$ is the needed extension vector. Therefore, the complete nested basis for $N(\mathbf{L})$ is

$$\mathbf{b}_1 = \begin{pmatrix} 6 \\ -6 \\ 0 \\ 0 \\ -6 \\ -6 \end{pmatrix} \in \mathcal{S}_2, \quad \mathbf{b}_2 = \begin{pmatrix} -5 \\ 7 \\ 2 \\ -2 \\ 3 \\ 1 \end{pmatrix} \in \mathcal{S}_1, \quad \text{and} \quad \mathbf{b}_3 = \begin{pmatrix} 2 \\ -4 \\ -1 \\ 3 \\ 0 \\ 0 \end{pmatrix} \in \mathcal{S}_0.$$

4. Complete the process by building a Jordan chain on top of each $\mathbf{b}_j \in \mathcal{S}_i$ by solving $\mathbf{L}^i\mathbf{x}_j = \mathbf{b}_j$ and by setting $\mathbf{J}_j = [\mathbf{L}^i\mathbf{x}_j \,|\, \cdots \,|\, \mathbf{Lx}_j \,|\, \mathbf{x}_j\,]$. Since $\mathbf{x}_1 = \mathbf{e}_1$ solves $\mathbf{L}^2\mathbf{x}_1 = \mathbf{b}_1$, we have $\mathbf{J}_1 = [\,\mathbf{L}^2\mathbf{e}_1 \,|\, \mathbf{Le}_1 \,|\, \mathbf{e}_1\,]$. Solving $\mathbf{Lx}_2 = \mathbf{b}_2$ yields $\mathbf{x}_2 = (-1, 0, 2, 0, 0, 0)^T$, so $\mathbf{J}_2 = [\,\mathbf{Lx}_2 \,|\, \mathbf{x}_2\,]$. Finally, $\mathbf{J}_3 = [\,\mathbf{b}_3\,]$. Putting these chains together produces

$$\mathbf{P} = [\,\mathbf{J}_1 \,|\, \mathbf{J}_2 \,|\, \mathbf{J}_3\,] = \begin{pmatrix} 6 & 1 & 1 & -5 & -1 & 2 \\ -6 & 3 & 0 & 7 & 0 & -4 \\ 0 & -2 & 0 & 2 & 2 & -1 \\ 0 & 2 & 0 & -2 & 0 & 3 \\ -6 & -5 & 0 & 3 & 0 & 0 \\ -6 & -3 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

It can be verified by direct multiplication that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$.

It's worthwhile to pay attention to how the results in this section translate into the language of direct sum decompositions of invariant subspaces as discussed in §4.9 (p. 259) and §5.9 (p. 383). For a linear nilpotent operator $\mathbf{L}$ of index k defined on a finite-dimensional vector space $\mathcal{V}$, statement (7.7.6) on p. 579 means that $\mathcal{V}$ can be decomposed as a direct sum $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2 \oplus \cdots \oplus \mathcal{V}_t$, where $\mathcal{V}_j = \text{span}(\mathcal{J}_{\mathbf{b}_j})$ is the space spanned by a Jordan chain emanating from the basis vector $\mathbf{b}_j \in N(\mathbf{L})$ and where $t = \dim N(\mathbf{L})$. Furthermore, each $\mathcal{V}_j$ is an

invariant subspace for $\mathbf{L}$, and the matrix representation of $\mathbf{L}$ with respect to the basis $\mathcal{J} = \mathcal{J}_{\mathbf{b}_1} \cup \mathcal{J}_{\mathbf{b}_2} \cup \cdots \cup \mathcal{J}_{\mathbf{b}_t}$ is

$$[\mathbf{L}]_{\mathcal{J}} = \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix} \quad \text{in which} \quad \mathbf{N}_j = \left[\mathbf{L}_{/\mathcal{V}_j}\right]_{\mathcal{J}_{\mathbf{b}_j}}. \tag{7.7.8}$$

Exercises for section 7.7

7.7.1. Can the index of an $n \times n$ nilpotent matrix ever exceed n?

7.7.2. Determine all possible Jordan forms $\mathbf{N}$ for a 4×4 nilpotent matrix.

7.7.3. Explain why the number of blocks of size $i \times i$ or larger in the Jordan form for a nilpotent matrix is given by $rank\left(\mathbf{L}^{i-1}\right) - rank\left(\mathbf{L}^{i}\right)$.

7.7.4. For a nilpotent matrix $\mathbf{L}$ of index k, let $\mathcal{M}_i = R\left(\mathbf{L}^i\right) \cap N\left(\mathbf{L}\right)$. Prove that $\mathcal{M}_i \subseteq \mathcal{M}_{i-1}$ for each $i = 0, 1, \ldots, k$.

7.7.5. Prove that $R\left(\mathbf{L}^{k-1}\right) \cap N\left(\mathbf{L}\right) = R\left(\mathbf{L}^{k-1}\right)$ for all nilpotent matrices $\mathbf{L}$ of index $k > 1$. In other words, prove $\mathcal{M}_{k-1} = R\left(\mathbf{L}^{k-1}\right)$.

7.7.6. Let $\mathbf{L}$ be a nilpotent matrix of index $k > 1$. Prove that if the columns of $\mathbf{B}$ are a basis for $R\left(\mathbf{L}^i\right)$ for $i \leq k-1$, and if $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s\}$ is a basis for $N\left(\mathbf{LB}\right)$, then $\{\mathbf{Bv}_1, \mathbf{Bv}_2, \ldots, \mathbf{Bv}_s\}$ is a basis for $\mathcal{M}_i$.

7.7.7. Find $\mathbf{P}$ and $\mathbf{N}$ such that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$ is in Jordan form, where

$$\mathbf{L} = \begin{pmatrix} 3 & 3 & 2 & 1 \\ -2 & -1 & -1 & -1 \\ 1 & -1 & 0 & 1 \\ -5 & -4 & -3 & -2 \end{pmatrix}.$$

7.7.8. Determine the Jordan form for the following 8×8 nilpotent matrix.

$$\mathbf{L} = \begin{pmatrix} 41 & 30 & 15 & 7 & 4 & 6 & 1 & 3 \\ -54 & -39 & -19 & -9 & -6 & -8 & -2 & -4 \\ 9 & 6 & 2 & 1 & 2 & 1 & 0 & 1 \\ -6 & -5 & -3 & -2 & 1 & -1 & 0 & 0 \\ -32 & -24 & -13 & -6 & -2 & -5 & -1 & -2 \\ -10 & -7 & -2 & 0 & -3 & 0 & 3 & -2 \\ -4 & -3 & -2 & -1 & 0 & -1 & -1 & 0 \\ 17 & 12 & 6 & 3 & 2 & 3 & 2 & 1 \end{pmatrix}.$$

7.7.9. Prove that if $\mathbf{N}$ is the Jordan form for a nilpotent matrix $\mathbf{L}$ as described in (7.7.5) and (7.7.6) on p. 579, then for any set of nonzero scalars $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_t\}$, the matrix $\mathbf{L}$ is similar to a matrix $\tilde{\mathbf{N}}$ of the form

$$\tilde{\mathbf{N}} = \begin{pmatrix} \epsilon_1\mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \epsilon_2\mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \epsilon_t\mathbf{N}_t \end{pmatrix}.$$

In other words, the 1's on the superdiagonal of the $\mathbf{N}_i$'s in (7.7.5) are artificial because any nonzero value can be forced onto the superdiagonal of any $\mathbf{N}_i$. What's important in the "Jordan structure" of $\mathbf{L}$ is the number and sizes of the nilpotent Jordan blocks (or chains) and not the values appearing on the superdiagonals of these blocks.

7.8 JORDAN FORM

The goal of this section is to do for general matrices $\mathbf{A} \in \mathcal{C}^{n\times n}$ what was done for nilpotent matrices in §7.7—reduce $\mathbf{A}$ by means of a similarity transformation to a block-diagonal matrix in which each block has a simple triangular form. The two major components for doing this are now in place—they are the core-nilpotent decomposition (p. 397) and the Jordan form for nilpotent matrices. All that remains is to connect these two ideas. To do so, it is convenient to adopt the following terminology.

Index of an Eigenvalue

The ***index of an eigenvalue*** λ for a matrix $\mathbf{A} \in \mathcal{C}^{n\times n}$ is defined to be the index of the matrix $(\mathbf{A} - \lambda\mathbf{I})$. In other words, from the characterizations of index given on p. 395, $index(\lambda)$ is the smallest positive integer k such that any one of the following statements is true.

- $rank\big((\mathbf{A}-\lambda\mathbf{I})^k\big) = rank\big((\mathbf{A}-\lambda\mathbf{I})^{k+1}\big)$.
- $R\big((\mathbf{A}-\lambda\mathbf{I})^k\big) = R\big((\mathbf{A}-\lambda\mathbf{I})^{k+1}\big)$.
- $N\big((\mathbf{A}-\lambda\mathbf{I})^k\big) = N\big((\mathbf{A}-\lambda\mathbf{I})^{k+1}\big)$.
- $R\big((\mathbf{A}-\lambda\mathbf{I})^k\big) \cap N\big((\mathbf{A}-\lambda\mathbf{I})^k\big) = \mathbf{0}$.
- $\mathcal{C}^n = R\big((\mathbf{A}-\lambda\mathbf{I})^k\big) \oplus N\big((\mathbf{A}-\lambda\mathbf{I})^k\big)$.

It is understood that $index(\mu) = 0$ if and only if $\mu \notin \sigma(\mathbf{A})$.

The Jordan form for $\mathbf{A} \in \mathcal{C}^{n\times n}$ is derived by digesting the distinct eigenvalues in $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ one at a time with a core-nilpotent decomposition as follows. If $index(\lambda_1) = k_1$, then there is a nonsingular matrix $\mathbf{X}_1$ such that

$$\mathbf{X}_1^{-1}(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{X}_1 = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{pmatrix}, \tag{7.8.1}$$

where $\mathbf{L}_1$ is nilpotent of index k_1 and $\mathbf{C}_1$ is nonsingular (it doesn't matter whether $\mathbf{C}_1$ or $\mathbf{L}_1$ is listed first, so, for the sake of convenience, the nilpotent block is listed first). We know from the results on nilpotent matrices (p. 579) that there is a nonsingular matrix $\mathbf{Y}_1$ such that

$$\mathbf{Y}_1^{-1}\mathbf{L}_1\mathbf{Y}_1 = \mathbf{N}(\lambda_1) = \begin{pmatrix} \mathbf{N}_1(\lambda_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2(\lambda_1) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_{t_1}(\lambda_1) \end{pmatrix}$$

is a block-diagonal matrix that is characterized by the following features.

▷ Every block in $\mathbf{N}(\lambda_1)$ has the form $\mathbf{N}_\star(\lambda_1) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}$.

▷ There are $t_1 = \dim N(\mathbf{L}_1) = \dim N(\mathbf{A} - \lambda_1\mathbf{I})$ such blocks in $\mathbf{N}(\lambda_1)$.

▷ The number of $i \times i$ blocks of the form $\mathbf{N}_\star(\lambda_1)$ contained in $\mathbf{N}(\lambda_1)$ is $\nu_i(\lambda_1) = rank\left(\mathbf{L}_1^{i-1}\right) - 2\,rank\left(\mathbf{L}_1^{i}\right) + rank\left(\mathbf{L}_1^{i+1}\right)$. But $\mathbf{C}_1$ in (7.8.1) is nonsingular, so $rank\left(\mathbf{L}_1^{p}\right) = rank\left((\mathbf{A} - \lambda_1\mathbf{I})^p\right) - rank\left(\mathbf{C}_1\right)$, and thus the number of $i \times i$ blocks $\mathbf{N}_\star(\lambda_1)$ contained in $\mathbf{N}(\lambda_1)$ can be expressed as

$$\nu_i(\lambda_1) = r_{i-1}(\lambda_1) - 2r_i(\lambda_1) + r_{i+1}(\lambda_1), \quad \text{where} \quad r_i(\lambda_1) = rank\left((\mathbf{A} - \lambda_1\mathbf{I})^i\right).$$

Now, $\mathbf{Q}_1 = \mathbf{X}_1 \begin{pmatrix} \mathbf{Y}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ is nonsingular, and $\mathbf{Q}_1^{-1}(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{Q}_1 = \begin{pmatrix} \mathbf{N}(\lambda_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{pmatrix}$ or, equivalently,

$$\mathbf{Q}_1^{-1}\mathbf{A}\mathbf{Q}_1 = \begin{pmatrix} \mathbf{N}(\lambda_1) + \lambda_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 + \lambda_1\mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{J}(\lambda_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1 \end{pmatrix}. \tag{7.8.2}$$

The upper-left-hand segment $\mathbf{J}(\lambda_1) = \mathbf{N}(\lambda_1) + \lambda_1\mathbf{I}$ has the block-diagonal form

$$\mathbf{J}(\lambda_1) = \begin{pmatrix} \mathbf{J}_1(\lambda_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2(\lambda_1) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_{t_1}(\lambda_1) \end{pmatrix} \quad \text{with} \quad \mathbf{J}_\star(\lambda_1) = \mathbf{N}_\star(\lambda_1) + \lambda_1\mathbf{I}.$$

The matrix $\mathbf{J}(\lambda_1)$ is called the ***Jordan segment*** associated with the eigenvalue λ_1, and the individual blocks $\mathbf{J}_\star(\lambda_1)$ contained in $\mathbf{J}(\lambda_1)$ are called ***Jordan blocks*** associated with the eigenvalue λ_1. The structure of the Jordan segment $\mathbf{J}(\lambda_1)$ is inherited from Jordan structure of the associated nilpotent matrix $\mathbf{L}_1$.

▷ Each Jordan block looks like $\mathbf{J}_\star(\lambda_1) = \mathbf{N}_\star(\lambda_1) + \lambda_1\mathbf{I} = \begin{pmatrix} \lambda_1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_1 \end{pmatrix}$.

▷ There are $t_1 = \dim N(\mathbf{A} - \lambda_1\mathbf{I})$ such Jordan blocks in the segment $\mathbf{J}(\lambda_1)$.

▷ The number of $i \times i$ Jordan blocks $\mathbf{J}_\star(\lambda_1)$ contained in $\mathbf{J}(\lambda_1)$ is

$$\nu_i(\lambda_1) = r_{i-1}(\lambda_1) - 2r_i(\lambda_1) + r_{i+1}(\lambda_1), \quad \text{where} \quad r_i(\lambda_1) = rank\left((\mathbf{A} - \lambda_1\mathbf{I})^i\right).$$

Since the distinct eigenvalues of $\mathbf{A}$ are $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$, the distinct eigenvalues of $\mathbf{A} - \lambda_1\mathbf{I}$ are

$$\sigma(\mathbf{A} - \lambda_1\mathbf{I}) = \{0,\, (\lambda_2 - \lambda_1),\, (\lambda_3 - \lambda_1), \ldots, (\lambda_s - \lambda_1)\}.$$

Couple this with the fact that the only eigenvalue for the nilpotent matrix $\mathbf{L}_1$ in (7.8.1) is zero to conclude that

$$\sigma(\mathbf{C}_1) = \{(\lambda_2 - \lambda_1),\ (\lambda_3 - \lambda_1), \ldots, (\lambda_s - \lambda_1)\}.$$

Therefore, the spectrum of $\mathbf{A}_1 = \mathbf{C}_1 + \lambda_1\mathbf{I}$ in (7.8.2) is $\sigma(\mathbf{A}_1) = \{\lambda_2, \lambda_3, \ldots, \lambda_s\}$. This means that the core-nilpotent decomposition process described above can be repeated on $\mathbf{A}_1 - \lambda_2\mathbf{I}$ to produce a nonsingular matrix $\mathbf{Q}_2$ such that

$$\mathbf{Q}_2^{-1}\mathbf{A}_1\mathbf{Q}_2 = \begin{pmatrix} \mathbf{J}(\lambda_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}, \quad \text{where} \quad \sigma(\mathbf{A}_2) = \{\lambda_3,\ \lambda_4, \ldots, \lambda_s\}, \tag{7.8.3}$$

and where $\mathbf{J}(\lambda_2) = \text{diag}\,(\mathbf{J}_1(\lambda_2), \mathbf{J}_2(\lambda_2), \ldots, \mathbf{J}_{t_2}(\lambda_2))$ is a Jordan segment composed of Jordan blocks $\mathbf{J}_\star(\lambda_2)$ with the following characteristics.

- ▷ Each Jordan block in $\mathbf{J}(\lambda_2)$ has the form $\mathbf{J}_\star(\lambda_2) = \begin{pmatrix} \lambda_2 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_2 \end{pmatrix}$.
- ▷ There are $t_2 = \dim N(\mathbf{A} - \lambda_2\mathbf{I})$ Jordan blocks in segment $\mathbf{J}(\lambda_2)$.
- ▷ The number of $i \times i$ Jordan blocks in segment $\mathbf{J}(\lambda_2)$ is $\nu_i(\lambda_2) = r_{i-1}(\lambda_2) - 2r_i(\lambda_2) + r_{i+1}(\lambda_2)$, where $r_i(\lambda_2) = rank\left((\mathbf{A} - \lambda_2\mathbf{I})^i\right)$.

If we set $\mathbf{P}_2 = \mathbf{Q}_1\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}$, then $\mathbf{P}_2$ is a nonsingular matrix such that

$$\mathbf{P}_2^{-1}\mathbf{A}\mathbf{P}_2 = \begin{pmatrix} \mathbf{J}(\lambda_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(\lambda_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 \end{pmatrix}, \quad \text{where} \quad \sigma(\mathbf{A}_2) = \{\lambda_3,\ \lambda_4, \ldots, \lambda_s\}.$$

Repeating this process until all eigenvalues have been depleted results in a nonsingular matrix $\mathbf{P}_s$ such that $\mathbf{P}_s^{-1}\mathbf{A}\mathbf{P}_s = \mathbf{J} = \text{diag}\,(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s))$ in which each $\mathbf{J}(\lambda_j)$ is a Jordan segment containing $t_j = \dim N(\mathbf{A} - \lambda_j\mathbf{I})$ Jordan blocks. The matrix $\mathbf{J}$ is called the ***Jordan form***[79] for $\mathbf{A}$ (some texts refer to $\mathbf{J}$ as the Jordan *canonical* form or the Jordan *normal* form). The ***Jordan structure*** of $\mathbf{A}$ is defined to be the number of Jordan segments in $\mathbf{J}$ along with the number and sizes of the Jordan blocks within each segment. The proof of uniqueness of the Jordan form for a nilpotent matrix (p. 580) can be extended to all $\mathbf{A} \in \mathcal{C}^{n\times n}$. In other words, the Jordan structure of a matrix is uniquely determined by its entries. Below is a formal summary of these developments.

[79] Marie Ennemond Camille Jordan (1838–1922) discussed this idea (not over the complex numbers but over a finite field) in 1870 in *Traité des substitutions et des équations algebraique* that earned him the Poncelet Prize of the Académie des Science. But Jordan may not have been the first to develop these concepts. It has been reported that the German mathematician Karl Theodor Wilhelm Weierstrass (1815–1897) had previously formulated results along these lines. However, Weierstrass did not publish his ideas because he was fanatical about rigor, and he would not release his work until he was sure it was on a firm mathematical foundation. Weierstrass once said that "a mathematician who is not also something of a poet will never be a perfect mathematician."

Jordan Form

For every $\mathbf{A} \in \mathcal{C}^{n\times n}$ with distinct eigenvalues $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$, there is a nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} = \begin{pmatrix} \mathbf{J}(\lambda_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(\lambda_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}(\lambda_s) \end{pmatrix}. \tag{7.8.4}$$

- $\mathbf{J}$ has one *Jordan segment* $\mathbf{J}(\lambda_j)$ for each eigenvalue $\lambda_j \in \sigma(\mathbf{A})$.
- Each segment $\mathbf{J}(\lambda_j)$ is made up of $t_j = \dim N(\mathbf{A} - \lambda_j\mathbf{I})$ *Jordan blocks* $\mathbf{J}_\star(\lambda_j)$ as described below.

$$\mathbf{J}(\lambda_j) = \begin{pmatrix} \mathbf{J}_1(\lambda_j) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2(\lambda_j) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_{t_j}(\lambda_j) \end{pmatrix} \text{ with } \mathbf{J}_\star(\lambda_j) = \begin{pmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{pmatrix}.$$

- The largest Jordan block in $\mathbf{J}(\lambda_j)$ is $k_j \times k_j$, where $k_j = index(\lambda_j)$.
- The number of $i \times i$ Jordan blocks in $\mathbf{J}(\lambda_j)$ is given by

$$\nu_i(\lambda_j) = r_{i-1}(\lambda_j) - 2r_i(\lambda_j) + r_{i+1}(\lambda_j) \text{ with } r_i(\lambda_j) = rank\left((\mathbf{A} - \lambda_j\mathbf{I})^i\right).$$

- Matrix $\mathbf{J}$ in (7.8.4) is called the ***Jordan form*** for $\mathbf{A}$. The *structure* of this form is unique in the sense that the number of Jordan segments in $\mathbf{J}$ as well as the number and sizes of the Jordan blocks in each segment is uniquely determined by the entries in $\mathbf{A}$. Furthermore, every matrix similar to $\mathbf{A}$ has the same Jordan structure—i.e., $\mathbf{A}, \mathbf{B} \in \mathcal{C}^{n\times n}$ are similar if and only if $\mathbf{A}$ and $\mathbf{B}$ have the same Jordan structure. The matrix $\mathbf{P}$ is not unique—see p. 594.

Example 7.8.1

Problem: Find the Jordan form for $\mathbf{A} = \begin{pmatrix} 5 & 4 & 0 & 0 & 4 & 3 \\ 2 & 3 & 1 & 0 & 5 & 1 \\ 0 & -1 & 2 & 0 & 2 & 0 \\ -8 & -8 & -1 & 2 & -12 & -7 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ -8 & -8 & -1 & 0 & -9 & -5 \end{pmatrix}$.

Solution: Computing the eigenvalues (which is the hardest part) reveals two distinct eigenvalues $\lambda_1 = 2$ and $\lambda_2 = -1$, so there are two Jordan segments in the Jordan form $\mathbf{J} = \begin{pmatrix} \mathbf{J}(2) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(-1) \end{pmatrix}$. Computing ranks $r_i(2) = rank\left((\mathbf{A} - 2\mathbf{I})^i\right)$ and $r_i(-1) = rank\left((\mathbf{A} + \mathbf{I})^i\right)$ until $r_k(\star) = r_{k+1}(\star)$ yields

$$\begin{aligned} r_1(2) &= rank\,(\mathbf{A} - 2\mathbf{I}) \quad = 4, & r_1(-1) &= rank\,(\mathbf{A} + \mathbf{I}) \quad = 4, \\ r_2(2) &= rank\left((\mathbf{A} - 2\mathbf{I})^2\right) = 3, & r_2(-1) &= rank\left((\mathbf{A} + \mathbf{I})^2\right) = 4, \\ r_3(2) &= rank\left((\mathbf{A} - 2\mathbf{I})^3\right) = 2, \\ r_4(2) &= rank\left((\mathbf{A} - 2\mathbf{I})^4\right) = 2, \end{aligned}$$

so $k_1 = index(\lambda_1) = 3$ and $k_2 = index(\lambda_2) = 1$. This tells us that the largest Jordan block in $\mathbf{J}(2)$ is 3×3, while the largest Jordan block in $\mathbf{J}(-1)$ is 1×1 so that $\mathbf{J}(-1)$ is a diagonal matrix (the associated eigenvalue is ***semisimple*** whenever this happens). Furthermore,

$$\begin{aligned} \nu_3(2) &= r_2(2) - 2r_3(2) + r_4(2) = 1 && \Longrightarrow \text{ one } 3 \times 3 \text{ block in } \mathbf{J}(2), \\ \nu_2(2) &= r_1(2) - 2r_2(2) + r_3(2) = 0 && \Longrightarrow \text{ no } 2 \times 2 \text{ blocks in } \mathbf{J}(2), \\ \nu_1(2) &= r_0(2) - 2r_1(2) + r_2(2) = 1 && \Longrightarrow \text{ one } 1 \times 1 \text{ block in } \mathbf{J}(2), \\ \nu_1(-1) &= r_0(-1) - 2r_1(-1) + r_2(-1) = 2 && \Longrightarrow \text{ two } 1 \times 1 \text{ blocks in } \mathbf{J}(-1). \end{aligned}$$

Therefore, $\mathbf{J}(2) = \left(\begin{array}{ccc|c} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ \hline 0 & 0 & 0 & 2 \end{array}\right)$ and $\mathbf{J}(-1) = \left(\begin{array}{c|c} -1 & 0 \\ \hline 0 & -1 \end{array}\right)$ so that

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}(2) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(-1) \end{pmatrix} = \left(\begin{array}{ccc|c||cc} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 0 & 0 \\ \hline\hline 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{array}\right).$$

The above example suggests that determining the Jordan form for $\mathbf{A}_{n\times n}$ is straightforward, and perhaps even easy. In theory, it is—just find $\sigma(\mathbf{A})$, and calculate some ranks. But, in practice, both of these tasks can be difficult. To begin with, the rank of a matrix is a discontinuous function of its entries, and rank computed with floating-point arithmetic can vary with the algorithm used and is often different than rank computed with exact arithmetic (recall Exercise 2.2.4).

Furthermore, computing higher-index eigenvalues with floating-point arithmetic is fraught with peril. To see why, consider the matrix

$$\mathbf{L}(\epsilon) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \epsilon & & & 0 \end{pmatrix}_{n\times n} \qquad \text{whose characteristic equation is} \quad \lambda^n - \epsilon = 0.$$

For $\epsilon = 0$, zero is the only eigenvalue (and it has index n), but for all $\epsilon > 0$, there are n distinct eigenvalues given by $\epsilon^{1/n}\mathrm{e}^{2k\pi \mathrm{i}/n}$ for $k = 0, 1, \ldots, n-1$. For example, if $n = 32$, and if ϵ changes from 0 to 10^{-16}, then the eigenvalues of $\mathbf{L}(\epsilon)$ change in magnitude from 0 to $10^{-1/2} \approx .316$, which is substantial for such a small perturbation. Sensitivities of this kind present significant problems for floating-point algorithms. In addition to showing that high-index eigenvalues are sensitive to small perturbations, this example also shows that the Jordan structure is highly discontinuous. $\mathbf{L}(0)$ is in Jordan form, and there is just one Jordan block of size n, but for all $\epsilon \neq 0$, the Jordan form of $\mathbf{L}(\epsilon)$ is a diagonal matrix—i.e., there are n Jordan blocks of size 1×1. Lest you think that this example somehow is an isolated case, recall from Example 7.3.6 (p. 532) that *every* matrix in $\mathcal{C}^{n\times n}$ is arbitrarily close to a diagonalizable matrix.

All of the above observations make it clear that it's hard to have faith in a Jordan form that has been computed with floating-point arithmetic. Consequently, numerical computation of Jordan forms is generally avoided.

Example 7.8.2

The Jordan form of $\mathbf{A}$ conveys complete information about the eigenvalues of $\mathbf{A}$. For example, if the Jordan form for $\mathbf{A}$ is

$$\mathbf{J} = \left(\begin{array}{ccc|cc|cc|c|c} 4 & 1 & 0 & & & & & & \\ & 4 & 1 & & & & & & \\ & & 4 & & & & & & \\ \hline & & & 4 & 1 & & & & \\ & & & 0 & 4 & & & & \\ \hline & & & & & 3 & 1 & & \\ & & & & & 0 & 3 & & \\ \hline & & & & & & & 2 & \\ \hline & & & & & & & & 2 \end{array}\right),$$

then we know that

- $\mathbf{A}_{9\times 9}$ has three distinct eigenvalues, namely $\sigma(\mathbf{A}) = \{4, 3, 2\}$;
- $alg\ mult(4) = 5$, $alg\ mult(3) = 2$, and $alg\ mult(2) = 2$;
- $geo\ mult(4) = 2$, $geo\ mult(3) = 1$, and $geo\ mult(2) = 2$;

- ▷ $index(4) = 3$, $index(3) = 2$, and $index(2) = 1$;
- ▷ $\lambda = 2$ is a semisimple eigenvalue, so, while $\mathbf{A}$ is not diagonalizable, part of it is; i.e., the restriction $\mathbf{A}_{/N(\mathbf{A}-2\mathbf{I})}$ is a diagonalizable linear operator.

Of course, if both $\mathbf{P}$ and $\mathbf{J}$ are known, then $\mathbf{A}$ can be completely reconstructed from (7.8.4), but the point being made here is that only $\mathbf{J}$ is needed to reveal the eigenstructure along with the other similarity invariants of $\mathbf{A}$.

Now that the structure of the Jordan form $\mathbf{J}$ is known, the structure of the similarity transformation $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is easily revealed. Focus on a single $p \times p$ Jordan block $\mathbf{J}_\star(\lambda)$ contained in the Jordan segment $\mathbf{J}(\lambda)$ associated with an eigenvalue λ, and let $\mathbf{P}_\star = [\,\mathbf{x}_1\,\mathbf{x}_2\,\cdots\,\mathbf{x}_p\,]$ be the portion of $\mathbf{P} = [\,\cdots\,|\,\mathbf{P}_\star\,|\,\cdots\,]$ that corresponds to the position of $\mathbf{J}_\star(\lambda)$ in $\mathbf{J}$. Notice that $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{J}$ implies $\mathbf{A}\mathbf{P}_\star = \mathbf{P}_\star\mathbf{J}_\star(\lambda)$ or, equivalently,

$$\mathbf{A}[\,\mathbf{x}_1\,\mathbf{x}_2\,\cdots\,\mathbf{x}_p\,] = [\,\mathbf{x}_1\,\mathbf{x}_2\,\cdots\,\mathbf{x}_p\,]\begin{pmatrix}\lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda\end{pmatrix}_{p\times p},$$

so equating columns on both sides of this equation produces

$$\begin{array}{lcll}
\mathbf{A}\mathbf{x}_1 = \lambda\mathbf{x}_1 & \Longrightarrow & \mathbf{x}_1 \text{ is an eigenvector} & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_1 = \mathbf{0}, \\
\mathbf{A}\mathbf{x}_2 = \mathbf{x}_1 + \lambda\mathbf{x}_2 & \Longrightarrow & (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_2 = \mathbf{x}_1 & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})^2\,\mathbf{x}_2 = \mathbf{0}, \\
\mathbf{A}\mathbf{x}_3 = \mathbf{x}_2 + \lambda\mathbf{x}_3 & \Longrightarrow & (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_3 = \mathbf{x}_2 & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})^3\,\mathbf{x}_3 = \mathbf{0}, \\
\quad\vdots & & \quad\vdots & \quad\vdots \\
\mathbf{A}\mathbf{x}_p = \mathbf{x}_{p-1} + \lambda\mathbf{x}_p & \Longrightarrow & (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_p = \mathbf{x}_{p-1} & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})^p\,\mathbf{x}_p = \mathbf{0}.
\end{array}$$

In other words, the first column $\mathbf{x}_1$ in $\mathbf{P}_\star$ is a eigenvector for $\mathbf{A}$ associated with λ. We already knew there had to be exactly one independent eigenvector for each Jordan block because there are $t = \dim N(\mathbf{A}-\lambda\mathbf{I})$ Jordan blocks $\mathbf{J}_\star(\lambda)$, but now we know precisely where these eigenvectors are located in $\mathbf{P}$.

Vectors $\mathbf{x}$ such that $\mathbf{x} \in N\big((\mathbf{A}-\lambda\mathbf{I})^g\big)$ but $\mathbf{x} \notin N\big((\mathbf{A}-\lambda\mathbf{I})^{g-1}\big)$ are called ***generalized eigenvectors of order g*** associated with λ. So $\mathbf{P}_\star$ consists of an eigenvector followed by generalized eigenvectors of increasing order. Moreover, the columns of $\mathbf{P}_\star$ form a ***Jordan chain*** analogous to (7.7.2) on p. 576; i.e., $\mathbf{x}_i = (\mathbf{A}-\lambda\mathbf{I})^{p-i}\,\mathbf{x}_p$ implies $\mathbf{P}_\star$ must have the form

$$\mathbf{P}_\star = \big[\,(\mathbf{A}-\lambda\mathbf{I})^{p-1}\,\mathbf{x}_p\,|\,(\mathbf{A}-\lambda\mathbf{I})^{p-2}\,\mathbf{x}_p\,|\,\cdots\,|\,(\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_p\,|\,\mathbf{x}_p\,\big]. \tag{7.8.5}$$

A complete set of Jordan chains associated with a given eigenvalue λ is determined in exactly the same way as Jordan chains for nilpotent matrices are

determined except that the nested subspaces $\mathcal{M}_i$ defined in (7.7.1) on p. 575 are redefined to be

$$\mathcal{M}_i = R\big((\mathbf{A}-\lambda\mathbf{I})^i\big) \cap N(\mathbf{A}-\lambda\mathbf{I}) \quad \text{for } i = 0, 1, \ldots, k, \tag{7.8.6}$$

where $k = index(\lambda)$. Just as in the case of nilpotent matrices, it follows that $\mathbf{0} = \mathcal{M}_k \subseteq \mathcal{M}_{k-1} \subseteq \cdots \subseteq \mathcal{M}_0 = N(\mathbf{A}-\lambda\mathbf{I})$ (see Exercise 7.8.8). Since $(\mathbf{A}-\lambda\mathbf{I})_{/\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^k)}$ is a nilpotent linear operator of index k (Example 5.10.4, p. 399), it can be argued that the same process used to build Jordan chains for nilpotent matrices can be used to build Jordan chains for a general eigenvalue λ. Below is a summary of the process adapted to the general case.

