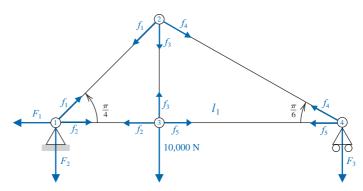
7

Iterative Techniques in Matrix Algebra

Introduction

Trusses are lightweight structures capable of carrying heavy loads. In bridge design, the individual members of the truss are connected with rotatable pin joints that permit forces to be transferred from one member of the truss to another. The accompanying figure shows a truss that is held stationary at the lower left endpoint ①, is permitted to move horizontally at the lower right endpoint ④, and has pin joints at ①, ②, ③, and ④. A load of 10,000 newtons (N) is placed at joint ③, and the resulting forces on the joints are given by f_1 , f_2 , f_3 , f_4 , and f_5 , as shown. When positive, these forces indicate tension on the truss elements, and when negative, compression. The stationary support member could have both a horizontal force component F_1 and a vertical force component F_2 , but the movable support member has only a vertical force component F_3 .



If the truss is in static equilibrium, the forces at each joint must add to the zero vector, so the sum of the horizontal and vertical components at each joint must be 0. This produces the system of linear equations shown in the accompanying table. An 8×8 matrix describing this system has 47 zero entries and only 17 nonzero entries. Matrices with a high percentage of zero entries are called *sparse* and are often solved using iterative, rather than direct, techniques. The iterative solution to this system is considered in Exercise 18 of Section 7.3 and Exercise 10 in Section 7.4.

| Joint | Horizontal Component | Vertical Component |
|-------|--|---|
| 1 | $-F_1 + \frac{\sqrt{2}}{2}f_1 + f_2 = 0$ | $\frac{\sqrt{2}}{2}f_1 - F_2 = 0$ |
| 2 | $-\frac{\sqrt{2}}{2}f_1 + \frac{\sqrt{3}}{2}f_4 = 0$ | $-\frac{\sqrt{2}}{2}f_1 - f_3 - \frac{1}{2}f_4 = 0$ |
| 3 | $-f_2 + f_5 = 0$ | $f_3 - 10,000 = 0$ |
| 4 | $-\frac{\sqrt{3}}{2}f_4 - f_5 = 0$ | $\frac{1}{2}f_4 - F_3 = 0$ |

The methods presented in Chapter 6 used direct techniques to solve a system of $n \times n$ linear equations of the form $A\mathbf{x} = \mathbf{b}$. In this chapter, we present iterative methods to solve a system of this type.

7.1 Norms of Vectors and Matrices

In Chapter 2 we described iterative techniques for finding roots of equations of the form f(x) = 0. An initial approximation (or approximations) was found, and new approximations are then determined based on how well the previous approximations satisfied the equation. The objective is to find a way to minimize the difference between the approximations and the exact solution.

To discuss iterative methods for solving linear systems, we first need to determine a way to measure the distance between n-dimensional column vectors. This will permit us to determine whether a sequence of vectors converges to a solution of the system.

In actuality, this measure is also needed when the solution is obtained by the direct methods presented in Chapter 6. Those methods required a large number of arithmetic operations, and using finite-digit arithmetic leads only to an approximation to an actual solution of the system.

A scalar is a real (or complex) number generally denoted using italic or Greek letters. Vectors are denoted using boldface letters.

Vector Norms

Let \mathbb{R}^n denote the set of all *n*-dimensional column vectors with real-number components. To define a distance in \mathbb{R}^n we use the notion of a norm, which is the generalization of the absolute value on \mathbb{R} , the set of real numbers.

Definition 7.1 A vector norm on \mathbb{R}^n is a function, $\|\cdot\|$, from \mathbb{R}^n into \mathbb{R} with the following properties:

- (i) $\|\mathbf{x}\| \ge 0$ for all $\mathbf{x} \in \mathbb{R}^n$,
- (ii) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$,
- (iii) $\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$ for all $\alpha \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$,
- (iv) $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Vectors in \mathbb{R}^n are column vectors, and it is convenient to use the transpose notation presented in Section 6.3 when a vector is represented in terms of its components. For example, the vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

will be written $\mathbf{x} = (x_1, x_2, \dots, x_n)^t$.

We will need only two specific norms on \mathbb{R}^n , although a third norm on \mathbb{R}^n is presented in Exercise 2.

Definition 7.2 The l_2 and l_{∞} norms for the vector $\mathbf{x} = (x_1, x_2, \dots, x_n)^t$ are defined by

$$\|\mathbf{x}\|_{2} = \left\{ \sum_{i=1}^{n} x_{i}^{2} \right\}^{1/2} \quad \text{and} \quad \|\mathbf{x}\|_{\infty} = \max_{1 \le i \le n} |x_{i}|.$$

Note that each of these norms reduces to the absolute value in the case n = 1.

The l_2 norm is called the **Euclidean norm** of the vector \mathbf{x} because it represents the usual notion of distance from the origin in case \mathbf{x} is in $\mathbb{R}^1 \equiv \mathbb{R}$, \mathbb{R}^2 , or \mathbb{R}^3 . For example, the l_2 norm of the vector $\mathbf{x} = (x_1, x_2, x_3)^t$ gives the length of the straight line joining the points (0,0,0) and (x_1,x_2,x_3) . Figure 7.1 shows the boundary of those vectors in \mathbb{R}^2 and \mathbb{R}^3 that have l_2 norm less than 1. Figure 7.2 is a similar illustration for the l_∞ norm.

Figure 7.1

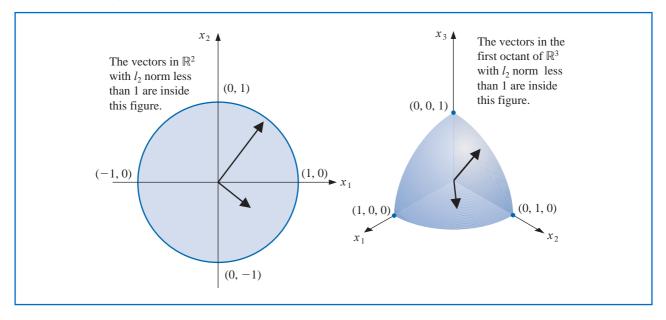
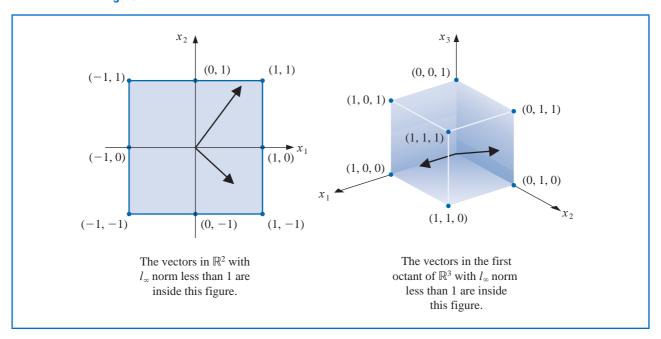


Figure 7.2



Example 1 Determine the l_2 norm and the l_{∞} norm of the vector $\mathbf{x} = (-1, 1, -2)^t$.

Solution The vector $\mathbf{x} = (-1, 1, -2)^t$ in \mathbb{R}^3 has norms

$$\|\mathbf{x}\|_2 = \sqrt{(-1)^2 + (1)^2 + (-2)^2} = \sqrt{6}$$

and

$$\|\mathbf{x}\|_{\infty} = \max\{|-1|, |1|, |-2|\} = 2.$$

It is easy to show that the properties in Definition 7.1 hold for the l_{∞} norm because they follow from similar results for absolute values. The only property that requires much demonstration is (iv), and in this case if $\mathbf{x} = (x_1, x_2, \dots, x_n)^t$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)^t$, then

$$\|\mathbf{x} + \mathbf{y}\|_{\infty} = \max_{1 \le i \le n} |x_i + y_i| \le \max_{1 \le i \le n} (|x_i| + |y_i|) \le \max_{1 \le i \le n} |x_i| + \max_{1 \le i \le n} |y_i| = \|\mathbf{x}\|_{\infty} + \|\mathbf{y}\|_{\infty}.$$

The first three conditions also are easy to show for the l_2 norm. But to show that

$$\|\mathbf{x} + \mathbf{y}\|_{2} \le \|\mathbf{x}\|_{2} + \|\mathbf{y}\|_{2}$$
, for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}_{n}$,

we need a famous inequality.

Theorem 7.3 (Cauchy-Bunyakovsky-Schwarz Inequality for Sums)

For each $\mathbf{x} = (x_1, x_2, \dots, x_n)^t$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)^t$ in \mathbb{R}^n ,

$$\mathbf{x}^{t}\mathbf{y} = \sum_{i=1}^{n} x_{i} y_{i} \le \left\{ \sum_{i=1}^{n} x_{i}^{2} \right\}^{1/2} \left\{ \sum_{i=1}^{n} y_{i}^{2} \right\}^{1/2} = \|\mathbf{x}\|_{2} \cdot \|\mathbf{y}\|_{2}.$$
 (7.1)

Proof If y = 0 or x = 0, the result is immediate because both sides of the inequality are zero.

Suppose $y \neq 0$ and $x \neq 0$. Note that for each $\lambda \in \mathbb{R}$ we have

$$0 \le ||\mathbf{x} - \lambda \mathbf{y}||_2^2 = \sum_{i=1}^n (x_i - \lambda y_i)^2 = \sum_{i=1}^n x_i^2 - 2\lambda \sum_{i=1}^n x_i y_i + \lambda^2 \sum_{i=1}^n y_i^2,$$

so that

$$2\lambda \sum_{i=1}^{n} x_i y_i \le \sum_{i=1}^{n} x_i^2 + \lambda^2 \sum_{i=1}^{n} y_i^2 = \|\mathbf{x}\|_2^2 + \lambda^2 \|\mathbf{y}\|_2^2.$$

However $\|\mathbf{x}\|_2 > 0$ and $\|\mathbf{y}\|_2 > 0$, so we can let $\lambda = \|\mathbf{x}\|_2 / \|\mathbf{y}\|_2$ to give

$$\left(2\frac{\|\mathbf{x}\|_{2}}{\|\mathbf{y}\|_{2}}\right)\left(\sum_{i=1}^{n}x_{i}y_{i}\right) \leq \|\mathbf{x}\|_{2}^{2} + \frac{\|\mathbf{x}\|_{2}^{2}}{\|\mathbf{y}\|_{2}^{2}}\|\mathbf{y}\|_{2}^{2} = 2\|\mathbf{x}\|_{2}^{2}.$$

Hence

$$2\sum_{i=1}^{n} x_i y_i \le 2\|\mathbf{x}\|_2^2 \frac{\|\mathbf{y}\|_2}{\|\mathbf{x}\|_2} = 2\|\mathbf{x}\|_2 \|\mathbf{y}\|_2,$$

There are many forms of this inequality, hence many discoverers. Augustin Louis Cauchy (1789-1857) describes this inequality in 1821 in Cours d'Analyse Algébrique, the first rigorous calculus book. An integral form of the equality appears in the work of Viktor Yakovlevich Bunyakovsky (1804-1889) in 1859, and Hermann Amandus Schwarz (1843-1921) used a double integral form of this inequality in 1885. More details on the history can be found in [Stee].

and

$$\mathbf{x}^{t}\mathbf{y} = \sum_{i=1}^{n} x_{i} y_{i} \le \|\mathbf{x}\|_{2} \|\mathbf{y}\|_{2} = \left\{ \sum_{i=1}^{n} x_{i}^{2} \right\}^{1/2} \left\{ \sum_{i=1}^{n} y_{i}^{2} \right\}^{1/2}.$$

With this result we see that for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{x} + \mathbf{y}\|_{2}^{2} = \sum_{i=1}^{n} (x_{i} + y_{i})^{2} = \sum_{i=1}^{n} x_{i}^{2} + 2 \sum_{i=1}^{n} x_{i} y_{i} + \sum_{i=1}^{n} y_{i}^{2} \le \|\mathbf{x}\|_{2}^{2} + 2\|\mathbf{x}\|_{2}\|\mathbf{y}\|_{2} + \|\mathbf{y}\|_{2}^{2},$$

which gives norm property (iv):

$$\|\mathbf{x} + \mathbf{y}\|_{2} \le (\|\mathbf{x}\|_{2}^{2} + 2\|\mathbf{x}\|_{2}\|\mathbf{y}\|_{2} + \|\mathbf{y}\|_{2}^{2})^{1/2} = \|\mathbf{x}\|_{2} + \|\mathbf{y}\|_{2}.$$

Distance between Vectors in \mathbb{R}^n

The norm of a vector gives a measure for the distance between an arbitrary vector and the zero vector, just as the absolute value of a real number describes its distance from 0. Similarly, the **distance between two vectors** is defined as the norm of the difference of the vectors just as distance between two real numbers is the absolute value of their difference.

Definition 7.4 If $\mathbf{x} = (x_1, x_2, \dots, x_n)^t$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)^t$ are vectors in \mathbb{R}^n , the l_2 and l_∞ distances between \mathbf{x} and \mathbf{y} are defined by

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left\{ \sum_{i=1}^n (x_i - y_i)^2 \right\}^{1/2} \text{ and } \|\mathbf{x} - \mathbf{y}\|_{\infty} = \max_{1 \le i \le n} |x_i - y_i|.$$

Example 2 The linear system

$$3.3330x_1 + 15920x_2 - 10.333x_3 = 15913,$$

 $2.2220x_1 + 16.710x_2 + 9.6120x_3 = 28.544,$
 $1.5611x_1 + 5.1791x_2 + 1.6852x_3 = 8.4254$

has the exact solution $\mathbf{x} = (x_1, x_2, x_3)^t = (1, 1, 1)^t$, and Gaussian elimination performed using five-digit rounding arithmetic and partial pivoting (Algorithm 6.2), produces the approximate solution

$$\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^t = (1.2001, 0.99991, 0.92538)^t.$$

Determine the l_2 and l_{∞} distances between the exact and approximate solutions.

Solution Measurements of $\mathbf{x} - \tilde{\mathbf{x}}$ are given by

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty} = \max\{|1 - 1.2001|, |1 - 0.99991|, |1 - 0.92538|\}$$
$$= \max\{0.2001, 0.00009, 0.07462\} = 0.2001$$

and

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_{2} = \left[(1 - 1.2001)^{2} + (1 - 0.99991)^{2} + (1 - 0.92538)^{2} \right]^{1/2}$$
$$= \left[(0.2001)^{2} + (0.00009)^{2} + (0.07462)^{2} \right]^{1/2} = 0.21356.$$

Although the components \tilde{x}_2 and \tilde{x}_3 are good approximations to x_2 and x_3 , the component \tilde{x}_1 is a poor approximation to x_1 , and $|x_1 - \tilde{x}_1|$ dominates both norms.

The concept of distance in \mathbb{R}^n is also used to define a limit of a sequence of vectors in this space.

Definition 7.5 A sequence $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ of vectors in \mathbb{R}^n is said to **converge** to \mathbf{x} with respect to the norm $\|\cdot\|$ if, given any $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| < \varepsilon$$
, for all $k \ge N(\varepsilon)$.

Theorem 7.6 The sequence of vectors $\{\mathbf{x}^{(k)}\}$ converges to \mathbf{x} in \mathbb{R}^n with respect to the l_{∞} norm if and only if $\lim_{k\to\infty} x_i^{(k)} = x_i$, for each $i = 1, 2, \dots, n$.

Proof Suppose $\{\mathbf{x}^{(k)}\}$ converges to \mathbf{x} with respect to the l_{∞} norm. Given any $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that for all $k \geq N(\varepsilon)$,

$$\max_{i=1,2,...,n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_{\infty} < \varepsilon.$$

This result implies that $|x_i^{(k)} - x_i| < \varepsilon$, for each i = 1, 2, ..., n, so $\lim_{k \to \infty} x_i^{(k)} = x_i$ for each i.

Conversely, suppose that $\lim_{k\to\infty} x_i^{(k)} = x_i$, for every $i = 1, 2, \dots, n$. For a given $\varepsilon > 0$, let $N_i(\varepsilon)$ for each i represent an integer with the property that

$$|x_i^{(k)} - x_i| < \varepsilon,$$

whenever $k \geq N_i(\varepsilon)$.

Define $N(\varepsilon) = \max_{i=1,2,\dots,n} N_i(\varepsilon)$. If $k \ge N(\varepsilon)$, then

$$\max_{i=1,2,\dots,n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_{\infty} < \varepsilon.$$

This implies that $\{\mathbf{x}^{(k)}\}$ converges to \mathbf{x} with respect to the l_{∞} norm.

Example 3 Show that

$$\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t = \left(1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k} \sin k\right)^t.$$

converges to $\mathbf{x} = (1, 2, 0, 0)^t$ with respect to the l_{∞} norm.

Solution Because

$$\lim_{k \to \infty} 1 = 1$$
, $\lim_{k \to \infty} (2 + 1/k) = 2$, $\lim_{k \to \infty} 3/k^2 = 0$ and $\lim_{k \to \infty} e^{-k} \sin k = 0$,

Theorem 7.6 implies that the sequence $\{\mathbf{x}^{(k)}\}$ converges to $(1, 2, 0, 0)^t$ with respect to the l_{∞} norm.

To show directly that the sequence in Example 3 converges to $(1, 2, 0, 0)^t$ with respect to the l_2 norm is quite complicated. It is better to prove the next result and apply it to this special case.

Theorem 7.7 For each $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_{2} \leq \sqrt{n} \|\mathbf{x}\|_{\infty}.$$

Proof Let x_j be a coordinate of \mathbf{x} such that $\|\mathbf{x}\|_{\infty} = \max_{1 \le i \le n} |x_i| = |x_j|$. Then

$$\|\mathbf{x}\|_{\infty}^2 = |x_j|^2 = x_j^2 \le \sum_{i=1}^n x_i^2 = \|\mathbf{x}\|_2^2,$$

and

$$\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_{2}.$$

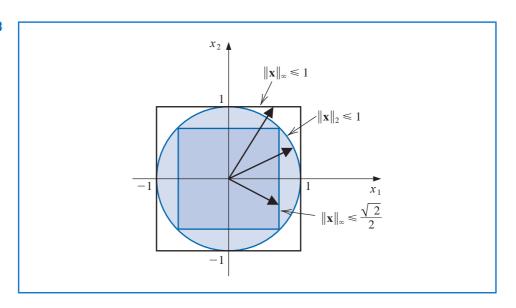
So

$$\|\mathbf{x}\|_{2}^{2} = \sum_{i=1}^{n} x_{i}^{2} \le \sum_{i=1}^{n} x_{j}^{2} = nx_{j}^{2} = n||\mathbf{x}||_{\infty}^{2},$$

and $\|\mathbf{x}\|_2 \leq \sqrt{n} \|\mathbf{x}\|_{\infty}$.

Figure 7.3 illustrates this result when n = 2.

Figure 7.3



Example 4 In Example 3, we found that the sequence $\{\mathbf{x}^{(k)}\}\$, defined by

$$\mathbf{x}^{(k)} = \left(1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k} \sin k\right)^t,$$

converges to $\mathbf{x} = (1, 2, 0, 0)^t$ with respect to the l_{∞} norm. Show that this sequence also converges to \mathbf{x} with respect to the l_2 norm.

Solution Given any $\varepsilon > 0$, there exists an integer $N(\varepsilon/2)$ with the property that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_{\infty} < \frac{\varepsilon}{2},$$

whenever $k \ge N(\varepsilon/2)$. By Theorem 7.7, this implies that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_2 \le \sqrt{4} \|\mathbf{x}^{(k)} - \mathbf{x}\|_{\infty} \le 2(\varepsilon/2) = \varepsilon,$$

when $k \ge N(\varepsilon/2)$. So $\{\mathbf{x}^{(k)}\}$ also converges to \mathbf{x} with respect to the l_2 norm.

It can be shown that all norms on \mathbb{R}^n are equivalent with respect to convergence; that is, if $\|\cdot\|$ and $\|\cdot\|'$ are any two norms on \mathbb{R}^n and $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ has the limit \mathbf{x} with respect to $\|\cdot\|$, then $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ also has the limit \mathbf{x} with respect to $\|\cdot\|'$. The proof of this fact for the general case can be found in [Or2], p. 8. The case for the l_2 and l_∞ norms follows from Theorem 7.7.

Matrix Norms and Distances

In the subsequent sections of this and later chapters, we will need methods for determining the distance between $n \times n$ matrices. This again requires the use of a norm.

- **Definition 7.8** A **matrix norm** on the set of all $n \times n$ matrices is a real-valued function, $\|\cdot\|$, defined on this set, satisfying for all $n \times n$ matrices A and B and all real numbers α :
 - (i) $||A|| \ge 0$;
 - (ii) ||A|| = 0, if and only if A is O, the matrix with all 0 entries;
 - **(iii)** $\|\alpha A\| = |\alpha| \|A\|$;
 - (iv) $||A + B|| \le ||A|| + ||B||$;
 - (v) $||AB|| \le ||A|| ||B||$.

The **distance between n \times n matrices** A and B with respect to this matrix norm is ||A - B||.

Although matrix norms can be obtained in various ways, the norms considered most frequently are those that are natural consequences of the vector norms l_2 and l_{∞} .

These norms are defined using the following theorem, whose proof is considered in Exercise 13.

Theorem 7.9 If $||\cdot||$ is a vector norm on \mathbb{R}^n , then

$$||A|| = \max_{\|\mathbf{x}\|=1} ||A\mathbf{x}|| \tag{7.2}$$

is a matrix norm.

Every vector norm produces an associated natural matrix norm.

Matrix norms defined by vector norms are called the **natural**, or *induced*, **matrix norm** associated with the vector norm. In this text, all matrix norms will be assumed to be natural matrix norms unless specified otherwise.

For any $\mathbf{z} \neq \mathbf{0}$, the vector $\mathbf{x} = \mathbf{z}/\|\mathbf{z}\|$ is a unit vector. Hence

$$\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \max_{\mathbf{z} \neq \mathbf{0}} \left\| A\left(\frac{\mathbf{z}}{\|\mathbf{z}\|}\right) \right\| = \max_{\mathbf{z} \neq \mathbf{0}} \frac{\|A\mathbf{z}\|}{\|\mathbf{z}\|},$$

and we can alternatively write

$$||A|| = \max_{\mathbf{z} \neq \mathbf{0}} \frac{||A\mathbf{z}||}{||\mathbf{z}||}.$$
 (7.3)

The following corollary to Theorem 7.9 follows from this representation of ||A||.

Corollary 7.10 For any vector $\mathbf{z} \neq \mathbf{0}$, matrix A, and any natural norm $\|\cdot\|$, we have

$$||A\mathbf{z}|| \le ||A|| \cdot ||\mathbf{z}||.$$

The measure given to a matrix under a natural norm describes how the matrix stretches unit vectors relative to that norm. The maximum stretch is the norm of the matrix. The matrix norms we will consider have the forms

$$\|A\|_{\infty} = \max_{\|\mathbf{x}\|_{\infty}=1} \|A\mathbf{x}\|_{\infty}, \quad \text{the } l_{\infty} \text{ norm,}$$

and

$$||A||_2 = \max_{\|\mathbf{x}\|_2=1} ||A\mathbf{x}||_2$$
, the l_2 norm.

An illustration of these norms when n=2 is shown in Figures 7.4 and 7.5 for the matrix

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

Figure 7.4

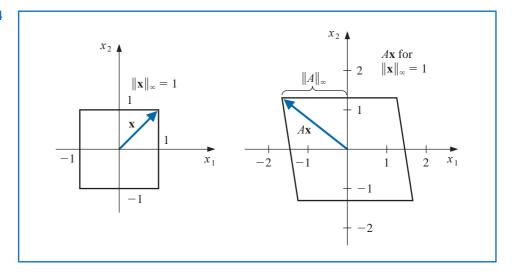
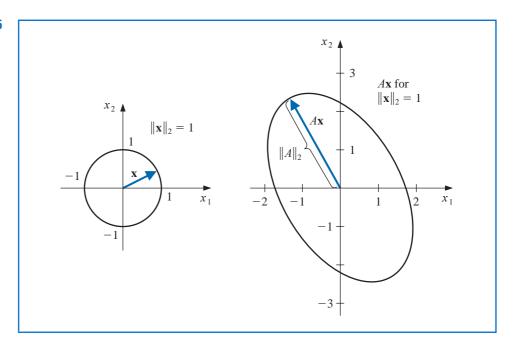


Figure 7.5



The l_{∞} norm of a matrix can be easily computed from the entries of the matrix.

Theorem 7.11 If $A = (a_{ij})$ is an $n \times n$ matrix, then

$$||A||_{\infty} = \max_{1 \le i \le n} \sum_{i=1}^{n} |a_{ij}|.$$

Proof First we show that $||A||_{\infty} \le \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|$.

Let **x** be an *n*-dimensional vector with $1 = \|\mathbf{x}\|_{\infty} = \max_{1 \le i \le n} |x_i|$. Since $A\mathbf{x}$ is also an *n*-dimensional vector,

$$||A\mathbf{x}||_{\infty} = \max_{1 \le i \le n} |(A\mathbf{x})_i| = \max_{1 \le i \le n} \left| \sum_{j=1}^n a_{ij} x_j \right| \le \max_{1 \le i \le n} \sum_{j=1}^n |a_{ij}| \max_{1 \le j \le n} |x_j|.$$

But $\max_{1 \le j \le n} |x_j| = ||\mathbf{x}||_{\infty} = 1$, so

$$||A\mathbf{x}||_{\infty} \le \max_{1 \le i \le n} \sum_{i=1}^{n} |a_{ij}|,$$

and consequently,

$$||A||_{\infty} = \max_{\|\mathbf{x}\|_{\infty} = 1} ||A\mathbf{x}||_{\infty} \le \max_{1 \le i \le n} \sum_{i=1}^{n} |a_{ij}|.$$
 (7.4)

Now we will show the opposite inequality. Let *p* be an integer with

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|,$$

and \mathbf{x} be the vector with components

$$x_j = \begin{cases} 1, & \text{if } a_{pj} \ge 0, \\ -1, & \text{if } a_{pj} < 0. \end{cases}$$

Then $\|\mathbf{x}\|_{\infty} = 1$ and $a_{pj}x_j = |a_{pj}|$, for all $j = 1, 2, \dots, n$, so

$$||A\mathbf{x}||_{\infty} = \max_{1 \le i \le n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \ge \left| \sum_{j=1}^{n} a_{pj} x_j \right| = \left| \sum_{j=1}^{n} |a_{pj}| \right| = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|.$$

This result implies that

$$||A||_{\infty} = \max_{\|\mathbf{x}\|_{\infty}=1} ||A\mathbf{x}||_{\infty} \ge \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|.$$

Putting this together with Inequality (7.4) gives $||A||_{\infty} = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|$.

Example 5 Determine $||A||_{\infty}$ for the matrix

$$A = \left[\begin{array}{rrr} 1 & 2 & -1 \\ 0 & 3 & -1 \\ 5 & -1 & 1 \end{array} \right].$$

Solution We have

$$\sum_{i=1}^{3} |a_{1i}| = |1| + |2| + |-1| = 4, \quad \sum_{i=1}^{3} |a_{2i}| = |0| + |3| + |-1| = 4,$$

and

$$\sum_{j=1}^{3} |a_{3j}| = |5| + |-1| + |1| = 7.$$

So Theorem 7.11 implies that $||A||_{\infty} = \max\{4, 4, 7\} = 7$.

In the next section, we will discover an alternative method for finding the l_2 norm of a matrix.

EXERCISE SET 7.1

1. Find l_{∞} and l_2 norms of the vectors.

a.
$$\mathbf{x} = (3, -4, 0, \frac{3}{2})^t$$

b.
$$\mathbf{x} = (2, 1, -3, 4)^t$$

c. $\mathbf{x} = (\sin k, \cos k, 2^k)^t$ for a fixed positive integer k

d. $\mathbf{x} = (4/(k+1), 2/k^2, k^2e^{-k})^t$ for a fixed positive integer k

2. a. Verify that the function $\|\cdot\|_1$, defined on \mathbb{R}^n by

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|,$$

is a norm on \mathbb{R}^n .

b. Find $\|\mathbf{x}\|_1$ for the vectors given in Exercise 1.

c. Prove that for all $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_1 \ge \|\mathbf{x}\|_2$.

3. Prove that the following sequences are convergent, and find their limits.

a.
$$\mathbf{x}^{(k)} = (1/k, e^{1-k}, -2/k^2)^t$$

b.
$$\mathbf{x}^{(k)} = (e^{-k}\cos k, k\sin(1/k), 3 + k^{-2})^t$$

c.
$$\mathbf{x}^{(k)} = (ke^{-k^2}, (\cos k)/k, \sqrt{k^2 + k} - k)^t$$

d.
$$\mathbf{x}^{(k)} = (e^{1/k}, (k^2 + 1)/(1 - k^2), (1/k^2)(1 + 3 + 5 + \dots + (2k - 1)))^t$$

4. Find the l_{∞} norm of the matrices.

a.
$$\begin{bmatrix} 10 & 15 \\ 0 & 1 \end{bmatrix}$$
c.
$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

b.
$$\begin{bmatrix} 10 & 0 \\ 15 & 1 \end{bmatrix}$$
d.
$$\begin{bmatrix} 4 & -1 & 7 \\ -1 & 4 & 0 \\ -7 & 0 & 4 \end{bmatrix}$$

- 5. The following linear systems $A\mathbf{x} = \mathbf{b}$ have \mathbf{x} as the actual solution and $\tilde{\mathbf{x}}$ as an approximate solution. Compute $\|\mathbf{x} \tilde{\mathbf{x}}\|_{\infty}$ and $\|A\tilde{\mathbf{x}} \mathbf{b}\|_{\infty}$.
 - **a.** $\frac{1}{2}x_1 + \frac{1}{3}x_2 = \frac{1}{63}$, $\frac{1}{3}x_1 + \frac{1}{4}x_2 = \frac{1}{168}$, $\mathbf{x} = (\frac{1}{7}, -\frac{1}{6})^t$, $\mathbf{x} = (0.142, -0.166)^t$. **b.** $x_1 + 2x_2 + 3x_3 = 1$, $2x_1 + 3x_2 + 4x_3 = -1$, $3x_1 + 4x_2 + 6x_3 = 2$, $\mathbf{x} = (0, -7, 5)^t$, $\tilde{\mathbf{x}} = (-0.33, -7.9, 5.8)^t$.
 - **c.** $x_1 + 2x_2 + 3x_3 = 1$, $2x_1 + 3x_2 + 4x_3 = -1$, $3x_1 + 4x_2 + 6x_3 = 2$, $3x_1 + 4x_2 + 6x_3 = 2$, $2x_1 + 3x_2 + 4x_3 = 11$, $2x_1 + 3x_2 + 4x_3 = 11$, $2x_2 + 3x_3 = 11$, $2x_3 = 11$, $2x_4 = 11$, $2x_4 = 11$, $2x_5 = 11$, 2x
- **6.** The matrix norm $\|\cdot\|_1$, defined by $\|A\|_1 = \max_{\|\mathbf{x}\|_1=1} \|A\mathbf{x}\|_1$, can be computed using the formula

$$||A||_1 = \max_{1 \le j \le n} \sum_{i=1}^n |a_{ij}|,$$

where the vector norm $\|\cdot\|_1$ is defined in Exercise 2. Find $\|\cdot\|_1$ for the matrices in Exercise 4.

- 7. Show by example that $\|\cdot\|_{\otimes}$, defined by $\|A\|_{\otimes} = \max_{1 \le i, j \le n} |a_{ij}|$, does not define a matrix norm.
- **8.** Show that $\|\cdot\|_{\widehat{\mathbb{Q}}}$, defined by

$$||A||_{\mathbb{O}} = \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|,$$

is a matrix norm. Find $\|\cdot\|_{\mathbb{D}}$ for the matrices in Exercise 4.

9. a. The Frobenius norm (which is not a natural norm) is defined for an $n \times n$ matrix A by

$$||A||_F = \left(\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2\right)^{1/2}.$$

Show that $\|\cdot\|_F$ is a matrix norm.