Constructing Jordan Chains

For each $\lambda \in \sigma(\mathbf{A}_{n\times n})$, set $\mathcal{M}_i = R\big((\mathbf{A}-\lambda\mathbf{I})^i\big) \cap N(\mathbf{A}-\lambda\mathbf{I})$ for $i = k-1, k-2, \ldots, 0$, where $k = index(\lambda)$.

- Construct a basis $\mathcal{B}$ for $N(\mathbf{A}-\lambda\mathbf{I})$.
  - ▷ Starting with any basis $\mathcal{S}_{k-1}$ for $\mathcal{M}_{k-1}$ (see p. 211), sequentially extend $\mathcal{S}_{k-1}$ with sets $\mathcal{S}_{k-2}$, $\mathcal{S}_{k-3}$, $\ldots$, $\mathcal{S}_0$ such that

$$\begin{array}{ll} \mathcal{S}_{k-1} & \text{is a basis for } \mathcal{M}_{k-1}, \\ \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} & \text{is a basis for } \mathcal{M}_{k-2}, \\ \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \mathcal{S}_{k-3} & \text{is a basis for } \mathcal{M}_{k-3}, \end{array}$$

    etc., until a basis $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$ for $\mathcal{M}_0 = N(\mathbf{A}-\lambda\mathbf{I})$ is obtained (see Example 7.7.3 on p. 582).
- Build a Jordan chain on top of each eigenvector $\mathbf{b}_\star \in \mathcal{B}$.
  - ▷ For each eigenvector $\mathbf{b}_\star \in \mathcal{S}_i$, solve $(\mathbf{A}-\lambda\mathbf{I})^i \mathbf{x}_\star = \mathbf{b}_\star$ (a necessarily consistent system) for $\mathbf{x}_\star$, and construct a Jordan chain on top of $\mathbf{b}_\star$ by setting

$$\mathbf{P}_\star = \Big[\, (\mathbf{A}-\lambda\mathbf{I})^i \mathbf{x}_\star \,\Big|\, (\mathbf{A}-\lambda\mathbf{I})^{i-1} \mathbf{x}_\star \,\Big|\, \cdots \,\Big|\, (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_\star \,\Big|\, \mathbf{x}_\star \,\Big]_{(i+1)\times n}.$$

  - ▷ Each such $\mathbf{P}_\star$ corresponds to one Jordan block $\mathbf{J}_\star(\lambda)$ in the Jordan segment $\mathbf{J}(\lambda)$ associated with λ.
  - ▷ The first column in $\mathbf{P}_\star$ is an eigenvector, and subsequent columns are generalized eigenvectors of increasing order.
- If all such $\mathbf{P}_\star$'s for a given $\lambda_j \in \sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ are put in a matrix $\mathbf{P}_j$, and if $\mathbf{P} = \big[\mathbf{P}_1 \,|\, \mathbf{P}_2 \,|\, \cdots \,|\, \mathbf{P}_s\big]$, then $\mathbf{P}$ is a nonsingular matrix such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} = \operatorname{diag}\big(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s)\big)$ is in Jordan form as described on p. 590.

Example 7.8.3

Caution! Not every basis for $N(\mathbf{A}-\lambda\mathbf{I})$ can be used to build Jordan chains associated with an eigenvalue $\lambda \in \sigma(\mathbf{A})$. For example, the Jordan form of

$$\mathbf{A} = \begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & -2 \\ -4 & 0 & -1 \end{pmatrix} \quad \text{is} \quad \mathbf{J} = \left(\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right)$$

because $\sigma(\mathbf{A}) = \{1\}$ and $index(1) = 2$. Consequently, if $\mathbf{P} = [\,\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \mathbf{x}_3\,]$ is a nonsingular matrix such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$, then the derivation beginning on p. 593 leading to (7.8.5) shows that $\{\mathbf{x}_1, \mathbf{x}_2\}$ must be a Jordan chain such that $(\mathbf{A}-\mathbf{I})\mathbf{x}_1 = \mathbf{0}$ and $(\mathbf{A}-\mathbf{I})\mathbf{x}_2 = \mathbf{x}_1$, while $\mathbf{x}_3$ is another eigenvector (not dependent on $\mathbf{x}_1$). Suppose we try to build the Jordan chains in $\mathbf{P}$ by starting with the eigenvectors

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{7.8.7}$$

obtained by solving $(\mathbf{A}-\mathbf{I})\mathbf{x} = \mathbf{0}$ with straightforward Gauss–Jordan elimination. This naive approach fails because $\mathbf{x}_1 \notin R(\mathbf{A}-\mathbf{I})$ means $(\mathbf{A}-\mathbf{I})\mathbf{x}_2 = \mathbf{x}_1$ is an inconsistent system, so $\mathbf{x}_2$ cannot be determined. Similarly, $\mathbf{x}_3 \notin R(\mathbf{A}-\mathbf{I})$ insures that the same difficulty occurs if $\mathbf{x}_3$ is used in place of $\mathbf{x}_1$. In other words, even though the vectors in (7.8.7) constitute an otherwise perfectly good basis for $N(\mathbf{A}-\mathbf{I})$, they are not suitable for building Jordan chains. You are asked in Exercise 7.8.2 to find the correct basis for $N(\mathbf{A}-\mathbf{I})$ that will yield the Jordan chains that constitute $\mathbf{P}$.

Example 7.8.4

Problem: What do the results concerning the Jordan form for $\mathbf{A} \in \mathcal{C}^{n\times n}$ say about the decomposition of $\mathcal{C}^n$ into invariant subspaces?

Solution: Consider $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} = \text{diag}\,(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s))$, where the $\mathbf{J}(\lambda_j)$'s are the Jordan segments and $\mathbf{P} = \left[\,\mathbf{P}_1 \,|\, \mathbf{P}_2 \,|\, \cdots \,|\, \mathbf{P}_s\,\right]$ is a matrix of Jordan chains as described in (7.8.5) and on p. 594. If $\mathbf{A}$ is considered as a linear operator on $\mathcal{C}^n$, and if the set of columns in $\mathbf{P}_i$ is denoted by $\mathcal{J}_i$, then the results in §4.9 (p. 259) concerning invariant subspaces together with those in §5.9 (p. 383) about direct sum decompositions guarantee that each $R(\mathbf{P}_i)$ is an invariant subspace for $\mathbf{A}$ such that

$$\mathcal{C}^n = R(\mathbf{P}_1) \oplus R(\mathbf{P}_2) \oplus \cdots \oplus R(\mathbf{P}_s) \quad \text{and} \quad \mathbf{J}(\lambda_i) = \left[\mathbf{A}_{/R(\mathbf{P}_i)}\right]_{\mathcal{J}_i}.$$

More can be said. If $alg\ mult(\lambda_i) = m_i$ and $index(\lambda_i) = k_i$, then $\mathcal{J}_i$ is a linearly independent set containing m_i vectors, and the discussion surrounding

(7.8.5) insures that each column in $\mathcal{J}_i$ belongs to $N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$. This coupled with the fact that $\dim N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big) = m_i$ (Exercise 7.8.7) implies that $\mathcal{J}_i$ is a basis for

$$R(\mathbf{P}_i) = N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big).$$

Consequently, each $N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$ is an invariant subspace for $\mathbf{A}$ such that

$$\mathcal{C}^n = N\big((\mathbf{A}-\lambda_1\mathbf{I})^{k_1}\big) \oplus N\big((\mathbf{A}-\lambda_2\mathbf{I})^{k_2}\big) \oplus \cdots \oplus N\big((\mathbf{A}-\lambda_s\mathbf{I})^{k_s}\big)$$

and

$$\mathbf{J}(\lambda_i) = \left[\mathbf{A}_{\Big/\mathcal{N}\left((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\right)}\right]_{\mathcal{J}_i}.$$

Of course, an even finer direct sum decomposition of $\mathcal{C}^n$ is possible because each Jordan segment is itself a block-diagonal matrix containing the individual Jordan blocks—the details are left to the interested reader.

Exercises for section 7.8

7.8.1. Find the Jordan form of the following matrix whose distinct eigenvalues are $\sigma(\mathbf{A}) = \{0, -1, 1\}$. Don't be frightened by the size of $\mathbf{A}$.

$$\mathbf{A} = \begin{pmatrix} -4 & -5 & -3 & 1 & -2 & 0 & 1 & -2 \\ 4 & 7 & 3 & -1 & 3 & 0 & -1 & 2 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 2 & -4 & 2 & 0 & -3 & 1 \\ -8 & -14 & -5 & 1 & -6 & 0 & 1 & -4 \\ 4 & 7 & 4 & -3 & 3 & -1 & -3 & 4 \\ 2 & -2 & -2 & 5 & -3 & 0 & 4 & -1 \\ 6 & 7 & 3 & 0 & 2 & 0 & 0 & 3 \end{pmatrix}.$$

7.8.2. For the matrix $\mathbf{A} = \begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & -2 \\ -4 & 0 & -1 \end{pmatrix}$ that was used in Example 7.8.3, use the technique described on p. 594 to construct a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is in Jordan form.

7.8.3. Explain why $index(\lambda) \leq alg\ mult(\lambda)$ for each $\lambda \in \sigma(\mathbf{A}_{n\times n})$.

7.8.4. Explain why $index(\lambda) = 1$ if and only if λ is a semisimple eigenvalue.

7.8.5. Prove that every square matrix is similar to its transpose. **Hint:** Consider the "reversal matrix" $\mathbf{R} = \begin{pmatrix} & & & 1 \\ & & \iddots & \\ & 1 & & \\ 1 & & & \end{pmatrix}$ obtained by reversing the order of the rows (or the columns) of the identity matrix $\mathbf{I}$.

7.8.6. Cayley–Hamilton Revisited. Prove the the Cayley–Hamilton theorem (pp. 509, 532) by means of the Jordan form; i.e., prove that every $\mathbf{A} \in \mathcal{C}^{n\times n}$ satisfies its own characteristic equation.

7.8.7. Prove that if λ is an eigenvalue of $\mathbf{A} \in \mathcal{C}^{n\times n}$ such that $index(\lambda) = k$ and $alg\ mult_{\mathbf{A}}(\lambda) = m$, then $\dim N\big((\mathbf{A} - \lambda\mathbf{I})^k\big) = m$. Is it also true that $\dim N\big((\mathbf{A} - \lambda\mathbf{I})^m\big) = m$?

7.8.8. Let λ_j be an eigenvalue of $\mathbf{A}$ with $index(\lambda_j) = k_j$. Prove that if $\mathcal{M}_i(\lambda_j) = R\big((\mathbf{A} - \lambda_j\mathbf{I})^i\big) \cap N(\mathbf{A} - \lambda_j\mathbf{I})$, then

$$\mathbf{0} = \mathcal{M}_{k_j}(\lambda_j) \subseteq \mathcal{M}_{k_{j-1}}(\lambda_j) \subseteq \cdots \subseteq \mathcal{M}_0(\lambda_j) = N(\mathbf{A} - \lambda_j\mathbf{I}).$$

7.8.9. Explain why $(\mathbf{A}-\lambda_j\mathbf{I})^i\mathbf{x} = \mathbf{b}(\lambda_j)$ must be a consistent system whenever $\lambda_j \in \sigma(\mathbf{A})$ and $\mathbf{b}(\lambda_j) \in \mathcal{S}_i(\lambda_j)$, where $\mathbf{b}(\lambda_j)$ and $\mathcal{S}_i(\lambda_j)$ are as defined on p. 594.

7.8.10. Does the result of Exercise 7.7.5 extend to nonnilpotent matrices? That is, if $\lambda \in \sigma(\mathbf{A})$ with $index(\lambda) = k > 1$, is $\mathcal{M}_{k-1} = R\big((\mathbf{A} - \lambda\mathbf{I})^{k-1}\big)$?

7.8.11. As defined in Exercise 5.8.15 (p. 380) and mentioned in Exercise 7.6.10 (p. 573), the ***Kronecker*** [80] ***product*** (sometimes called *tensor product*,

[80] Leopold Kronecker (1823–1891) was born in Liegnitz, Prussia (now Legnica, Poland), to a wealthy business family that hired private tutors to educate him until he enrolled at Gymnasium at Liegnitz where his mathematical talents were recognized by Eduard Kummer (1810–1893), who became his mentor and lifelong colleague. Kronecker went to Berlin University in 1841 to earn his doctorate, writing on algebraic number theory, under the supervision of Dirichlet (p. 563). Rather than pursuing a standard academic career, Kronecker returned to Liegnitz to marry his cousin and become involved in his uncle's banking business. But he never lost his enjoyment of mathematics. After estate and business interests were left to others in 1855, Kronecker joined Kummer in Berlin who had just arrived to occupy the position vacated by Dirichlet's move to Göttingen. Kronecker didn't need a salary, so he didn't teach or hold a university appointment, but his research activities led to his election to the Berlin Academy in 1860. He declined the offer of the mathematics chair in Göttingen in 1868, but he eventually accepted the chair in Berlin that was vacated upon Kummer's retirement in 1883. Kronecker held the unconventional view that mathematics should be reduced to arguments that involve only integers and a finite number of steps, and he questioned the validity of nonconstructive existence proofs, so he didn't like the use of irrational or transcendental numbers. Kronecker became famous for saying that "God created the integers, all else is the work of man." Kronecker's significant influence led to animosity with people of differing philosophies such as Georg Cantor (1845–1918), whose publications Kronecker tried to block. Kronecker's small physical size was another sensitive issue. After Hermann Schwarz (p. 271), who was Kummer's son-in-law and a student of Weierstrass (p. 589), tried to make a joke involving Weierstrass's large physique by stating that "he who does not honor the Smaller, is not worthy of the Greater," Kronecker had no further dealings with Schwarz.

or *direct product*) of $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{p\times q}$ is the $mp \times nq$ matrix

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}.$$

(a) Assuming conformability, establish the following properties.

- $\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C}) = (\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}$.
- $\mathbf{A}\otimes(\mathbf{B}+\mathbf{C}) = (\mathbf{A}\otimes\mathbf{B}) + (\mathbf{A}\otimes\mathbf{C})$.
- $(\mathbf{A}+\mathbf{B})\otimes\mathbf{C} = (\mathbf{A}\otimes\mathbf{C}) + (\mathbf{B}\otimes\mathbf{C})$.
- $(\mathbf{A}_1\otimes\mathbf{B}_1)(\mathbf{A}_2\otimes\mathbf{B}_2)\cdots(\mathbf{A}_k\otimes\mathbf{B}_k) = (\mathbf{A}_1\cdots\mathbf{A}_k)\otimes(\mathbf{B}_1\cdots\mathbf{B}_k)$.
- $(\mathbf{A}\otimes\mathbf{B})^* = \mathbf{A}^* + \mathbf{B}^*$.
- $rank\,(\mathbf{A}\otimes\mathbf{B}) = (rank\,(\mathbf{A}))(rank\,(\mathbf{B}))$.

Assume $\mathbf{A}$ is $m\times m$ and $\mathbf{B}$ is $n\times n$ for the following.

- $trace\,(\mathbf{A}\otimes\mathbf{B}) = (trace\,(\mathbf{A}))(trace\,(\mathbf{B}))$.
- $(\mathbf{A}\otimes\mathbf{I}_n)(\mathbf{I}_m\otimes\mathbf{B}) = \mathbf{A}\otimes\mathbf{B} = (\mathbf{I}_m\otimes\mathbf{B})(\mathbf{A}\otimes\mathbf{I}_n)$.
- $\det(\mathbf{A}\otimes\mathbf{B}) = (\det(\mathbf{A}))^m(\det(\mathbf{B}))^n$.
- $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$.

(b) Let the eigenvalues of $\mathbf{A}_{m\times m}$ be denoted by λ_i and let the eigenvalues of $\mathbf{B}_{n\times n}$ be denoted by μ_j. Prove the following.

- The eigenvalues of $\mathbf{A}\otimes\mathbf{B}$ are the mn numbers $\{\lambda_i\mu_j\}_{i=1\ j=1}^{m\ \ n}$.
- The eigenvalues of $(\mathbf{A}\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{B})$ are $\{\lambda_i+\mu_j\}_{i=1\ j=1}^{m\ \ n}$.

7.8.12. Use part (b) of Exercise 7.8.11 along with the result of Exercise 7.6.10 (p. 573) to construct an alternate derivation of (7.6.8) on p. 566. That is, show that the n^2 eigenvalues of the discrete Laplacian $\mathbf{L}_{n^2\times n^2}$ described in Example 7.6.2 (p. 563) are given by

$$\lambda_{ij} = 4\left[\sin^2\left(\frac{i\pi}{2(n+1)}\right) + \sin^2\left(\frac{j\pi}{2(n+1)}\right)\right], \quad i,j = 1,2,\ldots,n.$$

Hint: Recall Exercise 7.2.18 (p. 522).

7.8.13. Determine the eigenvalues of the three-dimensional discrete Laplacian by using the formula from Exercise 7.6.10 (p. 573) that states

$$\mathbf{L}_{n^3\times n^3} = (\mathbf{I}_n\otimes\mathbf{I}_n\otimes\mathbf{A}_n) + (\mathbf{I}_n\otimes\mathbf{A}_n\otimes\mathbf{I}_n) + (\mathbf{A}_n\otimes\mathbf{I}_n\otimes\mathbf{I}_n).$$

7.9 FUNCTIONS OF NONDIAGONALIZABLE MATRICES

The development of functions of nondiagonalizable matrices parallels the development for functions of diagonal matrices that was presented in §7.3 except that the Jordan form is used in place of the diagonal matrix of eigenvalues. Recall from the discussion surrounding (7.3.5) on p. 526 that if $\mathbf{A} \in \mathcal{C}^{n\times n}$ is diagonalizable, say $\mathbf{A} = \mathbf{PDP}^{-1}$, where $\mathbf{D} = \operatorname{diag}(\lambda_1\mathbf{I}, \lambda_2\mathbf{I}, \ldots, \lambda_s\mathbf{I})$, and if $f(\lambda_i)$ exists for each λ_i, then $f(\mathbf{A})$ is defined to be

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{D})\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} f(\lambda_1)\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & f(\lambda_2)\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & f(\lambda_s)\mathbf{I} \end{pmatrix}\mathbf{P}^{-1}.$$

The Jordan decomposition $\mathbf{A} = \mathbf{PJP}^{-1}$ described on p. 590 easily provides a generalization of this idea to nondiagonalizable matrices. If $\mathbf{J}$ is the Jordan form for $\mathbf{A}$, it's natural to define $f(\mathbf{A})$ by writing $f(\mathbf{A}) = \mathbf{P}f(\mathbf{J})\mathbf{P}^{-1}$. However, there are a couple of wrinkles that need to be ironed out before this notion actually makes sense. First, we have to specify what we mean by $f(\mathbf{J})$—this is not as clear as $f(\mathbf{D})$ is for diagonal matrices. And after this is taken care of we need to make sure that $\mathbf{P}f(\mathbf{J})\mathbf{P}^{-1}$ is a uniquely defined matrix. This also is not clear because, as mentioned on p. 590, the transforming matrix $\mathbf{P}$ is not unique—it would not be good if for a given $\mathbf{A}$ you used one $\mathbf{P}$, and I used another, and this resulted in your $f(\mathbf{A})$ being different than mine.

Let's first make sense of $f(\mathbf{J})$. Assume throughout that $\mathbf{A} = \mathbf{PJP}^{-1} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ and where $\mathbf{J} = \operatorname{diag}(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s))$ is the Jordan form for $\mathbf{A}$ in which each segment $\mathbf{J}(\lambda_j)$ is a block-diagonal matrix containing one or more Jordan blocks. That is,

$$\mathbf{J}(\lambda_j) = \begin{pmatrix} \mathbf{J}_1(\lambda_j) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2(\lambda_j) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_{t_j}(\lambda_j) \end{pmatrix} \quad \text{with} \quad \mathbf{J}_\star(\lambda_j) = \begin{pmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{pmatrix}.$$

We want to define $f(\mathbf{J})$ to be

$$f(\mathbf{J}) = \begin{pmatrix} f\big(\mathbf{J}(\lambda_1)\big) & & \\ & \ddots & \\ & & f\big(\mathbf{J}(\lambda_s)\big) \end{pmatrix} \quad \text{with} \quad f\big(\mathbf{J}(\lambda_j)\big) = \begin{pmatrix} \ddots & & \\ & f\big(\mathbf{J}_\star(\lambda_j)\big) & \\ & & \ddots \end{pmatrix},$$

but doing so requires that we give meaning to $f\big(\mathbf{J}_\star(\lambda_j)\big)$. To keep the notation from getting out of hand, let $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & \\ & \ddots & \ddots \\ & & \lambda \end{pmatrix}$ denote a generic $k \times k$ Jordan

block, and let's develop a definition of $f(\mathbf{J}_\star)$. Suppose for a moment that $f(z)$ is a function from $\mathcal{C}$ into $\mathcal{C}$ that has a Taylor series expansion about λ. That is, for some $r > 0$,

$$f(z) = f(\lambda) + f'(\lambda)(z-\lambda) + \frac{f''(\lambda)}{2!}(z-\lambda)^2 + \frac{f'''(\lambda)}{3!}(z-\lambda)^3 + \cdots \quad \text{for} \quad |z-\lambda| < r.$$

The representation (7.3.7) on p. 527 suggests that $f(\mathbf{J}_\star)$ should be defined as

$$f(\mathbf{J}_\star) = f(\lambda)\mathbf{I} + f'(\lambda)(\mathbf{J}_\star - \lambda\mathbf{I}) + \frac{f''(\lambda)}{2!}(\mathbf{J}_\star - \lambda\mathbf{I})^2 + \frac{f'''(\lambda)}{3!}(\mathbf{J}_\star - \lambda\mathbf{I})^3 + \cdots.$$

But since $\mathbf{N} = \mathbf{J}_\star - \lambda\mathbf{I}$ is nilpotent of index k, this series is just the finite sum

$$f(\mathbf{J}_\star) = \sum_{i=0}^{k-1} \frac{f^{(i)}(\lambda)}{i!}\mathbf{N}^i, \tag{7.9.1}$$

and this means that only $f(\lambda)$, $f'(\lambda)$, $\ldots$, $f^{(k-1)}(\lambda)$ are required to exist. Also,

$$\mathbf{N} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}, \quad \mathbf{N}^2 = \begin{pmatrix} 0 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 0 & 0 \\ & & & & 0 \end{pmatrix}, \ldots, \mathbf{N}^{k-1} = \begin{pmatrix} 0 & 0 & \cdots & 1 \\ & \ddots & & \vdots \\ & & \ddots & 0 \\ & & & 0 \end{pmatrix},$$

so the representation of $f(\mathbf{J}_\star)$ in (7.9.1) can be elegantly expressed as follows.

Functions of Jordan Blocks

For a $k \times k$ Jordan block $\mathbf{J}_\star$ with eigenvalue λ, and for a function $f(z)$ such that $f(\lambda)$, $f'(\lambda)$, $\ldots$, $f^{(k-1)}(\lambda)$ exist, $f(\mathbf{J}_\star)$ is defined to be

$$f(\mathbf{J}_\star) = f\begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} = \begin{pmatrix} f(\lambda) & f'(\lambda) & \dfrac{f''(\lambda)}{2!} & \cdots & \dfrac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & f(\lambda) & f'(\lambda) & \ddots & \vdots \\ & & \ddots & \ddots & \dfrac{f''(\lambda)}{2!} \\ & & & f(\lambda) & f'(\lambda) \\ & & & & f(\lambda) \end{pmatrix}. \tag{7.9.2}$$

Every Jordan form $\mathbf{J} = \begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star & \\ & & \ddots \end{pmatrix}$ is a block-diagonal matrix composed of various Jordan blocks $\mathbf{J}_\star$, so (7.9.2) allows us to define $f(\mathbf{J}) = \begin{pmatrix} \ddots & & \\ & f(\mathbf{J}_\star) & \\ & & \ddots \end{pmatrix}$ as long as we pay attention to the fact that a sufficient number of derivatives of f are required to exist at the various eigenvalues. More precisely, if the size of the largest Jordan block associated with an eigenvalue λ is k (i.e., if $index(\lambda) = k$), then $f(\lambda)$, $f'(\lambda)$, ..., $f^{(k-1)}(\lambda)$ must exist in order for $f(\mathbf{J})$ to make sense.

Matrix Functions

For $\mathbf{A} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$, let $k_i = index(\lambda_i)$.

- A function $f : \mathcal{C} \to \mathcal{C}$ is said to be defined (or to exist) at $\mathbf{A}$ when $f(\lambda_i)$, $f'(\lambda_i)$, ..., $f^{(k_i-1)}(\lambda_i)$ exist for each $\lambda_i \in \sigma(\mathbf{A})$.
- Suppose that $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, where $\mathbf{J} = \begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star & \\ & & \ddots \end{pmatrix}$ is in Jordan form with the $\mathbf{J}_\star$'s representing the various Jordan blocks described on p. 590. If f exists at $\mathbf{A}$, then the value of f at $\mathbf{A}$ is defined to be

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{J})\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \ddots & & \\ & f(\mathbf{J}_\star) & \\ & & \ddots \end{pmatrix}\mathbf{P}^{-1}, \tag{7.9.3}$$

where the $f(\mathbf{J}_\star)$'s are as defined in (7.9.2).

We still need to explain why (7.9.3) produces a uniquely defined matrix. The following argument will not only accomplish this purpose, but it will also establish an alternate expression for $f(\mathbf{A})$ that involves neither the Jordan form $\mathbf{J}$ nor the transforming matrix $\mathbf{P}$. Begin by partitioning $\mathbf{J}$ into its s Jordan segments as described on p. 590, and partition $\mathbf{P}$ and $\mathbf{P}^{-1}$ conformably as

$$\mathbf{P} = \Big(\mathbf{P}_1 \,|\, \cdots \,|\, \mathbf{P}_s\Big), \quad \mathbf{J} = \begin{pmatrix} \mathbf{J}(\lambda_1) & & \\ & \ddots & \\ & & \mathbf{J}(\lambda_s) \end{pmatrix}, \quad \text{and} \quad \mathbf{P}^{-1} = \begin{pmatrix} \mathbf{Q}_1 \\ \hline \vdots \\ \hline \mathbf{Q}_s \end{pmatrix}.$$

Define $\mathbf{G}_i = \mathbf{P}_i\mathbf{Q}_i$, and observe that if $k_i = index(\lambda_i)$, then $\mathbf{G}_i$ is the projector onto $N\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$ along $R\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$. To see this, notice that $\mathbf{L}_i = \mathbf{J}(\lambda_i) - \lambda_i\mathbf{I}$ is nilpotent of index k_i, but $\mathbf{J}(\lambda_j) - \lambda_i\mathbf{I}$ is nonsingular when

$i \neq j$, so

$$(\mathbf{A} - \lambda_i \mathbf{I}) = \mathbf{P}(\mathbf{J} - \lambda_i \mathbf{I})\mathbf{P}^{-1} = \mathbf{P} \begin{pmatrix} \mathbf{J}(\lambda_1) - \lambda_i \mathbf{I} & & & \\ & \ddots & & \\ & & \mathbf{L}_i & \\ & & & \ddots \\ & & & \mathbf{J}(\lambda_s) - \lambda_i \mathbf{I} \end{pmatrix} \mathbf{P}^{-1} \tag{7.9.4}$$

is a core-nilpotent decomposition as described on p. 397 (reordering the eigenvalues can put the nilpotent block $\mathbf{L}_i$ on the bottom to realize the form in (5.10.5)). Consequently, the results in Example 5.10.3 (p. 398) insure that $\mathbf{P}_i\mathbf{Q}_i = \mathbf{G}_i$ is the projector onto $N\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$ along $R\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$, and this is true for all similarity transformations that reduce $\mathbf{A}$ to $\mathbf{J}$. If $\mathbf{A}$ happens to be diagonalizable, then $k_i = 1$ for each i, and the matrices $\mathbf{G}_i = \mathbf{P}_i\mathbf{Q}_i$ are precisely the spectral projectors defined on p. 517. For this reason, there is no ambiguity in continuing to use the $\mathbf{G}_i$ notation, and we will continue to refer to the $\mathbf{G}_i$'s as ***spectral projectors.*** In the diagonalizable case, $\mathbf{G}_i$ projects onto the eigenspace associated with λ_i, and in the nondiagonalizable case $\mathbf{G}_i$ projects onto the generalized eigenspace associated with λ_i.

Now consider

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{J})\mathbf{P}^{-1} = \mathbf{P} \begin{pmatrix} f\big(\mathbf{J}(\lambda_1)\big) & & \\ & \ddots & \\ & & f\big(\mathbf{J}(\lambda_s)\big) \end{pmatrix} \mathbf{P}^{-1} = \sum_{i=1}^{s} \mathbf{P}_i f\big(\mathbf{J}(\lambda_i)\big)\mathbf{Q}_i. \tag{7.9.5}$$

Since $f\big(\mathbf{J}(\lambda_i)\big) = \begin{pmatrix} \ddots & & \\ & f\big(\mathbf{J}_\star(\lambda_i)\big) & \\ & & \ddots \end{pmatrix}$, where the $\mathbf{J}_\star(\lambda_i)$'s are the Jordan blocks associated with λ_i, (7.9.2) insures that if $k_i = index(\lambda_i)$, then

$$f\big(\mathbf{J}(\lambda_i)\big) = f(\lambda_i)\mathbf{I} + f'(\lambda_i)\mathbf{L}_i + \frac{f''(\lambda_i)}{2!}\mathbf{L}_i^2 + \cdots + \frac{f^{(k_i-1)}(\lambda_i)}{(k_i-1)!}\mathbf{L}_i^{k_i-1},$$

where $\mathbf{L}_i = \mathbf{J}(\lambda_i) - \lambda_i\mathbf{I}$, and thus (7.9.5) becomes

$$f(\mathbf{A}) = \sum_{i=1}^{s} \mathbf{P}_i f\big(\mathbf{J}(\lambda_i)\big)\mathbf{Q}_i = \sum_{i=1}^{s} \sum_{j=0}^{k_i-1} \frac{f^{(j)}(\lambda_i)}{j!}\mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i. \tag{7.9.6}$$

The terms $\mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i$ can be simplified by noticing that

$$\mathbf{P}^{-1}\mathbf{P} = \mathbf{I} \implies \mathbf{Q}_i\mathbf{P}_j = \begin{cases} \mathbf{I} & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \implies \mathbf{P}^{-1}\mathbf{G}_i = \begin{pmatrix} \mathbf{Q}_1 \\ \vdots \\ \mathbf{Q}_i \\ \vdots \\ \mathbf{Q}_s \end{pmatrix} \mathbf{P}_i\mathbf{Q}_i = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{Q}_i \\ \vdots \\ \mathbf{0} \end{pmatrix},$$

and by using this with (7.9.4) to conclude that

$$(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i = \mathbf{P}\begin{pmatrix} \big(\mathbf{J}(\lambda_1)-\lambda_i\mathbf{I}\big)^j & & & \\ & \ddots & & \\ & & \mathbf{L}_i^j & \\ & & & \ddots \\ & & & \big(\mathbf{J}(\lambda_s)-\lambda_i\mathbf{I}\big)^j \end{pmatrix}\mathbf{P}^{-1}\mathbf{G}_i = \mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i. \quad (7.9.7)$$

Thus (7.9.6) can be written as

$$f(\mathbf{A}) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i, \quad (7.9.8)$$

and this expression is independent of which similarity is used to reduce $\mathbf{A}$ to $\mathbf{J}$. Not only does (7.9.8) prove that $f(\mathbf{A})$ is uniquely defined, but it also provides a generalization of the spectral theorems for diagonalizable matrices given on pp. 517 and 526 because if $\mathbf{A}$ is diagonalizable, then each $k_i = 1$ so that (7.9.8) reduces to (7.3.6) on p. 526. Below is a formal summary along with some related properties.