- **b.** Find $\|\cdot\|_F$ for the matrices in Exercise 4.
- **c.** For any matrix A, show that $||A||_2 \le ||A||_F \le n^{1/2} ||A||_2$.
- 10. In Exercise 9 the Frobenius norm of a matrix was defined. Show that for any $n \times n$ matrix A and vector \mathbf{x} in \mathbb{R}^n , $\|A\mathbf{x}\|_2 \le \|A\|_F \|\mathbf{x}\|_2$.
- 11. Let S be a positive definite $n \times n$ matrix. For any \mathbf{x} in \mathbb{R}^n define $\|\mathbf{x}\| = (\mathbf{x}^t S \mathbf{x})^{1/2}$. Show that this defines a norm on \mathbb{R}^n . [Hint: Use the Cholesky factorization of S to show that $\mathbf{x}^t S \mathbf{y} = \mathbf{y}^t S \mathbf{x} \le (\mathbf{x}^t S \mathbf{x})^{1/2} (\mathbf{y}^t S \mathbf{y})^{1/2}$.]
- 12. Let *S* be a real and nonsingular matrix, and let $\|\cdot\|$ be any norm on \mathbb{R}^n . Define $\|\cdot\|'$ by $\|\mathbf{x}\|' = \|S\mathbf{x}\|$. Show that $\|\cdot\|'$ is also a norm on \mathbb{R}^n .
- **13.** Prove that if $\|\cdot\|$ is a vector norm on \mathbb{R}^n , then $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ is a matrix norm.
- **14.** The following excerpt from the *Mathematics Magazine* [Sz] gives an alternative way to prove the Cauchy-Buniakowsky-Schwarz Inequality.
 - a. Show that when $x \neq 0$ and $y \neq 0$, we have

$$\frac{\sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n} x_i^2\right)^{1/2} \left(\sum_{i=1}^{n} y_i^2\right)^{1/2}} = 1 - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i}{\left(\sum_{j=1}^{n} x_j^2\right)^{1/2}} - \frac{y_i}{\left(\sum_{j=1}^{n} y_j^2\right)^{1/2}} \right)^2.$$

b. Use the result in part (a) to show that

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

15. Show that the Cauchy-Buniakowsky-Schwarz Inequality can be strengthened to

$$\sum_{i=1}^{n} x_i y_i \le \sum_{i=1}^{n} |x_i y_i| \le \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} \left(\sum_{i=1}^{n} y_i^2\right)^{1/2}.$$

7.2 Eigenvalues and Eigenvectors

An $n \times m$ matrix can be considered as a function that uses matrix multiplication to take m-dimensional column vectors into n-dimensional column vectors. So an $n \times m$ matrix is actually a linear function from \mathbb{R}^m to \mathbb{R}^n . A square matrix A takes the set of n-dimensional vectors into itself, which gives a linear function from \mathbb{R}^n to \mathbb{R}^n . In this case, certain nonzero vectors \mathbf{x} might be parallel to $A\mathbf{x}$, which means that a constant λ exists with $A\mathbf{x} = \lambda \mathbf{x}$. For these vectors, we have $(A - \lambda I)\mathbf{x} = \mathbf{0}$. There is a close connection between these numbers λ and the likelihood that an iterative method will converge. We will consider this connection in this section.

Definition 7.12 If A is a square matrix, the **characteristic polynomial** of A is defined by

$$p(\lambda) = \det(A - \lambda I).$$

It is not difficult to show (see Exercise 13) that p is an nth-degree polynomial and, consequently, has at most n distinct zeros, some of which might be complex. If λ is a zero of p, then, since $\det(A - \lambda I) = 0$, Theorem 6.17 on page 398 implies that the linear system defined by $(A - \lambda I)\mathbf{x} = \mathbf{0}$ has a solution with $\mathbf{x} \neq \mathbf{0}$. We wish to study the zeros of p and the nonzero solutions corresponding to these systems.

Definition 7.13

If p is the characteristic polynomial of the matrix A, the zeros of p are **eigenvalues**, or characteristic values, of the matrix A. If λ is an eigenvalue of A and $\mathbf{x} \neq \mathbf{0}$ satisfies $(A - \lambda I)\mathbf{x} = \mathbf{0}$, then \mathbf{x} is an **eigenvector**, or characteristic vector, of A corresponding to the eigenvalue λ .

The prefix eigen comes from the German adjective meaning "to own", and is synonymous in English with the word characteristic. Each matrix has its own eigen- or characteristic equation, with corresponding eigen- or characteristic values and functions.

To determine the eigenvalues of a matrix, we can use the fact that

• λ is an eigenvalue of A if and only if $\det(A - \lambda I) = 0$.

Once an eigenvalue λ has been found a corresponding eigenvector $\mathbf{x} \neq \mathbf{0}$ is determined by solving the system

 $\bullet (A - \lambda I)\mathbf{x} = \mathbf{0}.$

Example 1 Show that there are no nonzero vectors \mathbf{x} in \mathbb{R}^2 with $A\mathbf{x}$ parallel to \mathbf{x} if

$$A = \left[\begin{array}{cc} 0 & 1 \\ -1 & 0 \end{array} \right].$$

Solution The eigenvalues of A are the solutions to the characteristic polynomial

$$0 = \det(A - \lambda I) = \det \begin{bmatrix} -\lambda & 1 \\ -1 & -\lambda \end{bmatrix} = \lambda^2 + 1,$$

so the eigenvalues of A are the complex numbers $\lambda_1 = i$ and $\lambda_2 = -i$. A corresponding eigenvector **x** for λ_1 needs to satisfy

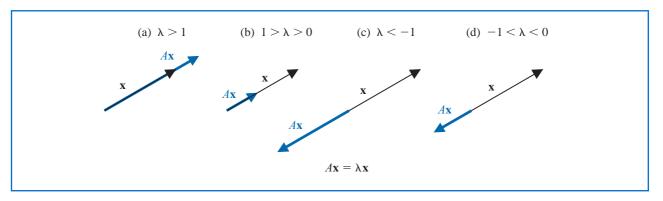
$$\left[\begin{array}{c} 0 \\ 0 \end{array}\right] = \left[\begin{array}{cc} -i & 1 \\ -1 & -i \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \left[\begin{array}{c} -ix_1 + x_2 \\ -x_1 - ix_2 \end{array}\right],$$

that is, $0 = -ix_1 + x_2$, so $x_2 = ix_1$, and $0 = -x_1 - ix_2$. Hence if **x** is an eigenvector of *A*, then exactly one of its components is real and the other is complex. As a consequence, there are no nonzero vectors **x** in \mathbb{R}^2 with A**x** parallel to **x**.

If **x** is an eigenvector associated with the real eigenvalue λ , then A**x** = λ **x**, so the matrix A takes the vector **x** into a scalar multiple of itself.

- If λ is real and $\lambda > 1$, then *A* has the effect of stretching **x** by a factor of λ , as illustrated in Figure 7.6(a).
- If $0 < \lambda < 1$, then A shrinks **x** by a factor of λ (see Figure 7.6(b)).
- If λ < 0, the effects are similar (see Figure 7.6(c) and (d)), although the direction of Ax is reversed.

Figure 7.6



Notice also that if \mathbf{x} is an eigenvector of A associated with the eigenvalue λ and α is any nonzero constant, then $\alpha \mathbf{x}$ is also an eigenvector since

$$A(\alpha \mathbf{x}) = \alpha(A\mathbf{x}) = \alpha(\lambda \mathbf{x}) = \lambda(\alpha \mathbf{x}).$$

An important consequence of this is that for any vector norm $||\cdot||$ we could choose the constant $\alpha = \pm ||\mathbf{x}||^{-1}$, which would result in $\alpha \mathbf{x}$ being an eigenvector with norm 1. So

• For every eigenvalue and any vector norm there are eigenvectors with norm 1.

Example 2 Determine the eigenvalues and eigenvectors for the matrix

$$A = \left[\begin{array}{rrr} 2 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & -1 & 4 \end{array} \right].$$

Solution The characteristic polynomial of A is

$$p(\lambda) = \det(A - \lambda I) = \det \begin{bmatrix} 2 - \lambda & 0 & 0 \\ 1 & 1 - \lambda & 2 \\ 1 & -1 & 4 - \lambda \end{bmatrix}$$
$$= -(\lambda^3 - 7\lambda^2 + 16\lambda - 12) = -(\lambda - 3)(\lambda - 2)^2,$$

so there are two eigenvalues of A: $\lambda_1 = 3$ and $\lambda_2 = 2$.

An eigenvector \mathbf{x}_1 corresponding to the eigenvalue $\lambda_1 = 3$ is a solution to the vector-matrix equation $(A - 3 \cdot I)\mathbf{x}_1 = \mathbf{0}$, so

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & -2 & 2 \\ 1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

which implies that $x_1 = 0$ and $x_2 = x_3$.

Any nonzero value of x_3 produces an eigenvector for the eigenvalue $\lambda_1 = 3$. For example, when $x_3 = 1$ we have the eigenvector $\mathbf{x}_1 = (0, 1, 1)^t$, and any eigenvector of A corresponding to $\lambda = 3$ is a nonzero multiple of \mathbf{x}_1 .

An eigenvector $\mathbf{x} \neq \mathbf{0}$ of A associated with $\lambda_2 = 2$ is a solution of the system $(A - 2 \cdot I)\mathbf{x} = \mathbf{0}$, so

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 2 \\ 1 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

In this case the eigenvector has only to satisfy the equation

$$x_1 - x_2 + 2x_3 = 0$$
,

which can be done in various ways. For example, when $x_1 = 0$ we have $x_2 = 2x_3$, so one choice would be $\mathbf{x}_2 = (0, 2, 1)^t$. We could also choose $x_2 = 0$, which requires that $x_1 = -2x_3$. Hence $\mathbf{x}_3 = (-2, 0, 1)^t$ gives a second eigenvector for the eigenvalue $\lambda_2 = 2$ that is not a multiple of \mathbf{x}_2 . The eigenvectors of A corresponding to the eigenvalue $\lambda_2 = 2$ generate an entire plane. This plane is described by all vectors of the form

$$\alpha \mathbf{x}_2 + \beta \mathbf{x}_3 = (-2\beta, 2\alpha, \alpha + \beta)^t$$

for arbitrary constants α and β , provided that at least one of the constants is nonzero.

The package *LinearAlgebra* in Maple provides the function *Eigenvalues* to compute eigenvalues. The function *Eigenvectors* gives both the eigenvalues and the corresponding eigenvectors of a matrix. To produce results for the matrix in Example 2, we first load the package with

with(LinearAlgebra)

Then we enter the matrix

$$A := ([[2,0,0],[1,1,2],[1,-1,4]])$$

giving

$$\left[\begin{array}{ccc}
2 & 0 & 0 \\
1 & 1 & 2 \\
1 & -1 & 4
\end{array}\right]$$

To determine the eigenvalues and eigenvectors we use

evalf(Eigenvectors(A))

which returns

$$\begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 & -2 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

implying that the eigenvalues are 3, 2, and 2 with corresponding eigenvectors given by the respective columns as $(0, 1, 1)^t$, $(-2, 0, 1)^t$, and $(1, 1, 0)^t$.

The *LinearAlgebra* package also contains the command *CharacteristicPolynomial*, so the eigenvalues could also be obtained with

 $p := CharacteristicPolynomial(A, \lambda); factor(p)$

This gives

$$-12 + \lambda^3 - 7\lambda^2 + 16\lambda$$
$$(\lambda - 3)(\lambda - 2)^2$$

The notions of eigenvalues and eigenvectors are introduced here for a specific computational convenience, but these concepts arise frequently in the study of physical systems. In fact, they are of sufficient interest that Chapter 9 is devoted to their numerical approximation.

Spectral Radius

Definition 7.14 The spectral radius $\rho(A)$ of a matrix A is defined by

 $\rho(A) = \max |\lambda|$, where λ is an eigenvalue of A.

(For complex
$$\lambda = \alpha + \beta i$$
, we define $|\lambda| = (\alpha^2 + \beta^2)^{1/2}$.)

For the matrix considered in Example 2, $\rho(A) = \max\{2, 3\} = 3$.

The spectral radius is closely related to the norm of a matrix, as shown in the following theorem.

Theorem 7.15 If A is an $n \times n$ matrix, then

(i)
$$||A||_2 = [\rho(A^t A)]^{1/2}$$
,

(ii)
$$\rho(A) \leq ||A||$$
, for any natural norm $||\cdot||$.

Proof The proof of part (i) requires more information concerning eigenvalues than we presently have available. For the details involved in the proof, see [Or2], p. 21.

To prove part (ii), suppose λ is an eigenvalue of A with eigenvector \mathbf{x} and $\|\mathbf{x}\| = 1$. Then $A\mathbf{x} = \lambda \mathbf{x}$ and

$$|\lambda| = |\lambda| \cdot ||\mathbf{x}|| = ||\lambda\mathbf{x}|| = ||A\mathbf{x}|| \le ||A|| ||\mathbf{x}|| = ||A||.$$

Thus

$$\rho(A) = \max |\lambda| \le ||A||.$$

Part (i) of Theorem 7.15 implies that if A is symmetric, then $||A||_2 = \rho(A)$ (see Exercise 14).

An interesting and useful result, which is similar to part (ii) of Theorem 7.15, is that for any matrix A and any $\varepsilon > 0$, there exists a natural norm $\|\cdot\|$ with the property that $\rho(A) < \|A\| < \rho(A) + \varepsilon$. Consequently, $\rho(A)$ is the greatest lower bound for the natural norms on A. The proof of this result can be found in [Or2], p. 23.

Example 3 Determine the l_2 norm of

$$A = \left[\begin{array}{rrr} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{array} \right].$$

Solution To apply Theorem 7.15 we need to calculate $\rho(A^tA)$, so we first need the eigenvalues of A^tA .

$$A^{t}A = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{bmatrix}.$$

If

$$0 = \det(A^{t}A - \lambda I) = \det\begin{bmatrix} 3 - \lambda & 2 & -1 \\ 2 & 6 - \lambda & 4 \\ -1 & 4 & 5 - \lambda \end{bmatrix}$$
$$= -\lambda^{3} + 14\lambda^{2} - 42\lambda = -\lambda(\lambda^{2} - 14\lambda + 42),$$

then $\lambda = 0$ or $\lambda = 7 \pm \sqrt{7}$. By Theorem 7.15 we have

$$||A||_2 = \sqrt{\rho(A^t A)} = \sqrt{\max\{0, 7 - \sqrt{7}, 7 + \sqrt{7}\}} = \sqrt{7 + \sqrt{7}} \approx 3.106.$$

The operations in Example 3 can also be performed using the *LinearAlgebra* package in Maple by first loading the package and then entering the matrix.

$$with(LinearAlgebra): A := Matrix([[1, 1, 0], [1, 2, 1], [-1, 1, 2]])$$

Maple will respond by showing the matrix that was entered. To determine the transpose of A we use

B := Transpose(A)

which gives

$$\left[\begin{array}{ccc}
1 & 1 & -1 \\
1 & 2 & 1 \\
0 & 1 & 2
\end{array}\right]$$

Then we can compute the product AB with

C := A.B

which produces

$$\left[\begin{array}{cccc}
3 & 2 & -1 \\
2 & 6 & 4 \\
-1 & 4 & 5
\end{array}\right]$$

The command

evalf(Eigenvalues(C))

gives the vector

Since $||A||_2 = \sqrt{\rho(A^t A)} = \sqrt{\rho(C)}$, we have

$$||A||_2 = \sqrt{9.645751311} = 3.105760987,$$

which we could also find with evalf(Norm(A, 2)).

To determine the l_{∞} norm of A, replace the last command with evalf(Norm(A, infinity)) which Maple gives as 4. This is seen to be correct because it is the sum of the magnitude of the entries in the second row.

Convergent Matrices

In studying iterative matrix techniques, it is of particular importance to know when powers of a matrix become small (that is, when all the entries approach zero). Matrices of this type are called *convergent*.

Definition 7.16 We call an $n \times n$ matrix A convergent if

$$\lim_{k \to \infty} (A^k)_{ij} = 0$$
, for each $i = 1, 2, ..., n$ and $j = 1, 2, ..., n$.

Example 4 Show that

$$A = \left[\begin{array}{cc} \frac{1}{2} & 0\\ \frac{1}{4} & \frac{1}{2} \end{array} \right]$$

is a convergent matrix.

Solution Computing powers of A, we obtain:

$$A^{2} = \begin{bmatrix} \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \quad A^{3} = \begin{bmatrix} \frac{1}{8} & 0 \\ \frac{3}{16} & \frac{1}{8} \end{bmatrix}, \quad A^{4} = \begin{bmatrix} \frac{1}{16} & 0 \\ \frac{1}{8} & \frac{1}{16} \end{bmatrix},$$

and, in general,

$$A^{k} = \begin{bmatrix} (\frac{1}{2})^{k} & 0\\ \frac{k}{2^{k+1}} & (\frac{1}{2})^{k} \end{bmatrix}.$$

So A is a convergent matrix because

$$\lim_{k \to \infty} \left(\frac{1}{2}\right)^k = 0 \quad \text{ and } \quad \lim_{k \to \infty} \frac{k}{2^{k+1}} = 0.$$

Notice that the convergent matrix A in Example 4 has $\rho(A) = \frac{1}{2}$, because $\frac{1}{2}$ is the only eigenvalue of A. This illustrates an important connection that exists between the spectral radius of a matrix and the convergence of the matrix, as detailed in the following result.

Theorem 7.17 The following statements are equivalent.

- (i) A is a convergent matrix.
- (ii) $\lim_{n\to\infty} ||A^n|| = 0$, for some natural norm.
- (iii) $\lim_{n\to\infty} ||A^n|| = 0$, for all natural norms.
- (iv) $\rho(A) < 1$.
- (v) $\lim_{n\to\infty} A^n \mathbf{x} = \mathbf{0}$, for every \mathbf{x} .

The proof of this theorem can be found in [IK], p. 14.

EXERCISE SET 7.2

1. Compute the eigenvalues and associated eigenvectors of the following matrices.

a.
$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$
 b. $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ **c.** $\begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$ **d.** $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ **e.** $\begin{bmatrix} -1 & 2 & 0 \\ 0 & 3 & 4 \\ 0 & 0 & 7 \end{bmatrix}$ **f.** $\begin{bmatrix} 2 & 1 & 1 \\ 2 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix}$

2. Compute the eigenvalues and associated eigenvectors of the following matrices.

a.
$$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix}$$
b.
$$\begin{bmatrix} -1 & -1 \\ \frac{1}{3} & \frac{1}{6} \end{bmatrix}$$
c.
$$\begin{bmatrix} 3 & 4 \\ 1 & 0 \end{bmatrix}$$
d.
$$\begin{bmatrix} 3 & 2 & -1 \\ 1 & -2 & 3 \\ 2 & 0 & 4 \end{bmatrix}$$
e.
$$\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ -1 & \frac{1}{2} & 0 \\ 2 & 2 & -\frac{1}{3} \end{bmatrix}$$
f.
$$\begin{bmatrix} 2 & -1 & 0 \\ 0 & 2 & 4 \\ 0 & 0 & 2 \end{bmatrix}$$

3. Find the complex eigenvalues and associated eigenvectors for the following matrices.

$$\begin{bmatrix} 2 & 2 \\ -1 & 2 \end{bmatrix}$$
b.
$$\begin{bmatrix} 1 & 2 \\ -1 & 2 \end{bmatrix}$$
d the complex eigenvalues and associated eigenvectors for the

4. Find the complex eigenvalues and associated eigenvectors for the following matrices.

$$\mathbf{b.} \quad \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix} \qquad \qquad \mathbf{b.} \quad \begin{bmatrix} 0 & 1 & -2 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

- 5. Find the spectral radius for each matrix in Exercise 1.
- **6.** Find the spectral radius for each matrix in Exercise 2.
- **7.** Which of the matrices in Exercise 1 are convergent?
- **8.** Which of the matrices in Exercise 2 are convergent?
- **9.** Find the l_2 norm for the matrices in Exercise 1.
- **10.** Find the l_2 norm for the matrices in Exercise 2.

11. Let
$$A_1 = \begin{bmatrix} 1 & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$
 and $A_2 = \begin{bmatrix} \frac{1}{2} & 0 \\ 16 & \frac{1}{2} \end{bmatrix}$. Show that A_1 is not convergent, but A_2 is convergent.

- 12. An $n \times n$ matrix A is called *nilpotent* if an integer m exists with $A^m = O_n$. Show that if λ is an eigenvalue of a nilpotent matrix, then $\lambda = 0$.
- 13. Show that the characteristic polynomial $p(\lambda) = \det(A \lambda I)$ for the $n \times n$ matrix A is an nth-degree polynomial. [*Hint*: Expand $\det(A \lambda I)$ along the first row, and use mathematical induction on n.]
- **14.** a. Show that if A is an $n \times n$ matrix, then

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

where $\lambda_i, \ldots, \lambda_n$ are the eigenvalues of A. [Hint: Consider p(0).]

b. Show that *A* is singular if and only if $\lambda = 0$ is an eigenvalue of *A*.

- **15.** Let λ be an eigenvalue of the $n \times n$ matrix A and $\mathbf{x} \neq \mathbf{0}$ be an associated eigenvector.
 - **a.** Show that λ is also an eigenvalue of A^t .
 - **b.** Show that for any integer $k \ge 1$, λ^k is an eigenvalue of A^k with eigenvector \mathbf{x} .
 - **c.** Show that if A^{-1} exists, then $1/\lambda$ is an eigenvalue of A^{-1} with eigenvector \mathbf{x} .
 - **d.** Generalize parts (b) and (c) to $(A^{-1})^k$ for integers k > 2.
 - **e.** Given the polynomial $q(x) = q_0 + q_1x + \cdots + q_kx^k$, define q(A) to be the matrix $q(A) = q_0I + q_1A + \cdots + q_kA^k$. Show that $q(\lambda)$ is an eigenvalue of q(A) with eigenvector \mathbf{x} .
 - **f.** Let $\alpha \neq \lambda$ be given. Show that if $A \alpha I$ is nonsingular, then $1/(\lambda \alpha)$ is an eigenvalue of $(A \alpha I)^{-1}$ with eigenvector \mathbf{x} .
- **16.** Show that if A is symmetric, then $||A||_2 = \rho(A)$.
- 17. In Exercise 15 of Section 6.3, we assumed that the contribution a female beetle of a certain type made to the future years' beetle population could be expressed in terms of the matrix

$$A = \left[\begin{array}{ccc} 0 & 0 & 6 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{array} \right],$$

where the entry in the *i*th row and *j*th column represents the probabilistic contribution of a beetle of age *j* onto the next year's female population of age *i*.

- a. Does the matrix A have any real eigenvalues? If so, determine them and any associated eigenvectors.
- **b.** If a sample of this species was needed for laboratory test purposes that would have a constant proportion in each age group from year to year, what criteria could be imposed on the initial population to ensure that this requirement would be satisfied?
- **18.** Find matrices *A* and *B* for which $\rho(A+B) > \rho(A) + \rho(B)$. (This shows that $\rho(A)$ cannot be a matrix norm.)
- 19. Show that if $||\cdot||$ is any natural norm, then $(||A^{-1}||)^{-1} \le |\lambda| \le ||A||$ for any eigenvalue λ of the nonsingular matrix A.

7.3 The Jacobi and Gauss-Siedel Iterative Techniques

In this section we describe the Jacobi and the Gauss-Seidel iterative methods, classic methods that date to the late eighteenth century. Iterative techniques are seldom used for solving linear systems of small dimension since the time required for sufficient accuracy exceeds that required for direct techniques such as Gaussian elimination. For large systems with a high percentage of 0 entries, however, these techniques are efficient in terms of both computer storage and computation. Systems of this type arise frequently in circuit analysis and in the numerical solution of boundary-value problems and partial-differential equations.

An iterative technique to solve the $n \times n$ linear system $A\mathbf{x} = \mathbf{b}$ starts with an initial approximation $\mathbf{x}^{(0)}$ to the solution \mathbf{x} and generates a sequence of vectors $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ that converges to \mathbf{x} .

Jacobi's Method

The **Jacobi iterative method** is obtained by solving the *i*th equation in $A\mathbf{x} = \mathbf{b}$ for x_i to obtain (provided $a_{ii} \neq 0$)

$$x_i = \sum_{\substack{j=1 \ j \neq i}}^n \left(-\frac{a_{ij}x_j}{a_{ii}} \right) + \frac{b_i}{a_{ii}}, \quad \text{for } i = 1, 2, \dots, n.$$

For each $k \ge 1$, generate the components $x_i^{(k)}$ of $\mathbf{x}^{(k)}$ from the components of $\mathbf{x}^{(k-1)}$ by

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[\sum_{\substack{j=1\\j\neq i}}^n \left(-a_{ij} x_j^{(k-1)} \right) + b_i \right], \quad \text{for } i = 1, 2, \dots, n.$$
 (7.5)

Example 1 The linear system $A\mathbf{x} = \mathbf{b}$ given by

Carl Gustav Jacob Jacobi (1804–1851) was initially recognized for his work in the area of number theory and elliptic functions, but his mathematical interests and abilities were very broad. He had a strong personality that was influential in establishing a research-oriented attitude that became the nucleus of a revival of mathematics at German universities in the 19th century.

$$E_1: 10x_1 - x_2 + 2x_3 = 6,$$

 $E_2: -x_1 + 11x_2 - x_3 + 3x_4 = 25,$
 $E_3: 2x_1 - x_2 + 10x_3 - x_4 = -11,$
 $E_4: 3x_2 - x_3 + 8x_4 = 15$

has the unique solution $\mathbf{x} = (1, 2, -1, 1)^t$. Use Jacobi's iterative technique to find approximations $\mathbf{x}^{(k)}$ to \mathbf{x} starting with $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$ until

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_{\infty}}{\|\mathbf{x}^{(k)}\|_{\infty}} < 10^{-3}.$$

Solution We first solve equation E_i for x_i , for each i = 1, 2, 3, 4, to obtain

$$x_{1} = \frac{1}{10}x_{2} - \frac{1}{5}x_{3} + \frac{3}{5},$$

$$x_{2} = \frac{1}{11}x_{1} + \frac{1}{11}x_{3} - \frac{3}{11}x_{4} + \frac{25}{11},$$

$$x_{3} = -\frac{1}{5}x_{1} + \frac{1}{10}x_{2} + \frac{1}{10}x_{4} - \frac{11}{10},$$

$$x_{4} = -\frac{3}{8}x_{2} + \frac{1}{8}x_{3} + \frac{15}{8},$$

From the initial approximation $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$ we have $\mathbf{x}^{(1)}$ given by

$$x_{1}^{(1)} = \frac{1}{10}x_{2}^{(0)} - \frac{1}{5}x_{3}^{(0)} + \frac{3}{5} = 0.6000,$$

$$x_{2}^{(1)} = \frac{1}{11}x_{1}^{(0)} + \frac{1}{11}x_{3}^{(0)} - \frac{3}{11}x_{4}^{(0)} + \frac{25}{11} = 2.2727,$$

$$x_{3}^{(1)} = -\frac{1}{5}x_{1}^{(0)} + \frac{1}{10}x_{2}^{(0)} + \frac{1}{10}x_{4}^{(0)} - \frac{11}{10} = -1.1000,$$

$$x_{4}^{(1)} = -\frac{3}{8}x_{2}^{(0)} + \frac{1}{8}x_{3}^{(0)} + \frac{15}{8} = 1.8750.$$

Additional iterates, $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t$, are generated in a similar manner and are presented in Table 7.1.

Table 7.1

| k | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------------|--------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $x_1^{(k)}$ | 0.0000 | 0.6000 | 1.0473 | 0.9326 | 1.0152 | 0.9890 | 1.0032 | 0.9981 | 1.0006 | 0.9997 | 1.0001 |
| $x_{2}^{(k)}$ | 0.0000 | 2.2727 | 1.7159 | 2.053 | 1.9537 | 2.0114 | 1.9922 | 2.0023 | 1.9987 | 2.0004 | 1.9998 |
| $x_{3}^{(k)}$ | 0.0000 | -1.1000 | -0.8052 | -1.0493 | -0.9681 | -1.0103 | -0.9945 | -1.0020 | -0.9990 | -1.0004 | -0.9998 |
| $x_{4}^{(k)}$ | 0.0000 | 1.8750 | 0.8852 | 1.1309 | 0.9739 | 1.0214 | 0.9944 | 1.0036 | 0.9989 | 1.0006 | 0.9998 |

We stopped after ten iterations because

$$\frac{\|\mathbf{x}^{(10)} - \mathbf{x}^{(9)}\|_{\infty}}{\|\mathbf{x}^{(10)}\|_{\infty}} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3}.$$

In fact, $\|\mathbf{x}^{(10)} - \mathbf{x}\|_{\infty} = 0.0002$.

In general, iterative techniques for solving linear systems involve a process that converts the system $A\mathbf{x} = \mathbf{b}$ into an equivalent system of the form $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ for some fixed matrix T and vector \mathbf{c} . After the initial vector $\mathbf{x}^{(0)}$ is selected, the sequence of approximate solution vectors is generated by computing

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}.$$

for each $k = 1, 2, 3, \dots$ This should be reminiscent of the fixed-point iteration studied in Chapter 2.

The Jacobi method can be written in the form $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$ by splitting A into its diagonal and off-diagonal parts. To see this, let D be the diagonal matrix whose diagonal entries are those of A, -L be the strictly lower-triangular part of A, and -U be the strictly upper-triangular part of A. With this notation,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

is split into

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ -a_{21} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

$$= D - L - U.$$

The equation $A\mathbf{x} = \mathbf{b}$, or $(D - L - U)\mathbf{x} = \mathbf{b}$, is then transformed into

$$D\mathbf{x} = (L + U)\mathbf{x} + \mathbf{b},$$

and, if D^{-1} exists, that is, if $a_{ii} \neq 0$ for each i, then

$$\mathbf{x} = D^{-1}(L+U)\mathbf{x} + D^{-1}\mathbf{b}.$$

This results in the matrix form of the Jacobi iterative technique:

$$\mathbf{x}^{(k)} = D^{-1}(L+U)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}, \quad k = 1, 2, \dots$$
 (7.6)

Introducing the notation $T_j = D^{-1}(L+U)$ and $\mathbf{c}_j = D^{-1}\mathbf{b}$ gives the Jacobi technique the form

$$\mathbf{x}^{(k)} = T_i \mathbf{x}^{(k-1)} + \mathbf{c}_i. \tag{7.7}$$

In practice, Eq. (7.5) is used in computation and Eq. (7.7) for theoretical purposes.

Example 2 Express the Jacobi iteration method for the linear system $A\mathbf{x} = \mathbf{b}$ given by

$$E_1: 10x_1 - x_2 + 2x_3 = 6,$$

 $E_2: -x_1 + 11x_2 - x_3 + 3x_4 = 25,$
 $E_3: 2x_1 - x_2 + 10x_3 - x_4 = -11,$
 $E_4: 3x_2 - x_3 + 8x_4 = 15$

in the form $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$.

Solution We saw in Example 1 that the Jacobi method for this system has the form

$$x_{1} = \frac{1}{10}x_{2} - \frac{1}{5}x_{3} + \frac{3}{5},$$

$$x_{2} = \frac{1}{11}x_{1} + \frac{1}{11}x_{3} - \frac{3}{11}x_{4} + \frac{25}{11},$$

$$x_{3} = -\frac{1}{5}x_{1} + \frac{1}{10}x_{2} + \frac{1}{10}x_{4} - \frac{11}{10},$$

$$x_{4} = -\frac{3}{8}x_{2} + \frac{1}{8}x_{3} + \frac{15}{8}.$$

Hence we have

$$T = \begin{bmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0\\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11}\\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10}\\ 0 & -\frac{3}{5} & \frac{1}{2} & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} \frac{3}{5}\\ \frac{25}{11}\\ -\frac{11}{10}\\ \frac{15}{25} \end{bmatrix}.$$

Algorithm 7.1 implements the Jacobi iterative technique.



Jacobi Iterative

To solve $A\mathbf{x} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$:

INPUT the number of equations and unknowns n; the entries a_{ij} , $1 \le i, j \le n$ of the matrix A; the entries b_i , $1 \le i \le n$ of \mathbf{b} ; the entries XO_i , $1 \le i \le n$ of $\mathbf{XO} = \mathbf{x}^{(0)}$; tolerance TOL; maximum number of iterations N.

OUTPUT the approximate solution x_1, \ldots, x_n or a message that the number of iterations was exceeded.

Step 1 Set k = 1.

Step 2 While $(k \le N)$ do Steps 3–6.

Step 3 For
$$i = 1, ..., n$$

set
$$x_i = \frac{1}{a_{ii}} \left[-\sum_{\substack{j=1\\i \neq i}}^{n} (a_{ij}XO_j) + b_i \right].$$

Step 4 If $||\mathbf{x} - \mathbf{XO}|| < TOL$ then OUTPUT (x_1, \dots, x_n) ;

(The procedure was successful.) STOP.