Spectral Resolution of $f(\mathbf{A})$

For $\mathbf{A}\in\mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1,\lambda_2,\ldots,\lambda_s\}$ such that $k_i = index(\lambda_i)$, and for a function $f:\mathcal{C}\to\mathcal{C}$ such that $f(\lambda_i)$, $f'(\lambda_i)$, $\ldots$, $f^{(k_i-1)}(\lambda_i)$ exist for each $\lambda_i\in\sigma(\mathbf{A})$, the value of $f(\mathbf{A})$ is

$$f(\mathbf{A}) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i, \quad (7.9.9)$$

where the ***spectral projectors*** $\mathbf{G}_i$'s have the following properties.

- $\mathbf{G}_i$ is the projector onto the generalized eigenspace $N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$ along $R\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$.
- $\mathbf{G}_1+\mathbf{G}_2+\cdots+\mathbf{G}_s = \mathbf{I}$. (7.9.10)
- $\mathbf{G}_i\mathbf{G}_j = \mathbf{0}$ when $i\neq j$. (7.9.11)
- $\mathbf{N}_i = (\mathbf{A}-\lambda_i\mathbf{I})\mathbf{G}_i = \mathbf{G}_i(\mathbf{A}-\lambda_i\mathbf{I})$ is nilpotent of index k_i. (7.9.12)
- If $\mathbf{A}$ is diagonalizable, then (7.9.9) reduces to (7.3.6) on p. 526, and the spectral projectors reduce to those described on p. 517.

Proof of (7.9.10)–(7.9.12). Property (7.9.10) results from using (7.9.9) with the function $f(z) = 1$, and property (7.9.11) is a consequence of

$$\mathbf{I} = \mathbf{P}^{-1}\mathbf{P} \implies \mathbf{Q}_i\mathbf{P}_j = \begin{cases} \mathbf{I} & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j. \end{cases} \tag{7.9.13}$$

To prove (7.9.12), establish that $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{G}_i = \mathbf{G}_i(\mathbf{A} - \lambda_i\mathbf{I})$ by noting that (7.9.13) implies $\mathbf{P}^{-1}\mathbf{G}_i = \begin{pmatrix}\mathbf{0}\cdots\mathbf{Q}_i\cdots\mathbf{0}\end{pmatrix}^T$ and $\mathbf{G}_i\mathbf{P} = \begin{pmatrix}\mathbf{0}\cdots\mathbf{P}_i\cdots\mathbf{0}\end{pmatrix}$. Use this with (7.9.4) to observe that $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{G}_i = \mathbf{P}_i\mathbf{L}_i\mathbf{Q}_i = \mathbf{G}_i(\mathbf{A} - \lambda_i\mathbf{I})$. Now

$$\mathbf{N}_i^j = (\mathbf{P}_i\mathbf{L}_i\mathbf{Q}_i)^j = \mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i \quad \text{for } j = 1, 2, 3, \ldots,$$

and thus $\mathbf{N}_i$ is nilpotent of index k_i because $\mathbf{L}_i$ is nilpotent of index k_i. ∎

Example 7.9.1

A coordinate-free version of the representation in (7.9.3) results by separating the first-order terms in (7.9.9) from the higher-order terms to write

$$f(\mathbf{A}) = \sum_{i=1}^{s}\left[f(\lambda_i)\mathbf{G}_i + \sum_{j=1}^{k_i-1}\frac{f^{(j)}(\lambda_i)}{j!}\mathbf{N}_i^j\right].$$

Using the identity function $f(z) = z$ produces a coordinate-free version of the Jordan decomposition of $\mathbf{A}$ in the form

$$\mathbf{A} = \sum_{i=1}^{s}\left[\lambda_i\mathbf{G}_i + \mathbf{N}_i\right],$$

and this is the extension of (7.2.7) on p. 517 to the nondiagonalizable case. Another version of (7.9.9) results from lumping things into one matrix to write

$$f(\mathbf{A}) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1} f^{(j)}(\lambda_i)\mathbf{Z}_{ij}, \quad \text{where} \quad \mathbf{Z}_{ij} = \frac{(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i}{j!}. \tag{7.9.14}$$

The $\mathbf{Z}_{ij}$'s are often called the *component matrices* or the *constituent matrices*.

Example 7.9.2

Problem: Describe $f(\mathbf{A})$ for functions f defined at $\mathbf{A} = \begin{pmatrix} 6 & 2 & 8 \\ -2 & 2 & -2 \\ 0 & 0 & 2 \end{pmatrix}$.

Solution: $\mathbf{A}$ is block triangular, so it's easy to see that $\lambda_1 = 2$ and $\lambda_2 = 4$ are the two distinct eigenvalues with $index(\lambda_1) = 1$ and $index(\lambda_2) = 2$. Thus $f(\mathbf{A})$ exists for all functions such that $f(2)$, $f(4)$, and $f'(4)$ exist, in which case

$$f(\mathbf{A}) = f(2)\mathbf{G}_1 + f(4)\mathbf{G}_2 + f'(4)(\mathbf{A} - 4\mathbf{I})\mathbf{G}_2.$$

The spectral projectors could be computed directly, but things are easier if some judicious choices of f are made. For example,

$$\left.\begin{array}{lcl} f(z) = 1 & \Rightarrow & \mathbf{I} = f(\mathbf{A}) = \mathbf{G}_1 + \mathbf{G}_2 \\ f(z) = (z-4)^2 & \Rightarrow & (\mathbf{A} - 4\mathbf{I})^2 = f(\mathbf{A}) = 4\mathbf{G}_1 \end{array}\right\} \implies \begin{array}{l} \mathbf{G}_1 = (\mathbf{A} - 4\mathbf{I})^2/4, \\ \mathbf{G}_2 = \mathbf{I} - \mathbf{G}_1. \end{array}$$

Now that the spectral projectors are known, any function defined at $\mathbf{A}$ can be evaluated. For example, if $f(z) = z^{1/2}$, then

$$f(\mathbf{A}) = \sqrt{\mathbf{A}} = \sqrt{2}\mathbf{G}_1 + \sqrt{4}\mathbf{G}_2 + (1/2\sqrt{4})(\mathbf{A} - 4\mathbf{I})\mathbf{G}_2 = \frac{1}{2}\begin{pmatrix} 5 & 1 & 7-2\sqrt{2} \\ -1 & 3 & 5-4\sqrt{2} \\ 0 & 0 & 2\sqrt{2} \end{pmatrix}.$$

This technique illustrated above is rather ad hoc, but it always works if a sufficient number of appropriate functions are used. For example, using $f(z) = z^p$ for $p = 0, 1, 2, \ldots$ will always produce a system of equations that will yield the component matrices $\mathbf{Z}_{ij}$ given in (7.9.14) because

$$\begin{aligned} &\text{for } f(z) = 1\text{:} && \mathbf{I} = \textstyle\sum \mathbf{Z}_{i0}, \\ &\text{for } f(z) = z\text{:} && \mathbf{A} = \textstyle\sum \lambda_i \mathbf{Z}_{i0} + \sum \mathbf{Z}_{i1}, \\ &\text{for } f(z) = z^2\text{:} && \mathbf{A}^2 = \textstyle\sum \lambda_i^2 \mathbf{Z}_{i0} + \sum 2\lambda_i \mathbf{Z}_{i1} + \sum 2\mathbf{Z}_{i2}, \\ &\qquad\vdots \end{aligned}$$

and this can be considered as a generalized Vandermonde linear system (p. 185)

$$\begin{pmatrix} 1 & \cdots & 1 & & & & & & & \\ \lambda_1 & \cdots & \lambda_s & 1 & \cdots & 1 & & & & \\ \lambda_1^2 & \cdots & \lambda_s^2 & 2\lambda_1 & \cdots & 2\lambda_s & 2 & \cdots & 2 & \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \\ & \cdots & & & \cdots & & & \cdots & & \cdots \end{pmatrix} \begin{pmatrix} \mathbf{Z}_{10} \\ \vdots \\ \mathbf{Z}_{s0} \\ \mathbf{Z}_{11} \\ \vdots \\ \mathbf{Z}_{s1} \\ \mathbf{Z}_{21} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{I} \\ \mathbf{A} \\ \mathbf{A}^2 \\ \mathbf{A}^3 \\ \vdots \end{pmatrix}$$

that can be solved for the $\mathbf{Z}_{ij}$'s. Other sets of polynomials such as

$$\{1,\ (z-\lambda_1)^{k_1},\ (z-\lambda_1)^{k_2}(z-\lambda_2)^{k_2},\ \ldots (z-\lambda_1)^{k_1}\cdots(z-\lambda_s)^{k_s}\}$$

will generate other linear systems that yield solutions containing the $\mathbf{Z}_{ij}$'s.

Example 7.9.3

Series Representations. Suppose that $\sum_{j=0}^{\infty} c_j(z-z_0)^j$ converges to $f(z)$ at each point inside a circle $|z - z_0| = r$, and suppose that $\mathbf{A}$ is a matrix such that $|\lambda_i - z_0| < r$ for each eigenvalue $\lambda_i \in \sigma(\mathbf{A})$.

Problem: Explain why $\sum_{j=0}^{\infty} c_j(\mathbf{A} - z_0\mathbf{I})^j$ converges to $f(\mathbf{A})$.

Solution: If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is in Jordan form as described on p. 601, then it's not difficult to argue that $\sum_{j=0}^{\infty} c_j(\mathbf{A} - z_0\mathbf{I})^j$ converges if and only if

$$\mathbf{P}^{-1}\Big(\sum_{j=0}^{\infty} c_j(\mathbf{A}-z_0\mathbf{I})^j\Big)\mathbf{P} = \sum_{j=0}^{\infty} c_j\mathbf{P}^{-1}(\mathbf{A}-z_0\mathbf{I})^j\mathbf{P} = \sum_{j=0}^{\infty} c_j(\mathbf{J}-z_0\mathbf{I})^j = \begin{pmatrix} \ddots & & \\ & \sum_{j=0}^{\infty} c_j(\mathbf{J}_\star - z_0\mathbf{I})^j & \\ & & \ddots \end{pmatrix}$$

converges. Consequently, it suffices to prove that $\sum_{j=0}^{\infty} c_j(\mathbf{J}_\star - z_0\mathbf{I})^j$ converges to $f(\mathbf{J}_\star)$ for a generic $k \times k$ Jordan block

$$\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix} = \lambda\mathbf{I} + \mathbf{N}, \quad \text{where} \quad \mathbf{N} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & & 0 \end{pmatrix}_{k\times k}.$$

A standard theorem from analysis states that if $\sum_{j=0}^{\infty} c_j(z - z_0)^j$ converges to $f(z)$ when $|z - z_0| < r$, then the series may be differentiated term by term to yield series that converge to derivatives of f at points inside the circle of convergence. Consequently, for each $i = 0, 1, 2, \ldots,$

$$\frac{f^{(i)}(z)}{i!} = \sum_{j=0}^{\infty} c_j \binom{j}{i} (z - z_0)^{j-i} \quad \text{when} \quad |z - z_0| < r. \tag{7.9.15}$$

We know from (7.9.1) (with $f(z) = z^j$) that

$$(\mathbf{J}_\star - z_0\mathbf{I})^j = (\lambda - z_0)^j\mathbf{I} + \binom{j}{1}(\lambda - z_0)^{j-1}\mathbf{N} + \cdots + \binom{j}{k-1}(\lambda - z_0)^{j-(k-1)}\mathbf{N}^{k-1},$$

so this together with (7.9.15) produces

$$\begin{aligned} \sum_{j=0}^{\infty} c_j(\mathbf{J}_\star - z_0\mathbf{I})^j &= \left(\sum_{j=0}^{\infty} c_j(\lambda - z_0)^j\right)\mathbf{I} + \left(\sum_{j=0}^{\infty} c_j\binom{j}{1}(\lambda - z_0)^{j-1}\right)\mathbf{N} \\ &\quad + \cdots + \left(\sum_{j=0}^{\infty} c_j\binom{j}{k-1}(\lambda - z_0)^{j-(k-1)}\right)\mathbf{N}^{k-1} \\ &= f(\lambda)\mathbf{I} + f'(\lambda)\mathbf{N} + \cdots + \frac{f^{(k-1)}}{(k-1)!}(\lambda)\mathbf{N}^{k-1} = f(\mathbf{J}_*). \end{aligned}$$

Note: The result of this example validates the statements made on p. 527.

Example 7.9.4

All Matrix Functions Are Polynomials. It was pointed out on p. 528 that if $\mathbf{A}$ is diagonalizable, and if $f(\mathbf{A})$ exists, then there is a polynomial $p(z)$ such that $f(\mathbf{A}) = p(\mathbf{A})$, and you were asked in Exercise 7.3.7 (p. 539) to use the Cayley–Hamilton theorem (pp. 509, 532) to extend this property to nondiagonalizable matrices for functions that have an infinite series expansion. We can now see why this is true in general.

Problem: For a function f defined at $\mathbf{A} \in \mathcal{C}^{n\times n}$, exhibit a polynomial $p(z)$ such that $f(\mathbf{A}) = p(\mathbf{A})$.

Solution: Suppose that $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $index(\lambda_i) = k_i$. The trick is to find a polynomial $p(z)$ such that for each $i = 1, 2, \ldots, s$,

$$p(\lambda_i) = f(\lambda_i), \quad p'(\lambda_i) = f'(\lambda_i), \quad \ldots, \quad p^{(k_i-1)}(\lambda_i) = f^{(k_i-1)}(\lambda_i) \tag{7.9.16}$$

because if such a polynomial exists, then (7.9.9) guarantees that

$$p(\mathbf{A}) = \sum_{i=1}^{s} \sum_{j=0}^{k_i-1} \frac{p^{(j)}(\lambda_i)}{j!}(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i = \sum_{i=1}^{s} \sum_{j=0}^{k_i-1} \frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i = f(\mathbf{A}).$$

Since there are $k = \sum_{i=1}^{s} k_i$ equations in (7.9.16) to be satisfied, let's look for a polynomial of the form

$$p(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_{k-1} z^{k-1}$$

by writing the equations in (7.9.16) as the following $k \times k$ linear system $\mathbf{Hx} = \mathbf{f}$:

$$\begin{array}{l}
p(\lambda_1) = f(\lambda_1) \\
\vdots \\
p(\lambda_s) = f(\lambda_s) \\
\hline
\vdots \\
p'(\lambda_i) = f'(\lambda_i) \\
\vdots \\
\hline
\vdots \\
p''(\lambda_i) = f''(\lambda_i) \\
\vdots \\
\hline
\vdots
\end{array}
\begin{array}{c}
\Rightarrow \\
\\
\Rightarrow \\
\\
\\
\Rightarrow \\
\\
\\
\Rightarrow \\
\\
\\
\\
\end{array}
\begin{pmatrix}
1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 & \cdots & \lambda_1^{k-1} \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
1 & \lambda_s & \lambda_s^2 & \lambda_s^3 & \cdots & \lambda_s^{k-1} \\
\hline
\vdots & \vdots & \vdots & \vdots & & \vdots \\
0 & 1 & 2\lambda_i & 3\lambda_i^2 & \cdots & (k-1)\lambda_i^{k-2} \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
\hline
\vdots & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & 2 & 6\lambda_i & \cdots & (k-1)(k-2)\lambda_i^{(k-3)} \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
\hline
\vdots & \vdots & \vdots & \vdots & & \vdots
\end{pmatrix}
\begin{pmatrix}
\alpha_0 \\ \\ \alpha_1 \\ \\ \alpha_2 \\ \\ \alpha_3 \\ \\ \vdots \\ \\ \vdots \\ \\ \vdots \\ \\ \alpha_{k-1}
\end{pmatrix}
=
\begin{pmatrix}
f(\lambda_1) \\ \vdots \\ f(\lambda_s) \\ \hline \vdots \\ f'(\lambda_i) \\ \vdots \\ \hline \vdots \\ f''(\lambda_i) \\ \vdots \\ \hline \vdots
\end{pmatrix}.$$

The coefficient matrix $\mathbf{H}$ can be proven to be nonsingular because the rows in each segment of $\mathbf{H}$ are linearly independent. The rows in the top segment of $\mathbf{H}$ are a subset of rows from a Vandermonde matrix (p. 185), while the nonzero portion of each succeeding segment has the form $\mathbf{VD}$, where the rows of $\mathbf{V}$ are a subset of rows from a Vandermonde matrix and $\mathbf{D}$ is a nonsingular diagonal matrix. Consequently, $\mathbf{Hx} = \mathbf{f}$ has a unique solution, and thus there is a unique polynomial $p(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_{k-1} z^{k-1}$ that satisfies the conditions in (7.9.16). This polynomial $p(z)$ is called the ***Hermite interpolation polynomial***, and it has the property that $f(\mathbf{A}) = p(\mathbf{A})$.

Example 7.9.5

Functional Identities. Scalar functional identities generally extend to the matrix case. For example, the scalar identity $\sin^2 z + \cos^2 z = 1$ extends to matrices as $\sin^2 \mathbf{Z} + \cos^2 \mathbf{Z} = \mathbf{I}$, and this is valid for all $\mathbf{Z} \in \mathcal{C}^{n\times n}$. While it's possible to prove such identities on a case-by-case basis by using (7.9.3) or (7.9.9), there is a more robust approach that is described below.

For two functions f_1 and f_2 from $\mathcal{C}$ into $\mathcal{C}$ and for a polynomial $p(x, y)$ in two variables, let h be the composition defined by $h(z) = p\big(f_1(z),\, f_2(z)\big)$. If $\mathbf{A}_{n\times n}$ has eigenvalues $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $index(\lambda_i) = k_i$, and if h is defined at $\mathbf{A}$, then we are allowed to assert that $h(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big)$ because Example 7.9.4 insures that there are polynomials $g(z)$ and $q(z)$ such that $h(\mathbf{A}) = g(\mathbf{A})$ and $p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big) = q(\mathbf{A})$, where for each $\lambda_i \in \sigma(\mathbf{A})$,

$$g^{(j)}(\lambda_i) = h^{(j)}(\lambda_i) = \left.\frac{d^j\big[p\big(f_1(z),\, f_2(z)\big)\big]}{dz^j}\right|_{z=\lambda_i} = q^{(j)}(\lambda_i) \quad \text{for } j = 0, 1, \ldots, k_i - 1,$$

so $g(\mathbf{A}) = q(\mathbf{A})$, and thus $h(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big)$. To build functional identities for $\mathbf{A}$, choose f_1 and f_2 in $h(z) = p\big(f_1(z),\, f_2(z)\big)$ that will make

$$h(\lambda_i) = h'(\lambda_i) = h''(\lambda_i) = \cdots = h^{(k_i-1)}(\lambda_i) = 0 \quad \text{for each} \quad \lambda_i \in \sigma(\mathbf{A}),$$

thereby insuring that $\mathbf{0} = \mathbf{h}(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big)$. This technique produces a plethora of functional identities. For example, using

$$\left\{\begin{array}{l} f_1(z) = \sin^2 z \\ f_2(z) = \cos^2 z \\ p(x, y) = x^2 + y^2 - 1 \end{array}\right\} \text{ produces } h(z) = p\big(f_1(z),\, f_2(z)\big) = \sin^2 z + \cos^2 z - 1.$$

Since $h(z) = 0$ for all $z \in \mathcal{C}$, it follows that $h(\mathbf{Z}) = \mathbf{0}$ for all $\mathbf{Z} \in \mathcal{C}^{n\times n}$, and thus $\sin^2 \mathbf{Z} + \cos^2 \mathbf{Z} = \mathbf{I}$ for all $\mathbf{Z} \in \mathcal{C}^{n\times n}$. It's evident that this technique can be extended to include any number of functions $f_1, f_2, \ldots, f_m$ with a polynomial $p(x_1, x_2, \ldots, x_m)$ to produce even more complicated relationships.

Example 7.9.6

Systems of Differential Equations Revisited. The purpose here is to extend the discussion in §7.4 to cover the nondiagonalizable case. Write the system of differential equations in (7.4.1) on p. 541 in matrix form as

$$\mathbf{u}'(t) = \mathbf{A}_{n\times n}\mathbf{u}(t) \quad \text{with} \quad \mathbf{u}(0) = \mathbf{c}, \tag{7.9.17}$$

but this time don't assume that $\mathbf{A}_{n\times n}$ is diagonalizable—suppose instead that $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $index(\lambda_i) = k_i$. The development parallels that

for the diagonalizable case, but $e^{\mathbf{A}t}$ is now a little more complicated than (7.4.2). Using $f(z) = e^{zt}$ in (7.9.3) and (7.9.2) yields

$$e^{\mathbf{A}t} = \mathbf{P}\begin{pmatrix} \ddots & & \\ & e^{\mathbf{J}_\star t} & \\ & & \ddots \end{pmatrix}\mathbf{P}^{-1} \text{ with } e^{\mathbf{J}_\star t} = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} & \frac{t^2e^{\lambda t}}{2!} & \cdots & \frac{t^{k-1}e^{\lambda t}}{(k-1)!} \\ & e^{\lambda t} & te^{\lambda t} & \ddots & \vdots \\ & & \ddots & \ddots & \frac{t^2e^{\lambda t}}{2!} \\ & & & e^{\lambda t} & te^{\lambda t} \\ & & & & e^{\lambda t} \end{pmatrix}, \quad (7.9.18)$$

while setting $f(z) = e^{zt}$ in (7.9.9) produces

$$e^{\mathbf{A}t} = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{t^j e^{\lambda_i t}}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i. \quad (7.9.19)$$

Either of these can be used to show that the three properties (7.4.3)–(7.4.5) on p. 541 still hold. In particular, $d\,e^{\mathbf{A}t}/dt = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}$, so, just as in the diagonalizable case, $\mathbf{u}(t) = e^{\mathbf{A}t}\mathbf{c}$ is the unique solution of (7.9.17) (the uniqueness argument given in §7.4 remains valid). In the diagonalizable case, the solution of (7.9.17) involves only the eigenvalues and eigenvectors of $\mathbf{A}$ as described in (7.4.7) on p. 542, but generalized eigenvectors are needed for the nondiagonalizable case. Using (7.9.19) yields the solution to (7.9.17) as

$$\mathbf{u}(t) = e^{\mathbf{A}t}\mathbf{c} = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{t^j e^{\lambda_i t}}{j!}\mathbf{v}_j(\lambda_i), \quad \text{where } \mathbf{v}_j(\lambda_i) = (\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i\mathbf{c}. \quad (7.9.20)$$

Each $\mathbf{v}_{k_i-1}(\lambda_i)$ is an eigenvector associated with λ_i because $(\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\mathbf{G}_i = \mathbf{0}$, and $\{\mathbf{v}_{k_i-2}(\lambda_i), \ldots, \mathbf{v}_1(\lambda_i), \mathbf{v}_0(\lambda_i)\}$ is an associated chain of generalized eigenvectors. The behavior of the solution (7.9.20) as $t \to \infty$ is similar but not identical to that discussed on p. 544 because for $\lambda = x + iy$ and $t > 0$,

$$t^j e^{\lambda t} = t^j e^{xt}(\cos yt + i\sin yt) \to \begin{cases} 0 & \text{if } x < 0, \\ \text{unbounded} & \text{if } x \geq 0 \text{ and } j > 0, \\ \text{oscillates indefinitely} & \text{if } x = j = 0 \text{ and } y \neq 0, \\ 1 & \text{if } x = y = j = 0. \end{cases}$$

In particular, if $\mathrm{Re}\,(\lambda_i) < 0$ for every $\lambda_i \in \sigma(\mathbf{A})$, then $\mathbf{u}(t) \to \mathbf{0}$ for every initial vector $\mathbf{c}$, in which case the system is said to be ***stable***.

- **Nonhomogeneous Systems.** It can be verified by direct manipulation that the solution of $\mathbf{u}'(t) = \mathbf{A}\mathbf{u}(t) + \mathbf{f}(t)$ with $\mathbf{u}(t_0) = \mathbf{c}$ is given by

$$\mathbf{u}(t) = e^{\mathbf{A}(t-t_0)}\mathbf{c} + \int_{t_0}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{f}(\tau)d\tau.$$

Example 7.9.7

Nondiagonalizable Mixing Problem. To make the point that even simple problems in nature can be nondiagonalizable, consider three V gallon tanks as shown in Figure 7.9.1 that are initially full of polluted water in which the i^{th} tank contains c_i lbs of a pollutant. In an attempt to flush the pollutant out, all spigots are opened at once allowing fresh water at the rate of r gal/sec to flow into the top of tank #3, while r gal/sec flow from its bottom into the top of tank #2, and so on.

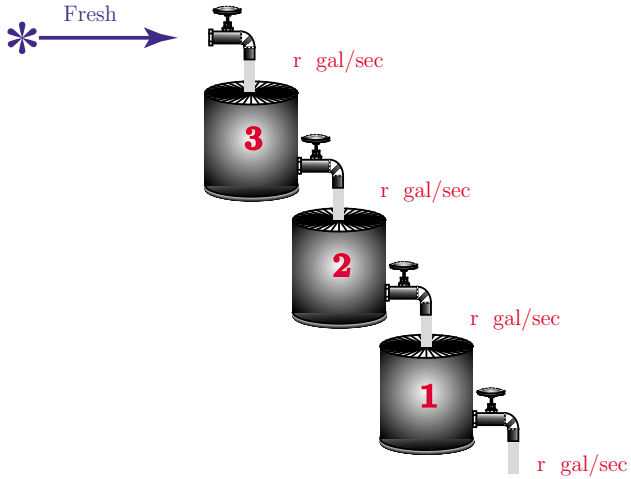


FIGURE 7.9.1

Problem: How many pounds of the pollutant are in each tank at any finite time $t > 0$ when instantaneous and continuous mixing occurs?

Solution: If $u_i(t)$ denotes the number of pounds of pollutant in tank i at time $t > 0$, then the concentration of pollutant in tank i at time t is $u_i(t)/V$ lbs/gal, so the model $u_i'(t) = (\text{lbs/sec})$ coming in $-(\text{lbs/sec})$ going out produces the nondiagonalizable system:

$$\begin{pmatrix} u_1'(t) \\ u_2'(t) \\ u_3'(t) \end{pmatrix} = \frac{r}{V} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix}, \text{ or } \mathbf{u}' = \mathbf{A}\mathbf{u} \text{ with } \mathbf{u}(0) = \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}.$$

This setup is almost the same as that in Exercise 3.5.11 (p. 104). Notice that $\mathbf{A}$ is simply a scalar multiple of a single Jordan block $\mathbf{J}_\star = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}$, so $e^{\mathbf{A}t}$ is easily determined by replacing t by rt/V and λ by -1 in the second equation of (7.9.18) to produce

$$e^{\mathbf{A}t} = e^{(rt/V)\mathbf{J}_\star} = e^{-rt/V} \begin{pmatrix} 1 & rt/V & (rt/V)^2/2 \\ 0 & 1 & rt/V \\ 0 & 0 & 1 \end{pmatrix}.$$

Therefore,

$$\mathbf{u}(t) = e^{\mathbf{A}t}\mathbf{c} = e^{-rt/V}\begin{pmatrix} c_1 + c_2(rt/V) + c_3\left(rt/V\right)^2/2 \\ c_2 + c_3(rt/V) \\ c_3 \end{pmatrix},$$

and, just as common sense dictates, the pollutant is never completely flushed from the tanks in finite time. Only in the limit does each $u_i \to 0$, and it's clear that the rate at which $u_1 \to 0$ is slower than the rate at which $u_2 \to 0$, which in turn is slower than the rate at which $u_3 \to 0$.

Example 7.9.8

The Cauchy integral formula is an elegant result from complex analysis stating that if $f : \mathcal{C} \to \mathcal{C}$ is analytic in and on a simple closed contour $\Gamma \subset \mathcal{C}$ with positive (counterclockwise) orientation, and if ξ_0 is interior to Γ, then

$$f(\xi_0) = \frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \xi_0}d\xi \quad \text{and} \quad f^{(j)}(\xi_0) = \frac{j!}{2\pi i}\int_\Gamma \frac{f(\xi)}{(\xi - \xi_0)^{j+1}}d\xi. \tag{7.9.21}$$

These formulas produce analogous representations of matrix functions. Suppose that $\mathbf{A} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ and $index(\lambda_i) = k_i$. For a complex variable ξ, the ***resolvent of*** $\mathbf{A} \in \mathcal{C}^{n\times n}$ is defined to be the matrix

$$\mathbf{R}(\xi) = (\xi\mathbf{I} - \mathbf{A})^{-1}.$$

If $\xi \notin \sigma(\mathbf{A})$, then $r(z) = (\xi - z)^{-1}$ is defined at $\mathbf{A}$ with $r(\mathbf{A}) = \mathbf{R}(\xi)$, so the spectral resolution theorem (p. 603) can be used to write

$$\mathbf{R}(\xi) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{r^{(j)}(\lambda_i)}{j!}(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{1}{(\xi - \lambda_i)^{j+1}}(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i.$$

If $\sigma(\mathbf{A})$ is in the interior of a simple closed contour Γ, and if the contour integral of a matrix is defined by entrywise integration, then (7.9.21) produces

$$\begin{aligned}
\frac{1}{2\pi i}\int_\Gamma f(\xi)(\xi\mathbf{I} - \mathbf{A})^{-1}d\xi &= \frac{1}{2\pi i}\int_\Gamma f(\xi)\mathbf{R}(\xi)d\xi \\
&= \frac{1}{2\pi i}\int_\Gamma \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f(\xi)}{(\xi - \lambda_i)^{j+1}}(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i d\xi \\
&= \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\left[\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{(\xi - \lambda_i)^{j+1}}d\xi\right](\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i \\
&= \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i = f(\mathbf{A}).
\end{aligned}$$

- In other words, if Γ is a simple closed contour containing $\sigma(\mathbf{A})$ in its interior, then
$$f(\mathbf{A}) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma} f(\xi)(\xi \mathbf{I} - \mathbf{A})^{-1} d\xi \tag{7.9.22}$$
whenever f is analytic in and on Γ. Since this formula makes sense for general linear operators, it is often adopted as a definition for $f(\mathbf{A})$ in more general settings.

- Furthermore, if Γ_i is a simple closed contour enclosing λ_i but excluding all other eigenvalues of $\mathbf{A}$, then the i^{th} spectral projector is given by
$$\mathbf{G}_i = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} \mathbf{R}(\xi) d\xi = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi \quad \text{(Exercise 7.9.19).}$$

Exercises for section 7.9

7.9.1. Lake $\#i$ in a closed system of three lakes of equal volume V initially contains c_i lbs of a pollutant. If the water in the system is circulated at rates (gal/sec) as indicated in Figure 7.9.2, find the amount of pollutant in each lake at time $t > 0$ (assume continuous mixing), and then determine the pollution in each lake in the long run.

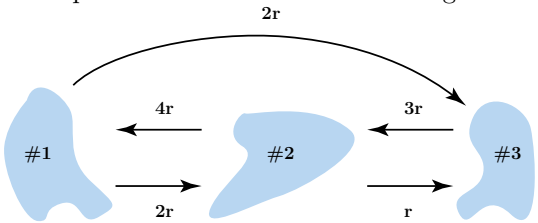


FIGURE 7.9.2

7.9.2. Suppose that $\mathbf{A} \in \mathcal{C}^{n \times n}$ has eigenvalues λ_i with $index(\lambda_i) = k_i$. Explain why the i^{th} spectral projector is given by
$$\mathbf{G}_i = f_i(\mathbf{A}), \quad \text{where} \quad f_i(z) = \begin{cases} 1 & \text{when } z = \lambda_i, \\ 0 & \text{otherwise.} \end{cases}$$

7.9.3. Explain why each spectral projector $\mathbf{G}_i$ can be expressed as a polynomial in $\mathbf{A}$.

7.9.4. If $\sigma(\mathbf{A}_{n \times n}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $k_i = index(\lambda_i)$, explain why
$$\mathbf{A}^k = \sum_{i=1}^{s} \sum_{j=0}^{k_i - 1} \binom{k}{j} \lambda_i^{k-j} (\mathbf{A} - \lambda_i \mathbf{I})^j \mathbf{G}_i.$$

7.9.5. With the convention that $\binom{k}{j} = 0$ for $j > k$, explain why

$$\begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}^k_{m\times m} = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \cdots & \binom{k}{m-1}\lambda^{k-m+1} \\ & \lambda^k & \binom{k}{1}\lambda^{k-1} & \ddots & \vdots \\ & & \ddots & \ddots & \binom{k}{2}\lambda^{k-2} \\ & & & \lambda^k & \binom{k}{1}\lambda^{k-1} \\ & & & & \lambda^k \end{pmatrix}.$$

7.9.6. Determine $e^{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 6 & 2 & 8 \\ -2 & 2 & -2 \\ 0 & 0 & 2 \end{pmatrix}$.