Step 5 Set k = k + 1.



Step 6 For
$$i = 1, \ldots, n$$
 set $XO_i = x_i$.

Step 7 OUTPUT ('Maximum number of iterations exceeded'); (The procedure was successful.) STOP.

Step 3 of the algorithm requires that $a_{ii} \neq 0$, for each i = 1, 2, ..., n. If one of the a_{ii} entries is 0 and the system is nonsingular, a reordering of the equations can be performed so that no $a_{ii} = 0$. To speed convergence, the equations should be arranged so that a_{ii} is as large as possible. This subject is discussed in more detail later in this chapter.

Another possible stopping criterion in Step 4 is to iterate until

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|}$$

is smaller than some prescribed tolerance. For this purpose, any convenient norm can be used, the usual being the l_{∞} norm.

The *NumericalAnalysis* subpackage of the Maple *Student* package implements the Jacobi iterative method. To illustrate this with our example we first enter both *Numerical-Analysis* and *LinearAlgebra*.

with(Student[NumericalAnalysis]): with(LinearAlgebra):

Colons are used at the end of the commands to suppress output for both packages. Enter the matrix with

$$A := Matrix([[10, -1, 2, 0, 6], [-1, 11, -1, 3, 25], [2, -1, 10, -1, -11], [0, 3, -1, 8, 15]])$$

The following command gives a collection of output that is in agreement with the results in Table 7.1.

 $Iterative Approximate (A, initial approx = Vector([0., 0., 0., 0.]), tolerance = 10^{-3}, \\ maxiterations = 20, stopping criterion = relative (infinity), method = jacobi, \\ output = approximates)$

If the option output = approximates is omitted, then only the final approximation result is output. Notice that the initial approximations was specified by [0., 0., 0., 0.], with decimal points placed after the entries. This was done so that Maple will give the results as 10-digit decimals. If the specification had simply been [0, 0, 0, 0], the output would have been given in fractional form.

The Gauss-Seidel Method

A possible improvement in Algorithm 7.1 can be seen by reconsidering Eq. (7.5). The components of $\mathbf{x}^{(k-1)}$ are used to compute all the components $x_i^{(k)}$ of $\mathbf{x}^{(k)}$. But, for i>1, the components $x_1^{(k)},\ldots,x_{i-1}^{(k)}$ of $\mathbf{x}^{(k)}$ have already been computed and are expected to be better approximations to the actual solutions x_1,\ldots,x_{i-1} than are $x_1^{(k-1)},\ldots,x_{i-1}^{(k-1)}$. It seems reasonable, then, to compute $x_i^{(k)}$ using these most recently calculated values. That is, to use

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} (a_{ij} x_j^{(k)}) - \sum_{j=i+1}^{n} (a_{ij} x_j^{(k-1)}) + b_i \right], \tag{7.8}$$

for each i = 1, 2, ..., n, instead of Eq. (7.5). This modification is called the **Gauss-Seidel iterative technique** and is illustrated in the following example.

Phillip Ludwig Seidel (1821-1896) worked as an assistant to Jacobi solving problems on systems of linear equations that resulted from Gauss's work on least squares. These equations generally had off-diagonal elements that were much smaller than those on the diagonal, so the iterative methods were particularly effective. The iterative techniques now known as Jacobi and Gauss-Seidel were both known to Gauss before being applied in this situation, but Gauss's results were not often widely communicated.

Example 3 Use the Gauss-Seidel iterative technique to find approximate solutions to

$$10x_1 - x_2 + 2x_3 = 6,$$

$$-x_1 + 11x_2 - x_3 + 3x_4 = 25,$$

$$2x_1 - x_2 + 10x_3 - x_4 = -11,$$

$$3x_2 - x_3 + 8x_4 = 15$$

starting with $\mathbf{x} = (0, 0, 0, 0)^t$ and iterating until

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_{\infty}}{\|\mathbf{x}^{(k)}\|_{\infty}} < 10^{-3}.$$

Solution The solution $\mathbf{x} = (1, 2, -1, 1)^t$ was approximated by Jacobi's method in Example 1. For the Gauss-Seidel method we write the system, for each k = 1, 2, ... as

$$\begin{split} x_1^{(k)} &= \frac{1}{10} x_2^{(k-1)} - \frac{1}{5} x_3^{(k-1)} + \frac{3}{5}, \\ x_2^{(k)} &= \frac{1}{11} x_1^{(k)} + \frac{1}{11} x_3^{(k-1)} - \frac{3}{11} x_4^{(k-1)} + \frac{25}{11}, \\ x_3^{(k)} &= -\frac{1}{5} x_1^{(k)} + \frac{1}{10} x_2^{(k)} + \frac{1}{10} x_4^{(k-1)} - \frac{11}{10}, \\ x_4^{(k)} &= -\frac{3}{8} x_2^{(k)} + \frac{1}{8} x_3^{(k)} + \frac{15}{8}. \end{split}$$

When $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$, we have $\mathbf{x}^{(1)} = (0.6000, 2.3272, -0.9873, 0.8789)^t$. Subsequent iterations give the values in Table 7.2.

Table 7.2

| k | 0 | 1 | 2 | 3 | 4 | 5 |
|--------------------------|--------|---------|--------|---------|---------|---------|
| $x_1^{(k)}$ | 0.0000 | 0.6000 | 1.030 | 1.0065 | 1.0009 | 1.0001 |
| $x_2^{(k)} \\ x_3^{(k)}$ | 0.0000 | 2.3272 | 2.037 | 2.0036 | 2.0003 | 2.0000 |
| | 0.0000 | -0.9873 | -1.014 | -1.0025 | -1.0003 | -1.0000 |
| $x_4^{(k)}$ | 0.0000 | 0.8789 | 0.9844 | 0.9983 | 0.9999 | 1.0000 |

Because

$$\frac{\|\mathbf{x}^{(5)} - \mathbf{x}^{(4)}\|_{\infty}}{\|\mathbf{x}^{(5)}\|_{\infty}} = \frac{0.0008}{2.000} = 4 \times 10^{-4},$$

 $\mathbf{x}^{(5)}$ is accepted as a reasonable approximation to the solution. Note that Jacobi's method in Example 1 required twice as many iterations for the same accuracy.

To write the Gauss-Seidel method in matrix form, multiply both sides of Eq. (7.8) by a_{ii} and collect all kth iterate terms, to give

$$a_{i1}x_1^{(k)} + a_{i2}x_2^{(k)} + \dots + a_{ii}x_i^{(k)} = -a_{i,i+1}x_{i+1}^{(k-1)} - \dots - a_{in}x_n^{(k-1)} + b_i,$$

for each i = 1, 2, ..., n. Writing all n equations gives

$$a_{11}x_1^{(k)} = -a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \dots - a_{1n}x_n^{(k-1)} + b_1,$$

$$a_{21}x_1^{(k)} + a_{22}x_2^{(k)} = -a_{23}x_3^{(k-1)} - \dots - a_{2n}x_n^{(k-1)} + b_2,$$

$$\vdots$$

$$a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \dots + a_{nn}x_n^{(k)} = b_n;$$

with the definitions of D, L, and U given previously, we have the Gauss-Seidel method represented by

$$(D-L)\mathbf{x}^{(k)} = U\mathbf{x}^{(k-1)} + \mathbf{b}$$

and

$$\mathbf{x}^{(k)} = (D - L)^{-1} U \mathbf{x}^{(k-1)} + (D - L)^{-1} \mathbf{b}, \text{ for each } k = 1, 2, \dots$$
 (7.9)

Letting $T_g = (D-L)^{-1}U$ and $\mathbf{c}_g = (D-L)^{-1}\mathbf{b}$, gives the Gauss-Seidel technique the form

$$\mathbf{x}^{(k)} = T_g \mathbf{x}^{(k-1)} + \mathbf{c}_g. \tag{7.10}$$

For the lower-triangular matrix D-L to be nonsingular, it is necessary and sufficient that $a_{ii} \neq 0$, for each i = 1, 2, ..., n.

Algorithm 7.2 implements the Gauss-Seidel method.



Gauss-Seidel Iterative

To solve $A\mathbf{x} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$:

INPUT the number of equations and unknowns n; the entries a_{ij} , $1 \le i, j \le n$ of the matrix A; the entries b_i , $1 \le i \le n$ of \mathbf{b} ; the entries XO_i , $1 \le i \le n$ of $\mathbf{XO} = \mathbf{x}^{(0)}$; tolerance TOL; maximum number of iterations N.

OUTPUT the approximate solution x_1, \ldots, x_n or a message that the number of iterations was exceeded.

Step 1 Set k = 1.

Step 2 While $(k \le N)$ do Steps 3–6.

Step 3 For $i = 1, \ldots, n$

$$set x_i = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} a_{ij} x_j - \sum_{j=i+1}^{n} a_{ij} X O_j + b_i \right].$$

Step 4 If $||\mathbf{x} - \mathbf{XO}|| < TOL$ then OUTPUT (x_1, \dots, x_n) ;

(The procedure was successful.)

Step 5 Set k = k + 1.

Step 6 For $i = 1, \ldots, n$ set $XO_i = x_i$.

Step 7 OUTPUT ('Maximum number of iterations exceeded'); (The procedure was successful.)
STOP.

The comments following Algorithm 7.1 regarding reordering and stopping criteria also apply to the Gauss-Seidel Algorithm 7.2.

The results of Examples 1 and 2 appear to imply that the Gauss-Seidel method is superior to the Jacobi method. This is almost always true, but there are linear systems for which the Jacobi method converges and the Gauss-Seidel method does not (see Exercises 9 and 10).

The *NumericalAnalysis* subpackage of the Maple *Student* package implements the Gauss-Siedel method in a manner similar to that of the Jacobi iterative method. The results in Table 7.2 are obtained by loading both *NumericalAnalysis* and *LinearAlgebra*, the matrix *A*, and then using the command

IterativeApproximate(A, initialapprox = Vector([0., 0., 0., 0.]), tolerance = 10^{-3} , maxiterations = 20, stoppingcriterion = relative(infinity), method = gaussseidel, output = approximates)

If we change the final option to output = [approximates, distances], the output also includes the l_{∞} distances between the approximations and the actual solution.

General Iteration Methods

To study the convergence of general iteration techniques, we need to analyze the formula

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$$
, for each $k = 1, 2, ...$

where $\mathbf{x}^{(0)}$ is arbitrary. The next lemma and Theorem 7.17 on page 449 provide the key for this study.

Lemma 7.18 If the spectral radius satisfies $\rho(T) < 1$, then $(I - T)^{-1}$ exists, and

$$(I-T)^{-1} = I + T + T^2 + \dots = \sum_{i=0}^{\infty} T^i.$$

Proof Because $T\mathbf{x} = \lambda \mathbf{x}$ is true precisely when $(I - T)\mathbf{x} = (1 - \lambda)\mathbf{x}$, we have λ as an eigenvalue of T precisely when $1 - \lambda$ is an eigenvalue of I - T. But $|\lambda| \le \rho(T) < 1$, so $\lambda = 1$ is not an eigenvalue of T, and 0 cannot be an eigenvalue of T. Hence, $(I - T)^{-1}$

Let
$$S_m = I + T + T^2 + \cdots + T^m$$
. Then

$$(I-T)S_m = (1+T+T^2+\cdots+T^m)-(T+T^2+\cdots+T^{m+1})=I-T^{m+1},$$

and, since T is convergent, Theorem 7.17 implies that

$$\lim_{m\to\infty} (I-T)S_m = \lim_{m\to\infty} (I-T^{m+1}) = I.$$

Thus,
$$(I-T)^{-1} = \lim_{m \to \infty} S_m = I + T + T^2 + \dots = \sum_{j=0}^{\infty} T^j$$
.

Theorem 7.19 For any $\mathbf{x}^{(0)} \in \mathbb{R}^n$, the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ defined by

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}, \quad \text{for each } k \ge 1, \tag{7.11}$$

converges to the unique solution of $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ if and only if $\rho(T) < 1$.

Proof First assume that $\rho(T) < 1$. Then,

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$$

$$= T(T\mathbf{x}^{(k-2)} + \mathbf{c}) + \mathbf{c}$$

$$= T^2\mathbf{x}^{(k-2)} + (T+I)\mathbf{c}$$

$$\vdots$$

$$= T^k\mathbf{x}^{(0)} + (T^{k-1} + \dots + T+I)\mathbf{c}.$$

Because $\rho(T) < 1$, Theorem 7.17 implies that T is convergent, and

$$\lim_{k\to\infty} T^k \mathbf{x}^{(0)} = \mathbf{0}.$$

Lemma 7.18 implies that

$$\lim_{k \to \infty} \mathbf{x}^{(k)} = \lim_{k \to \infty} T^k \mathbf{x}^{(0)} + \left(\sum_{j=0}^{\infty} T^j \right) \mathbf{c} = \mathbf{0} + (I - T)^{-1} \mathbf{c} = (I - T)^{-1} \mathbf{c}.$$

Hence, the sequence $\{\mathbf{x}^{(k)}\}$ converges to the vector $\mathbf{x} \equiv (I - T)^{-1}\mathbf{c}$ and $\mathbf{x} = T\mathbf{x} + \mathbf{c}$.

To prove the converse, we will show that for any $\mathbf{z} \in \mathbb{R}^n$, we have $\lim_{k \to \infty} T^k \mathbf{z} = \mathbf{0}$. By Theorem 7.17, this is equivalent to $\rho(T) < 1$.

Let **z** be an arbitrary vector, and **x** be the unique solution to $\mathbf{x} = T\mathbf{x} + \mathbf{c}$. Define $\mathbf{x}^{(0)} = \mathbf{x} - \mathbf{z}$, and, for $k \ge 1$, $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$. Then $\{\mathbf{x}^{(k)}\}$ converges to **x**. Also,

$$\mathbf{x} - \mathbf{x}^{(k)} = (T\mathbf{x} + \mathbf{c}) - (T\mathbf{x}^{(k-1)} + \mathbf{c}) = T(\mathbf{x} - \mathbf{x}^{(k-1)}),$$

so

$$\mathbf{x} - \mathbf{x}^{(k)} = T(\mathbf{x} - \mathbf{x}^{(k-1)}) = T^2(\mathbf{x} - \mathbf{x}^{(k-2)}) = \dots = T^k(\mathbf{x} - \mathbf{x}^{(0)}) = T^k \mathbf{z}.$$

Hence $\lim_{k\to\infty} T^k \mathbf{z} = \lim_{k\to\infty} T^k \left(\mathbf{x} - \mathbf{x}^{(0)}\right) = \lim_{k\to\infty} \left(\mathbf{x} - \mathbf{x}^{(k)}\right) = \mathbf{0}$. But $\mathbf{z} \in \mathbb{R}^n$ was arbitrary, so by Theorem 7.17, T is convergent and $\rho(T) < 1$.

The proof of the following corollary is similar to the proofs in Corollary 2.5 on page 62. It is considered in Exercise 13.

If ||T|| < 1 for any natural matrix norm and \mathbf{c} is a given vector, then the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ defined by $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$ converges, for any $\mathbf{x}^{(0)} \in \mathbb{R}^n$, to a vector $\mathbf{x} \in \mathbb{R}^n$, with Corollary 7.20 $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, and the following error bounds hold:

(i)
$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \le \|T\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\|;$$
 (ii) $\|\mathbf{x} - \mathbf{x}^{(k)}\| \le \frac{\|T\|^k}{1 - \|T\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|.$

We have seen that the Jacobi and Gauss-Seidel iterative techniques can be written

$$\mathbf{x}^{(k)} = T_j \mathbf{x}^{(k-1)} + \mathbf{c}_j$$
 and $\mathbf{x}^{(k)} = T_g \mathbf{x}^{(k-1)} + \mathbf{c}_g$,

using the matrices

$$T_i = D^{-1}(L+U)$$
 and $T_g = (D-L)^{-1}U$.

If $\rho(T_j)$ or $\rho(T_g)$ is less than 1, then the corresponding sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ will converge to the solution \mathbf{x} of $A\mathbf{x} = \mathbf{b}$. For example, the Jacobi scheme has

$$\mathbf{x}^{(k)} = D^{-1}(L+U)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b},$$

and, if $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converges to \mathbf{x} , then

$$\mathbf{x} = D^{-1}(L+U)\mathbf{x} + D^{-1}\mathbf{b}.$$

This implies that

$$D\mathbf{x} = (L+U)\mathbf{x} + \mathbf{b}$$
 and $(D-L-U)\mathbf{x} = \mathbf{b}$.

Since D - L - U = A, the solution **x** satisfies A**x** = **b**.

We can now give easily verified sufficiency conditions for convergence of the Jacobi and Gauss-Seidel methods. (To prove convergence for the Jacobi scheme see Exercise 14, and for the Gauss-Seidel scheme see [Or2], p. 120.)

Theorem 7.21 If A is strictly diagonally dominant, then for any choice of $\mathbf{x}^{(0)}$, both the Jacobi and Gauss-Seidel methods give sequences $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ that converge to the unique solution of $A\mathbf{x} = \mathbf{b}$.

The relationship of the rapidity of convergence to the spectral radius of the iteration matrix T can be seen from Corollary 7.20. The inequalities hold for any natural matrix norm, so it follows from the statement after Theorem 7.15 on page 446 that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \approx \rho(T)^k \|\mathbf{x}^{(0)} - \mathbf{x}\|.$$
 (7.12)

Thus we would like to select the iterative technique with minimal $\rho(T) < 1$ for a particular system $A\mathbf{x} = \mathbf{b}$. No general results exist to tell which of the two techniques, Jacobi or Gauss-Seidel, will be most successful for an arbitrary linear system. In special cases, however, the answer is known, as is demonstrated in the following theorem. The proof of this result can be found in [Y], pp. 120–127.

Theorem 7.22 (Stein-Rosenberg)

If $a_{ij} \le 0$, for each $i \ne j$ and $a_{ii} > 0$, for each i = 1, 2, ..., n, then one and only one of the following statements holds:

(i)
$$0 \le \rho(T_g) < \rho(T_i) < 1$$
;

(ii)
$$1 < \rho(T_j) < \rho(T_g);$$

(iii)
$$\rho(T_i) = \rho(T_g) = 0;$$

(iv)
$$\rho(T_i) = \rho(T_g) = 1$$
.

For the special case described in Theorem 7.22, we see from part (i) that when one method gives convergence, then both give convergence, and the Gauss-Seidel method converges faster than the Jacobi method. Part (ii) indicates that when one method diverges then both diverge, and the divergence is more pronounced for the Gauss-Seidel method.

EXERCISE SET 7.3

1. Find the first two iterations of the Jacobi method for the following linear systems, using $\mathbf{x}^{(0)} = \mathbf{0}$:

a.
$$3x_1 - x_2 + x_3 = 1$$
, $3x_1 + 6x_2 + 2x_3 = 0$, $3x_1 + 3x_2 + 7x_3 = 4$.

b.
$$10x_1 - x_2 = 9$$
, $-x_1 + 10x_2 - 2x_3 = 7$, $-2x_2 + 10x_3 = 6$.

c.
$$10x_1 + 5x_2 = 6$$
,
 $5x_1 + 10x_2 - 4x_3 = 25$,
 $-4x_2 + 8x_3 - x_4 = -11$,
 $-x_3 + 5x_4 = -11$.

d.
$$4x_1 + x_2 + x_3 + x_5 = 6,$$

 $-x_1 - 3x_2 + x_3 + x_4 = 6,$
 $2x_1 + x_2 + 5x_3 - x_4 - x_5 = 6,$
 $-x_1 - x_2 - x_3 + 4x_4 = 6,$
 $2x_2 - x_3 + x_4 + 4x_5 = 6.$

2. Find the first two iterations of the Jacobi method for the following linear systems, using $\mathbf{x}^{(0)} = \mathbf{0}$:

a.
$$4x_1 + x_2 - x_3 = 5,$$

 $-x_1 + 3x_2 + x_3 = -4,$
 $2x_1 + 2x_2 + 5x_3 = 1.$

b.
$$-2x_1 + x_2 + \frac{1}{2}x_3 = 4,$$

 $x_1 - 2x_2 - \frac{1}{2}x_3 = -4,$
 $x_2 + 2x_3 = 0.$

c.
$$4x_1 + x_2 - x_3 + x_4 = -2$$
,
 $x_1 + 4x_2 - x_3 - x_4 = -1$,
 $-x_1 - x_2 + 5x_3 + x_4 = 0$,
 $x_1 - x_2 + x_3 + 3x_4 = 1$.

d.
$$4x_1 - x_2 - x_4 = 0,$$

 $-x_1 + 4x_2 - x_3 - x_5 = 5,$
 $-x_2 + 4x_3 - x_6 = 0,$
 $-x_1 + 4x_4 - x_5 = 6,$
 $-x_2 - x_4 + 4x_5 - x_6 = -2,$
 $-x_3 - x_5 + 4x_6 = 6.$

- 3. Repeat Exercise 1 using the Gauss-Seidel method.
- 4. Repeat Exercise 2 using the Gauss-Seidel method.
- 5. Use the Jacobi method to solve the linear systems in Exercise 1, with $TOL = 10^{-3}$ in the l_{∞} norm.
- **6.** Use the Jacobi method to solve the linear systems in Exercise 2, with $TOL = 10^{-3}$ in the l_{∞} norm.
- 7. Use the Gauss-Seidel method to solve the linear systems in Exercise 1, with $TOL = 10^{-3}$ in the l_{∞} norm
- 8. Use the Gauss-Seidel method to solve the linear systems in Exercise 2, with $TOL = 10^{-3}$ in the l_{∞} norm.
- **9.** The linear system

$$2x_1 - x_2 + x_3 = -1,$$

$$2x_1 + 2x_2 + 2x_3 = 4,$$

$$-x_1 - x_2 + 2x_3 = -5$$

has the solution $(1, 2, -1)^t$.

- **a.** Show that $\rho(T_j) = \frac{\sqrt{5}}{2} > 1$.
- **b.** Show that the Jacobi method with $\mathbf{x}^{(0)} = \mathbf{0}$ fails to give a good approximation after 25 iterations.
- **c.** Show that $\rho(T_g) = \frac{1}{2}$.
- **d.** Use the Gauss-Seidel method with $\mathbf{x}^{(0)} = \mathbf{0}$ to approximate the solution to the linear system to within 10^{-5} in the l_{∞} norm.
- 10. The linear system

$$x_1 + 2x_2 - 2x_3 = 7,$$

 $x_1 + x_2 + x_3 = 2,$
 $2x_1 + 2x_2 + x_3 = 5$

has the solution $(1, 2, -1)^t$.

- **a.** Show that $\rho(T_i) = 0$.
- **b.** Use the Jacobi method with $\mathbf{x}^{(0)} = \mathbf{0}$ to approximate the solution to the linear system to within 10^{-5} in the l_{∞} norm.
- c. Show that $\rho(T_g) = 2$.
- d. Show that the Gauss-Seidel method applied as in part (b) fails to give a good approximation in 25 iterations.
- 11. The linear system

$$x_1 - x_3 = 0.2,$$

$$-\frac{1}{2}x_1 + x_2 - \frac{1}{4}x_3 = -1.425,$$

$$x_1 - \frac{1}{2}x_2 + x_3 = 2.$$

has the solution $(0.9, -0.8, 0.7)^t$.

a. Is the coefficient matrix

$$A = \left[\begin{array}{rrr} 1 & 0 & -1 \\ -\frac{1}{2} & 1 & -\frac{1}{4} \\ 1 & -\frac{1}{2} & 1 \end{array} \right]$$

strictly diagonally dominant?

- Compute the spectral radius of the Gauss-Seidel matrix T_g .
- c. Use the Gauss-Seidel iterative method to approximate the solution to the linear system with a tolerance of 10⁻² and a maximum of 300 iterations.
- **d.** What happens in part (c) when the system is changed to

$$x_1 - 2x_3 = 0.2,$$

$$-\frac{1}{2}x_1 + x_2 - \frac{1}{4}x_3 = -1.425,$$

$$x_1 - \frac{1}{2}x_2 + x_3 = 2.$$

- **12.** Repeat Exercise 11 using the Jacobi method.
- **13.** a. Prove that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \|T\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\| \quad \text{and} \quad \|\mathbf{x}^{(k)} - \mathbf{x}\| \le \frac{\|T\|^k}{1 - \|T\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|,$$

where T is an $n \times n$ matrix with ||T|| < 1 and

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}, \quad k = 1, 2, \dots$$

with $\mathbf{x}^{(0)}$ arbitrary, $\mathbf{c} \in \mathbb{R}^n$, and $\mathbf{x} = T\mathbf{x} + \mathbf{c}$.

- **b.** Apply the bounds to Exercise 1, when possible, using the l_{∞} norm.
- **14.** Show that if *A* is strictly diagonally dominant, then $||T_j||_{\infty} < 1$.
- 15. Use (a) the Jacobi and (b) the Gauss-Seidel methods to solve the linear system $A\mathbf{x} = \mathbf{b}$ to within 10^{-5} in the l_{∞} norm, where the entries of A are

$$a_{i,j} = \begin{cases} 2i, & \text{when } j = i \text{ and } i = 1, 2, \dots, 80, \\ 0.5i, & \text{when } \begin{cases} j = i + 2 \text{ and } i = 1, 2, \dots, 78, \\ j = i - 2 \text{ and } i = 3, 4, \dots, 80, \end{cases} \\ 0.25i, & \text{when } \begin{cases} j = i + 4 \text{ and } i = 1, 2, \dots, 76, \\ j = i - 4 \text{ and } i = 5, 6, \dots, 80, \end{cases} \\ 0, & \text{otherwise,} \end{cases}$$

and those of **b** are $b_i = \pi$, for each i = 1, 2, ..., 80.

16. Suppose that an object can be at any one of n+1 equally spaced points x_0, x_1, \ldots, x_n . When an object is at location x_i , it is equally likely to move to either x_{i-1} or x_{i+1} and cannot directly move to any other location. Consider the probabilities $\{P_i\}_{i=0}^n$ that an object starting at location x_i will reach the left endpoint x_0 before reaching the right endpoint x_n . Clearly, $P_0 = 1$ and $P_n = 0$. Since the object can move to x_i only from x_{i-1} or x_{i+1} and does so with probability $\frac{1}{2}$ for each of these locations,

$$P_i = \frac{1}{2}P_{i-1} + \frac{1}{2}P_{i+1}$$
, for each $i = 1, 2, ..., n - 1$.

a. Show that

$$\begin{bmatrix} 1 & -\frac{1}{2} & 0 & \cdots & \cdots & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} & \ddots & \ddots & \vdots \\ 0 & -\frac{1}{2} & 1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_{n-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- **b.** Solve this system using n = 10, 50, and 100.
- c. Change the probabilities to α and 1α for movement to the left and right, respectively, and derive the linear system similar to the one in part (a).
- **d.** Repeat part (b) with $\alpha = \frac{1}{3}$.
- **17.** Suppose that *A* is a positive definite.
 - **a.** Show that we can write $A = D L L^t$, where *D* is diagonal with $d_{ii} > 0$ for each $1 \le i \le n$ and *L* is lower triangular. Further, show that D L is nonsingular.
 - **b.** Let $T_g = (D L)^{-1} L^t$ and $P = A T_g^t A T_g$. Show that P is symmetric.
 - **c.** Show that T_g can also be written as $T_g = I (D L)^{-1}A$.

- **d.** Let $Q = (D L)^{-1}A$. Show that $T_g = I Q$ and $P = Q'[AQ^{-1} A + (Q')^{-1}A]Q$.
- **e.** Show that $P = Q^t DQ$ and P is positive definite.
- **f.** Let λ be an eigenvalue of T_g with eigenvector $\mathbf{x} \neq \mathbf{0}$. Use part (b) to show that $\mathbf{x}'P\mathbf{x} > 0$ implies that $|\lambda| < 1$.
- **g.** Show that T_g is convergent and prove that the Gauss-Seidel method converges.
- **18.** The forces on the bridge truss described in the opening to this chapter satisfy the equations in the following table:

| Joint | Horizontal Component | Vertical Component |
|-------|--|---|
| 1 | $-F_1 + \frac{\sqrt{2}}{2}f_1 + f_2 = 0$ | $\frac{\sqrt{2}}{2}f_1 - F_2 = 0$ |
| 2 | $-\frac{\sqrt{2}}{2}f_1 + \frac{\sqrt{3}}{2}f_4 = 0$ | $-\frac{\sqrt{2}}{2}f_1 - f_3 - \frac{1}{2}f_4 = 0$ |
| 3 | $-f_2 + f_5 = 0$ | $f_3 - 10,000 = 0$ |
| 4 | $-\frac{\sqrt{3}}{2}f_4 - f_5 = 0$ | $\frac{1}{2}f_4 - F_3 = 0$ |

This linear system can be placed in the matrix form

$$\begin{bmatrix} -1 & 0 & 0 & \frac{\sqrt{2}}{2} & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & \frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & 0 & -1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & 0 & 0 & \frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\sqrt{3}}{2} & -1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 10,000 \\ 0 \\ 0 \end{bmatrix}.$$

- **a.** Explain why the system of equations was reordered.
- **b.** Approximate the solution of the resulting linear system to within 10^{-2} in the l_{∞} norm using as initial approximation the vector all of whose entries are 1s with (i) the Jacobi method and (ii) the Gauss-Seidel method.

7.4 Relaxation Techniques for Solving Linear Systems

We saw in Section 7.3 that the rate of convergence of an iterative technique depends on the spectral radius of the matrix associated with the method. One way to select a procedure to accelerate convergence is to choose a method whose associated matrix has minimal spectral radius. Before describing a procedure for selecting such a method, we need to introduce a new means of measuring the amount by which an approximation to the solution to a linear system differs from the true solution to the system. The method makes use of the vector described in the following definition.

Definition 7.23

Suppose $\tilde{\mathbf{x}} \in \mathbb{R}^n$ is an approximation to the solution of the linear system defined by $A\mathbf{x} = \mathbf{b}$. The **residual vector** for $\tilde{\mathbf{x}}$ with respect to this system is $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$.

The word residual means what is left over, which is an appropriate name for this vector.

In procedures such as the Jacobi or Gauss-Seidel methods, a residual vector is associated with each calculation of an approximate component to the solution vector. The true objective is to generate a sequence of approximations that will cause the residual vectors to converge rapidly to zero. Suppose we let

$$\mathbf{r}_{i}^{(k)} = (r_{1i}^{(k)}, r_{2i}^{(k)}, \dots, r_{ni}^{(k)})^{t}$$

denote the residual vector for the Gauss-Seidel method corresponding to the approximate solution vector $\mathbf{x}_i^{(k)}$ defined by

$$\mathbf{x}_{i}^{(k)} = (x_{1}^{(k)}, x_{2}^{(k)}, \dots, x_{i-1}^{(k)}, x_{i}^{(k-1)}, \dots, x_{n}^{(k-1)})^{t}.$$

The *m*th component of $\mathbf{r}_{i}^{(k)}$ is

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^n a_{mj} x_j^{(k-1)},$$
(7.13)

or, equivalently,

$$r_{mi}^{(k)} = b_m - \sum_{i=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{i=i+1}^n a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)},$$

for each m = 1, 2, ..., n.

In particular, the *i*th component of $\mathbf{r}_{i}^{(k)}$ is

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^n a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k-1)},$$

so

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{i=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{i=i+1}^{n} a_{ij}x_j^{(k-1)}.$$
 (7.14)

Recall, however, that in the Gauss-Seidel method, $x_i^{(k)}$ is chosen to be

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^n a_{ij} x_j^{(k-1)} \right], \tag{7.15}$$

so Eq. (7.14) can be rewritten as

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}x_i^{(k)}.$$

Consequently, the Gauss-Seidel method can be characterized as choosing $x_i^{(k)}$ to satisfy

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}. (7.16)$$

We can derive another connection between the residual vectors and the Gauss-Seidel technique. Consider the residual vector $\mathbf{r}_{i+1}^{(k)}$, associated with the vector $\mathbf{x}_{i+1}^{(k)} = (x_1^{(k)}, \dots, x_i^{(k)}, x_{i+1}^{(k-1)}, \dots, x_n^{(k-1)})^t$. By Eq. (7.13) the ith component of $\mathbf{r}_{i+1}^{(k)}$ is

$$\begin{split} r_{i,i+1}^{(k)} &= b_i - \sum_{j=1}^i a_{ij} x_j^{(k)} - \sum_{j=i+1}^n a_{ij} x_j^{(k-1)} \\ &= b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^n a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k)}. \end{split}$$

By the manner in which $x_i^{(k)}$ is defined in Eq. (7.15) we see that $r_{i,i+1}^{(k)} = 0$. In a sense, then, the Gauss-Seidel technique is characterized by choosing each $x_{i+1}^{(k)}$ in such a way that the *i*th component of $\mathbf{r}_{i+1}^{(k)}$ is zero.