7.9.7. For $f(z) = 4\sqrt{z} - 1$, determine $f(\mathbf{A})$ when $\mathbf{A} = \begin{pmatrix} -3 & -8 & -9 \\ 5 & 11 & 9 \\ -1 & -2 & 1 \end{pmatrix}$.

7.9.8. (a) Explain why every nonsingular $\mathbf{A} \in \mathcal{C}^{n\times n}$ has a square root.

(b) Give necessary and sufficient conditions for the existence of $\sqrt{\mathbf{A}}$ when $\mathbf{A}$ is singular.

7.9.9. Spectral Mapping Property. Prove that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(f(\lambda), \mathbf{x})$ is an eigenpair for $f(\mathbf{A})$ whenever $f(\mathbf{A})$ exists. Does it also follow that $alg\ mult_{\mathbf{A}}(\lambda) = alg\ mult_{f(\mathbf{A})}(f(\lambda))$?

7.9.10. Let f be defined at $\mathbf{A}$, and let $\lambda \in \sigma(\mathbf{A})$. Give an example or an explanation of why the following statements are *not* necessarily true.

(a) $f(\mathbf{A})$ is similar to $\mathbf{A}$.

(b) $geo\ mult_{\mathbf{A}}(\lambda) = geo\ mult_{f(\mathbf{A})}(f(\lambda))$.

(c) $index_{\mathbf{A}}(\lambda) = index_{f(\mathbf{A})}(f(\lambda))$.

7.9.11. Explain why $\mathbf{A}f(\mathbf{A}) = f(\mathbf{A})\mathbf{A}$ whenever $f(\mathbf{A})$ exists.

7.9.12. Explain why a function f is defined at $\mathbf{A} \in \mathcal{C}^{n\times n}$ if and only if f is defined at $\mathbf{A}^T$, and then prove that $f(\mathbf{A}^T) = \big[f(\mathbf{A})\big]^T$. Why can't $(\star)^*$ be used in place of $(\star)^T$?

7.9.13. Use the technique of Example 7.9.5 (p. 608) to establish the following identities.

(a) $e^{\mathbf{A}}e^{-\mathbf{A}} = \mathbf{I}$ for all $\mathbf{A} \in \mathcal{C}^{n\times n}$.

(b) $e^{\alpha\mathbf{A}} = \left(e^{\mathbf{A}}\right)^{\alpha}$ for all $\alpha \in \mathcal{C}$ and $\mathbf{A} \in \mathcal{C}^{n\times n}$.

(c) $e^{i\mathbf{A}} = \cos\mathbf{A} + i\sin\mathbf{A}$ for all $\mathbf{A} \in \mathcal{C}^{n\times n}$.

7.9.14. (a) Show that if $\mathbf{AB} = \mathbf{BA}$, then $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$.

(b) Give an example to show that $e^{\mathbf{A}+\mathbf{B}} \neq e^{\mathbf{A}}e^{\mathbf{B}}$ in general.

7.9.15. Find the Hermite interpolation polynomial $p(z)$ as described in Example 7.9.4 such that $p(\mathbf{A}) = e^{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 \\ -3 & -2 & -1 \\ -3 & -2 & -1 \end{pmatrix}$.

7.9.16. The Cayley–Hamilton theorem (pp. 509, 532) says that every $\mathbf{A} \in \mathcal{C}^{n\times n}$ satisfies its own characteristic equation, and this guarantees that $\mathbf{A}^{n+j}$ $(j = 0, 1, 2, \ldots)$ can be expressed as a polynomial in $\mathbf{A}$ of at most degree $n-1$. Since $f(\mathbf{A})$ is always a polynomial in $\mathbf{A}$, the Cayley–Hamilton theorem insures that $f(\mathbf{A})$ can be expressed as a polynomial in $\mathbf{A}$ of at most degree $n-1$. Such a polynomial can be determined whenever $f^{(j)}(\lambda_i)$, $j = 0, 1, \ldots, a_i - 1$ exists for each $\lambda_i \in \sigma(\mathbf{A})$, where $a_i = \textit{alg mult}(\lambda_i)$. The strategy is the same as that in Example 7.9.4 except that a_i is used in place of k_i. If we can find a polynomial $p(z) = \alpha_0 + \alpha_1 z + \cdots + \alpha_{n-1}z^{n-1}$ such that for each $\lambda_i \in \sigma(\mathbf{A})$,

$$p(\lambda_i) = f(\lambda_i), \quad p'(\lambda_i) = f'(\lambda_i), \quad \ldots, \quad p^{(a_i-1)}(\lambda_i) = f^{(a_i-1)}(\lambda_i),$$

then $p(\mathbf{A}) = f(\mathbf{A})$. Why? These equations are an $n \times n$ linear system with the α_i's as the unknowns, and, for the same reason outlined in Example 7.9.4, a solution is always possible.

(a) What advantages and disadvantages does this approach have with respect to the approach in Example 7.9.4?

(b) Use this method to find a polynomial $p(z)$ such that $p(\mathbf{A}) = e^{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 \\ -3 & -2 & -1 \\ -3 & -2 & -1 \end{pmatrix}$. Compare with Exercise 7.9.15.

7.9.17. Show that if f is a function defined at

$$\mathbf{A} = \begin{pmatrix} \alpha & \beta & \gamma \\ 0 & \alpha & \beta \\ 0 & 0 & \alpha \end{pmatrix} = \alpha\mathbf{I} + \beta\mathbf{N} + \gamma\mathbf{N}^2, \quad \text{where} \quad \mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

then $f(\mathbf{A}) = f(\alpha)\mathbf{I} + \beta f'(\alpha)\mathbf{N} + \left[\gamma f'(\alpha) + \dfrac{\beta^2 f''(\alpha)}{2!}\right]\mathbf{N}^2$.

7.9.18. Composition of Matrix Functions. If $h(z) = f(g(z))$, where f and g are functions such that $g(\mathbf{A})$ and $f\big(g(\mathbf{A})\big)$ each exist, then $h(\mathbf{A}) = f\big(g(\mathbf{A})\big)$. However, it's not legal to prove this simply by saying "replace z by $\mathbf{A}$." One way to prove that $h(\mathbf{A}) = f\big(g(\mathbf{A})\big)$ is to demonstrate that $h(\mathbf{J}_\star) = f\big(g(\mathbf{J}_\star)\big)$ for a generic Jordan block and then invoke (7.9.3). Do this for a 3×3 Jordan block—the generalization to $k \times k$ blocks is similar. That is, let $h(z) = f(g(z))$, and use Exercise 7.9.17 to prove that if $g(\mathbf{J}_\star)$ and $f\big(g(\mathbf{J}_\star)\big)$ each exist, then

$$h(\mathbf{J}_\star) = f\big(g(\mathbf{J}_\star)\big) \quad \text{for} \quad \mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}.$$

7.9.19. Prove that if Γ_i is a simple closed contour enclosing $\lambda_i \in \sigma(\mathbf{A})$ but excluding all other eigenvalues of $\mathbf{A}$, then the i^{th} spectral projector is

$$\mathbf{G}_i = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} \mathbf{R}(\xi) d\xi.$$

7.9.20. For $f(z) = z^{-1}$, verify that $f(\mathbf{A}) = \mathbf{A}^{-1}$ for every nonsingular $\mathbf{A}$.

7.9.21. If Γ is a simple closed contour enclosing all eigenvalues of a nonsingular matrix $\mathbf{A}$, what is the value of $\dfrac{1}{2\pi \mathrm{i}} \displaystyle\int_{\Gamma} \xi^{-1} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi$?

7.9.22. Generalized Inverses. The inverse function $f(z) = z^{-1}$ is not defined at singular matrices, but the *generalized inverse function*

$$g(z) = \begin{cases} z^{-1} & \text{if } z \neq 0, \\ 0 & \text{if } z = 0, \end{cases}$$

is defined on all square matrices. It's clear from Exercise 7.9.20 that if $\mathbf{A}$ is nonsingular, then $g(\mathbf{A}) = \mathbf{A}^{-1}$, so $g(\mathbf{A})$ is a natural way to extend the concept of inversion to include singular matrices. Explain why $g(\mathbf{A}) = \mathbf{A}^D$ is the Drazin inverse of Example 5.10.5 (p. 399) and not necessarily the Moore–Penrose pseudoinverse $\mathbf{A}^\dagger$ described on p. 423.

7.9.23. Drazin Is "Natural." Suppose that $\mathbf{A}$ is a singular matrix, and let Γ be a simple closed contour that contains all eigenvalues of $\mathbf{A}$ except $\lambda_1 = 0$, which is neither in nor on Γ. Prove that

$$\frac{1}{2\pi \mathrm{i}} \int_{\Gamma} \xi^{-1} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi = \mathbf{A}^D$$

is the Drazin inverse for $\mathbf{A}$ as defined in Example 5.10.5 (p. 399). **Hint:** The *Cauchy–Goursat theorem* states that if a function f is analytic at all points inside and on a simple closed contour Γ, then $\int_\Gamma f(z) dz = 0$.

7.10 DIFFERENCE EQUATIONS, LIMITS, AND SUMMABILITY

A *linear difference equation* of order m with constant coefficients has the form

$$y(k+1) = \alpha_m y(k) + \alpha_{m-1} y(k-1) \cdots + \alpha_1 y(k-m+1) + \alpha_0 \tag{7.10.1}$$

in which $\alpha_0, \alpha_1, \ldots, \alpha_m$ along with initial conditions $y(0), y(1), \ldots, y(m-1)$ are known constants, and $y(m), y(m+1), y(m+2)\ldots$ are unknown. Difference equations are the discrete analogs of differential equations, and, among other ways, they arise by discretizing differential equations. For example, discretizing a second-order linear differential equation results in a system of second-order difference equations as illustrated in Example 1.4.1, p 19. The theory of linear difference equations parallels the theory for linear differential equations, and a technique similar to the one used to solve linear differential equations with constant coefficients produces the solution of (7.10.1) as

$$y(k) = \frac{\alpha_0}{1 - \alpha_1 - \cdots - \alpha_m} + \sum_{i=1}^{m} \beta_i \lambda_i^k, \qquad \text{for } k = 0, 1, \ldots \tag{7.10.2}$$

in which the λ_i's are the roots of $\lambda^m - \alpha_m \lambda^{m-1} - \cdots - \alpha_0 = 0$, and the β_i's are constants determined by the initial conditions $y(0), y(1), \ldots, y(m-1)$. The first term on the right-hand side of (7.10.2) is a particular solution of (7.10.1), and the summation term in (7.10.2) is the general solution of the associated homogeneous equation defined by setting $\alpha_0 = 0$.

This section focuses on systems of first-order linear difference equations with constant coefficients, and such systems can be written in matrix form as

$$\begin{aligned} &\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) && \text{(a homogeneous system)} \\ \text{or} \quad &\\ &\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{b}(k) && \text{(a nonhomogeneous system),} \end{aligned} \tag{7.10.3}$$

where matrix $\mathbf{A}_{n\times n}$, the initial vector $\mathbf{x}(0)$, and vectors $\mathbf{b}(k)$, $k = 0, 1, \ldots,$ are known. The problem is to determine the unknown vectors $\mathbf{x}(k)$, $k = 1, 2, \ldots,$ along with an expression for the limiting vector $\lim_{k\to\infty} \mathbf{x}(k)$. Such systems are used to model linear discrete-time evolutionary processes, and the goal is usually to predict how (or to where) the process eventually evolves given the initial state of the process. For example, the population migration problem in Example 7.3.5 (p. 531) produces a 2×2 system of homogeneous linear difference equations (7.3.14), and the long-run (or steady-state) population distribution is obtained by finding the limiting solution. More sophisticated applications are given in Example 7.10.8 (p. 635) and Example 8.3.7 (p. 683).

Solving the equations in (7.10.3) is easy. Direct substitution verifies that

$$\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) \quad \text{for} \quad k = 1, 2, 3, \ldots$$

and

$$\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) + \sum_{j=0}^{k-1} \mathbf{A}^{k-j-1}\mathbf{b}(j) \quad \text{for} \quad k = 1, 2, 3, \ldots \tag{7.10.4}$$

are respective solutions to (7.10.3). So rather than finding $\mathbf{x}(k)$ for any finite k, the real problem is to understand the nature of the limiting solution $\lim_{k\to\infty} \mathbf{x}(k)$, and this boils down to analyzing $\lim_{k\to\infty} \mathbf{A}^k$. We begin this analysis by establishing conditions under which $\mathbf{A}^k \to \mathbf{0}$.

For scalars α we know that $\alpha^k \to 0$ if and only if $|\alpha| < 1$, so it's natural to ask if there is an analogous statement for matrices. The first inclination is to replace $|\star|$ by a matrix norm $\|\star\|$, but this doesn't work for the standard norms. For example, if $\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$, then $\mathbf{A}^k \to \mathbf{0}$ but $\|\mathbf{A}\| = 2$ for all of the standard matrix norms. Although it's possible to construct a rather goofy-looking matrix norm $\|\star\|_g$ such that $\|\mathbf{A}\|_g < 1$ when $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$, the underlying mechanisms governing convergence to zero are better understood and analyzed by using eigenvalues and the Jordan form rather than norms. In particular, the ***spectral radius*** of $\mathbf{A}$ defined as $\rho(\mathbf{A}) = \max_{\lambda\in\sigma(\mathbf{A})} |\lambda|$ (Example 7.1.4, p. 497) plays a central role.

Convergence to Zero

$$\text{For } \mathbf{A} \in \mathcal{C}^{n\times n}, \quad \lim_{k\to\infty} \mathbf{A}^k = \mathbf{0} \quad \text{if and only if} \quad \rho(\mathbf{A}) < 1. \tag{7.10.5}$$

Proof. If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is the Jordan form for $\mathbf{A}$, then

$$\mathbf{A}^k = \mathbf{P}\mathbf{J}^k\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star^k & \\ & & \ddots \end{pmatrix}\mathbf{P}^{-1}, \quad \text{where} \quad \mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} \tag{7.10.6}$$

denotes a generic Jordan block in $\mathbf{J}$. Clearly, $\mathbf{A}^k \to \mathbf{0}$ if and only if $\mathbf{J}_\star^k \to \mathbf{0}$ for each Jordan block, so it suffices to prove that $\mathbf{J}_\star^k \to \mathbf{0}$ if and only if $|\lambda| < 1$. Using the function $f(z) = z^n$ in formula (7.9.2) on p. 600 along with the convention that $\binom{k}{j} = 0$ for $j > k$ produces

$$\mathbf{J}_\star^k = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \cdots & \binom{k}{m-1}\lambda^{k-m+1} \\ & \lambda^k & \binom{k}{1}\lambda^{k-1} & \ddots & \vdots \\ & & \ddots & \ddots & \binom{k}{2}\lambda^{k-2} \\ & & & \lambda^k & \binom{k}{1}\lambda^{k-1} \\ & & & & \lambda^k \end{pmatrix}_{m\times m}. \tag{7.10.7}$$

It's clear from the diagonal entries that if $\mathbf{J}_\star^k \to \mathbf{0}$, then $\lambda^k \to 0$, so $|\lambda| < 1$. Conversely, if $|\lambda| < 1$ then $\lim_{k\to\infty}\binom{k}{j}\lambda^{k-j} = 0$ for each fixed value of j because

$$\binom{k}{j} = \frac{k(k-1)\cdots(k-j+1)}{j!} \le \frac{k^j}{j!} \implies \left|\binom{k}{j}\lambda^{k-j}\right| \le \frac{k^j}{j!}|\lambda|^{k-j} \to 0.$$

You can see that the last term on the right-hand side goes to zero as $k \to \infty$ either by applying l'Hopital's rule or by realizing that k^j goes to infinity with polynomial speed while $|\lambda|^{k-j}$ is going to zero with exponential speed. Therefore, if $|\lambda| < 1$, then $\mathbf{J}_\star^k \to \mathbf{0}$, and thus (7.10.5) is proven. ∎

Intimately related to the question of convergence to zero is the convergence of the ***Neumann series*** $\sum_{k=0}^{\infty}\mathbf{A}^k$. It was demonstrated in (3.8.5) on p. 126 that if $\lim_{n\to\infty}\mathbf{A}^n = \mathbf{0}$, then the Neumann series converges, and it was argued in Example 7.3.1 (p. 527) that the converse holds for diagonalizable matrices. Now we are in a position to prove that the converse is true for *all* square matrices and thereby produce the following complete statement regarding the convergence of the Neumann series.

Neumann Series

For $\mathbf{A} \in \mathcal{C}^{n\times n}$, the following statements are equivalent.

- The Neumann series $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots$ converges. (7.10.8)
- $\rho(\mathbf{A}) < 1$. (7.10.9)
- $\lim_{k\to\infty}\mathbf{A}^k = \mathbf{0}$. (7.10.10)

In which case, $(\mathbf{I}-\mathbf{A})^{-1}$ exists and $\sum_{k=0}^{\infty}\mathbf{A}^k = (\mathbf{I}-\mathbf{A})^{-1}$. (7.10.11)

Proof. We know from (7.10.5) that (7.10.9) and (7.10.10) are equivalent, and it was argued on p. 126 that (7.10.10) implies (7.10.8), so the theorem can be established by proving that (7.10.8) implies (7.10.9). If $\sum_{k=0}^{\infty}\mathbf{A}^k$ converges, it follows that $\sum_{k=0}^{\infty}\mathbf{J}_*^k$ must converge for each Jordan block $\mathbf{J}_*$ in the Jordan form for $\mathbf{A}$. This together with (7.10.7) implies that $\left[\sum_{k=0}^{\infty}\mathbf{J}_*^k\right]_{ii} = \sum_{k=0}^{\infty}\lambda^k$ converges for

each $\lambda \in \sigma(\mathbf{A})$, and this scalar geometric series converges if and only if $|\lambda| < 1$. Thus the convergence of $\sum_{k=0}^{\infty} \mathbf{A}^k$ implies $\rho(\mathbf{A}) < 1$. When it converges, $\sum_{k=0}^{\infty} \mathbf{A}^k = (\mathbf{I}-\mathbf{A})^{-1}$ because $(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{k-1}) = \mathbf{I}-\mathbf{A}^k \to \mathbf{I}$ as $k \to \infty$. ∎

The following examples illustrate the utility of the previous results for establishing some useful (and elegant) statements concerning spectral radius.

Example 7.10.1

Spectral Radius as a Limit. It was shown in Example 7.1.4 (p. 497) that if $\mathbf{A} \in \mathcal{C}^{n\times n}$, then $\rho(\mathbf{A}) \le \|\mathbf{A}\|$ for every matrix norm. But this was just the precursor to the following elegant relationship between spectral radius and norm.

Problem: Prove that for every matrix norm,

$$\rho(\mathbf{A}) = \lim_{k\to\infty} \left\|\mathbf{A}^k\right\|^{1/k}. \tag{7.10.12}$$

Solution: First note that $\rho(\mathbf{A})^k = \rho\left(\mathbf{A}^k\right) \le \left\|\mathbf{A}^k\right\| \implies \rho(\mathbf{A}) \le \left\|\mathbf{A}^k\right\|^{1/k}$. Next, observe that $\rho\big(\mathbf{A}/(\rho(\mathbf{A})+\epsilon)\big) < 1$ for every $\epsilon > 0$, so, by (7.10.5),

$$\lim_{k\to\infty} \left(\frac{\mathbf{A}}{\rho(\mathbf{A})+\epsilon}\right)^k = 0 \implies \lim_{k\to\infty} \frac{\left\|\mathbf{A}^k\right\|}{(\rho(\mathbf{A})+\epsilon)^k} = 0.$$

Consequently, there is a positive integer K_ϵ such that $\left\|\mathbf{A}^k\right\|/(\rho(\mathbf{A})+\epsilon)^k < 1$ for all $k \ge K_\epsilon$, so $\left\|\mathbf{A}^k\right\|^{1/k} < \rho(\mathbf{A})+\epsilon$ for all $k \ge K_\epsilon$, and thus

$$\rho(\mathbf{A}) \le \left\|\mathbf{A}^k\right\|^{1/k} < \rho(\mathbf{A})+\epsilon \quad \text{for } k \ge K_\epsilon.$$

Because this holds for each $\epsilon > 0$, it follows that $\lim_{k\to\infty} \left\|\mathbf{A}^k\right\|^{1/k} = \rho(\mathbf{A})$.

Example 7.10.2

For $\mathbf{A} \in \mathcal{C}^{n\times n}$ let $|\mathbf{A}|$ denote the matrix having entries $|a_{ij}|$, and for matrices $\mathbf{B}, \mathbf{C} \in \Re^{n\times n}$ define $\mathbf{B} \le \mathbf{C}$ to mean $b_{ij} \le c_{ij}$ for each i and j.

Problem: Prove that if $|\mathbf{A}| \le \mathbf{B}$, then

$$\rho(\mathbf{A}) \le \rho(|\mathbf{A}|) \le \rho(\mathbf{B}). \tag{7.10.13}$$

Solution: The triangle inequality yields $|\mathbf{A}^k| \le |\mathbf{A}|^k$ for every positive integer k. Furthermore, $|\mathbf{A}| \le \mathbf{B}$ implies that $|\mathbf{A}|^k \le \mathbf{B}^k$. This with (7.10.12) produces

$$\begin{aligned}
\left\|\mathbf{A}^k\right\|_\infty = \left\| \left|\mathbf{A}^k\right| \right\|_\infty &\le \left\| |\mathbf{A}|^k \right\|_\infty \le \left\|\mathbf{B}^k\right\|_\infty \\
&\implies \left\|\mathbf{A}^k\right\|_\infty^{1/k} \le \left\| |\mathbf{A}|^k \right\|_\infty^{1/k} \le \left\|\mathbf{B}^k\right\|_\infty^{1/k} \\
&\implies \lim_{k\to\infty} \left\|\mathbf{A}^k\right\|_\infty^{1/k} \le \lim_{k\to\infty} \le \left\| |\mathbf{A}|^k \right\|_\infty^{1/k} \le \lim_{k\to\infty} \le \left\|\mathbf{B}^k\right\|_\infty^{1/k} \\
&\implies \rho(\mathbf{A}) \le \rho(|\mathbf{A}|) \le \rho(\mathbf{B}).
\end{aligned}$$

Example 7.10.3

Problem: Prove that if $\mathbf{0} \le \mathbf{B}_{n\times n}$, then

$$\rho(\mathbf{B}) < r \text{ if and only if } (r\mathbf{I}-\mathbf{B})^{-1} \text{ exists and } (r\mathbf{I}-\mathbf{B})^{-1} \ge \mathbf{0}. \tag{7.10.14}$$

Solution: If $\rho(\mathbf{B}) < r$, then $\rho(\mathbf{B}/r) < 1$, so (7.10.8)–(7.10.11) imply that

$$r\mathbf{I}-\mathbf{B} = r\Big(\mathbf{I}-\frac{\mathbf{B}}{r}\Big) \text{ is nonsingular and } (r\mathbf{I}-\mathbf{B})^{-1} = \frac{1}{r}\sum_{k=0}^{\infty}\Big(\frac{\mathbf{B}}{r}\Big)^k \ge \mathbf{0}.$$

To prove the converse, it's convenient to adopt the following notation. For any $\mathbf{P} \in \Re^{m\times n}$, let $|\mathbf{P}| = \big[\,|p_{ij}|\,\big]$ denote the matrix of absolute values, and notice that the triangle inequality insures that $|\mathbf{PQ}| \le |\mathbf{P}|\,|\mathbf{Q}|$ for all conformable $\mathbf{P}$ and $\mathbf{Q}$. Now assume that $r\mathbf{I}-\mathbf{B}$ is nonsingular and $(r\mathbf{I}-\mathbf{B})^{-1} \ge \mathbf{0}$, and prove $\rho(\mathbf{B}) < r$. Let $(\lambda, \mathbf{x})$ be any eigenpair for $\mathbf{B}$, and use $\mathbf{B} \ge \mathbf{0}$ together with $(r\mathbf{I}-\mathbf{B})^{-1} \ge \mathbf{0}$ to write

$$\begin{aligned}
\lambda\mathbf{x} = \mathbf{B}\mathbf{x} &\implies |\lambda|\,|\mathbf{x}| = |\lambda\mathbf{x}| = |\mathbf{B}\mathbf{x}| \le |\mathbf{B}|\,|\mathbf{x}| = \mathbf{B}\,|\mathbf{x}| \\
&\implies (r\mathbf{I}-\mathbf{B})|\mathbf{x}| \le (r-|\lambda|)\,|\mathbf{x}| \\
&\implies \mathbf{0} \le |\mathbf{x}| \le (r-|\lambda|)\,(r\mathbf{I}-\mathbf{B})^{-1}|\mathbf{x}| \qquad (7.10.15) \\
&\implies r-|\lambda| \ge 0.
\end{aligned}$$

But $|\lambda| \ne r$; otherwise (7.10.15) would imply that $|\mathbf{x}|$ (and hence $\mathbf{x}$) is zero, which is impossible. Thus $|\lambda| < r$ for all $\lambda \in \sigma(\mathbf{B})$, which means $\rho(\mathbf{B}) < r$.

Iterative algorithms are often used in lieu of direct methods to solve large sparse systems of linear equations, and some of the traditional iterative schemes fall into the following class of nonhomogeneous linear difference equations.

Linear Stationary Iterations

Let $\mathbf{Ax} = \mathbf{b}$ be a linear system that is square but otherwise arbitrary.

- A ***splitting*** of $\mathbf{A}$ is a factorization $\mathbf{A} = \mathbf{M}-\mathbf{N}$, where $\mathbf{M}^{-1}$ exists.
- Let $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ (called the ***iteration matrix***), and set $\mathbf{d} = \mathbf{M}^{-1}\mathbf{b}$.
- For an initial vector $\mathbf{x}(0)_{n\times 1}$, a ***linear stationary iteration*** is

$$\mathbf{x}(k) = \mathbf{H}\mathbf{x}(k-1) + \mathbf{d}, \qquad k = 1, 2, 3, \ldots . \tag{7.10.16}$$

- If $\rho(\mathbf{H}) < 1$, then $\mathbf{A}$ is nonsingular and

$$\lim_{k\to\infty} \mathbf{x}(k) = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \quad \text{for every initial vector } \mathbf{x}(0). \tag{7.10.17}$$

Proof. To prove (7.10.17), notice that if $\mathbf{A} = \mathbf{M} - \mathbf{N} = \mathbf{M}(\mathbf{I} - \mathbf{H})$ is a splitting for which $\rho(\mathbf{H}) < 1$, then (7.10.11) guarantees that $(\mathbf{I} - \mathbf{H})^{-1}$ exists, and thus $\mathbf{A}$ is nonsingular. Successive substitution applied to (7.10.16) yields

$$\mathbf{x}(k) = \mathbf{H}^k\mathbf{x}(0) + (\mathbf{I} + \mathbf{H} + \mathbf{H}^2 + \cdots + \mathbf{H}^{k-1})\mathbf{d},$$

so if $\rho(\mathbf{H}) < 1$, then (7.10.9)–(7.10.11) insures that for all $\mathbf{x}(0)$,

$$\lim_{k\to\infty} \mathbf{x}(k) = (\mathbf{I} - \mathbf{H})^{-1}\mathbf{d} = (\mathbf{I} - \mathbf{H})^{-1}\mathbf{M}^{-1}\mathbf{b} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{x}. \quad \blacksquare \tag{7.10.18}$$

It's clear that the convergence rate of (7.10.16) is governed by the size of $\rho(\mathbf{H})$ along with the index of its associated eigenvalue (go back and look at (7.10.7)). But what really is needed is an indication of how many digits of accuracy can be expected to be gained per iteration. So as not to obscure the simple underlying idea, assume that $\mathbf{H}_{n\times n}$ is diagonalizable with

$$\sigma(\mathbf{H}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}, \quad \text{where} \quad 1 > |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_s|$$

(which is frequently the case in applications), and let $\boldsymbol{\epsilon}(k) = \mathbf{x}(k) - \mathbf{x}$ denote the error after the k^{th} iteration. Subtracting $\mathbf{x} = \mathbf{H}\mathbf{x} + \mathbf{d}$ (a consequence of (7.10.18)) from $\mathbf{x}(k) = \mathbf{H}\mathbf{x}(k-1) + \mathbf{d}$ produces (for large k)

$$\boldsymbol{\epsilon}(k) = \mathbf{H}\boldsymbol{\epsilon}(k-1) = \mathbf{H}^k\boldsymbol{\epsilon}(0) = (\lambda_1^k\mathbf{G}_1 + \lambda_2^k\mathbf{G}_2 + \cdots + \lambda_s^k\mathbf{G}_s)\boldsymbol{\epsilon}(0) \approx \lambda_1^k\mathbf{G}_1\boldsymbol{\epsilon}(0),$$

where the $\mathbf{G}_i$'s are the spectral projectors occurring in the spectral decomposition (pp. 517 and 520) of $\mathbf{H}^k$. Similarly, $\boldsymbol{\epsilon}(k-1) \approx \lambda_1^{k-1}\mathbf{G}_1\boldsymbol{\epsilon}(0)$, so comparing the i^{th} components of $\boldsymbol{\epsilon}(k-1)$ and $\boldsymbol{\epsilon}(k)$ reveals that after several iterations,

$$\left|\frac{\boldsymbol{\epsilon}_i(k-1)}{\boldsymbol{\epsilon}_i(k)}\right| \approx \frac{1}{|\lambda_1|} = \frac{1}{\rho(\mathbf{H})} \quad \text{for each} \quad i = 1, 2, \ldots, n.$$

To understand the significance of this, suppose for example that

$$|\boldsymbol{\epsilon}_i(k-1)| = 10^{-q} \quad \text{and} \quad |\boldsymbol{\epsilon}_i(k)| = 10^{-p} \quad \text{with} \quad p \geq q > 0,$$

so that the error in each entry is reduced by $p - q$ digits per iteration. Since

$$p - q = \log_{10}\left|\frac{\boldsymbol{\epsilon}_i(k-1)}{\boldsymbol{\epsilon}_i(k)}\right| \approx -\log_{10}\rho(\mathbf{H}),$$

we see that $-\log_{10}\rho(\mathbf{H})$ provides us with an indication of the number of digits of accuracy that can be expected to be eventually gained on each iteration. For this reason, the number $R = -\log_{10}\rho(\mathbf{H})$ (or, alternately, $R = -\ln\rho(\mathbf{H})$) is called the ***asymptotic rate of convergence***, and this is the primary tool for comparing different linear stationary iterative algorithms.

The trick is to find splittings that guarantee rapid convergence while insuring that $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ and $\mathbf{d} = \mathbf{M}^{-1}\mathbf{b}$ can be computed easily. The following three examples present the classical splittings.

Example 7.10.4

Jacobi's method [81] is produced by splitting $\mathbf{A} = \mathbf{D} - \mathbf{N}$, where $\mathbf{D}$ is the diagonal part of $\mathbf{A}$ (we assume each $a_{ii} \neq 0$), and $-\mathbf{N}$ is the matrix containing the off-diagonal entries of $\mathbf{A}$. Clearly, both $\mathbf{H} = \mathbf{D}^{-1}\mathbf{N}$ and $\mathbf{d} = \mathbf{D}^{-1}\mathbf{b}$ can be formed with little effort. Notice that the i^{th} component in the Jacobi iteration $\mathbf{x}(k) = \mathbf{D}^{-1}\mathbf{N}\mathbf{x}(k-1) + \mathbf{D}^{-1}\mathbf{b}$ is given by

$$x_i(k) = \big(b_i - \textstyle\sum_{j \neq i} a_{ij}x_j(k-1)\big)/a_{ii}. \tag{7.10.19}$$

This shows that the order in which the equations are considered is irrelevant and that the algorithm can process equations independently (or in parallel). For this reason, Jacobi's method was referred to in the 1940s as the *method of simultaneous displacements*.

Problem: Explain why Jacobi's method is guaranteed to converge for all initial vectors $\mathbf{x}(0)$ and for all right-hand sides $\mathbf{b}$ when $\mathbf{A}$ is diagonally dominant as defined and discussed in Examples 4.3.3 (p. 184) and 7.1.6 (p. 499).