Choosing $x_{i+1}^{(k)}$ so that one coordinate of the residual vector is zero, however, is not necessarily the most efficient way to reduce the norm of the vector $\mathbf{r}_{i+1}^{(k)}$. If we modify the

Gauss-Seidel procedure, as given by Eq. (7.16), to

$$x_i^{(k)} = x_i^{(k-1)} + \omega \frac{r_{ii}^{(k)}}{a_{ii}},\tag{7.17}$$

then for certain choices of positive ω we can reduce the norm of the residual vector and obtain significantly faster convergence.

Methods involving Eq. (7.17) are called **relaxation methods**. For choices of ω with $0 < \omega < 1$, the procedures are called **under-relaxation methods**. We will be interested in choices of ω with $1 < \omega$, and these are called **over-relaxation methods**. They are used to accelerate the convergence for systems that are convergent by the Gauss-Seidel technique. The methods are abbreviated **SOR**, for **Successive Over-Relaxation**, and are particularly useful for solving the linear systems that occur in the numerical solution of certain partial-differential equations.

Before illustrating the advantages of the SOR method, we note that by using Eq. (7.14), we can reformulate Eq. (7.17) for calculation purposes as

$$x_i^{(k)} = (1 - \omega)x_i^{(k-1)} + \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^n a_{ij} x_j^{(k-1)} \right].$$

To determine the matrix form of the SOR method, we rewrite this as

$$a_{ii}x_i^{(k)} + \omega \sum_{i=1}^{i-1} a_{ij}x_j^{(k)} = (1 - \omega)a_{ii}x_i^{(k-1)} - \omega \sum_{i=i+1}^n a_{ij}x_j^{(k-1)} + \omega b_i,$$

so that in vector form, we have

$$(D - \omega L)\mathbf{x}^{(k)} = [(1 - \omega)D + \omega U]\mathbf{x}^{(k-1)} + \omega \mathbf{b}.$$

That is,

$$\mathbf{x}^{(k)} = (D - \omega L)^{-1} [(1 - \omega)D + \omega U] \mathbf{x}^{(k-1)} + \omega (D - \omega L)^{-1} \mathbf{b}.$$
 (7.18)

Letting $T_{\omega} = (D - \omega L)^{-1}[(1 - \omega)D + \omega U]$ and $\mathbf{c}_{\omega} = \omega(D - \omega L)^{-1}\mathbf{b}$, gives the SOR technique the form

$$\mathbf{x}^{(k)} = T_{\omega}\mathbf{x}^{(k-1)} + \mathbf{c}_{\omega}. \tag{7.19}$$

The linear system $A\mathbf{x} = \mathbf{b}$ given by

$$4x_1 + 3x_2 = 24,$$

$$3x_1 + 4x_2 - x_3 = 30,$$

$$-x_2 + 4x_3 = -24,$$

has the solution $(3, 4, -5)^t$. Compare the iterations from the Gauss-Seidel method and the SOR method with $\omega = 1.25$ using $\mathbf{x}^{(0)} = (1, 1, 1)^t$ for both methods.

Solution For each k = 1, 2, ..., the equations for the Gauss-Seidel method are

$$x_1^{(k)} = -0.75x_2^{(k-1)} + 6,$$

$$x_2^{(k)} = -0.75x_1^{(k)} + 0.25x_3^{(k-1)} + 7.5,$$

$$x_3^{(k)} = 0.25x_2^{(k)} - 6,$$

and the equations for the SOR method with $\omega=1.25$ are

$$\begin{split} x_1^{(k)} &= -0.25x_1^{(k-1)} - 0.9375x_2^{(k-1)} + 7.5, \\ x_2^{(k)} &= -0.9375x_1^{(k)} - 0.25x_2^{(k-1)} + 0.3125x_3^{(k-1)} + 9.375, \\ x_3^{(k)} &= 0.3125x_2^{(k)} - 0.25x_3^{(k-1)} - 7.5. \end{split}$$

The first seven iterates for each method are listed in Tables 7.3 and 7.4. For the iterates to be accurate to seven decimal places, the Gauss-Seidel method requires 34 iterations, as opposed to 14 iterations for the SOR method with $\omega = 1.25$.

Table 7.3

| k | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---------------|---|-----------|------------|------------|------------|------------|------------|------------|
| $x_1^{(k)}$ | 1 | 5.250000 | 3.1406250 | 3.0878906 | 3.0549316 | 3.0343323 | 3.0214577 | 3.0134110 |
| $x_{2}^{(k)}$ | 1 | 3.812500 | 3.8828125 | 3.9267578 | 3.9542236 | 3.9713898 | 3.9821186 | 3.9888241 |
| $x_{3}^{(k)}$ | 1 | -5.046875 | -5.0292969 | -5.0183105 | -5.0114441 | -5.0071526 | -5.0044703 | -5.0027940 |

Table 7.4

| k | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---------------|---|------------|------------|------------|------------|------------|------------|------------|
| $x_1^{(k)}$ | 1 | 6.312500 | 2.6223145 | 3.1333027 | 2.9570512 | 3.0037211 | 2.9963276 | 3.0000498 |
| $x_{2}^{(k)}$ | 1 | 3.5195313 | 3.9585266 | 4.0102646 | 4.0074838 | 4.0029250 | 4.0009262 | 4.0002586 |
| $x_{3}^{(k)}$ | 1 | -6.6501465 | -4.6004238 | -5.0966863 | -4.9734897 | -5.0057135 | -4.9982822 | -5.0003486 |

An obvious question to ask is how the appropriate value of ω is chosen when the SOR method is used. Although no complete answer to this question is known for the general $n \times n$ linear system, the following results can be used in certain important situations.

Theorem 7.24 (Kahan)

If $a_{ii} \neq 0$, for each i = 1, 2, ..., n, then $\rho(T_{\omega}) \geq |\omega - 1|$. This implies that the SOR method can converge only if $0 < \omega < 2$.

The proof of this theorem is considered in Exercise 9. The proof of the next two results can be found in [Or2], pp. 123–133. These results will be used in Chapter 12.

Theorem 7.25 (Ostrowski-Reich)

If A is a positive definite matrix and $0 < \omega < 2$, then the SOR method converges for any choice of initial approximate vector $\mathbf{x}^{(0)}$.

Theorem 7.26 If A is positive definite and tridiagonal, then $\rho(T_g) = [\rho(T_j)]^2 < 1$, and the optimal choice of ω for the SOR method is

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}}.$$

With this choice of ω , we have $\rho(T_{\omega}) = \omega - 1$.

Example 2 Find the optimal choice of ω for the SOR method for the matrix

$$A = \left[\begin{array}{rrr} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{array} \right].$$

Solution This matrix is clearly tridiagonal, so we can apply the result in Theorem 7.26 if we can also who that it is positive definite. Because the matrix is symmetric, Theorem 6.24 on page 416 states that it is positive definite if and only if all its leading principle submatrices has a positive determinant. This is easily seen to be the case because

$$det(A) = 24$$
, $det \left(\begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix} \right) = 7$, and $det([4]) = 4$.

Because

$$T_{j} = D^{-1}(L+U) = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & -3 & 0 \\ -3 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -0.75 & 0 \\ -0.75 & 0 & 0.25 \\ 0 & 0.25 & 0 \end{bmatrix},$$

we have

$$T_j - \lambda I = \begin{bmatrix} -\lambda & -0.75 & 0 \\ -0.75 & -\lambda & 0.25 \\ 0 & 0.25 & -\lambda \end{bmatrix},$$

so

$$\det(T_i - \lambda I) = -\lambda(\lambda^2 - 0.625).$$

Thus

$$\rho(T_i) = \sqrt{0.625}$$

and

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}} = \frac{2}{1 + \sqrt{1 - 0.625}} \approx 1.24.$$

This explains the rapid convergence obtained in Example 1 when using $\omega = 1.25$.

We close this section with Algorithm 7.3 for the SOR method.



SOR

To solve $A\mathbf{x} = \mathbf{b}$ given the parameter ω and an initial approximation $\mathbf{x}^{(0)}$:

INPUT the number of equations and unknowns n; the entries a_{ij} , $1 \le i, j \le n$, of the matrix A; the entries b_i , $1 \le i \le n$, of **b**; the entries XO_i , $1 \le i \le n$, of $XO = \mathbf{x}^{(0)}$; the parameter ω ; tolerance TOL; maximum number of iterations N.

OUTPUT the approximate solution x_1, \ldots, x_n or a message that the number of iterations was exceeded.

Step 1 Set k = 1.

Step 2 While $(k \le N)$ do Steps 3–6.

Step 3 For $i = 1, \ldots, n$

set
$$x_i = (1 - \omega)XO_i + \frac{1}{a_{ii}} \left[\omega \left(-\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^n a_{ij}XO_j + b_i \right) \right].$$

Step 4 If $||\mathbf{x} - \mathbf{XO}|| < TOL$ then OUTPUT (x_1, \dots, x_n) ;

(The procedure was successful.)

Step 5 Set k = k + 1.

Step 6 For i = 1, ..., n set $XO_i = x_i$.

Step 7 OUTPUT ('Maximum number of iterations exceeded'); (The procedure was successful.) STOP.

The *NumericalAnalysis* subpackage of the Maple *Student* package implements the SOR method in a manner similar to that of the Jacobi and Gauss-Seidel methods. The SOR results in Table 7.4 are obtained by loading both *NumericalAnalysis* and *LinearAlgebra*, the matrix A, the vector $\mathbf{b} = [24, 30, -24]^t$, and then using the command

IterativeApproximate(A, **b**, initialapprox = Vector([1., 1., 1., 1.]), tolerance = 10^{-3} , maxiterations = 20, stoppingcriterion = relative(infinity), method = SOR(1.25), output = approximates)

The input method = SOR(1.25) indicates that the SOR method should use the value $\omega = 1.25$.

EXERCISE SET 7.4

1. Find the first two iterations of the SOR method with $\omega = 1.1$ for the following linear systems, using $\mathbf{x}^{(0)} = \mathbf{0}$:

a.
$$3x_1 - x_2 + x_3 = 1$$
, $3x_1 + 6x_2 + 2x_3 = 0$, $3x_1 + 3x_2 + 7x_3 = 4$.

c.
$$10x_1 + 5x_2 = 6$$
,
 $5x_1 + 10x_2 - 4x_3 = 25$,
 $-4x_2 + 8x_3 - x_4 = -11$,
 $-x_3 + 5x_4 = -11$.

b.
$$10x_1 - x_2 = 9$$
,
 $-x_1 + 10x_2 - 2x_3 = 7$,
 $-2x_2 + 10x_3 = 6$.

d.
$$4x_1 + x_2 + x_3 + x_5 = 6,$$

 $-x_1 - 3x_2 + x_3 + x_4 = 6,$
 $2x_1 + x_2 + 5x_3 - x_4 - x_5 = 6,$
 $-x_1 - x_2 - x_3 + 4x_4 = 6,$
 $2x_2 - x_3 + x_4 + 4x_5 = 6.$

- 2. Find the first two iterations of the SOR method with $\omega = 1.1$ for the following linear systems, using $\mathbf{x}^{(0)} = \mathbf{0}$:
 - **a.** $4x_1 + x_2 x_3 = 5,$ $-x_1 + 3x_2 + x_3 = -4,$ $2x_1 + 2x_2 + 5x_3 = 1.$
- **b.** $-2x_1 + x_2 + \frac{1}{2}x_3 = 4,$ $x_1 - 2x_2 - \frac{1}{2}x_3 = -4,$ $x_2 + 2x_3 = 0.$
- c. $4x_1 + x_2 x_3 + x_4 = -2$, $x_1 + 4x_2 - x_3 - x_4 = -1$, $-x_1 - x_2 + 5x_3 + x_4 = 0$, $x_1 - x_2 + x_3 + 3x_4 = 1$.
- **d.** $4x_1 x_2 = 0,$ $-x_1 + 4x_2 - x_3 = 5,$ $-x_2 + 4x_3 = 0,$ $+4x_4 - x_5 = 6,$ $-x_4 + 4x_5 - x_6 = -2,$ $-x_5 + 4x_6 = 6.$
- 3. Repeat Exercise 1 using $\omega = 1.3$.
- **4.** Repeat Exercise 2 using $\omega = 1.3$.
- 5. Use the SOR method with $\omega=1.2$ to solve the linear systems in Exercise 1 with a tolerance $TOL=10^{-3}$ in the l_{∞} norm.
- **6.** Use the SOR method with $\omega=1.2$ to solve the linear systems in Exercise 2 with a tolerance $TOL=10^{-3}$ in the l_{∞} norm.
- 7. Determine which matrices in Exercise 1 are tridiagonal and positive definite. Repeat Exercise 1 for these matrices using the optimal choice of ω .
- 8. Determine which matrices in Exercise 2 are tridiagonal and positive definite. Repeat Exercise 2 for these matrices using the optimal choice of ω .
- 9. Prove Kahan's Theorem 7.24. [Hint: If $\lambda_1, \ldots, \lambda_n$ are eigenvalues of T_{ω} , then det $T_{\omega} = \prod_{i=1}^{n} \lambda_i$. Since det $D^{-1} = \det(D \omega L)^{-1}$ and the determinant of a product of matrices is the product of the determinants of the factors, the result follows from Eq. (7.18).]
- **10.** The forces on the bridge truss described in the opening to this chapter satisfy the equations in the following table:

| Joint | Horizontal Component | Vertical Component |
|-------|--|---|
| 1 | $-F_1 + \frac{\sqrt{2}}{2}f_1 + f_2 = 0$ | $\frac{\sqrt{2}}{2}f_1 - F_2 = 0$ |
| 2 | $-\frac{\sqrt{2}}{2}f_1 + \frac{\sqrt{3}}{2}f_4 = 0$ | $-\frac{\sqrt{2}}{2}f_1 - f_3 - \frac{1}{2}f_4 = 0$ |
| 3 | $-f_2 + f_5 = 0$ | $f_3 - 10,000 = 0$ |
| 4 | $-\frac{\sqrt{3}}{2}f_4 - f_5 = 0$ | $\frac{1}{2}f_4 - F_3 = 0$ |

This linear system can be placed in the matrix form

$$\begin{bmatrix} -1 & 0 & 0 & \frac{\sqrt{2}}{2} & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & \frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & 0 & -1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & 0 & 0 & \frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\sqrt{3}}{2} & -1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 10,000 \\ 0 \\ 0 \end{bmatrix}.$$

- a. Explain why the system of equations was reordered.
- **b.** Approximate the solution of the resulting linear system to within 10^{-2} in the l_{∞} norm using as initial approximation the vector all of whose entries are 1s and the SOR method with $\omega = 1.25$.

11. Use the SOR method to solve the linear system $A\mathbf{x} = \mathbf{b}$ to within 10^{-5} in the l_{∞} norm, where the entries of A are

$$a_{i,j} = \begin{cases} 2i, & \text{when } j = i \text{ and } i = 1, 2, \dots, 80, \\ 0.5i, & \text{when } \begin{cases} j = i + 2 \text{ and } i = 1, 2, \dots, 78, \\ j = i - 2 \text{ and } i = 3, 4, \dots, 80, \end{cases} \\ 0.25i, & \text{when } \begin{cases} j = i + 4 \text{ and } i = 1, 2, \dots, 76, \\ j = i - 4 \text{ and } i = 5, 6, \dots, 80, \end{cases} \\ 0, & \text{otherwise,} \end{cases}$$

and those of **b** are $b_i = \pi$, for each i = 1, 2, ..., 80.

12. In Exercise 17 of Section 7.3 a technique was outlined to prove that the Gauss-Seidel method converges when A is a positive definite matrix. Extend this method of proof to show that in this case there is also convergence for the SOR method with $0 < \omega < 2$.

7.5 Error Bounds and Iterative Refinement

It seems intuitively reasonable that if $\tilde{\mathbf{x}}$ is an approximation to the solution \mathbf{x} of $A\mathbf{x} = \mathbf{b}$ and the residual vector $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ has the property that $\|\mathbf{r}\|$ is small, then $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ would be small as well. This is often the case, but certain systems, which occur frequently in practice, fail to have this property.