Solution: According to (7.10.17), it suffices to show that $\rho(\mathbf{H}) < 1$. This follows by combining $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ for each i with the fact that $\rho(\mathbf{H}) \leq \|H\|_\infty$ (Example 7.1.4, p. 497) to write

$$\rho(\mathbf{H}) \leq \|H\|_\infty = \max_i \sum_j \frac{|a_{ij}|}{|a_{ii}|} = \max_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Example 7.10.5

The Gauss–Seidel method [82] is the result of splitting $\mathbf{A} = (\mathbf{D} - \mathbf{L}) - \mathbf{U}$, where $\mathbf{D}$ is the diagonal part of $\mathbf{A}$ ($a_{ii} \neq 0$ is assumed) and where $-\mathbf{L}$ and $-\mathbf{U}$ contain the entries occurring below and above the diagonal of $\mathbf{A}$, respectively. The iteration matrix is $\mathbf{H} = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}$, and $\mathbf{d} = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{b}$. The i^{th} entry in the Gauss–Seidel iteration $\mathbf{x}(k) = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}\mathbf{x}(k-1) + (\mathbf{D} - \mathbf{L})^{-1}\mathbf{b}$ is

$$x_i(k) = \big(b_i - \textstyle\sum_{j<i} a_{ij}x_j(k) - \sum_{j>i} a_{ij}x_j(k-1)\big)/a_{ii}. \tag{7.10.20}$$

This shows that Gauss–Seidel determines $x_i(k)$ by using the newest possible information—namely, $x_1(k), x_2(k), \ldots, x_{i-1}(k)$ in the current iterate in conjunction with $x_{i+1}(k-1), x_{i+2}(k-1), \ldots, x_n(k-1)$ from the previous iterate.

---

[81] Karl Jacobi (p. 353) considered this method in 1845, but it seems to have been independently discovered by others. In addition to being called the *method of simultaneous displacements* in 1945, Jacobi's method was referred to as the *Richardson iterative method* in 1958.

[82] Ludwig Philipp von Seidel (1821–1896) studied with Dirichlet in Berlin in 1840 and with Jacobi (and others) in Königsberg. Seidel's involvement in transforming Jacobi's method into the Gauss–Seidel scheme is natural, but the reason for attaching Gauss's name is unclear. Seidel went on to earn his doctorate (1846) in Munich, where he stayed as a professor for the rest of his life. In addition to mathematics, Seidel made notable contributions in the areas of optics and astronomy, and in 1970 a lunar crater was named for Seidel.

This differs from Jacobi's method because Jacobi relies strictly on the old data in $\mathbf{x}(k-1)$. The Gauss–Seidel algorithm was known in the 1940s as the *method of successive displacements* (as opposed to the method of *simultaneous displacements*, which is Jacobi's method). Because Gauss–Seidel computes $x_i(k)$ with newer data than that used by Jacobi, it appears at first glance that Gauss–Seidel should be the superior algorithm. While this is often the case, it is not universally true—see Exercise 7.10.7.

Other Comparisons. Another major difference between Gauss–Seidel and Jacobi is that the order in which the equations are processed is irrelevant for Jacobi's method, but the value (not just the position) of the components $x_i(k)$ in the Gauss–Seidel iterate can change when the order of the equations is changed. Since this ordering feature can affect the performance of the algorithm, it was the object of much study at one time. Furthermore, when core memory is a concern, Gauss–Seidel enjoys an advantage because as soon as a new component $x_i(k)$ is computed, it can immediately replace the old value $x_i(k-1)$, whereas Jacobi requires all old values in $\mathbf{x}(k-1)$ to be retained until all new values in $\mathbf{x}(k)$ have been determined. Something that both algorithms have in common is that diagonal dominance in $\mathbf{A}$ guarantees global convergence of each method.

Problem: Explain why diagonal dominance in $\mathbf{A}$ is sufficient to guarantee convergence of the Gauss–Seidel method for all initial vectors $\mathbf{x}(0)$ and for all right-hand sides $\mathbf{b}$.

Solution: Show $\rho(\mathbf{H}) < 1$. Let $(\lambda, \mathbf{z})$ be any eigenpair for $\mathbf{H}$, and suppose that the component of maximal magnitude in $\mathbf{z}$ occurs in position m. Write $(\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\mathbf{z} = \lambda\mathbf{z}$ as $\lambda(\mathbf{D}-\mathbf{L})\mathbf{z} = \mathbf{U}\mathbf{z}$, and write the m^{th} row of this latter equation as $\lambda(d-l) = u$, where

$$d = a_{mm}z_m, \quad l = -\sum_{j<m} a_{mj}z_j, \quad \text{and} \quad u = -\sum_{j>m} a_{mj}z_j.$$

Diagonal dominance $|a_{mm}| > \sum_{j\neq m}|a_{mj}|$ and $|z_j| \leq |z_m|$ for all j yields

$$\begin{aligned} |u| + |l| &= \Big|\sum_{j<m} a_{mj}z_j\Big| + \Big|\sum_{j>m} a_{mj}z_j\Big| \leq |z_m|\Big(\sum_{j<m}|a_{mj}| + \sum_{j>m}|a_{mj}|\Big) \\ &< |z_m||a_{mm}| = |d| \quad \Longrightarrow \quad |u| < |d| - |l|. \end{aligned}$$

This together with $\lambda(d-l) = u$ and the backward triangle inequality (Example 5.1.1, p. 273) produces the conclusion that

$$|\lambda| = \frac{|u|}{|d-l|} \leq \frac{|u|}{|d|-|l|} < 1, \quad \text{and thus} \quad \rho(\mathbf{H}) < 1.$$

Note: Diagonal dominance in $\mathbf{A}$ guarantees convergence for both Jacobi and Gauss–Seidel, but diagonal dominance is a rather severe condition that is often

not present in applications. For example the linear system in Example 7.6.2 (p. 563) that results from discretizing Laplace's equation on a square is not diagonally dominant (e.g., look at the fifth row in the 9×9 system on p. 564). But such systems are always positive definite (Example 7.6.2), and there is a classical theorem stating that *if* $\mathbf{A}$ *is positive definite, then the Gauss–Seidel iteration converges to the solution of* $\mathbf{Ax} = \mathbf{b}$ *for every initial vector* $\mathbf{x}(0)$. The same cannot be said for Jacobi's method, but there are matrices (the *M-matrices* of Example 7.10.7, p. 626) having properties resembling positive definiteness for which Jacobi's method is guaranteed to converge—see (7.10.29).

Example 7.10.6

The successive overrelaxation (SOR) method improves on Gauss–Seidel by introducing a real number $\omega \neq 0$, called a ***relaxation parameter,*** to form the splitting $\mathbf{A} = \mathbf{M} - \mathbf{N}$, where $\mathbf{M} = \omega^{-1}\mathbf{D} - \mathbf{L}$ and $\mathbf{N} = (\omega^{-1} - 1)\mathbf{D} + \mathbf{U}$. As before, $\mathbf{D}$ is the diagonal part of $\mathbf{A}$ ($a_{ii} \neq 0$ is assumed) and $-\mathbf{L}$ and $-\mathbf{U}$ contain the entries occurring below and above the diagonal of $\mathbf{A}$, respectively. Since $\mathbf{M}^{-1} = \omega(\mathbf{D} - \omega\mathbf{L})^{-1} = \omega(\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}$, the SOR iteration matrix is

$$\mathbf{H}_\omega = \mathbf{M}^{-1}\mathbf{N} = (\mathbf{D} - \omega\mathbf{L})^{-1}\big[(1-\omega)\mathbf{D} + \omega\mathbf{U}\big] = (\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}\big[(1-\omega)\mathbf{I} + \omega\mathbf{D}^{-1}\mathbf{U}\big],$$

and the k^{th} SOR iterate emanating from (7.10.16) is

$$\mathbf{x}(k) = \mathbf{H}_\omega\mathbf{x}(k-1) + \omega(\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\mathbf{b}. \tag{7.10.21}$$

This is the Gauss–Seidel iteration when $\omega = 1$. Using $\omega > 1$ is called *overrelaxation*, while taking $\omega < 1$ is referred to as *underrelaxation*. Writing (7.10.21) in the form $(\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})\mathbf{x}(k) = \big[(1-\omega)\mathbf{I} + \omega\mathbf{D}^{-1}\mathbf{U}\big]\mathbf{x}(k-1) + \omega\mathbf{D}^{-1}\mathbf{b}$ and considering the i^{th} component on both sides of this equality produces

$$x_i(k) = (1-\omega)x_i(k-1) + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}x_j(k) - \sum_{j>i} a_{ij}x_j(k-1)\Big). \tag{7.10.22}$$

The matrix splitting approach is elegant and unifying, but it obscures the simple idea behind SOR. To understand the original motivation, write the Gauss–Seidel iterate in (7.10.20) as $\widetilde{x}_i(k) = \widetilde{x}_i(k-1) + c_k$, where c_k is the "correction term"

$$c_k = \frac{1}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}\widetilde{x}_j(k) - \sum_{j=i}^{n} a_{ij}\widetilde{x}_j(k-1)\Big).$$

This clearly suggests that the performance of the iteration can be affected by adjusting (or "relaxing") the correction term—i.e., by replacing c_k with ωc_k. The resulting algorithm, $\widetilde{x}_i(k) = \widetilde{x}_i(k-1) + \omega c_k$, is in fact (7.10.22), which produces (7.10.21). Moreover, it was observed early on that Gauss–Seidel applied to finite difference approximations for elliptic partial differential equations, such

as the one in Example 7.6.2 (p. 563), often produces successive corrections c_k that have the same sign, so it was reasoned that convergence might be accelerated for these applications by increasing the magnitude of the correction factor at each step (i.e., by setting $\omega > 1$). Thus the technique became known as "successive overrelaxation" rather than simply "successive relaxation." It's not hard to see that $\rho(\mathbf{H}_\omega) < 1$ only if $0 < \omega < 2$ (Exercise 7.10.9), and it can be proven that positive definiteness of $\mathbf{A}$ is sufficient to guarantee $\rho(\mathbf{H}_\omega) < 1$ whenever $0 < \omega < 2$. But determining ω to minimize $\rho(\mathbf{H}_\omega)$ is generally a difficult task.

Nevertheless, there is one famous special case[83] for which the optimal value of ω can be explicitly given. If $\det(\alpha\mathbf{D} - \mathbf{L} - \mathbf{U}) = \det\left(\alpha\mathbf{D} - \beta\mathbf{L} - \beta^{-1}\mathbf{U}\right)$ for all real α and $\beta \neq 0$, and if the iteration matrix $\mathbf{H}_J$ for Jacobi's method has real eigenvalues with $\rho(\mathbf{H}_J) < 1$, then the eigenvalues λ_J for $\mathbf{H}_J$ are related to the eigenvalues λ_ω of $\mathbf{H}_\omega$ by

$$(\lambda_\omega + \omega - 1)^2 = \omega^2 \lambda_J^2 \lambda_\omega. \tag{7.10.23}$$

From this it can be proven that the optimum value of ω for SOR is

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathbf{H}_J)}} \quad \text{and} \quad \rho\left(\mathbf{H}_{\omega_{\text{opt}}}\right) = \omega_{\text{opt}} - 1. \tag{7.10.24}$$

Furthermore, setting $\omega = 1$ in (7.10.23) yields $\rho(\mathbf{H}_{GS}) = \rho^2(\mathbf{H}_J)$, where $\mathbf{H}_{GS}$ is the Gauss–Seidel iteration matrix. For example, the discrete Laplacian $\mathbf{L}_{n^2 \times n^2}$ in Example 7.6.2 (p. 563) satisfies the special case conditions, and the spectral radii of the iteration matrices associated with $\mathbf{L}$ are

$$\begin{aligned}
\text{Jacobi:} \quad & \rho(\mathbf{H}_J) = \cos \pi h \approx 1 - (\pi^2 h^2/2) \quad \text{(see Exercise 7.10.10)}, \\
\text{Gauss–Seidel:} \quad & \rho(\mathbf{H}_{GS}) = \cos^2 \pi h \approx 1 - \pi^2 h^2, \\
\text{SOR:} \quad & \rho\left(\mathbf{H}_{\omega_{\text{opt}}}\right) = \frac{1 - \sin \pi h}{1 + \sin \pi h} \approx 1 - 2\pi h,
\end{aligned}$$

where we have set $h = 1/(n+1)$. Examining asymptotic rates of convergence reveals that Gauss–Seidel is twice as fast as Jacobi on the discrete Laplacian because $R_{GS} = -\log_{10} \cos^2 \pi h = -2 \log_{10} \cos \pi h = 2R_J$. However, optimal SOR is much better because $1 - 2\pi h$ is significantly smaller than $1 - \pi^2 h^2$ for even moderately small h. The point is driven home by looking at the asymptotic rates of convergence for $h = .02$ $(n = 49)$ as shown below:

$$\begin{aligned}
\text{Jacobi:} \quad & R_J \approx .000858, \\
\text{Gauss–Seidel:} \quad & R_{GS} = 2R_J \approx .001716, \\
\text{SOR:} \quad & R_{\text{opt}} \approx .054611 \approx 32 R_{GS} = 64 R_J.
\end{aligned}$$

[83] This special case was developed by the contemporary numerical analyst David M. Young, Jr., who produced much of the SOR theory in his 1950 Ph.D. dissertation that was directed by Garrett Birkhoff at Harvard University. The development of SOR is considered to be one of the major computational achievements of the first half of the twentieth century, and it motivated at least two decades of intense effort in matrix computations.

In other words, after things settle down, a single SOR step on $\mathbf{L}$ (for $h = .02$) is equivalent to about 32 Gauss–Seidel steps and 64 Jacobi steps!

Note: In spite of the preceding remarks, SOR has limitations. Special cases for which the optimum ω can be explicitly determined are rare, so adaptive computational procedures are generally necessary to approximate a good ω, and the results are often not satisfying. While SOR was a big step forward over the algorithms of the nineteenth century, the second half of the twentieth century saw the development of more robust methods—such as the preconditioned conjugate gradient method (p. 657) and GMRES (p. 655)—that have relegated SOR to a secondary role.

Example 7.10.7

M-matrices[84] are real nonsingular matrices $\mathbf{A}_{n\times n}$ such that $a_{ij} \leq 0$ for all $i \neq j$ and $\mathbf{A}^{-1} \geq \mathbf{0}$ (each entry of $\mathbf{A}^{-1}$ is nonnegative). They arise naturally in a broad variety of applications ranging from economics (Example 8.3.6, p. 681) to hard-core engineering problems, and, as shown in (7.10.29), they are particularly relevant in formulating and analyzing iterative methods. Some important properties of M-matrices are developed below.

- $\mathbf{A}$ is an M-matrix if and only if there exists a matrix $\mathbf{B} \geq \mathbf{0}$ and a real number $r > \rho(\mathbf{B})$ such that $\mathbf{A} = r\mathbf{I} - \mathbf{B}$. (7.10.25)

- If $\mathbf{A}$ is an M-matrix, then $\mathrm{Re}\,(\lambda) > 0$ for all $\lambda \in \sigma(\mathbf{A})$. Conversely, all matrices with nonpositive off-diagonal entries whose spectrums are in the right-hand halfplane are M-matrices. (7.10.26)

- Principal submatrices of M-matrices are also M-matrices. (7.10.27)

- If $\mathbf{A}$ is an M-matrix, then all principal minors in $\mathbf{A}$ are positive. Conversely, all matrices with nonpositive off-diagonal entries whose principal minors are positive are M-matrices. (7.10.28)

- If $\mathbf{A} = \mathbf{M} - \mathbf{N}$ is a splitting of an M-matrix for which $\mathbf{M}^{-1} \geq \mathbf{0}$, then the linear stationary iteration (7.10.16) is convergent for all initial vectors $\mathbf{x}(0)$ and for all right-hand sides $\mathbf{b}$. In particular, Jacobi's method in Example 7.10.4 (p. 622) converges for all M-matrices. (7.10.29)

Proof of (7.10.25). Suppose that $\mathbf{A}$ is an M-matrix, and let $r = \max_i |a_{ii}|$ so that $\mathbf{B} = r\mathbf{I} - \mathbf{A} \geq \mathbf{0}$. Since $\mathbf{A}^{-1} = (r\mathbf{I} - \mathbf{B})^{-1} \geq \mathbf{0}$, it follows from (7.10.14) in Example 7.10.3 (p. 620) that $r > \rho(\mathbf{B})$. Conversely, if $\mathbf{A}$ is any matrix of

[84] This terminology was introduced in 1937 by the twentieth-century mathematician Alexander Markowic Ostrowski, who made several contributions to the analysis of classical iterative methods. The "M" is short for "Minkowski" (p. 278).

the form $\mathbf{A} = r\mathbf{I} - \mathbf{B}$, where $\mathbf{B} \geq \mathbf{0}$ and $r > \rho(\mathbf{B})$, then (7.10.14) guarantees that $\mathbf{A}^{-1}$ exists and $\mathbf{A}^{-1} \geq \mathbf{0}$, and it's clear that $a_{ij} \leq 0$ for each $i \neq j$, so $\mathbf{A}$ must be an M-matrix. ∎

Proof of (7.10.26). If $\mathbf{A}$ is an M-matrix, then, by (7.10.25), $\mathbf{A} = r\mathbf{I} - \mathbf{B}$, where $r > \rho(\mathbf{B})$. This means that if $\lambda_{\mathbf{A}} \in \sigma(\mathbf{A})$, then $\lambda_{\mathbf{A}} = r - \lambda_{\mathbf{B}}$ for some $\lambda_{\mathbf{B}} \in \sigma(\mathbf{B})$. If $\lambda_{\mathbf{B}} = \alpha + \mathrm{i}\beta$, then $r > \rho(\mathbf{B}) \geq |\lambda_{\mathbf{B}}| = \sqrt{\alpha^2 + \beta^2} \geq |\alpha| \geq \alpha$ implies that $\mathrm{Re}(\lambda_{\mathbf{A}}) = r - \alpha \geq 0$. Now suppose that $\mathbf{A}$ is any matrix such that $a_{ij} \leq 0$ for all $i \neq j$ and $\mathrm{Re}(\lambda_{\mathbf{A}}) > 0$ for all $\lambda_{\mathbf{A}} \in \sigma(\mathbf{A})$. This means that there is a real number γ such that the circle centered at γ and having radius equal to γ contains $\sigma(\mathbf{A})$—see Figure 7.10.1. Let r be any real number such that $r > \max\{2\gamma, \max_i |a_{ii}|\}$, and set $\mathbf{B} = r\mathbf{I} - \mathbf{A}$. It's apparent that $\mathbf{B} \geq \mathbf{0}$, and, as can be seen from Figure 7.10.1, the distance $|r - \lambda_{\mathbf{A}}|$ between r and every point in $\sigma(\mathbf{A})$ is less than r.

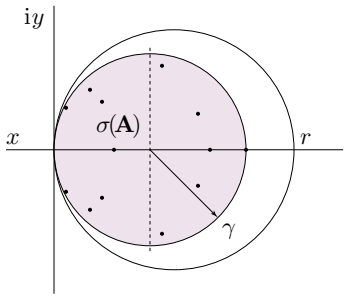


Figure 7.10.1

All eigenvalues of $\mathbf{B}$ look like $\lambda_{\mathbf{B}} = r - \lambda_{\mathbf{A}}$, and $|\lambda_{\mathbf{B}}| = |r - \lambda_{\mathbf{A}}| < r$, so $\rho(\mathbf{B}) < r$. Since $\mathbf{A} = r\mathbf{I} - \mathbf{B}$ is nonsingular (because $0 \notin \sigma(\mathbf{A})$) with $\mathbf{B} \geq \mathbf{0}$ and $r > \rho(\mathbf{B})$, it follows from (7.10.14) in Example 7.10.3 (p. 620) that $\mathbf{A}^{-1} \geq \mathbf{0}$, and thus $\mathbf{A}$ is an M-matrix. ∎

Proof of (7.10.27). If $\widetilde{\mathbf{A}}_{k\times k}$ is the principal submatrix lying on the intersection of rows and columns $i_1, \ldots, i_k$ in an M-matrix $\mathbf{A} = r\mathbf{I} - \mathbf{B}$, where $\mathbf{B} \geq \mathbf{0}$ and $r > \rho(\mathbf{B})$, then $\widetilde{\mathbf{A}} = r\mathbf{I} - \widetilde{\mathbf{B}}$, where $\widetilde{\mathbf{B}} \geq \mathbf{0}$ is the corresponding principal submatrix of $\mathbf{B}$. Let $\mathbf{P}$ be a permutation matrix such that

$$\mathbf{P}^T\mathbf{B}\mathbf{P} = \begin{pmatrix} \widetilde{\mathbf{B}} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}, \text{ or } \mathbf{B} = \mathbf{P}\begin{pmatrix} \widetilde{\mathbf{B}} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}\mathbf{P}^T, \text{ and let } \mathbf{C} = \mathbf{P}\begin{pmatrix} \widetilde{\mathbf{B}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^T.$$

Clearly, $\mathbf{0} \leq \mathbf{C} \leq \mathbf{B}$, so, by (7.10.13) on p. 619, $\rho(\widetilde{\mathbf{B}}) = \rho(\mathbf{C}) \leq \rho(\mathbf{B}) < r$. Consequently, (7.10.25) insures that $\widetilde{\mathbf{A}}$ is an M-matrix. ∎

Proof of (7.10.28). If $\mathbf{A}$ is an M-matrix, then $\det(\mathbf{A}) > 0$ because the eigenvalues of a real matrix appear in complex conjugate pairs, so (7.10.26) and (7.1.8),

p. 494, guarantee that $\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i > 0$. It follows that each principal minor is positive because each submatrix of an M-matrix is again an M-matrix. Now prove that if $\mathbf{A}_{n\times n}$ is a matrix such that $a_{ij} \leq 0$ for $i \neq j$ and each principal minor is positive, then $\mathbf{A}$ must be an M-matrix. Proceed by induction on n. For $n = 1$, the assumption of positive principal minors implies that $\mathbf{A} = [\rho]$ with $\rho > 0$, so $\mathbf{A}^{-1} = 1/\rho > 0$. Suppose the result is true for $n = k$, and consider the LU factorization

$$\mathbf{A}_{(k+1)\times(k+1)} = \begin{pmatrix} \widetilde{\mathbf{A}}_{k\times k} & \mathbf{c} \\ \mathbf{d}^T & \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{d}^T\widetilde{\mathbf{A}}^{-1} & 1 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{A}} & \mathbf{c} \\ \mathbf{0} & \alpha - \mathbf{d}^T\widetilde{\mathbf{A}}^{-1}\mathbf{c} \end{pmatrix} = \mathbf{LU}.$$

We know that $\mathbf{A}$ is nonsingular ($\det(\mathbf{A})$ is a principal minor) and $\alpha > 0$ (it's a 1×1 principal minor), and the induction hypothesis insures that $\widetilde{\mathbf{A}}^{-1} \geq \mathbf{0}$. Combining these facts with $\mathbf{c} \leq \mathbf{0}$ and $\mathbf{d}^T \leq \mathbf{0}$ produces

$$\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1} = \begin{pmatrix} \widetilde{\mathbf{A}}^{-1} & \dfrac{-\widetilde{\mathbf{A}}^{-1}\mathbf{c}}{\alpha - \mathbf{d}^T\widetilde{\mathbf{A}}^{-1}\mathbf{c}} \\ \mathbf{0} & \dfrac{1}{\alpha - \mathbf{d}^T\widetilde{\mathbf{A}}^{-1}\mathbf{c}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{d}^T\widetilde{\mathbf{A}}^{-1} & 1 \end{pmatrix} \geq \mathbf{0},$$

and thus the induction argument is completed. ∎

Proof of (7.10.29). If $\mathbf{A} = \mathbf{M} - \mathbf{N}$ is an M-matrix, and if $\mathbf{M}^{-1} \geq \mathbf{0}$ and $\mathbf{N} \geq \mathbf{0}$, then the iteration matrix $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ is clearly nonnegative. Furthermore,

$$(\mathbf{I} - \mathbf{H})^{-1} - \mathbf{I} = (\mathbf{I} - \mathbf{H})^{-1}\mathbf{H} = \mathbf{A}^{-1}\mathbf{N} \geq \mathbf{0} \implies (\mathbf{I} - \mathbf{H})^{-1} \geq \mathbf{I} \geq \mathbf{0},$$

so (7.10.14) in Example 7.10.3 (p. 620) insures that $\rho(\mathbf{H}) < 1$. Convergence of Jacobi's method is a special case because the Jacobi splitting is $\mathbf{A} = \mathbf{D} - \mathbf{N}$, where $\mathbf{D} = \text{diag}\,(a_{11}, a_{22}, \ldots, a_{nn})$, and (7.10.28) implies that each $a_{ii} > 0$. ∎

Note: Comparing properties of M-matrices with those of positive definite matrices reveals many parallels, and, in a rough sense, an M-matrix often plays the role of "a poor man's positive definite matrix." Only a small sample of M-matrix theory has been presented here, but there is in fact enough to fill a monograph on the subject. For example, there are at least 50 known equivalent conditions that can be imposed on a real matrix with nonpositive off-diagonal entries (often called a *Z-matrix*) to guarantee that it is an M-matrix—see Exercise 7.10.12 for a sample of such conditions in addition to those listed above.

We now focus on broader issues concerning when $\lim_{k\to\infty}\mathbf{A}^k$ exists but may be nonzero. Start from the fact that $\lim_{k\to\infty}\mathbf{A}^k$ exists if and only if $\lim_{k\to\infty}\mathbf{J}_\star^k$ exists for each Jordan block in (7.10.6). It's clear from (7.10.7) that $\lim_{k\to\infty}\mathbf{J}_\star^k$ cannot exist when $|\lambda|>1$, and we already know the story for $|\lambda|<1$, so we only have to examine the case when $|\lambda|=1$. If $|\lambda|=1$ with $\lambda\neq 1$ (i.e., $\lambda=e^{i\theta}$ with $0<\theta<2\pi$), then the diagonal terms λ^k oscillate indefinitely, and this prevents $\mathbf{J}_\star^k$ (and $\mathbf{A}^k$) from having a limit. When $\lambda=1$,

$$\mathbf{J}_\star^k=\begin{pmatrix}1 & \binom{k}{1} & \cdots & \binom{k}{m-1}\\ & \ddots & \ddots & \vdots\\ & & \ddots & \binom{k}{1}\\ & & & 1\end{pmatrix}_{m\times m} \tag{7.10.30}$$

has a limiting value if and only if $m=1$, which is equivalent to saying that $\lambda=1$ is a semisimple eigenvalue. But $\lambda=1$ may be repeated p times so that there are p Jordan blocks of the form $\mathbf{J}_\star=[1]_{1\times 1}$. Consequently, $\lim_{k\to\infty}\mathbf{A}^k$ exists if and only if the Jordan form for $\mathbf{A}$ has the structure

$$\mathbf{J}=\mathbf{P}^{-1}\mathbf{A}\mathbf{P}=\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}\end{pmatrix},\ \text{where } p=\textit{alg mult}(1)\ \text{ and }\ \rho(\mathbf{K})<1. \tag{7.10.31}$$

Now that we know when $\lim_{k\to\infty}\mathbf{A}^k$ exists, let's describe what $\lim_{k\to\infty}\mathbf{A}^k$ looks like. We already know the answer when $p=0$—it's $\mathbf{0}$ (because $\rho(\mathbf{A})<1$). But when p is nonzero, $\lim_{k\to\infty}\mathbf{A}^k\neq\mathbf{0}$, and it can be evaluated in a couple of different ways. One way is to partition $\mathbf{P}=\begin{pmatrix}\mathbf{P}_1\,|\,\mathbf{P}_2\end{pmatrix}$ and $\mathbf{P}^{-1}=\begin{pmatrix}\mathbf{Q}_1\\ \mathbf{Q}_2\end{pmatrix}$, and use (7.10.5) and (7.10.31) to write

$$\begin{aligned}\lim_{k\to\infty}\mathbf{A}^k_{n\times n}&=\lim_{k\to\infty}\mathbf{P}\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}^k\end{pmatrix}\mathbf{P}^{-1}=\mathbf{P}\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{P}^{-1}\\ &=\begin{pmatrix}\mathbf{P}_1\,|\,\mathbf{P}_2\end{pmatrix}\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Q}_1\\ \mathbf{Q}_2\end{pmatrix}=\mathbf{P}_1\mathbf{Q}_1=\mathbf{G}.\end{aligned} \tag{7.10.32}$$

Another way is to use $f(z)=z^k$ in the spectral resolution theorem on p. 603. If $\sigma(\mathbf{A})=\{\lambda_1,\lambda_2,\ldots,\lambda_s\}$ with $1=\lambda_1>|\lambda_2|\geq\cdots\geq|\lambda_s|$, and if $\textit{index}(\lambda_i)=k_i$, where $k_1=1$, then $\lim_{k\to\infty}\binom{k}{j}\lambda_i^{k-j}=0$ for $i\geq 2$ (see p. 618), and

$$\begin{aligned}\mathbf{A}^k&=\sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\binom{k}{j}\lambda_i^{k-j}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i\\ &=\mathbf{G}_1+\sum_{i=2}^{s}\sum_{j=0}^{k_i-1}\binom{k}{j}\lambda_i^{k-j}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i\to\mathbf{G}_1\quad\text{as}\quad k\to\infty.\end{aligned}$$

In other words, $\lim_{k\to\infty}\mathbf{A}^k = \mathbf{G}_1 = \mathbf{G}$ is the spectral projector associated with $\lambda_1 = 1$. Since $index(\lambda_1) = 1$, we know from the discussion on p. 603 that $R(\mathbf{G}) = N(\mathbf{I}-\mathbf{A})$ and $N(\mathbf{G}) = R(\mathbf{I}-\mathbf{A})$. Notice that if $\rho(\mathbf{A}) < 1$, then $\mathbf{I}-\mathbf{A}$ is nonsingular, and $N(\mathbf{I}-\mathbf{A}) = \{\mathbf{0}\}$. So regardless of whether the limit is zero or nonzero, $\lim_{k\to\infty}\mathbf{A}^k$ is always the projector onto $N(\mathbf{I}-\mathbf{A})$ along $R(\mathbf{I}-\mathbf{A})$. Below is a summary of the above observations.

Limits of Powers

For $\mathbf{A}\in\mathcal{C}^{n\times n}$, $\lim_{k\to\infty}\mathbf{A}^k$ exists if and only if

$$\rho(\mathbf{A}) < 1$$

or else

$$\rho(\mathbf{A}) = 1, \quad \text{where } \lambda = 1 \text{ is the only eigenvalue on the unit circle, and } \lambda = 1 \text{ is semisimple.} \tag{7.10.33}$$

When it exists,

$$\lim_{k\to\infty}\mathbf{A}^k = \text{the projector onto } N(\mathbf{I}-\mathbf{A}) \text{ along } R(\mathbf{I}-\mathbf{A}). \tag{7.10.34}$$

With each scalar sequence $\{\alpha_1,\alpha_2,\alpha_3,\ldots\}$ there is an associated sequence of averages $\{\mu_1,\mu_2,\mu_3,\ldots\}$ in which

$$\mu_1 = \alpha_1, \quad \mu_2 = \frac{\alpha_1+\alpha_2}{2}, \quad \ldots, \quad \mu_n = \frac{\alpha_1+\alpha_2+\cdots+\alpha_n}{n}.$$

This sequence of averages is called the associated ***Cesàro sequence***,[85] and when $\lim_{n\to\infty}\mu_n = \alpha$, we say that $\{\alpha_n\}$ is ***Cesàro summable*** (or merely *summable*) to α. It can be proven (Exercise 7.10.11) that if $\{\alpha_n\}$ converges to α, then $\{\mu_n\}$ converges to α, but not conversely. In other words, convergence implies summability, but summability doesn't insure convergence. To see that a sequence can be summable without being convergent, notice that the oscillatory sequence $\{0,1,0,1,\ldots\}$ doesn't converge, but it is Cesàro summable to $1/2$, which is the mean value of $\{0,1\}$. This is typical because averaging has a smoothing effect so that oscillations that prohibit convergence of the original sequence tend to be smoothed away or averaged out in the Cesàro sequence.

[85] Ernesto Cesàro (1859–1906) was an Italian mathematician who worked mainly in differential geometry but also contributed to number theory, divergent series, and mathematical physics. After studying in Naples, Liège, and Paris, Cesàro received his doctorate from the University of Rome in 1887, and he went on to occupy the chair of mathematics at Palermo. Cesàro's most important contribution is considered to be his 1890 book *Lezione di geometria intrinseca*, but, in large part, his name has been perpetuated because of its attachment to the concept of Cesàro summability.

Similar statements hold for general sequences of vectors and matrices (Exercise 7.10.11), but Cesàro summability is particularly interesting when it is applied to the sequence $\mathcal{P} = \{\mathbf{A}^k\}_{k=0}^{\infty}$ of powers of a square matrix $\mathbf{A}$. We know from (7.10.33) and (7.10.34) under what conditions sequence $\mathcal{P}$ converges as well as the nature of the limit, so let's now suppose that $\mathcal{P}$ doesn't converge, and decide when $\mathcal{P}$ is summable, and what $\mathcal{P}$ is summable to.

From now on, we will say that $\mathbf{A}_{n\times n}$ is a ***convergent matrix*** when $\lim_{k\to\infty}\mathbf{A}^k$ exists, and we will say that $\mathbf{A}$ is a ***summable matrix*** when $\lim_{k\to\infty}(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{k-1})/k$ exists. As in the scalar case, if $\mathbf{A}$ is convergent to $\mathbf{G}$, then $\mathbf{A}$ is summable to $\mathbf{G}$, but not conversely (Exercise 7.10.11). To analyze the summability of $\mathbf{A}$ in the absence of convergence, begin with the observation that $\mathbf{A}$ is summable if and only if the Jordan form $\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ for $\mathbf{A}$ is summable, which in turn is equivalent to saying that each Jordan block $\mathbf{J}_\star$ in $\mathbf{J}$ is summable. Consequently, $\mathbf{A}$ cannot be summable whenever $\rho(\mathbf{A}) > 1$ because if $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix}$ is a Jordan block in which $|\lambda| > 1$, then each diagonal entry of $\left(\mathbf{I}+\mathbf{J}_\star+\cdots+\mathbf{J}_\star^{k-1}\right)/k$ is

$$\delta(\lambda,k) = \frac{1+\lambda+\cdots+\lambda^{k-1}}{k} = \frac{1}{k}\left(\frac{1-\lambda^k}{1-\lambda}\right) = \frac{1}{1-\lambda}\left(\frac{1}{k}-\frac{\lambda^k}{k}\right), \quad (7.10.35)$$

and this becomes unbounded as $k \to \infty$. In other words, it's necessary that $\rho(\mathbf{A}) \le 1$ for $\mathbf{A}$ to be summable. Since we already know that $\mathbf{A}$ is convergent (and hence summable) to $\mathbf{0}$ when $\rho(\mathbf{A}) < 1$, we need only consider the case when $\mathbf{A}$ has eigenvalues on the unit circle.

If $\lambda \in \sigma(\mathbf{A})$ such that $|\lambda| = 1$, $\lambda \neq 1$, and if $index(\lambda) > 1$, then there is an associated Jordan block $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix}$ that is larger than 1×1. Each entry on the first superdiagonal of $\left(\mathbf{I}+\mathbf{J}_\star+\cdots+\mathbf{J}_\star^{k-1}\right)/k$ is the derivative $\partial\delta/\partial\lambda$ of the expression in (7.10.35), and it's not hard to see that $\partial\delta/\partial\lambda$ oscillates indefinitely as $k \to \infty$. In other words, $\mathbf{A}$ cannot be summable if there are eigenvalues $\lambda \neq 1$ on the unit circle such that $index(\lambda) > 1$.

Similarly, if $\lambda = 1$ is an eigenvalue of index greater than one, then $\mathbf{A}$ can't be summable because each entry on the first superdiagonal of

$$\frac{\mathbf{I}+\mathbf{J}_\star+\cdots+\mathbf{J}_\star^{k-1}}{k} \quad \text{is} \quad \frac{1+2+\cdots+(k-1)}{k} = \frac{k(k-1)}{2k} = \frac{k-1}{2} \to \infty.$$

Therefore, if $\mathbf{A}$ is summable and has eigenvalues λ such that $|\lambda| = 1$, then it's necessary that $index(\lambda) = 1$. The condition also is sufficient—i.e., if $\rho(\mathbf{A}) = 1$ and each eigenvalue on the unit circle is semisimple, then $\mathbf{A}$ is summable. This follows because each Jordan block associated with an eigenvalue μ such that $|\mu| < 1$ is convergent (and hence summable) to $\mathbf{0}$ by (7.10.5), and for semisimple

eigenvalues λ such that $|\lambda| = 1$, the associated Jordan blocks are 1×1 and hence summable because (7.10.35) implies

$$\frac{1+\lambda+\cdots+\lambda^{k-1}}{k} = \begin{cases} \dfrac{1}{1-\lambda}\left(\dfrac{1}{k} - \dfrac{\lambda^k}{k}\right) \to 0 & \text{for } |\lambda| = 1,\ \lambda \neq 1, \\ 1 & \text{for } \lambda = 1. \end{cases}$$

In addition to providing a necessary and sufficient condition for $\mathbf{A}$ to be Cesàro summable, the preceding analysis also reveals the nature of the Cesàro limit because if $\mathbf{A}$ is summable, then each Jordan block $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}$ in the Jordan form for $\mathbf{A}$ is summable, in which case we have established that

$$\lim_{k\to\infty} \frac{\mathbf{I}+\mathbf{J}_\star+\cdots+\mathbf{J}_\star^{k-1}}{k} = \begin{cases} \big[1\big]_{1\times 1} & \text{if } \lambda = 1 \text{ and } index(\lambda) = 1, \\ \big[0\big]_{1\times 1} & \text{if } |\lambda| = 1,\ \lambda \neq 1, \text{ and } index(\lambda) = 1, \\ \mathbf{0} & \text{if } |\lambda| < 1. \end{cases}$$

Consequently, if $\mathbf{A}$ is summable, then the Jordan form for $\mathbf{A}$ must look like

$$\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}, \quad \text{where} \quad p = alg\ mult_{\mathbf{A}}\,(\lambda = 1)\,,$$

and the eigenvalues of $\mathbf{C}$ are such that $|\lambda| < 1$ or else $|\lambda| = 1$, $\lambda \neq 1$, $index(\lambda) = 1$. So $\mathbf{C}$ is summable to $\mathbf{0}$, $\mathbf{J}$ is summable to $\begin{pmatrix} \mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, and

$$\frac{\mathbf{I}+\mathbf{A}+\cdots+\mathbf{A}^{k-1}}{k} = \mathbf{P}\left(\frac{\mathbf{I}+\mathbf{J}+\cdots+\mathbf{J}^{k-1}}{k}\right)\mathbf{P}^{-1} \to \mathbf{P}\begin{pmatrix} \mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^{-1} = \mathbf{G}.$$

Comparing this expression with that in (7.10.32) reveals that *the Cesàro limit is exactly the same as the ordinary limit, had it existed.* In other words, if $\mathbf{A}$ is summable, then regardless of whether or not $\mathbf{A}$ is convergent, $\mathbf{A}$ is summable to the projector onto $N(\mathbf{I}-\mathbf{A})$ along $R(\mathbf{I}-\mathbf{A})$. Below is a formal summary of our observations concerning Cesàro summability.

Cesàro Summability

- $\mathbf{A} \in \mathcal{C}^{n\times n}$ is Cesàro summable if and only if $\rho(\mathbf{A}) < 1$ or else $\rho(\mathbf{A}) = 1$ with each eigenvalue on the unit circle being semisimple.
- When it exists, the Cesàro limit

$$\lim_{k\to\infty} \frac{\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1}}{k} = \mathbf{G} \tag{7.10.36}$$

  is the projector onto $N(\mathbf{I}-\mathbf{A})$ along $R(\mathbf{I}-\mathbf{A})$.
- $\mathbf{G} \neq \mathbf{0}$ if and only if $1 \in \sigma(\mathbf{A})$, in which case $\mathbf{G}$ is the spectral projector associated with $\lambda = 1$.
- If $\mathbf{A}$ is convergent to $\mathbf{G}$, then $\mathbf{A}$ is summable to $\mathbf{G}$, but not conversely.

Since the projector $\mathbf{G}$ onto $N(\mathbf{I}-\mathbf{A})$ along $R(\mathbf{I}-\mathbf{A})$ plays a prominent role, let's consider how $\mathbf{G}$ might be computed. Of course, we could just iterate on $\mathbf{A}^k$ or $(\mathbf{I}+\mathbf{A}+\cdots+\mathbf{A}^{k-1})/k$, but this is inefficient and, depending on the proximity of the eigenvalues relative to the unit circle, convergence can be slow—averaging in particular can be extremely slow. The Jordan form is the basis for the theoretical development, but using it to compute $\mathbf{G}$ would be silly (see p. 592). The formula for a projector given in (5.9.12) on p. 386 is a possibility, but using a full-rank factorization of $\mathbf{I}-\mathbf{A}$ is an attractive alternative.

A ***full-rank factorization*** of a matrix $\mathbf{M}_{m\times n}$ of rank r is a factorization

$$\mathbf{M} = \mathbf{B}_{m\times r}\mathbf{C}_{r\times n}, \quad \text{where} \quad rank(\mathbf{B}) = rank(\mathbf{C}) = r = rank(\mathbf{M}). \tag{7.10.37}$$

All of the standard reduction techniques produce full-rank factorizations. For example, Gaussian elimination can be used because if $\mathbf{B}$ is the matrix of basic columns of $\mathbf{M}$, and if $\mathbf{C}$ is the matrix containing the nonzero rows in the reduced row echelon form $\mathbf{E}_{\mathbf{M}}$, then $\mathbf{M} = \mathbf{BC}$ is a full-rank factorization (Exercise 3.9.8, p. 140). If orthogonal reduction (p. 341) is used to produce a unitary matrix $\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 \\ \hline \mathbf{P}_2 \end{pmatrix}$ and an upper-trapezoidal matrix $\mathbf{T} = \begin{pmatrix} \mathbf{T}_1 \\ \hline \mathbf{0} \end{pmatrix}$ such that $\mathbf{PA} = \mathbf{T}$, where $\mathbf{P}_1$ is $r \times m$ and $\mathbf{T}_1$ contains the nonzero rows, then $\mathbf{M} = \mathbf{P}_1^*\mathbf{T}_1$ is a full-rank factorization. If

$$\mathbf{M} = \mathbf{U}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^* = (\mathbf{U}_1 \,|\, \mathbf{U}_2)\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{V}_1^* \\ \hline \mathbf{V}_2^* \end{pmatrix} = \mathbf{U}_1\mathbf{D}\mathbf{V}_1^* \tag{7.10.38}$$

is the singular value decomposition (5.12.2) on p. 412 (a URV factorization (p. 407) could also be used), then $\mathbf{M} = \mathbf{U}_1(\mathbf{D}\mathbf{V}_1^*) = (\mathbf{U}_1\mathbf{D})\mathbf{V}_1^*$ are full-rank factorizations. Projectors, in general, and limiting projectors, in particular, are nicely described in terms of full-rank factorizations.

Projectors

If $\mathbf{M}_{n\times n} = \mathbf{B}_{n\times r}\mathbf{C}_{r\times n}$ is any full-rank factorization as described in (7.10.37), and if $R(\mathbf{M})$ and $N(\mathbf{M})$ are complementary subspaces of $\mathcal{C}^n$, then the projector onto $R(\mathbf{M})$ along $N(\mathbf{M})$ is given by

$$\mathbf{P} = \mathbf{B}(\mathbf{C}\mathbf{B})^{-1}\mathbf{C} \tag{7.10.39}$$

or

$$\mathbf{P} = \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-1}\mathbf{V}_1^* \quad \text{when (7.10.38) is used.} \tag{7.10.40}$$

If $\mathbf{A}$ is convergent or summable to $\mathbf{G}$ as described in (7.10.34) and (7.10.36), and if $\mathbf{I} - \mathbf{A} = \mathbf{B}\mathbf{C}$ is a full-rank factorization, then

$$\mathbf{G} = \mathbf{I} - \mathbf{B}(\mathbf{C}\mathbf{B})^{-1}\mathbf{C} \tag{7.10.41}$$

or

$$\mathbf{G} = \mathbf{I} - \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-1}\mathbf{V}_1^* \quad \text{when (7.10.38) is used.} \tag{7.10.42}$$

Note: Formulas (7.10.39) and (7.10.40) are extensions of (5.13.3) on p. 430.

Proof. It's always true (Exercise 4.5.12, p. 220) that

$$\begin{aligned} R(\mathbf{X}_{m\times n}\mathbf{Y}_{n\times p}) &= R(\mathbf{X}) \quad \text{when} \quad rank(\mathbf{Y}) = n, \\ N(\mathbf{X}_{m\times n}\mathbf{Y}_{n\times p}) &= N(\mathbf{Y}) \quad \text{when} \quad rank(\mathbf{X}) = n. \end{aligned} \tag{7.10.43}$$

If $\mathbf{M}_{n\times n} = \mathbf{B}_{n\times r}\mathbf{C}_{r\times n}$ is a full-rank factorization, and if $R(\mathbf{M})$ and $N(\mathbf{M})$ are complementary subspaces of $\mathcal{C}^N$, then $rank(\mathbf{M}) = rank(\mathbf{M}^2)$ (Exercise 5.10.12, p. 402), so combining this with the first part of (7.10.43) produces

$$r = rank(\mathbf{B}\mathbf{C}) = rank(\mathbf{B}\mathbf{C}\mathbf{B}\mathbf{C}) = rank(\mathbf{C}\mathbf{B})_{r\times r} \implies (\mathbf{C}\mathbf{B})^{-1} \text{ exists.}$$

$\mathbf{P} = \mathbf{B}(\mathbf{C}\mathbf{B})^{-1}\mathbf{C}$ is a projector because $\mathbf{P}^2 = \mathbf{P}$ (recall (5.9.8), p. 386), and (7.10.43) insures that $R(\mathbf{P}) = R(\mathbf{B}) = R(\mathbf{M})$ and $N(\mathbf{P}) = N(\mathbf{C}) = N(\mathbf{M})$. Thus (7.10.39) is proved. If (7.10.38) is used to produce a full-rank factorization $\mathbf{M} = \mathbf{U}_1(\mathbf{D}\mathbf{V}_1^*)$, then, because $\mathbf{D}$ is nonsingular,

$$\mathbf{P} = (\mathbf{U}_1\mathbf{D})(\mathbf{V}_1^*(\mathbf{U}_1\mathbf{D}))^{-1}\mathbf{V}_1^* = \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-1}\mathbf{V}_1^*.$$

Equations (7.10.41) and (7.10.42) follow from (5.9.11), p. 386. ∎

Formulas (7.10.40) and (7.10.42) are useful because all good matrix computation packages contain numerically stable SVD implementations from which $\mathbf{U}_1$ and $\mathbf{V}_1^*$ can be obtained. But, of course, the singular values are not needed in this application.

Example 7.10.8

Shell Game. As depicted in Figure 7.10.2, a pea is placed under one of four shells, and an agile manipulator quickly rearranges them by a sequence of discrete moves. At the end of each move the shell containing the pea has been shifted either to the left or right by only one position according to the following rules.

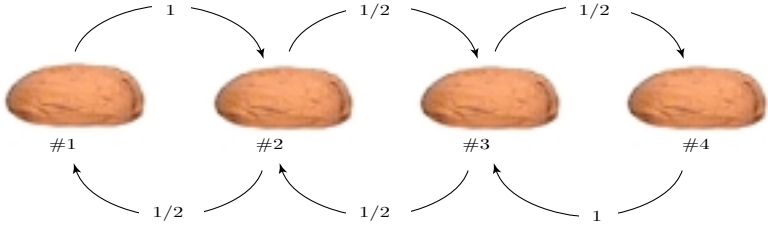


FIGURE 7.10.2

When the pea is under shell #1, it is moved to position #2, and if the pea is under shell #4, it is moved to position #3. When the pea is under shell #2 or #3, it is equally likely to be moved one position to the left or to the right.

Problem 1: Given that we know something about where the pea starts, what is the probability of finding the pea in any given position after k moves?

Problem 2: In the long run, what proportion of time does the pea occupy each of the four positions?

Solution to Problem 1: Let $p_j(k)$ denote the probability that the pea is in position j after the k^{th} move, and translate the given information into four difference equations by writing

$$\begin{aligned} p_1(k) &= \frac{p_2(k-1)}{2} \\ p_2(k) &= p_1(k-1) + \frac{p_3(k-1)}{2} \\ p_3(k) &= \frac{p_2(k-1)}{2} + p_4(k-1) \\ p_4(k) &= \frac{p_3(k-1)}{2} \end{aligned} \quad \text{or} \quad \begin{pmatrix} p_1(k) \\ p_2(k) \\ p_3(k) \\ p_4(k) \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1 \\ 0 & 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} p_1(k-1) \\ p_2(k-1) \\ p_3(k-1) \\ p_4(k-1) \end{pmatrix}.$$

The matrix equation on the right-hand side is a homogeneous difference equation $\mathbf{p}(k) = \mathbf{A}\mathbf{p}(k-1)$ whose solution, from (7.10.4), is $\mathbf{p}(k) = \mathbf{A}^k\mathbf{p}(0)$, and thus Problem 1 is solved. For example, if you know that the pea is initially under shell #2, then $\mathbf{p}(0) = \mathbf{e}_2$, and after six moves the probability that the pea is in the fourth position is $p_4(6) = \left[\mathbf{A}^6\mathbf{e}_2\right]_4 = 21/64$. If you don't know exactly where the pea starts, but you assume that it is equally likely to start under any one of the four shells, then $\mathbf{p}(0) = (1/4, 1/4, 1/4, 1/4)^T$, and the probabilities

for occupying the four positions after six moves are given by $\mathbf{p}(6) = \mathbf{A}^6\mathbf{p}(0)$, or

$$\begin{pmatrix} p_1(6) \\ p_2(6) \\ p_3(6) \\ p_4(6) \end{pmatrix} = \begin{pmatrix} 11/32 & 0 & 21/64 & 0 \\ 0 & 43/64 & 0 & 21/32 \\ 21/32 & 0 & 43/64 & 0 \\ 0 & 21/64 & 0 & 11/32 \end{pmatrix} \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \frac{1}{256} \begin{pmatrix} 43 \\ 85 \\ 85 \\ 43 \end{pmatrix}.$$

Solution to Problem 2: There is a straightforward solution when $\mathbf{A}$ is a convergent matrix because if $\mathbf{A}^k \to \mathbf{G}$ as $k \to \infty$, then $\mathbf{p}(k) \to \mathbf{G}\mathbf{p}(0) = \mathbf{p}$, and the components in this limiting (or steady-state) vector $\mathbf{p}$ provide the answer. Intuitively, if $\mathbf{p}(k) \to \mathbf{p}$, then after awhile $\mathbf{p}(k)$ is practically constant, so the probability that the pea occupies a particular position remains essentially the same move after move. Consequently, the components in $\lim_{k\to\infty} \mathbf{p}(k)$ reveal the proportion of time spent in each position over the long run. For example, if $\lim_{k\to\infty} \mathbf{p}(k) = (1/6, 1/3, 1/3, 1/6)^T$, then, as the game runs on indefinitely, the pea is expected to be under shell #1 for about 16.7% of the time, under shell #2 for about 33.3% of the time, etc.

A Fly in the Ointment: Everything above rests on the assumption that $\mathbf{A}$ is convergent. But $\mathbf{A}$ is *not* convergent for the shell game because a bit of computation reveals that $\sigma(\mathbf{A}) = \{\pm 1, \pm(1/2)\}$. That is, there is an eigenvalue other than 1 on the unit circle, so (7.10.33) guarantees that $\lim_{k\to\infty} \mathbf{A}^k$ does not exist. Consequently, there's no limiting solution $\mathbf{p}$ to the difference equation $\mathbf{p}(k) = \mathbf{A}\mathbf{p}(k-1)$, and the intuitive analysis given above does not apply.

Cesàro to the Rescue: However, $\mathbf{A}$ is summable because $\rho(\mathbf{A}) = 1$, and every eigenvalue on the unit circle is semisimple—these are the conditions in (7.10.36). So as $k \to \infty$,

$$\left(\frac{\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1}}{k}\right)\mathbf{p}(0) \to \mathbf{G}\mathbf{p}(0) = \mathbf{p}.$$

The job now is to interpret the meaning of this Cesàro limit in the context of the shell game. To do so, focus on a particular position—say the j^{th} one—and set up "counting functions" (random variables) defined as

$$X(0) = \begin{cases} 1 & \text{if the pea starts under shell } j, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$X(i) = \begin{cases} 1 & \text{if the pea is under shell } j \text{ after the } i^{th} \text{ move,} \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, 2, 3, \ldots.$$

Notice that $X(0) + X(1) + \cdots + X(k-1)$ counts the number of times the pea occupies position j before the k^{th} move, so $\big(X(0) + X(1) + \cdots + X(k-1)\big)/k$

represents the *fraction* of times that the pea is under shell j before the k^{th} move. Since the expected (or mean) value of $X(i)$ is, by definition,

$$E[X(i)] = 1 \times P\big(X(i) = 1\big) \; + \; 0 \times P\big(X(i) = 0\big) = p_j(i),$$

and since expectation is linear ($E[\alpha X(i) + X(h)] = \alpha E[X(i)] + E[X(h)]$), the expected fraction of times that the pea occupies position j before move k is

$$\begin{aligned} E\left[\frac{X(0) + X(1) + \cdots + X(k-1)}{k}\right] &= \frac{E[X(0)] + E[X(1)] + \cdots + E[X(k-1)]}{k} \\ &= \frac{p_j(0) + p_j(1) + \cdots + p_j(k-1)}{k} = \left[\frac{\mathbf{p}(0) + \mathbf{p}(1) + \cdots + \mathbf{p}(k-1)}{k}\right]_j \\ &= \left[\frac{\mathbf{p}(0) + \mathbf{A}\mathbf{p}(0) + \cdots + \mathbf{A}^{k-1}\mathbf{p}(0)}{k}\right]_j = \left[\left(\frac{\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1}}{k}\right)\mathbf{p}(0)\right]_j \\ &\to [\mathbf{G}\mathbf{p}(0)]_j \, . \end{aligned}$$

In other words, as the game progresses indefinitely, the components of the Cesàro limit $\mathbf{p} = \mathbf{G}\mathbf{p}(0)$ provide the expected proportion of times that the pea is under each shell, and this is exactly what we wanted to know.

Computing the Limiting Vector. Of course, $\mathbf{p}$ can be determined by first computing $\mathbf{G}$ with a full-rank factorization of $\mathbf{I} - \mathbf{A}$ as described in (7.10.41), but there is some special structure in this problem that can be exploited to make the task easier. Recall from (7.2.12) on p. 518 that if λ is a simple eigenvalue for $\mathbf{A}$, and if $\mathbf{x}$ and $\mathbf{y}^*$ are respective right-hand and left-hand eigenvectors associated with λ, then $\mathbf{x}\mathbf{y}^*/\mathbf{y}^*\mathbf{x}$ is the projector onto $N(\lambda\mathbf{I} - \mathbf{A})$ along $R(\lambda\mathbf{I} - \mathbf{A})$. We can use this because, for the shell game, $\lambda = 1$ is a simple eigenvalue for $\mathbf{A}$. Furthermore, we get an associated left-hand eigenvector for free—namely, $\mathbf{e}^T = (1, 1, 1, 1)$ —because each column sum of $\mathbf{A}$ is one, so $\mathbf{e}^T\mathbf{A} = \mathbf{e}^T$. Consequently, if $\mathbf{x}$ is any right-hand eigenvector of $\mathbf{A}$ associated with $\lambda = 1$, then (by noting that $\mathbf{e}^T\mathbf{p}(0) = p_1(0) + p_2(0) + p_3(0) + p_4(0) = 1$) the limiting vector is given by

$$\mathbf{p} = \mathbf{G}\mathbf{p}(0) = \frac{\mathbf{x}\mathbf{e}^T\mathbf{p}(0)}{\mathbf{e}^T\mathbf{x}} = \frac{\mathbf{x}}{\mathbf{e}^T\mathbf{x}} = \frac{\mathbf{x}}{\sum x_i}. \tag{7.10.44}$$

In other words, the limiting vector is obtained by normalizing any nonzero solution of $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$ to make the components sum to one. Not only does (7.10.44) show how to compute the limiting proportions, it also shows that *the limiting proportions are independent of the initial values in* $\mathbf{p}(0)$. For example, a simple calculation reveals that $\mathbf{x} = (1, 2, 2, 1)^T$ is one solution of $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$, so the vector of limiting proportions is $\mathbf{p} = (1/6, 1/3, 1/3, 1/6)^T$. Therefore, if many moves are made, then, regardless of where the pea starts, we expect the pea to end up under shell #1 in about 16.7% of the moves, under #2 for about

33.3% of the moves, under #3 for about 33.3% of the moves, and under shell #4 for about 16.7% of the moves.

Note: The shell game (and its analysis) is a typical example of *a random walk with reflecting barriers,* and these problems belong to a broader classification of stochastic processes known as *irreducible, periodic Markov chains.* (Markov chains are discussed in detail in §8.4 on p. 687.) The shell game is irreducible in the sense of Exercise 4.4.20 (p. 209), and it is periodic because the pea can return to given position only at definite periods, as reflected in the periodicity of the powers of $\mathbf{A}$. More details are given in Example 8.4.3 on p. 694.

Exercises for section 7.10

7.10.1. Which of the following are convergent, and which are summable?

$$\mathbf{A}=\begin{pmatrix} -1/2 & 3/2 & -3/2 \\ 1 & 0 & -1/2 \\ 1 & -1 & 1/2 \end{pmatrix}. \quad \mathbf{B}=\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \quad \mathbf{C}=\begin{pmatrix} -1 & -2 & -3/2 \\ 1 & 2 & 1 \\ 1 & 1 & 3/2 \end{pmatrix}.$$

7.10.2. For the matrices in Exercise 7.10.1, evaluate the limit of each convergent matrix, and evaluate the Cesàro limit for each summable matrix.

7.10.3. Verify that the expressions in (7.10.4) are indeed the solutions to the difference equations in (7.10.3).

7.10.4. Determine the limiting vector for the shell game in Example 7.10.8 by first computing the Cesàro limit $\mathbf{G}$ with a full-rank factorization.

7.10.5. Verify that the expressions in (7.10.4) are indeed the solutions to the difference equations in (7.10.3).

7.10.6. Prove that if there exists a matrix norm such that $\|\mathbf{A}\| < 1$, then $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$.

7.10.7. By examining the iteration matrix, compare the convergence of Jacobi's method and the Gauss–Seidel method for each of the following coefficient matrices with an arbitrary right-hand side. Explain why this shows that neither method can be universally favored over the other.

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}. \qquad \mathbf{A}_2 = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}.$$

7.10.8. Let $\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ (the finite-difference Example 1.4.1, p. 19).

(a) Verify that $\mathbf{A}$ satisfies the special case conditions given in Example 7.10.6 that guarantee the validity of (7.10.24).

(b) Determine the optimum SOR relaxation parameter.

(c) Find the asymptotic rates of convergence for Jacobi, Gauss–Seidel, and optimum SOR.

(d) Use $\mathbf{x}(0) = (1, 1, 1)^T$ and $\mathbf{b} = (2, 4, 6)^T$ to run through several steps of Jacobi, Gauss–Seidel, and optimum SOR to solve $\mathbf{Ax} = \mathbf{b}$ until you can see a convergence pattern.

7.10.9. Prove that if $\rho(\mathbf{H}_\omega) < 1$, where $\mathbf{H}_\omega$ is the iteration matrix for the SOR method, then $0 < \omega < 2$. **Hint:** Use $\det(\mathbf{H}_\omega)$ to show $|\lambda_k| \geq |1 - \omega|$ for some $\lambda_k \in \sigma(\mathbf{H}_\omega)$.

7.10.10. Show that the spectral radius of the Jacobi iteration matrix for the discrete Laplacian $\mathbf{L}_{n^2 \times n^2}$ described in Example 7.6.2 (p. 563) is $\rho(\mathbf{H}_J) = \cos \pi/(n+1)$.

7.10.11. Consider a scalar sequence $\{\alpha_1, \alpha_2, \alpha_3, \ldots\}$ and the associated Cesàro sequence of averages $\{\mu_1, \mu_2, \mu_3, \ldots\}$, where $\mu_n = (\alpha_1 + \alpha_2 + \cdots + \alpha_n)/n$. Prove that if $\{\alpha_n\}$ converges to α, then $\{\mu_n\}$ also converges to α.

Note: Like scalars, a vector sequence $\{\mathbf{v}_n\}$ in a finite-dimensional space converges to $\mathbf{v}$ if and only if for each $\epsilon > 0$ there is a natural number $N = N(\epsilon)$ such that $\|\mathbf{v}_n - \mathbf{v}\| < \epsilon$ for all $n \geq N$, and, by virtue of Example 5.1.3 (p. 276), it doesn't matter which norm is used. Therefore, your proof should also be valid for vectors (and matrices).

7.10.12. M-matrices Revisited. For matrices with nonpositive off-diagonal entries (Z-matrices), prove that the following statements are equivalent.

(a) $\mathbf{A}$ is an M-matrix.

(b) All *leading* principal minors of $\mathbf{A}$ are positive.

(c) $\mathbf{A}$ has an LU factorization, and both $\mathbf{L}$ and $\mathbf{U}$ are M-matrices.

(d) There exists a vector $\mathbf{x} > \mathbf{0}$ such that $\mathbf{Ax} > \mathbf{0}$.

(e) Each $a_{ii} > 0$ and $\mathbf{AD}$ is diagonally dominant for some diagonal matrix $\mathbf{D}$ with positive diagonal entries.

(f) $\mathbf{Ax} \geq \mathbf{0}$ implies $\mathbf{x} \geq \mathbf{0}$.

7.10.13. Index by Full-Rank Factorization. Suppose that $\lambda \in \sigma(\mathbf{A})$, and let $\mathbf{M}_1 = \mathbf{A} - \lambda\mathbf{I}$. The following procedure yields the value of $index(\lambda)$.

Factor $\mathbf{M}_1 = \mathbf{B}_1\mathbf{C}_1$ as a full-rank factorization.
Set $\mathbf{M}_2 = \mathbf{C}_1\mathbf{B}_1$.
Factor $\mathbf{M}_2 = \mathbf{B}_2\mathbf{C}_2$ as a full-rank factorization.
Set $\mathbf{M}_3 = \mathbf{C}_2\mathbf{B}_2$.
$\vdots$

In general, $\mathbf{M}_i = \mathbf{C}_{i-1}\mathbf{B}_{i-1}$, where $\mathbf{M}_{i-1} = \mathbf{B}_{i-1}\mathbf{C}_{i-1}$ is a full-rank factorization.

(a) Explain why this procedure must eventually produce a matrix $\mathbf{M}_k$ that is either nonsingular or zero.

(b) Prove that if k is the smallest positive integer such that $\mathbf{M}_k^{-1}$ exists or $\mathbf{M}_k = \mathbf{0}$, then

$$index(\lambda) = \begin{cases} k-1 & \text{if } \mathbf{M}_k \text{ is nonsingular,} \\ k & \text{if } \mathbf{M}_k = \mathbf{0}. \end{cases}$$

7.10.14. Use the procedure in Exercise 7.10.13 to find the index of each eigenvalue of $\mathbf{A} = \begin{pmatrix} -3 & -8 & -9 \\ 5 & 11 & 9 \\ -1 & -2 & 1 \end{pmatrix}$. **Hint:** $\sigma(\mathbf{A}) = \{4, 1\}$.

7.10.15. Let $\mathbf{A}$ be the matrix given in Exercise 7.10.14.

(a) Find the Jordan form for $\mathbf{A}$.

(b) For any function f defined at $\mathbf{A}$, find the Hermite interpolation polynomial that is described in Example 7.9.4 (p. 606), and describe $f(\mathbf{A})$.

7.10.16. Limits and Group Inversion. Given a matrix $\mathbf{B}_{n\times n}$ of rank r such that $index(\mathbf{B}) \leq 1$ (i.e., $index(\lambda = 0) \leq 1$), the Jordan form for $\mathbf{B}$ looks like $\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{r\times r} \end{pmatrix} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}$, so $\mathbf{B} = \mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}\mathbf{P}^{-1}$, where $\mathbf{C}$ is nonsingular. This implies that $\mathbf{B}$ belongs to an algebraic group $\mathcal{G}$ with respect to matrix multiplication, and the inverse of $\mathbf{B}$ in $\mathcal{G}$ is $\mathbf{B}^{\#} = \mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{-1} \end{pmatrix}\mathbf{P}^{-1}$. Naturally, $\mathbf{B}^{\#}$ is called the ***group inverse*** of $\mathbf{B}$. The group inverse is a special case of the Drazin inverse discussed in Example 5.10.5 on p. 399, and properties of group inversion are developed in Exercises 5.10.11–5.10.13 on p. 402. Prove that if $\lim_{k\to\infty} \mathbf{A}^k$ exists, and if $\mathbf{B} = \mathbf{I} - \mathbf{A}$, then

$$\lim_{k\to\infty} \mathbf{A}^k = \mathbf{I} - \mathbf{B}\mathbf{B}^{\#}.$$

In other words, the limiting matrix can be characterized as the difference of two identity elements—$\mathbf{I}$ is the identity in the multiplicative group of nonsingular matrices, and $\mathbf{B}\mathbf{B}^{\#}$ is the identity element in the multiplicative group containing $\mathbf{B}$.

7.10.17. If $\mathbf{M}_{n\times n}$ is a group matrix (i.e., if $index(\mathbf{M}) \leq 1$), then the group inverse of $\mathbf{M}$ can be characterized as the unique solution $\mathbf{M}^{\#}$ of the equations $\mathbf{M}\mathbf{M}^{\#}\mathbf{M} = \mathbf{M}$, $\mathbf{M}^{\#}\mathbf{M}\mathbf{M}^{\#} = \mathbf{M}^{\#}$, and $\mathbf{M}\mathbf{M}^{\#} = \mathbf{M}^{\#}\mathbf{M}$. In fact, some authors use these equations to define $\mathbf{M}^{\#}$. Use this characterization to show that if $\mathbf{M} = \mathbf{B}\mathbf{C}$ is any full-rank factorization of $\mathbf{M}$, then $\mathbf{M}^{\#} = \mathbf{B}(\mathbf{C}\mathbf{B})^{-2}\mathbf{C}$. In particular, if $\mathbf{M} = \mathbf{U}_1\mathbf{D}\mathbf{V}_1^*$ is the full-rank factorization derived from the singular value decomposition as described in (7.10.38), then

$$\begin{aligned}\mathbf{M}^{\#} &= \mathbf{U}_1\mathbf{D}^{-1/2}(\mathbf{V}_1^*\mathbf{U}_1)^{-2}\mathbf{D}^{-1/2}\mathbf{V}_1^* \\ &= \mathbf{U}_1\mathbf{D}^{-1}(\mathbf{V}_1^*\mathbf{U}_1)^{-2}\mathbf{V}_1^* \\ &= \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-2}\mathbf{D}^{-1}\mathbf{V}_1^*.\end{aligned}$$

7.11 MINIMUM POLYNOMIALS AND KRYLOV METHODS

The characteristic polynomial plays a central role in the theoretical development of linear algebra and matrix analysis, but it is not alone in this respect. There are other polynomials that occur naturally, and the purpose of this section is to explore some of them.

In this section it is convenient to consider the characteristic polynomial of $\mathbf{A} \in \mathcal{C}^{n\times n}$ to be $c(x) = \det(x\mathbf{I} - \mathbf{A})$. This differs from the definition given on p. 492 only in the sense that the coefficients of $c(x) = \det(x\mathbf{I} - \mathbf{A})$ have different signs than the coefficients of $\hat{c}(x) = \det(\mathbf{A} - x\mathbf{I})$. In particular, $c(x)$ is a *monic polynomial* (i.e., its leading coefficient is 1), whereas the leading coefficient of $\hat{c}(x)$ is $(-1)^n$. (Of course, the roots of c and $\hat{c}$ are identical.)

Monic polynomials $p(x)$ such that $p(\mathbf{A}) = \mathbf{0}$ are said to be ***annihilating polynomials*** for $\mathbf{A}$. For example, the Cayley–Hamilton theorem (pp. 509, 532) guarantees that $c(x)$ is an annihilating polynomial of degree n.

Minimum Polynomial for a Matrix

There is a unique annihilating polynomial for $\mathbf{A}$ of minimal degree, and this polynomial, denoted by $m(x)$, is called the ***minimum polynomial*** for $\mathbf{A}$. The Cayley–Hamilton theorem guarantees that $\deg[m(x)] \leq n$.

Proof. Only uniqueness needs to be proven. Let k be the smallest degree of any annihilating polynomial for $\mathbf{A}$. There is a unique annihilating polynomial for $\mathbf{A}$ of degree k because if there were two different annihilating polynomials $p_1(x)$ and $p_2(x)$ of degree k, then $d(x) = p_1(x) - p_2(x)$ would be a nonzero polynomial such that $d(\mathbf{A}) = \mathbf{0}$ and $\deg[d(x)] < k$. Dividing $d(x)$ by its leading coefficient would produce an annihilating polynomial of degree less than k, the minimal degree, and this is impossible. ∎

The first problem is to describe what the minimum polynomial $m(x)$ for $\mathbf{A} \in \mathcal{C}^{n\times n}$ looks like, and the second problem is to uncover the relationship between $m(x)$ and the characteristic polynomial $c(x)$. The Jordan form for $\mathbf{A}$ reveals everything. Suppose that $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, where $\mathbf{J}$ is in Jordan form. Since $p(\mathbf{A}) = \mathbf{0}$ if and only if $p(\mathbf{J}) = \mathbf{0}$ or, equivalently, $p(\mathbf{J}_\star) = \mathbf{0}$ for each Jordan block $\mathbf{J}_\star$, it's clear that $m(x)$ is the monic polynomial of smallest degree that annihilates all Jordan blocks. If $\mathbf{J}_\star$ is a $k \times k$ Jordan block associated with an eigenvalue λ, then (7.9.2) on p. 600 insures that $p(\mathbf{J}_\star) = \mathbf{0}$ if and only if $p^{(i)}(\lambda) = 0$ for $i = 0, 2, \ldots, k-1$, and this happens if and only if $p(x) = (x-\lambda)^k q(x)$ for some polynomial $q(x)$. Since this must be true for all Jordan blocks associated with λ, it must be true for the *largest* Jordan block associated with λ, and thus the minimum degree monic polynomial that

annihilates all Jordan blocks associated with λ is

$$p_\lambda(x) = (x-\lambda)^{k_\lambda}, \quad \text{where} \quad k_\lambda = index(\lambda).$$

Since the minimum polynomial for $\mathbf{A}$ must annihilate the largest Jordan block associated with each $\lambda_j \in \sigma(\mathbf{A})$, it follows that

$$m(x) = (x-\lambda_1)^{k_1}(x-\lambda_2)^{k_2}\cdots(x-\lambda_s)^{k_s}, \quad \text{where} \quad k_j = index(\lambda_j) \quad (7.11.1)$$

is the minimum polynomial for $\mathbf{A}$.

Example 7.11.1

Minimum Polynomial, Gram–Schmidt, and QR. If you are willing to compute the eigenvalues λ_j and their indicies k_j for a given $\mathbf{A} \in \mathcal{C}^{n\times n}$, then, as shown in (7.11.1), the minimum polynomial for $\mathbf{A} \in \mathcal{C}^{n\times n}$ is obtained by setting $m(x) = (x-\lambda_1)^{k_1}(x-\lambda_2)^{k_2}\cdots(x-\lambda_s)^{k_s}$. But finding the eigenvalues and their indicies can be a substantial task, so let's consider how we might construct $m(x)$ without computing eigenvalues. An approach based on first principles is to determine the first matrix $\mathbf{A}^k$ for which $\{\mathbf{I}, \mathbf{A}, \mathbf{A}^2, \ldots, \mathbf{A}^k\}$ is linearly dependent. In other words, if k is the smallest positive integer such that $\mathbf{A}^k = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j$, then the minimum polynomial for $\mathbf{A}$ is

$$m(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j.$$

The Gram–Schmidt orthogonalization procedure (p. 309) with the standard inner product $\langle \mathbf{A} | \mathbf{B} \rangle = trace(\mathbf{A}^*\mathbf{B})$ (p. 286) is the perfect theoretical tool for determining k and the α_j's. Gram–Schmidt applied to $\{\mathbf{I}, \mathbf{A}, \mathbf{A}^2, \ldots\}$ begins by setting $\mathbf{U}_0 = \mathbf{I}/\|\mathbf{I}\|_F = \mathbf{I}/\sqrt{n}$, and it proceeds by sequentially computing

$$\mathbf{U}_j = \frac{\mathbf{A}^j - \sum_{i=0}^{j-1} \langle \mathbf{U}_i | \mathbf{A}^j \rangle \mathbf{U}_i}{\|\mathbf{A}^j - \sum_{i=0}^{j-1} \langle \mathbf{U}_i | \mathbf{A}^j \rangle \mathbf{U}_i\|_F} \quad \text{for} \quad j = 1, 2, \ldots \quad (7.11.2)$$

until $\mathbf{A}^k - \sum_{i=0}^{k-1} \langle \mathbf{U}_i | \mathbf{A}^k \rangle \mathbf{U}_i = \mathbf{0}$. The first such k is the smallest positive integer such that $\mathbf{A}^k \in span\{\mathbf{U}_0, \mathbf{U}_1, \ldots, \mathbf{U}_{k-1}\} = span\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{k-1}\}$. The coefficients α_j such that $\mathbf{A}^k = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j$ are easily determined from the upper-triangular matrix $\mathbf{R}$ in the QR factorization produced by the Gram–Schmidt process. To see how, extend the notation in the discussion on p. 311 in an obvious way to write (7.11.2) in matrix form as

$$\left[\mathbf{I} \,|\, \mathbf{A} \,|\cdots|\, \mathbf{A}^k\right] = \left[\mathbf{U}_0 \,|\, \mathbf{U}_1 \,|\cdots|\, \mathbf{U}_k\right] \begin{pmatrix} \nu_0 & r_{01} & \cdots & r_{0k-1} & r_{0k} \\ 0 & \nu_1 & \cdots & r_{1k-1} & r_{1k} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & & \nu_{k-1} & r_{k-1k} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \quad (7.11.3)$$

where $\nu_0 = \|\mathbf{I}\|_F = \sqrt{n}$, $\nu_j = \left\| \mathbf{A}^j - \sum_{i=0}^{j-1} \langle \mathbf{U}_i | \mathbf{A}^j \rangle \mathbf{U}_i \right\|_F$, and $r_{ij} = \langle \mathbf{U}_i | \mathbf{A}^j \rangle$. If we set $\mathbf{R} = \begin{pmatrix} \nu_0 & \cdots & r_{0k-1} \\ & \ddots & \vdots \\ & & \nu_{k-1} \end{pmatrix}$ and $\mathbf{c} = \begin{pmatrix} r_{0k} \\ \vdots \\ r_{k-1k} \end{pmatrix}$, then (7.11.3) implies that $\mathbf{A}^k = \begin{bmatrix} \mathbf{U}_0 | \cdots | \mathbf{U}_{k-1} \end{bmatrix} \mathbf{c} = \begin{bmatrix} \mathbf{I} | \cdots | \mathbf{A}^{k-1} \end{bmatrix} \mathbf{R}^{-1}\mathbf{c}$, so $\mathbf{R}^{-1}\mathbf{c} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{k-1} \end{pmatrix}$ contains the coefficients such that $\mathbf{A}^k = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j$, and thus the coefficients in the minimum polynomial are determined.

Caution! While Gram–Schmidt works fine to produce $m(x)$ in exact arithmetic, things are not so nice in floating-point arithmetic. For example, if $\mathbf{A}$ has a dominant eigenvalue, then, as explained in the power method (Example 7.3.7, p. 533), $\mathbf{A}^k$ asymptotically approaches the dominant spectral projector $\mathbf{G}_1$, so, as k grows, $\mathbf{A}^k$ becomes increasingly close to $span\left\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{k-1}\right\}$. Consequently, finding the first $\mathbf{A}^k$ that is truly in $span\left\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{k-1}\right\}$ is an ill-conditioned problem, and Gram–Schmidt may not work well in floating-point arithmetic—the modified Gram–Schmidt algorithm (p. 316), or a version of Householder reduction (p. 341), or Arnoldi's method (p. 653) works better. Fortunately, explicit knowledge of the minimum polynomial often is not needed in applied work.

The relationship between the characteristic polynomial $c(x)$ and the minimum polynomial $m(x)$ for $\mathbf{A}$ is now transparent. Since

$$c(x) = (x - \lambda_1)^{a_1}(x - \lambda_2)^{a_2} \cdots (x - \lambda_s)^{a_s}, \quad \text{where} \quad a_j = alg\ mult\,(\lambda_j),$$

and

$$m(x) = (x - \lambda_1)^{k_1}(x - \lambda_2)^{k_2} \cdots (x - \lambda_s)^{k_s}, \quad \text{where} \quad k_j = index\,(\lambda_j),$$

it's clear that $m(x)$ divides $c(x)$. Furthermore, $m(x) = c(x)$ if and only if $alg\ mult\,(\lambda_j) = index\,(\lambda_j)$ for each $\lambda_j \in \sigma(\mathbf{A})$. Matrices for which $m(x) = c(x)$ are said to be ***nonderogatory matrices,*** and they are precisely the ones for which $geo\ mult\,(\lambda_j) = 1$ for each eigenvalue λ_j because

$$\begin{aligned} m(x) = c(x) &\iff alg\ mult\,(\lambda_j) = index\,(\lambda_j) \text{ for each } j \\ &\iff \text{there is only one Jordan block for each } \lambda_j \\ &\iff \text{there is only one independent eigenvector for each } \lambda_j \\ &\iff geo\ mult\,(\lambda_j) = 1 \text{ for each } \lambda_j. \end{aligned}$$

In addition to dividing the characteristic polynomial $c(x)$, the minimum polynomial $m(x)$ divides all other annihilating polynomials $p(x)$ for $\mathbf{A}$ because $\deg[m(x)] \leq \deg[p(x)]$ insures the existence of polynomials $q(x)$ and $r(x)$ (quotient and remainder) such that

$$p(x) = m(x)q(x) + r(x), \quad \text{where} \quad \deg[r(x)] < \deg[m(x)].$$

Since

$$\mathbf{0} = p(\mathbf{A}) = m(\mathbf{A})q(\mathbf{A}) + r(\mathbf{A}) = r(\mathbf{A}),$$

it follows that $r(x) = 0$; otherwise $r(x)$, when normalized to be monic, would be an annihilating polynomial having degree smaller than the degree of the minimum polynomial.

The structure of the minimum polynomial for $\mathbf{A}$ is related to the diagonalizability of $\mathbf{A}$. By combining the fact that $k_j = index(\lambda_j)$ is the size of the largest Jordan block for λ_j with the fact that $\mathbf{A}$ is diagonalizable if and only if all Jordan blocks are 1×1, it follows that $\mathbf{A}$ is diagonalizable if and only if $k_j = 1$ for each j, which, by (7.11.1), is equivalent to saying that $m(x) = (x - \lambda_1)(x - \lambda_2)\cdots(x - \lambda_s)$. In other words, $\mathbf{A}$ is diagonalizable if and only if its minimum polynomial is the product of distinct linear factors.

Below is a summary of the preceding observations about properties of $m(x)$.

Properties of the Minimum Polynomial

Let $\mathbf{A} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$.

- The minimum polynomial of $\mathbf{A}$ is the unique monic polynomial $m(x)$ of minimal degree such that $m(\mathbf{A}) = \mathbf{0}$.
- $m(x) = (x - \lambda_1)^{k_1}(x - \lambda_2)^{k_2}\cdots(x - \lambda_s)^{k_s}$, where $k_j = index(\lambda_j)$.
- $m(x)$ divides every polynomial $p(x)$ such that $p(\mathbf{A}) = \mathbf{0}$. In particular, $m(x)$ divides the characteristic polynomial $c(x)$. (7.11.4)
- $m(x) = c(x)$ if and only if $geo\ mult(\lambda_j) = 1$ for each λ_j or, equivalently, $alg\ mult(\lambda_j) = index(\lambda_j)$ for each j, in which case $\mathbf{A}$ is called a ***nonderogatory matrix.***
- $\mathbf{A}$ is diagonalizable if and only if $m(x) = (x-\lambda_1)(x-\lambda_2)\cdots(x-\lambda_s)$ (i.e., if and only if $m(x)$ is a product of distinct linear factors).

The next immediate aim is to extend the concept of the minimum polynomial for a matrix to formulate the notion of a minimum polynomial for a vector. To do so, it's helpful to introduce Krylov[86] sequences, subspaces, and matrices.

[86] Aleksei Nikolaevich Krylov (1863–1945) showed in 1931 how to use sequences of the form $\{\mathbf{b}, \mathbf{Ab}, \mathbf{A^2b}, \ldots\}$ to construct the characteristic polynomial of a matrix (see Example 7.11.3 on p. 649). Krylov was a Russian applied mathematician whose scientific interests arose from his early training in naval science that involved the theories of buoyancy, stability, rolling and pitching, vibrations, and compass theories. Krylov served as the director of the Physics–Mathematics Institute of the Soviet Academy of Sciences from 1927 until 1932, and in 1943 he was awarded a "state prize" for his work on compass theory. Krylov was made a "hero of

Krylov Sequences, Subspaces, and Matrices

For $\mathbf{A} \in \mathcal{C}^{n\times n}$ and $\mathbf{0} \neq \mathbf{b} \in \mathcal{C}^{n\times 1}$, we adopt the following terminology.

- $\{\mathbf{b}, \mathbf{Ab}, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{j-1}\mathbf{b}\}$ is called a ***Krylov sequence.***
- $\mathcal{K}_j = span\left\{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{j-1}\mathbf{b}\right\}$ is called a ***Krylov subspace.***
- $\mathbf{K}_{n\times j} = \left(\mathbf{b} \,|\, \mathbf{Ab} \,|\cdots|\, \mathbf{A}^{j-1}\mathbf{b}\right)$ is called a ***Krylov matrix.***

Since $\dim(\mathcal{K}_j) \leq n$ (because $\mathcal{K}_j \subseteq \mathcal{C}^{n\times 1}$), there is a first vector $\mathbf{A}^k\mathbf{b}$ in the Krylov sequence that is a linear combination of preceding Krylov vectors. If

$$\mathbf{A}^k\mathbf{b} = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j\mathbf{b}, \quad \text{then we define} \quad v(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j,$$

and we say that $v(x)$ is an ***annihilating polynomial for*** $\mathbf{b}$ ***relative to*** $\mathbf{A}$ because $v(x)$ is a monic polynomial such that $v(\mathbf{A})\mathbf{b} = \mathbf{0}$. The argument on p. 642 that establishes uniqueness of the minimum polynomial for matrices can be reapplied to prove that for each matrix–vector pair $(\mathbf{A}, \mathbf{b})$ there is a unique annihilating polynomial of $\mathbf{b}$ relative to $\mathbf{A}$ that has minimal degree. These observations are formalized below.

Minimum Polynomial for a Vector

- The ***minimum polynomial for*** $\mathbf{b} \in \mathcal{C}^{n\times 1}$ ***relative to*** $\mathbf{A} \in \mathcal{C}^{n\times n}$ is defined to be the monic polynomial $v(x)$ of minimal degree such that $v(\mathbf{A})\mathbf{b} = \mathbf{0}$.
- If $\mathbf{A}^k\mathbf{b}$ is the first vector in the Krylov sequence $\{\mathbf{b}, \mathbf{Ab}, \mathbf{A}^3\mathbf{b}, \ldots\}$ that is a linear combination of preceding Krylov vectors (say $\mathbf{A}^k\mathbf{b} = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j\mathbf{b}$), then $v(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j$ (or $v(x) = 1$ when $\mathbf{b} = \mathbf{0}$) is the minimum polynomial for $\mathbf{b}$ relative to $\mathbf{A}$.

socialist labor," and he is one of a few mathematicians to have a lunar feature named in his honor—on the moon there is the "Crater Krylov."

So is the minimum polynomial for a matrix related to minimum polynomials for vectors? It seems intuitive that knowing the minimum polynomial of $\mathbf{b}$ relative to $\mathbf{A}$ for enough different vectors $\mathbf{b}$ should somehow lead to the minimum polynomial for $\mathbf{A}$. This is indeed the case, and here is how it's done. Recall that the least common multiple (LCM) of polynomials $v_1(x), \ldots, v_n(x)$ is the unique monic polynomial $l(x)$ such that

(i) each $v_i(x)$ divides $l(x)$;

(ii) if each $v_i(x)$ also divides $q(x)$, then $l(x)$ divides $q(x)$.

Minimum Polynomial as LCM

Let $\mathbf{A} \in \mathcal{C}^{n\times n}$, and let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ be any basis for $\mathcal{C}^{n\times 1}$. If $v_i(x)$ is the minimum polynomial for $\mathbf{b}_i$ relative to $\mathbf{A}$, then the minimum polynomial $m(x)$ for $\mathbf{A}$ is the least common multiple of $v_1(x)$, $v_2(x)$, $\ldots$, $v_n(x)$. (7.11.5)

Proof. The strategy first is to prove that if $l(x)$ is the LCM of the $v_i(x)$'s, then $m(x)$ divides $l(x)$. Then prove the reverse by showing that $l(x)$ also divides $m(x)$. Since each $v_i(x)$ divides $l(x)$, it follows that $l(\mathbf{A})\mathbf{b}_i = \mathbf{0}$ for each i. In other words, $\mathcal{B} \subset N(l(\mathbf{A}))$, so $\dim N(l(\mathbf{A})) = n$ or, equivalently, $l(\mathbf{A}) = \mathbf{0}$. Therefore, by property (7.11.4) on p. 645, $m(x)$ divides $l(x)$. Now show that $l(x)$ divides $m(x)$. Since $m(\mathbf{A})\mathbf{b}_i = \mathbf{0}$ for every $\mathbf{b}_i$, it follows that $\deg[v_i(x)] < \deg[m(x)]$ for each i, and hence there exist polynomials $q_i(x)$ and $r_i(x)$ such that $m(x) = q_i(x)v_i(x) + r_i(x)$, where $\deg[r_i(x)] < \deg[v_i(x)]$. But

$$\mathbf{0} = m(\mathbf{A})\mathbf{b}_i = q_i(\mathbf{A})v_i(\mathbf{A})\mathbf{b}_i + r_i(\mathbf{A})\mathbf{b}_i = r_i(\mathbf{A})\mathbf{b}_i$$

insures $r_i(x) = 0$, for otherwise $r_i(x)$ (when normalized to be monic) would be an annihilating polynomial for $\mathbf{b}_i$ of degree smaller than the minimum polynomial for $\mathbf{b}_i$, which is impossible. In other words, each $v_i(x)$ divides $m(x)$, and this implies $l(x)$ must also divide $m(x)$. Therefore, since $m(x)$ and $l(x)$ are divisors of each other, it must be the case that $m(x) = l(x)$. ∎

The utility of this result is illustrated in the following development. We already know that associated with $n \times n$ matrix $\mathbf{A}$ is an n^{th}-degree monic polynomial—namely, the characteristic polynomial $c(x) = \det(x\mathbf{I} - \mathbf{A})$. But the reverse is also true. That is, every n^{th}-degree monic polynomial is the characteristic polynomial of some $n \times n$ matrix.

Companion Matrix of a Polynomial

For each monic polynomial $p(x) = x^n + \alpha_{n-1}x^{n-1} + \cdots + \alpha_1 x + \alpha_0$, the ***companion matrix*** of $p(x)$ is defined (by G. Frobenius) to be

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & \cdots & 0 & -\alpha_0 \\ 1 & 0 & \cdots & 0 & -\alpha_1 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 1 & 0 & -\alpha_{n-2} \\ 0 & 0 & \cdots & 1 & -\alpha_{n-1} \end{pmatrix}_{n\times n}. \tag{7.11.6}$$

- The polynomial $p(x)$ is both the characteristic and minimum polynomial for $\mathbf{C}$ (i.e., $\mathbf{C}$ is nonderogatory).

Proof. To prove that $\det(x\mathbf{I} - \mathbf{C}) = p(x)$, write $\mathbf{C} = \mathbf{N} - \mathbf{c}\mathbf{e}_n^T$, where

$$\mathbf{N} = \begin{pmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix},$$

and use (6.2.3) on p. 475 to conclude that

$$\begin{aligned} \det(x\mathbf{I} - \mathbf{C}) &= \det(x\mathbf{I} - \mathbf{N})(1 + \mathbf{e}_n^T \det(x\mathbf{I} - \mathbf{N})^{-1}\mathbf{c}) \\ &= x^n \left(1 + \mathbf{e}_n^T \left(\frac{\mathbf{I}}{x} + \frac{\mathbf{N}}{x^2} + \frac{\mathbf{N}^2}{x^3} + \cdots + \frac{\mathbf{N}^{n-1}}{x^n}\right)\mathbf{c}\right) \\ &= x^n + \alpha_{n-1}x^{n-1} + \alpha_{n-2}x^{n-2} + \cdots + \alpha_0 \\ &= p(x). \end{aligned}$$

The fact that $p(x)$ is also the minimum polynomial for $\mathbf{C}$ is a consequence of (7.11.5). Set $\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, and let $v_i(x)$ be the minimum polynomial of $\mathbf{e}_i$ with respect to $\mathbf{C}$. Observe that $v_1(x) = p(x)$ because $\mathbf{C}\mathbf{e}_j = \mathbf{e}_{j+1}$ for $j = 1, \ldots, n-1$, so

$$\{\mathbf{e}_1, \mathbf{C}\mathbf{e}_1, \mathbf{C}^2\mathbf{e}_1, \ldots, \mathbf{C}^{n-1}\mathbf{e}_1\} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_n\}$$

and

$$\mathbf{C}^n\mathbf{e}_1 = \mathbf{C}\mathbf{e}_n = \mathbf{C}_{*n} = -\sum_{j=0}^{n-1} \alpha_j \mathbf{e}_{j+1} = -\sum_{j=0}^{n-1} \alpha_j \mathbf{C}^j\mathbf{e}_1 \implies v_1(x) = p(x).$$

Since $v_1(x)$ divides the LCM of all $v_i(x)$'s (which we know from (7.11.5) to be the minimum polynomial $m(x)$ for $\mathbf{C}$), we conclude that $p(x)$ divides $m(x)$. But $m(x)$ always divides $p(x)$—recall (7.11.4)—so $m(x) = p(x)$. ∎

Example 7.11.2

Poor Man's Root Finder. The companion matrix is the source of what is often called the *poor man's root finder* because any general purpose algorithm designed to compute eigenvalues (e.g., the QR iteration on p. 535) can be applied to the companion matrix for a polynomial $p(x)$ to compute the roots of $p(x)$. When used in conjunction with (7.1.12) on p. 497, the companion matrix is also a *poor man's root bounder*. For example, it follows that if λ is a root of $p(x)$, then

$$|\lambda| \leq \|\mathbf{C}\|_\infty = \max\{|\alpha_0|, 1+|\alpha_1|, \ldots, 1+|\alpha_{n-1}|\} \leq 1 + \max|\alpha_i|.$$

The results on p. 647 insure that the minimum polynomial $v(x)$ for every nonzero vector $\mathbf{b}$ relative to $\mathbf{A} \in \mathcal{C}^{n\times n}$ divides the minimum polynomial $m(x)$ for $\mathbf{A}$, which in turn divides the characteristic polynomial $c(x)$ for $\mathbf{A}$, so it follows that every $v(x)$ divides $c(x)$. This suggests that it might be possible to construct $c(x)$ as a product of $v_i(x)$'s. In fact, this is what Krylov did in 1931, and the following example shows how he did it.

Example 7.11.3

Krylov's method for constructing the characteristic polynomial for $\mathbf{A} \in \mathcal{C}^{n\times n}$ as a product of minimum polynomials for vectors is as follows.

Starting with any nonzero vector $\mathbf{b}_{n\times 1}$, let $v_1(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j$ be the minimum polynomial for $\mathbf{b}$ relative to $\mathbf{A}$, and let $\mathbf{K}_1 = \left(\mathbf{b} \,|\, \mathbf{A}\mathbf{b} \,|\, \cdots \,|\, \mathbf{A}^{k-1}\mathbf{b}\right)_{n\times k}$ be the associated Krylov matrix. Notice that $rank\,(\mathbf{K}_1) = k$ (by definition of the minimum polynomial for $\mathbf{b}$). If $\mathbf{C}_1$ is the $k \times k$ companion matrix of $v(x)$ as described in (7.11.6), then direct multiplication shows that

$$\mathbf{K}_1\mathbf{C}_1 = \mathbf{A}\mathbf{K}_1. \tag{7.11.7}$$

If $k = n$, then $\mathbf{K}_1^{-1}\mathbf{A}\mathbf{K}_1 = \mathbf{C}_1$, so $v_1(x)$ must be the characteristic polynomial for $\mathbf{A}$, and there is nothing more to do. If $k < n$, then use any $n \times (n-k)$ matrix $\widetilde{\mathbf{K}}_1$ such that $\mathbf{K}_2 = \left(\mathbf{K}_1 \,|\, \widetilde{\mathbf{K}}_1\right)_{n\times n}$ is nonsingular, and use (7.11.7) to write

$$\mathbf{A}\mathbf{K}_2 = \left(\mathbf{A}\mathbf{K}_1 \,|\, \mathbf{A}\widetilde{\mathbf{K}}_1\right) = \left(\mathbf{K}_1 \,|\, \widetilde{\mathbf{K}}_1\right)\begin{pmatrix} \mathbf{C}_1 & \mathbf{X} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} \mathbf{X} \\ \mathbf{A}_2 \end{pmatrix} = \mathbf{K}_2^{-1}\mathbf{A}\widetilde{\mathbf{K}}_1.$$

Therefore, $\mathbf{K}_2^{-1}\mathbf{A}\mathbf{K}_2 = \begin{pmatrix} \mathbf{C}_1 & \mathbf{X} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}$, and hence

$$c(x) = \det(x\mathbf{I} - \mathbf{A}) = \det(x\mathbf{I} - \mathbf{C}_1)\det(x\mathbf{I} - \mathbf{A}_2) = v_1(x)\det(x\mathbf{I} - \mathbf{A}_2).$$

Repeat the process on $\mathbf{A}_2$. If the Krylov matrix on the second time around is nonsingular, then $c(x) = v_1(x)v_2(x)$; otherwise $c(x) = v_1(x)v_2(x)\det(x\mathbf{I} - \mathbf{A}_3)$ for some matrix $\mathbf{A}_3$. Continuing in this manner until a nonsingular Krylov matrix is obtained—say at the m^{th} step—produces a nonsingular matrix $\mathbf{K}$ such that

$$\mathbf{K}^{-1}\mathbf{A}\mathbf{K} = \begin{pmatrix} \mathbf{C}_1 & \cdots & \star \\ & \ddots & \vdots \\ & & \mathbf{C}_m \end{pmatrix} = \mathbf{H}, \tag{7.11.8}$$

where the $\mathbf{C}_j$'s are companion matrices, and thus $c(x) = v_1(x)v_2(x)\cdots v_m(x)$.

Note: All companion matrices are upper-Hessenberg matrices as described in Example 5.7.4 (p. 350)—e.g., a 5×5 Hessenberg form is

$$\mathbf{H}_5 = \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}.$$

Since the matrix $\mathbf{H}$ in (7.11.8) is upper Hessenberg, we see that Krylov's method boils down to a recipe for using Krylov sequences to build a similarity transformation that will reduce $\mathbf{A}$ to upper-Hessenberg form. In effect, this means that most information about $\mathbf{A}$ can be derived from Krylov sequences and the associated Hessenberg form $\mathbf{H}$. This is the real message of this example.

Deriving information about $\mathbf{A}$ by using a Hessenberg form and a Krylov similarity transformation as shown in (7.11.8) has some theoretical appeal, but it's not a practical idea as far as computation is concerned. Krylov sequences tend to be nearly linearly dependent sets because, as the power method of Example 7.3.7 (p. 533) indicates, the directions of the vectors $\mathbf{A}^k\mathbf{b}$ want to converge to the direction of an eigenvector for $\mathbf{A}$, so, as k grows, the vectors in a Krylov sequence become ever closer to being multiples of each other. This means that Krylov matrices tend to be ill conditioned. Putting conditioning issues aside, there is still a problem with computational efficiency because $\mathbf{K}$ is usually a dense matrix (one with a preponderance of nonzero entries) even when $\mathbf{A}$ is sparse (which it often is in applied work), so the amount of arithmetic involved in the reduction (7.11.8) is prohibitive.

However, these objections often can be overcome by replacing a Krylov matrix $\mathbf{K} = \big(\mathbf{b} \,|\, \mathbf{A}\mathbf{b} \,|\, \cdots \,|\, \mathbf{A}^{k-1}\mathbf{b}\big)$ with its QR factorization $\mathbf{K} = \mathbf{Q}_{n\times k}\mathbf{R}_{k\times k}$. Doing so in (7.11.7) (and dropping the subscript) produces

$$\mathbf{A}\mathbf{K} = \mathbf{K}\mathbf{C} \implies \mathbf{A}\mathbf{Q}\mathbf{R} = \mathbf{Q}\mathbf{R}\mathbf{C} \implies \mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{R}\mathbf{C}\mathbf{R}^{-1} = \mathbf{H}. \tag{7.11.9}$$

While $\mathbf{H} = \mathbf{R}\mathbf{C}\mathbf{R}^{-1}$ is no longer a companion matrix, it's still in upper-Hessenberg form (convince yourself by writing out the pattern for the 4×4 case). In other words, an orthonormal basis for a Krylov subspace can reduce a

matrix to upper-Hessenberg form. Since matrices with orthonormal columns are perfectly conditioned, the first objection raised above is overcome. The second objection concerning computational efficiency is dealt with in Examples 7.11.4 and 7.11.5.

If $k < n$, then $\mathbf{Q}$ is not square, and $\mathbf{Q}^*\mathbf{AQ} = \mathbf{H}$ is not a similarity transformation, so it would be wrong to conclude that $\mathbf{A}$ and $\mathbf{H}$ have the same spectral properties. Nevertheless, it's often the case that the eigenvalues of $\mathbf{H}$, which are called the ***Ritz values*** for $\mathbf{A}$, are remarkably good approximations to the extreme eigenvalues of $\mathbf{A}$, especially when $\mathbf{A}$ is hermitian. This is somewhat intuitive because $\mathbf{Q}^*\mathbf{AQ}$ can be viewed as a generalization of (7.5.4) on p. 549 that says $\lambda_{\max} = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{Ax}$ and $\lambda_{\min} = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{Ax}$. The results of Exercise 5.9.15 (p. 392) can be used to argue the point further.

Example 7.11.4

Lanczos[87] Tridiagonalization Algorithm. The fact that the matrix $\mathbf{H}$ in (7.11.9) is upper Hessenberg is particularly nice when $\mathbf{A}$ is real and symmetric because $\mathbf{A}^T = \mathbf{A}$ implies $\mathbf{H}^T = (\mathbf{Q}^T\mathbf{AQ})^T = \mathbf{H}$, and symmetric Hessenberg matrices are tridiagonal in structure. That is,

$$\mathbf{H} = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{pmatrix} \quad \text{when } \mathbf{A} = \mathbf{A}^T. \tag{7.11.10}$$

This makes $\mathbf{Q}$ particularly easy to determine. While the matrix $\mathbf{Q}$ in (7.11.9) was only $n \times k$, let's be greedy and look for an $n \times n$ orthogonal matrix $\mathbf{Q}$ such that $\mathbf{AQ} = \mathbf{QH}$, where $\mathbf{H}$ is tridiagonal as depicted in (7.11.10). If we set $\mathbf{Q} = \begin{pmatrix} \mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n \end{pmatrix}$, and if we agree to let $\beta_0 = 0$ and $\mathbf{q}_{n+1} = \mathbf{0}$, then

[87] Cornelius Lanczos (1893–1974) was born Kornél Löwy in Budapest, Hungary, to Jewish parents, but he changed his name to avoid trouble during the dangerous times preceding World War II. After receiving his doctorate from the University of Budapest in 1921, Lanczos moved to Germany where he became Einstein's assistant in Berlin in 1928. After coming home to Germany from a visit to Purdue University in Lafayette, Indiana, in 1931, Lanczos decided that the political climate in Germany was unacceptable, and he returned to Purdue in 1932 to continue his work in mathematical physics. The development of electronic computers stimulated Lanczos's interest in numerical analysis, and this led to positions at the Boeing Company in Seattle and at the Institute for Numerical Analysis of the National Bureau of Standards in Los Angeles. When senator Joseph R. McCarthy led a crusade against communism in the 1950s, Lanczos again felt threatened, so he left the United States to accept an offer from the famous Nobel physicist Erwin Schrödinger (1887–1961) to head the Theoretical Physics Department at the Dublin Institute for Advanced Study in Ireland where Lanczos returned to his first love—the theory of relativity. Lanczos was aware of the fast Fourier transform algorithm (p. 373) 25 years before the heralded work of J. W. Cooley and J. W. Tukey (p. 368) in 1965, but 1940 was too early for applications of the FFT to be realized. This is yet another instance where credit and fame are accorded to those who first make good use of an idea rather than to those who first conceive it.

equating the j^{th} column of $\mathbf{AQ}$ to the j^{th} column of $\mathbf{QH}$ tells us that we must have

$$\mathbf{A}\mathbf{q}_j = \beta_{j-1}\mathbf{q}_{j-1} + \alpha_j\mathbf{q}_j + \beta_j\mathbf{q}_{j+1} \quad \text{for } j = 1, 2, \ldots, n$$

or, equivalently,

$$\beta_j\mathbf{q}_{j+1} = \mathbf{v}_j, \quad \text{where} \quad \mathbf{v}_j = \mathbf{A}\mathbf{q}_j - \alpha_j\mathbf{q}_j - \beta_{j-1}\mathbf{q}_{j-1} \quad \text{for } j = 1, 2, \ldots, n.$$

By observing that $\alpha_j = \mathbf{q}_j^T\mathbf{A}\mathbf{q}_j$ and $\beta_j = \|\mathbf{v}_j\|_2$, we are led to ***Lanczos's algorithm.***

- Start with an arbitrary $\mathbf{b} \neq \mathbf{0}$, set $\beta_0 = 0$, $\mathbf{q}_0 = \mathbf{0}$, $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$, and iterate as indicated below.

For $j = 1$ *to* n
$\quad \mathbf{v} \leftarrow \mathbf{A}\mathbf{q}_j$
$\quad \alpha_j \leftarrow \mathbf{q}_j^T\mathbf{v}$
$\quad \mathbf{v} \leftarrow \mathbf{v} - \alpha_j\mathbf{q}_j - \beta_{j-1}\mathbf{q}_{j-1}$
$\quad \beta_j \leftarrow \|\mathbf{v}\|_2$
$\quad$ *If* $\beta_j = 0$, *then quit*
$\quad \mathbf{q}_{j+1} \leftarrow \mathbf{v}/\beta_j$
End

After the k^{th} step we have an $n \times (k+1)$ matrix $\mathbf{Q}_{k+1} = \left(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_{k+1}\right)$ of orthonormal columns such that

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\begin{pmatrix} \mathbf{T}_k \\ \beta_k\mathbf{e}_k^T \end{pmatrix}, \text{ where } \mathbf{T}_k \text{ is the } k \times k \text{ tridiagonal form (7.11.10).}$$

If the iteration terminates prematurely because $\beta_j = 0$ for $j < n$, then restart the algorithm with a new initial vector $\mathbf{b}$ that is orthogonal to $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_j$. When a full orthonormal set $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ has been computed and turned into an orthogonal matrix $\mathbf{Q}$, we will have

$$\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{T}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{T}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{T}_m \end{pmatrix}, \quad \text{where each } \mathbf{T}_i \text{ is tridiagonal} \qquad (7.11.11)$$

with the splits occurring at rows where the β_j's are zero. Of course, having these splits is generally a desirable state of affairs, especially when the objective is to compute the eigenvalues of $\mathbf{A}$.

Note: The Lanczos algorithm is computationally efficient because if each row of $\mathbf{A}$ has ν nonzero entries, then each matrix–vector product uses νn multiplications, so each step of the process uses only $\nu n + 4n$ multiplications (and about

the same number of additions). This can be a tremendous savings over what is required by Householder (or Givens) reduction as discussed in Example 5.7.4 (p. 350). Once the form (7.11.11) has been determined, spectral properties of $\mathbf{A}$ usually can be extracted by a variety of standard methods such as the QR iteration (p. 535). An alternative to computing the full tridiagonal decomposition is to stop the Lanczos iteration before completion, accept the Ritz values (the eigenvalues $\mathbf{H}_{k\times k} = \mathbf{Q}_{k\times n}^T \mathbf{A}\mathbf{Q}_{n\times k}$) as approximations to a portion of $\sigma(\mathbf{A})$, deflate the problem, and repeat the process on the smaller result.

Even when $\mathbf{A}$ is not symmetric, the same logic that produces the Lanczos algorithm can be applied to obtain an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{H}$ is upper Hessenberg. But we can't expect to obtain the efficiency that Lanczos provides because the tridiagonal structure is lost. The more general algorithm is called ***Arnoldi's*** [88] ***method,*** and it's presented below.

Example 7.11.5

Arnoldi Orthogonalization Algorithm. Given $\mathbf{A} \in \mathcal{C}^{n\times n}$, the goal is to compute an orthogonal matrix $\mathbf{Q} = \begin{pmatrix}\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n\end{pmatrix}$ such that $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{H}$ is upper Hessenberg. Proceed in the manner that produced the Lanczos algorithm by equating the j^{th} column of $\mathbf{AQ}$ to the j^{th} column of $\mathbf{QH}$ to obtain

$$\mathbf{A}\mathbf{q}_j = \sum_{i=1}^{j+1} \mathbf{q}_i h_{ij} \quad \Longrightarrow \quad \mathbf{q}_k^T\mathbf{A}\mathbf{q}_j = \sum_{i=1}^{j+1} \mathbf{q}_k^T\mathbf{q}_i h_{ij} = h_{kj} \quad \text{for each } 1 \le k \le j$$

$$\Longrightarrow \quad h_{j+1,j}\mathbf{q}_{j+1} = \mathbf{A}\mathbf{q}_j - \sum_{i=1}^{j} \mathbf{q}_i h_{ij}.$$

By observing that $h_{j+1,j} = \|\mathbf{v}_j\|_2$ for $\mathbf{v}_j = \mathbf{A}\mathbf{q}_j - \sum_{i=1}^{j} \mathbf{q}_i h_{ij}$, we are led to ***Arnoldi's algorithm.***

- Start with an arbitrary $\mathbf{b} \neq \mathbf{0}$, set $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$, and then iterate as indicated below.

---

[88] Walter Edwin Arnoldi (1917–1995) was an American engineer who published this technique in 1951, not far from the time that Lanczos's algorithm emerged. Arnoldi received his undergraduate degree in mechanical engineering from Stevens Institute of Technology, Hoboken, New Jersey, in 1937 and his MS degree at Harvard University in 1939. He spent his career working as an engineer in the Hamilton Standard Division of the United Aircraft Corporation where he eventually became the division's chief researcher. He retired in 1977. While his research concerned mechanical and aerodynamic properties of aircraft and aerospace structures, Arnoldi's name is kept alive by his orthogonalization procedure.

$$
\begin{array}{l}
\textit{For } j = 1 \textit{ to } n \\
\quad \mathbf{v} \quad \leftarrow \quad \mathbf{A}\mathbf{q}_j \\
\quad \textit{For } i = 1 \textit{ to } j \\
\quad\quad h_{ij} \leftarrow \mathbf{q}_i^T\mathbf{v} \\
\quad\quad \mathbf{v} \leftarrow \mathbf{v} - h_{ij}\mathbf{q}_i \\
\quad \textit{End For} \\
\quad h_{j+1,j} \leftarrow \|\mathbf{v}\|_2 \\
\quad \textit{If } h_{j+1,j} = 0, \textit{ then quit} \\
\quad \mathbf{q}_{j+1} \leftarrow \mathbf{v}/h_{j+1,j} \\
\textit{End For}
\end{array}
\tag{7.11.12}
$$

After the k^{th} step we have an $n \times (k+1)$ matrix $\mathbf{Q}_{k+1} = \left(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_{k+1}\right)$ of orthonormal columns such that

$$
\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\begin{pmatrix} \mathbf{H}_k \\ h_{k+1,k}\mathbf{e}_k^T \end{pmatrix}, \tag{7.11.13}
$$

where $\mathbf{H}_k$ is a $k \times k$ upper-Hessenberg matrix.

Note: Remarks similar to those made about the Lanczos algorithm also hold for Arnoldi's algorithm, but the computational efficiency of Arnoldi is not as great as that of Lanczos. Close examination of Arnoldi's method reveals that it amounts to a modified Gram–Schmidt process (p. 316).

Krylov methods are a natural way to solve systems of linear equations. To see why, suppose that $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ is a nonsingular system, and let $v(x) = x^k - \sum_{j=0}^{k-1}\alpha_j x^j$ be the minimum polynomial of $\mathbf{b}$ with respect to $\mathbf{A}$. Since $\alpha_0 \neq 0$ (otherwise $v(x)/x$ would be an annihilating polynomial for $\mathbf{b}$ of degree less than $\deg v$), we have

$$
\mathbf{A}^k\mathbf{b} - \sum_{j=0}^{k-1}\alpha_j\mathbf{A}^j\mathbf{b} = \mathbf{0} \implies \mathbf{A}\left[\frac{\mathbf{A}^{k-1}\mathbf{b} - \alpha_{k-1}\mathbf{A}^{k-2}\mathbf{b} - \cdots - \alpha_1\mathbf{b}}{\alpha_0}\right] = \mathbf{b}.
$$

In other words, the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is somewhere in the Krylov space $\mathcal{K}_k$.

A technique for sorting through $\mathcal{K}_k$ to find the solution (or at least an acceptable approximate solution) of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is to sequentially consider the subspaces $\mathbf{A}(\mathcal{K}_1)$, $\mathbf{A}(\mathcal{K}_2)$, $\ldots$, $\mathbf{A}(\mathcal{K}_k)$, where at the j^{th} step of the process the vector $\mathbf{x}_j \in \mathbf{A}(\mathcal{K}_j)$ that is closest to $\mathbf{b}$ is used as an approximation to $\mathbf{x}$. If $\mathbf{Q}_j$ is an $n \times j$ orthogonal matrix whose columns constitute a basis for $\mathcal{K}_j$, then $R(\mathbf{A}\mathbf{Q}_j) = \mathbf{A}(\mathcal{K}_j)$, so the vector $\mathbf{x}_j \in \mathbf{A}(\mathcal{K}_j)$ that is closest to $\mathbf{b}$ is the orthogonal projection of $\mathbf{b}$ onto $R(\mathbf{A}\mathbf{Q}_j)$. This means that $\mathbf{x}_j$ is the least squares solution of $\mathbf{A}\mathbf{Q}_j\mathbf{z} = \mathbf{b}$ (p. 439). If the solution of this least squares problem yields a vector $\mathbf{x}_j$ such that the residual $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{Q}_j\mathbf{x}_j$ is zero (or satisfactorily small), then set $\mathbf{x} = \mathbf{Q}_j\mathbf{x}_j$, and quit. Otherwise move up one

dimension, and compute the least squares solution $\mathbf{x}_{j+1}$ of $\mathbf{AQ}_{j+1}\mathbf{z} = \mathbf{b}$. Since $\mathbf{x} \in \mathcal{K}_k$, the process is guaranteed to terminate in $k \leq n$ steps or less (when exact arithmetic is used). When Arnoldi's method is used to implement this idea, the resulting algorithm is known as ***GMRES*** (an acronym for the *generalized minimal residual* algorithm that was formulated by Yousef Saad and Martin H. Schultz in 1986).

Example 7.11.6

GMRES Algorithm. To implement the idea discussed above by employing Arnoldi's algorithm, recall from (7.11.13) that after j steps of the Arnoldi process we have matrices $\mathbf{Q}_j$ and $\mathbf{Q}_{j+1}$ with orthonormal columns that span $\mathcal{K}_j$ and $\mathcal{K}_{j+1}$, respectively, along with a $j \times j$ upper-Hessenberg matrix $\mathbf{H}_j$ such that

$$\mathbf{AQ}_j = \mathbf{Q}_{j+1}\widetilde{\mathbf{H}}_j, \quad \text{where} \quad \widetilde{\mathbf{H}}_j = \begin{pmatrix} \mathbf{H}_j \\ h_{j+1,j}\mathbf{e}_j^T \end{pmatrix}.$$

Consequently the least squares solution of $\mathbf{AQ}_j\mathbf{z} = \mathbf{b}$ is the same as the least squares solution of $\mathbf{Q}_{j+1}\widetilde{\mathbf{H}}_j\mathbf{z} = \mathbf{b}$, which in turn is the same as the least squares solution of $\widetilde{\mathbf{H}}_j\mathbf{z} = \mathbf{Q}_{j+1}^T\mathbf{b}$. But $\mathbf{Q}_{j+1}^T\mathbf{b} = \|\mathbf{b}\|_2\,\mathbf{e}_1$ (because the first column in $\mathbf{Q}_{j+1}$ is $\mathbf{b}/\|\mathbf{b}\|_2$), so the GMRES algorithm is as follows.

- To compute the solution to a nonsingular linear system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b} \neq \mathbf{0}$, start with $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$, and iterate as indicated below.

For $j = 1$ *to* n
  execute the j^{th} Arnoldi step in (7.11.12)
  compute the least squares solution of $\widetilde{\mathbf{H}}_j\mathbf{z} = \|\mathbf{b}\|_2\,\mathbf{e}_1$ by using a QR factorization of $\widetilde{\mathbf{H}}_j$ (see **Note** at the end of the example)
  If $\|\mathbf{b} - \mathbf{AQ}_j\mathbf{z}\|_2 = 0$ (or is satisfactorily small)
    set $\mathbf{x} = \mathbf{Q}_j\mathbf{z}$, and quit (see **Note** at the end of the example)
  End If
End For

The structure of the $\widetilde{\mathbf{H}}_j$'s allows us to update the QR factors of $\widetilde{\mathbf{H}}_j$ to produce the QR factors of $\widetilde{\mathbf{H}}_{j+1}$ with a single plane rotation (p. 333). To see how this is done, consider what happens when moving from the third step to the fourth step of the process. Let $\mathbf{U}_3 = \begin{pmatrix} \mathbf{Q}^T \\ \mathbf{v}^T \end{pmatrix}$ be the 4×4 orthogonal matrix that was previously accumulated (as a product of plane rotations) to give $\mathbf{U}_3\widetilde{\mathbf{H}}_3 = \begin{pmatrix} \mathbf{R}_3 \\ \mathbf{0} \end{pmatrix}$ with $\mathbf{R}_3$ being upper triangular so that $\widetilde{\mathbf{H}}_3 = \mathbf{QR}_3$. Since

$$\begin{pmatrix} \mathbf{U}_3 & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}\widetilde{\mathbf{H}}_4 = \begin{pmatrix} \mathbf{U}_3 & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}\left(\begin{array}{ccc|c} & \widetilde{\mathbf{H}}_3 & & \begin{matrix}\star\\\star\\\star\\\star\end{matrix} \\ \hline 0 & 0 & 0 & \star \end{array}\right) = \left(\begin{array}{ccc|c} & \mathbf{U}_3\widetilde{\mathbf{H}}_3 & & \begin{matrix}\star\\\star\\\star\\\star\end{matrix} \\ \hline 0 & 0 & 0 & \star \end{array}\right) = \left(\begin{array}{cccc} \star & \star & \star & \star \\ 0 & \star & \star & \star \\ 0 & 0 & \star & \star \\ \hline 0 & 0 & 0 & \star \end{array}\right),$$

a plane rotation of the form $\mathbf{P}_{45} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & \begin{matrix} c & s \\ -s & c \end{matrix} \end{pmatrix}$ will annihilate the entry in the lower-right-hand corner of this last array. Consequently, $\mathbf{U}_4 = \mathbf{P}_{45}\begin{pmatrix} \mathbf{U}_3 & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}$ is an orthogonal matrix such that $\mathbf{U}_4\widetilde{\mathbf{H}}_4 = \begin{pmatrix} \mathbf{R}_4 \\ \mathbf{0} \end{pmatrix}$, where $\mathbf{R}_4$ is upper triangular, and this produces the QR factors of $\widetilde{\mathbf{H}}_4$.

Note: The value of the residual norm $\|\mathbf{b} - \mathbf{A}\mathbf{Q}_j\mathbf{z}\|_2$ at each step of GMRES is available at almost no cost. To see why, notice that the previous discussion shows that at the j^{th} step there is a $(j+1)\times(j+1)$ orthogonal matrix $\mathbf{U} = \begin{pmatrix} \mathbf{Q}^T \\ \mathbf{v}^T \end{pmatrix}$ (that exists as an accumulation of plane rotations) such that $\mathbf{U}\widetilde{\mathbf{H}}_j = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}$, and this produces $\widetilde{\mathbf{H}}_j = \mathbf{Q}\mathbf{R}$. The least squares solution of $\widetilde{\mathbf{H}}_j\mathbf{z} = \|\mathbf{b}\|_2\,\mathbf{e}_1$ is obtained by solving $\mathbf{R}\mathbf{z} = \mathbf{Q}^T\|\mathbf{b}\|_2\,\mathbf{e}_1$ (p. 314), so

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\mathbf{Q}_j\mathbf{z}\|_2 &= \left\|\|\mathbf{b}\|_2\,\mathbf{e}_1 - \widetilde{\mathbf{H}}_j\mathbf{z}\right\|_2 = \left\|\|\mathbf{b}\|_2\,\mathbf{U}\mathbf{e}_1 - \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}\mathbf{z}\right\|_2 \\ &= \left\|\|\mathbf{b}\|_2\begin{pmatrix} \mathbf{Q}^T \\ \mathbf{v}^T \end{pmatrix}\mathbf{e}_1 - \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}\mathbf{z}\right\|_2 = \left\|\begin{pmatrix} \mathbf{0} \\ \|\mathbf{b}\|_2\,\mathbf{v}^T\mathbf{e}_1 \end{pmatrix}\right\|_2 \\ &= \|\mathbf{b}\|_2\,|u_{j+1,1}|. \end{aligned}$$

Since $u_{j+1,1}$ is just the last entry in the accumulation of the various plane rotations applied to $\mathbf{e}_1$, the cost of producing these values as the algorithm proceeds is small, so deciding on the acceptability of an approximate solution at each step in the GMRES algorithm is cheap.

When solving nonsingular symmetric systems $\mathbf{A}\mathbf{x} = \mathbf{b}$, a strategy similar to the one that produced the GMRES algorithm can be adopted except that the Lanczos procedure (p. 651) is used in place of the Arnoldi process (p. 653). When this is done, the resulting algorithm is called ***MINRES*** (an acronym for *minimal residual algorithm*), and, as you might guess, there is an increase in computational efficiency when Lanczos replaces Arnoldi. Historically, MINRES preceded GMRES.

Another Krylov method that deserves mention is the ***conjugate gradient algorithm***, presented by Magnus R. Hestenes and Eduard Stiefel in 1952, that is used to solve positive definite systems.

Example 7.11.7

Conjugate Gradient Algorithm. Suppose that $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b} \neq \mathbf{0}$ is a (real) positive definite system, and suppose that the minimum polynomial of $\mathbf{b}$ with respect to $\mathbf{A}$ is $v(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j$ so that the solution $\mathbf{x}$ is somewhere in the Krylov space $\mathcal{K}_k$ (p. 654). The conjugate gradient algorithm emanated from the observation that if $\mathbf{A}$ is positive definite, then the quadratic function

$$f(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{2} - \mathbf{x}^T\mathbf{b}$$

has as its gradient

$$\nabla f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b},$$

and there is a unique minimizer for f that happens to be the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$. Consequently, any technique that attempts to minimize f is a technique that attempts to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$. Since the $\mathbf{x}$ is somewhere in $\mathcal{K}_k$, it makes sense to try to minimize f over $\mathcal{K}_k$. One approach for doing this is the ***method of steepest descent*** in which a current approximation $\mathbf{x}_j$ is updated by adding a correction term directed along the negative gradient $-\nabla f(\mathbf{x}_j) = \mathbf{b} - \mathbf{A}\mathbf{x}_j = \mathbf{r}_j$ (the j^{th} residual). In other words, let

$$\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j\mathbf{r}_j, \quad \text{and set} \quad \alpha_j = \frac{\mathbf{r}_j^T\mathbf{r}_j}{\mathbf{r}_j^T\mathbf{A}\mathbf{r}_j}$$

because this α_j minimizes $f(\mathbf{x}_{j+1})$. In spite of the fact that successive residuals are orthogonal ($\mathbf{r}_{j+1}^T\mathbf{r}_j = \mathbf{0}$), the rate of convergence can be slow because as the ratio of eigenvalues $\lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A})$ becomes larger, the surface defined by f becomes more distorted, and a negative gradient $\mathbf{r}_j$ need not point in a direction aimed anywhere near the lowest point on the surface. An ingenious mechanism for overcoming this difficulty is to replace the search directions $\mathbf{r}_j$ by directions defined by vectors $\mathbf{q}_1, \mathbf{q}_2, \ldots$ that are ***conjugate*** to each other in the sense that $\mathbf{q}_i^T\mathbf{A}\mathbf{q}_j = 0$ for all $i \neq j$ (some authors say "A-orthogonal"). Starting with $\mathbf{x}_0 = \mathbf{0}$, the idea is to begin by moving in the direction of steepest descent with

$$\mathbf{x}_1 = \alpha_1\mathbf{q}_1, \quad \text{where } \mathbf{q}_1 = \mathbf{r}_0 = \mathbf{b} \text{ and } \alpha_1 = \frac{\mathbf{r}_0^T\mathbf{r}_0}{\mathbf{r}_0^T\mathbf{A}\mathbf{r}_0},$$

but at the second step use a direction vector

$$\mathbf{q}_2 = \mathbf{r}_1 + \beta_1\mathbf{q}_1, \quad \text{where } \beta_1 \text{ is chosen to force } \mathbf{q}_2^T\mathbf{A}\mathbf{q}_1 = 0.$$

With a bit of effort you can see that $\beta_1 = \mathbf{r}_1^T\mathbf{r}_1/\mathbf{r}_0^T\mathbf{r}_0$ does the job. Then set $\mathbf{x}_2 = \mathbf{x}_1 + \alpha_2\mathbf{q}_2$, and recycle the process. The formal algorithm is as follows.

Formal Conjugate Gradient Algorithm. To compute the solution to a positive definite linear system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$, start with $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$, and $\mathbf{q}_1 = \mathbf{b}$, and iterate as indicated below.

For $j = 1$ *to* n

α_j	$\leftarrow$	$\mathbf{r}_{j-1}^T\mathbf{r}_{j-1}/\mathbf{q}_j^T\mathbf{A}\mathbf{q}_j$	(step size)
$\mathbf{x}_j$	$\leftarrow$	$\mathbf{x}_{j-1} + \alpha_j\mathbf{q}_j$	(approximate solution)
$\mathbf{r}_j$	$\leftarrow$	$\mathbf{r}_{j-1} - \alpha_j\mathbf{A}\mathbf{q}_j$	(residual)

If $\|\mathbf{r}_j\|_2 = 0$ (or is satisfactorily small)
  set $\mathbf{x} = \mathbf{x}_j$, and quit
End If

β_j	$\leftarrow$	$\mathbf{r}_j^T\mathbf{r}_j/\mathbf{r}_{j-1}^T\mathbf{r}_{j-1}$	(conjugation factor)
$\mathbf{q}_{j+1}$	$\leftarrow$	$\mathbf{r}_j + \beta_j\mathbf{q}_j$	(search direction)

End For

It can be shown that vectors produced by this algorithm after j steps are such that (in exact arithmetic)

$$span\{\mathbf{x}_1, \ldots, \mathbf{x}_j\} = span\{\mathbf{q}_1, \ldots, \mathbf{q}_j\} = span\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{j-1}\} = \mathcal{K}_j,$$

and, in addition to having $\mathbf{q}_i\mathbf{A}\mathbf{q}_j = 0$ for $i < j$, the residuals are orthogonal—i.e., $\mathbf{r}_i^T\mathbf{r}_j = 0$ for $i < j$. Furthermore, the algorithm will find the solution in $k \leq n$ steps.

As mentioned earlier, Krylov solvers such as GMRES and the conjugate gradient algorithm produce the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ in $k \leq n$ steps (in exact arithmetic), so, at first glance, this looks like good news. But in practice n can be prohibitively large, and it's not rare to have $k = n$. Consequently, Krylov algorithms are often viewed as iterative methods that are terminated long before n steps have been completed. The challenge in applying Krylov solvers (as well as iterative methods in general) revolves around the issue of how to replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ with an equivalent ***preconditioned system*** $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$ that requires only a small number of iterations to deliver a reasonably accurate approximate solution. Building effective preconditioners $\mathbf{M}^{-1}$ is part science and part art, and the techniques vary from algorithm to algorithm.

Classical linear stationary iterative methods (p. 620) are formed by splitting $\mathbf{A} = \mathbf{M} - \mathbf{N}$ and setting $\mathbf{x}(k) = \mathbf{H}\mathbf{x}(k-1) + \mathbf{d}$, where $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ and $\mathbf{d} = \mathbf{M}^{-1}\mathbf{b}$. This is a preconditioning technique because the effect is to replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ by $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$, where $\mathbf{M}^{-1}\mathbf{A} = \mathbf{I} - \mathbf{H}$ such that $\rho(\mathbf{H}) < 1$. The goal is to find an easily inverted $\mathbf{M}$ (in the sense that $\mathbf{M}\mathbf{d} = \mathbf{b}$ is easily solved) that drives the value of $\rho(\mathbf{H})$ down far enough to insure a satisfactory rate of convergence, and this is a delicate balancing act.

The goal in preconditioning Krylov solvers is somewhat different. For example, if $k = \deg v(x)$ is the degree of the minimum polynomial of $\mathbf{b}$ with respect to $\mathbf{A}$, then GMRES sorts through $\mathcal{K}_k$ to find the solution of $\mathbf{Ax} = \mathbf{b}$ in k steps. So the aim of preconditioning GMRES might be to manipulate the interplay between $\mathbf{M}^{-1}\mathbf{b}$ and $\mathbf{M}^{-1}\mathbf{A}$ to insure that the degree of minimum polynomial $\widetilde{v}(x)$ of $\mathbf{M}^{-1}\mathbf{b}$ with respect to $\mathbf{M}^{-1}\mathbf{A}$ is significantly smaller than k. Since this is difficult to do, an alternate goal is to try to reduce the degree of the minimum polynomial $\widetilde{m}(x)$ for $\mathbf{M}^{-1}\mathbf{A}$ because driving down $\deg \widetilde{m}(x)$ also drives down $\deg \widetilde{v}(x)$—remember, $\widetilde{v}(x)$ is a divisor of $\widetilde{m}(x)$ (p. 647). If a preconditioner $\mathbf{M}^{-1}$ can be found to force $\mathbf{M}^{-1}\mathbf{A}$ to be diagonalizable with only a few distinct eigenvalues (say j of them), then $\deg \widetilde{m}(x) = j$ (p. 645), and GMRES will find the solution in no more than j steps. But this too is an overly ambitious goal for practical problems. In reality this objective is compromised by looking for a preconditioner such that $\mathbf{M}^{-1}\mathbf{A}$ is diagonalizable whose eigenvalues fall into a few small clusters—say j of them. The hope is that if $\mathbf{M}^{-1}\mathbf{A}$ is diagonalizable, and if the diameters of the clusters are small enough, then $\mathbf{M}^{-1}\mathbf{A}$ will behave numerically like a diagonalizable matrix with j distinct eigenvalues, so GMRES is inclined to produce reasonably accurate approximations in no more than j steps. While the intuition is simple, subtleties involving the magnitudes of eigenvalues, separation of clusters, and the meaning of "small diameter" complicate the picture to make definitive statements and rigorous arguments difficult to formulate. Constructing good preconditioners and proving they actually work as advertised remains an active area of research in the field of numerical analysis.

Only the tip of the iceberg concerning practical applications of Krylov methods is revealed in this section. The analysis required to more fully understand the numerical behavior of various Krylov methods can be found in several excellent advanced texts specializing in matrix computations.

Exercises for section 7.11

7.11.1. Determine the minimum polynomial for $\mathbf{A} = \begin{pmatrix} 5 & 1 & 2 \\ -4 & 0 & -2 \\ -4 & -1 & -1 \end{pmatrix}$.

7.11.2. Find the minimum polynomial of $\mathbf{b} = (-1, 1, 1)^T$ with respect to the matrix $\mathbf{A}$ given in Exercise 7.11.1.

7.11.3. Use Krylov's method to determine the characteristic polynomial for the matrix $\mathbf{A}$ given in Exercise 7.11.1.

7.11.4. What is the Jordan form for a matrix whose minimum polynomial is $m(x) = (x - \lambda)(x - \mu)^2$ and whose characteristic polynomial is $c(x) = (x - \lambda)^2(x - \mu)^4$?

7.11.5. Use the technique described in Example 7.11.1 (p. 643) to determine the minimum polynomial for $\mathbf{A} = \begin{pmatrix} -7 & -4 & 8 & -8 \\ -4 & -1 & 4 & -4 \\ -16 & -8 & 17 & -16 \\ -6 & -3 & 6 & -5 \end{pmatrix}$.

7.11.6. Explain why similar matrices have the same minimum and characteristic polynomials.

7.11.7. Show that two matrices can have the same minimum and characteristic polynomials without being similar by considering $\mathbf{A} = \begin{pmatrix} \mathbf{N} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} \mathbf{N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, where $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

7.11.8. Prove that if $\mathbf{A}$ and $\mathbf{B}$ are nonderogatory matrices that have the same characteristic polynomial, then $\mathbf{A}$ is similar to $\mathbf{B}$.

7.11.9. Use the Lanczos algorithm to find an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{AP} = \mathbf{T}$ is tridiagonal, where $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$.

7.11.10. Starting with $\mathbf{x}_0 = \mathbf{0}$, apply the conjugate gradient algorithm to solve $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}$.

7.11.11. Use Arnoldi's algorithm to find an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^T\mathbf{AQ} = \mathbf{H}$ is upper Hessenberg, where $\mathbf{A} = \begin{pmatrix} 5 & 1 & 2 \\ -4 & 0 & -2 \\ -4 & -1 & -1 \end{pmatrix}$.

7.11.12. Use GMRES to solve $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{A} = \begin{pmatrix} 5 & 1 & 2 \\ -4 & 0 & -2 \\ -4 & -1 & -1 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$.