Example 1 The linear system $A\mathbf{x} = \mathbf{b}$ given by

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

has the unique solution $\mathbf{x} = (1, 1)^t$. Determine the residual vector for the poor approximation $\tilde{\mathbf{x}} = (3, -0.0001)^t$.

Solution We have

$$\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}} = \begin{bmatrix} 3\\3.0001 \end{bmatrix} - \begin{bmatrix} 1\\1.0001 & 2 \end{bmatrix} \begin{bmatrix} 3\\-0.0001 \end{bmatrix} = \begin{bmatrix} 0.0002\\0 \end{bmatrix},$$

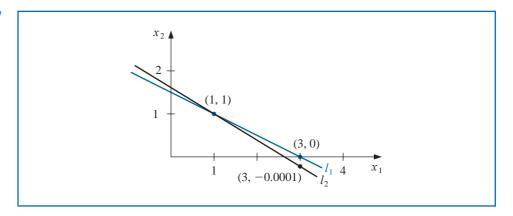
so $\|\mathbf{r}\|_{\infty} = 0.0002$. Although the norm of the residual vector is small, the approximation $\tilde{\mathbf{x}} = (3, -0.0001)^t$ is obviously quite poor; in fact, $\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty} = 2$.

The difficulty in Example 1 is explained quite simply by noting that the solution to the system represents the intersection of the lines

$$l_1: x_1 + 2x_2 = 3$$
 and $l_2: 1.0001x_1 + 2x_2 = 3.0001$.

The point (3, -0.0001) lies on l_2 , and the lines are nearly parallel. This implies that (3, -0.0001) also lies close to l_1 , even though it differs significantly from the solution of the system, given by the intersection point (1, 1). (See Figure 7.7.)

Figure 7.7



Example 1 was clearly constructed to show the difficulties that can—and, in fact, do—arise. Had the lines not been nearly coincident, we would expect a small residual vector to imply an accurate approximation.

In the general situation, we cannot rely on the geometry of the system to give an indication of when problems might occur. We can, however, obtain this information by considering the norms of the matrix A and its inverse.

Theorem 7.27 Suppose that $\tilde{\mathbf{x}}$ is an approximation to the solution of $A\mathbf{x} = \mathbf{b}$, A is a nonsingular matrix, and \mathbf{r} is the residual vector for $\tilde{\mathbf{x}}$. Then for any natural norm,

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \le \|\mathbf{r}\| \cdot \|A^{-1}\|$$

and if $x \neq 0$ and $b \neq 0$,

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \|A\| \cdot \|A^{-1}\| \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$
 (7.20)

Proof Since $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}} = A\mathbf{x} - A\tilde{\mathbf{x}}$ and A is nonsingular, we have $\mathbf{x} - \tilde{\mathbf{x}} = A^{-1}\mathbf{r}$. Theorem 7.11 on page 440 implies that

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| \le \|A^{-1}\| \cdot \|\mathbf{r}\|.$$

Moreover, since $\mathbf{b} = A\mathbf{x}$, we have $\|\mathbf{b}\| \le \|A\| \cdot \|\mathbf{x}\|$. So $1/\|\mathbf{x}\| \le \|A\|/\|\mathbf{b}\|$ and

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \frac{\|A\| \cdot \|A^{-1}\|}{\|\mathbf{b}\|} \|\mathbf{r}\|.$$

Condition Numbers

The inequalities in Theorem 7.27 imply that $\|A^{-1}\|$ and $\|A\| \cdot \|A^{-1}\|$ provide an indication of the connection between the residual vector and the accuracy of the approximation. In general, the relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\| / \|\mathbf{x}\|$ is of most interest, and, by Inequality (7.20), this error is bounded by the product of $\|A\| \cdot \|A^{-1}\|$ with the relative residual for this approximation, $\|\mathbf{r}\| / \|\mathbf{b}\|$. Any convenient norm can be used for this approximation; the only requirement is that it be used consistently throughout.

Definition 7.28 The condition number of the nonsingular matrix A relative to a norm $\|\cdot\|$ is

$$K(A) = ||A|| \cdot ||A^{-1}||.$$

With this notation, the inequalities in Theorem 7.27 become

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \le K(A) \frac{\|\mathbf{r}\|}{\|A\|}$$

and

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le K(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

For any nonsingular matrix A and natural norm $\|\cdot\|$,

$$1 = ||I|| = ||A \cdot A^{-1}|| < ||A|| \cdot ||A^{-1}|| = K(A).$$

A matrix A is **well-conditioned** if K(A) is close to 1, and is **ill-conditioned** when K(A) is significantly greater than 1. Conditioning in this context refers to the relative security that a small residual vector implies a correspondingly accurate approximate solution.

Example 2 Determine the condition number for the matrix

$$A = \left[\begin{array}{cc} 1 & 2 \\ 1.0001 & 2 \end{array} \right].$$

Solution We saw in Example 1 that the very poor approximation $(3, -0.0001)^t$ to the exact solution $(1, 1)^t$ had a residual vector with small norm, so we should expect the condition number of A to be large. We have $||A||_{\infty} = \max\{|1| + |2|, |1.001| + |2|\} = 3.0001$, which would not be considered large. However,

$$A^{-1} = \begin{bmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{bmatrix}$$
, so $||A^{-1}||_{\infty} = 20000$,

and for the infinity norm, K(A) = (20000)(3.0001) = 60002. The size of the condition number for this example should certainly keep us from making hasty accuracy decisions based on the residual of an approximation.

The condition number K_{∞} can be computed in Maple by first loading the *LinearAlgebra* package and the matrix. Then the command *ConditionNumber(A)* gives the condition number in the l_{∞} norm. For example, we can obtain the condition number of the matrix A in Example 2 with

$$A := Matrix([[1, 2], [1.0001, 2]]): ConditionNumber(A)$$

60002.00000

Although the condition number of a matrix depends totally on the norms of the matrix and its inverse, the calculation of the inverse is subject to roundoff error and is dependent on the accuracy with which the calculations are performed. If the operations involve arithmetic with t digits of accuracy, the approximate condition number for the matrix A is the norm of the matrix times the norm of the approximation to the inverse of A, which is obtained using t-digit arithmetic. In fact, this condition number also depends on the method used to calculate the inverse of A. In addition, because of the number of calculations needed to compute the inverse, we need to be able to estimate the condition number without directly determining the inverse.

If we assume that the approximate solution to the linear system $A\mathbf{x} = \mathbf{b}$ is being determined using *t*-digit arithmetic and Gaussian elimination, it can be shown (see [FM], pp. 45–47) that the residual vector \mathbf{r} for the approximation $\tilde{\mathbf{x}}$ has

$$\|\mathbf{r}\| \approx 10^{-t} \|A\| \cdot \|\tilde{\mathbf{x}}\|. \tag{7.21}$$

From this approximation, an estimate for the effective condition number in t-digit arithmetic can be obtained without the need to invert the matrix A. In actuality, this approximation assumes that all the arithmetic operations in the Gaussian elimination technique are performed using t-digit arithmetic but that the operations needed to determine the residual are done in double-precision (that is, 2t-digit) arithmetic. This technique does not add significantly to the computational effort and eliminates much of the loss of accuracy involved with the subtraction of the nearly equal numbers that occur in the calculation of the residual.

The approximation for the t-digit condition number K(A) comes from consideration of the linear system

$$Ay = r$$

The solution to this system can be readily approximated because the multipliers for the Gaussian elimination method have already been calculated. So A can be factored in the form P^tLU as described in Section 5 of Chapter 6. In fact $\tilde{\mathbf{y}}$, the approximate solution of $A\mathbf{y} = \mathbf{r}$, satisfies

$$\tilde{\mathbf{y}} \approx A^{-1}\mathbf{r} = A^{-1}(\mathbf{b} - A\tilde{\mathbf{x}}) = A^{-1}\mathbf{b} - A^{-1}A\tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}};$$
 (7.22)

and

$$\mathbf{x} \approx \tilde{\mathbf{x}} + \tilde{\mathbf{y}}$$
.

So $\tilde{\mathbf{y}}$ is an estimate of the error produced when $\tilde{\mathbf{x}}$ approximates the solution \mathbf{x} to the original system. Equations (7.21) and (7.22) imply that

$$\|\tilde{\mathbf{y}}\| \approx \|\mathbf{x} - \tilde{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| \le \|A^{-1}\| \cdot \|\mathbf{r}\| \approx \|A^{-1}\| \left(10^{-t}\|A\| \cdot \|\tilde{\mathbf{x}}\|\right) = 10^{-t}\|\tilde{\mathbf{x}}\|K(A).$$

This gives an approximation for the condition number involved with solving the system $A\mathbf{x} = \mathbf{b}$ using Gaussian elimination and the *t*-digit type of arithmetic just described:

$$K(A) \approx \frac{\|\tilde{\mathbf{y}}\|}{\|\tilde{\mathbf{x}}\|} 10^t. \tag{7.23}$$

Illustration The linear system given by

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$$

has the exact solution $\mathbf{x} = (1, 1, 1)^t$.

Using Gaussian elimination and five-digit rounding arithmetic leads successively to the augmented matrices

and

$$\left[\begin{array}{ccccc} 3.3330 & 15920 & -10.333 & 15913 \\ 0 & -10596 & 16.501 & -10580 \\ 0 & 0 & -5.0790 & -4.7000 \end{array} \right].$$

The approximate solution to this system is

$$\tilde{\mathbf{x}} = (1.2001, 0.99991, 0.92538)^t.$$

The residual vector corresponding to $\tilde{\mathbf{x}}$ is computed in double precision to be

$$\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$$

$$= \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix} - \begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} 1.2001 \\ 0.99991 \\ 0.92538 \end{bmatrix}$$

$$= \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix} - \begin{bmatrix} 15913.00518 \\ 28.26987086 \\ 8.611560367 \end{bmatrix} = \begin{bmatrix} -0.00518 \\ 0.27412914 \\ -0.186160367 \end{bmatrix},$$

so

$$\|\mathbf{r}\|_{\infty} = 0.27413.$$

The estimate for the condition number given in the preceding discussion is obtained by first solving the system $A\mathbf{y} = \mathbf{r}$ for $\tilde{\mathbf{y}}$:

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -0.00518 \\ 0.27413 \\ -0.18616 \end{bmatrix}.$$

This implies that $\tilde{\mathbf{y}} = (-0.20008, 8.9987 \times 10^{-5}, 0.074607)^t$. Using the estimate in Eq. (7.23) gives

$$K(A) \approx \frac{\|\tilde{\mathbf{y}}\|_{\infty}}{\|\tilde{\mathbf{x}}\|_{\infty}} 10^5 = \frac{0.20008}{1.2001} 10^5 = 16672.$$
 (7.24)

To determine the *exact* condition number of A, we first must find A^{-1} . Using five-digit rounding arithmetic for the calculations gives the approximation:

$$A^{-1} \approx \left[\begin{array}{cccc} -1.1701 \times 10^{-4} & -1.4983 \times 10^{-1} & 8.5416 \times 10^{-1} \\ 6.2782 \times 10^{-5} & 1.2124 \times 10^{-4} & -3.0662 \times 10^{-4} \\ -8.6631 \times 10^{-5} & 1.3846 \times 10^{-1} & -1.9689 \times 10^{-1} \end{array} \right].$$

Theorem 7.11 on page 440 implies that $||A^{-1}||_{\infty} = 1.0041$ and $||A||_{\infty} = 15934$. As a consequence, the ill-conditioned matrix A has

$$K(A) = (1.0041)(15934) = 15999.$$

The estimate in (7.24) is quite close to K(A) and requires considerably less computational effort.

Since the actual solution $\mathbf{x} = (1, 1, 1)^t$ is known for this system, we can calculate both

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty} = 0.2001$$
 and $\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty}}{\|\mathbf{x}\|_{\infty}} = \frac{0.2001}{1} = 0.2001.$

The error bounds given in Theorem 7.27 for these values are

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty} \le K(A) \frac{\|\mathbf{r}\|_{\infty}}{\|A\|_{\infty}} = \frac{(15999)(0.27413)}{15934} = 0.27525$$

and

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty}}{\|\mathbf{x}\|_{\infty}} \le K(A) \frac{\|\mathbf{r}\|_{\infty}}{\|\mathbf{b}\|_{\infty}} = \frac{(15999)(0.27413)}{15913} = 0.27561.$$

Iterative Refinement

In Eq. (7.22), we used the estimate $\tilde{\mathbf{y}} \approx \mathbf{x} - \tilde{\mathbf{x}}$, where $\tilde{\mathbf{y}}$ is the approximate solution to the system $A\mathbf{y} = \mathbf{r}$. In general, $\tilde{\mathbf{x}} + \tilde{\mathbf{y}}$ is a more accurate approximation to the solution of the linear system $A\mathbf{x} = \mathbf{b}$ than the original approximation $\tilde{\mathbf{x}}$. The method using this assumption is called **iterative refinement**, or *iterative improvement*, and consists of performing iterations on the system whose right-hand side is the residual vector for successive approximations until satisfactory accuracy results.

If the process is applied using t-digit arithmetic and if $K_{\infty}(A) \approx 10^q$, then after k iterations of iterative refinement the solution has approximately the smaller of t and k(t-q) correct digits. If the system is well-conditioned, one or two iterations will indicate that the solution is accurate. There is the possibility of significant improvement on ill-conditioned systems unless the matrix A is so ill-conditioned that $K_{\infty}(A) > 10^t$. In that situation, increased precision should be used for the calculations.



Iterative Refinement

To approximate the solution to the linear system $A\mathbf{x} = \mathbf{b}$:

INPUT the number of equations and unknowns n; the entries a_{ij} , $1 \le i, j \le n$ of the matrix A; the entries b_i , $1 \le i \le n$ of **b**; the maximum number of iterations N; tolerance TOL; number of digits of precision t.

OUTPUT the approximation $\mathbf{x}\mathbf{x} = (xx_i, \dots, xx_n)^t$ or a message that the number of iterations was exceeded, and an approximation COND to $K_{\infty}(A)$.

Step 0 Solve the system $A\mathbf{x} = \mathbf{b}$ for x_1, \dots, x_n by Gaussian elimination saving the multipliers $m_{ji}, j = i+1, i+2, \dots, n, i = 1, 2, \dots, n-1$ and noting row interchanges.

Step 1 Set k = 1.

Step 2 While $(k \le N)$ do Steps 3–9.

Step 3 For
$$i = 1, 2, ..., n$$
 (Calculate \mathbf{r} .)

$$\operatorname{set} r_i = b_i - \sum_{j=1}^n a_{ij} x_j.$$

(Perform the computations in double-precision arithmetic.)

Step 4 Solve the linear system $A\mathbf{y} = \mathbf{r}$ by using Gaussian elimination in the same order as in Step 0.

Step 5 For
$$i = 1, ..., n$$
 set $xx_i = x_i + y_i$.

Step 6 If
$$k = 1$$
 then set $COND = \frac{\|\mathbf{y}\|_{\infty}}{\|\mathbf{x}\mathbf{x}\|_{\infty}} 10^{t}$.

Step 7 If
$$\|\mathbf{x} - \mathbf{x}\mathbf{x}\|_{\infty} < TOL$$
 then OUTPUT $(\mathbf{x}\mathbf{x})$; OUTPUT $(COND)$; (The procedure was successful.) STOP.

Step 8 Set
$$k = k + 1$$
.

Step 9 For
$$i = 1, ..., n$$
 set $x_i = xx_i$.



Step 10 OUTPUT ('Maximum number of iterations exceeded'); OUTPUT (COND); (The procedure was unsuccessful.) STOP.

If t-digit arithmetic is used, a recommended stopping procedure in Step 7 is to iterate until $|y_i^{(k)}| \le 10^{-t}$, for each $i = 1, 2, \dots, n$.

Illustration In our earlier illustration we found the approximation to the linear system

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$$

using five-digit arithmetic and Gaussian elimination, to be

$$\tilde{\mathbf{x}}^{(1)} = (1.2001, 0.99991, 0.92538)^t$$

and the solution to $A\mathbf{y} = \mathbf{r}^{(1)}$ to be

$$\tilde{\mathbf{y}}^{(1)} = (-0.20008, 8.9987 \times 10^{-5}, 0.074607)^t.$$

By Step 5 in this algorithm,

$$\tilde{\mathbf{x}}^{(2)} = \tilde{\mathbf{x}}^{(1)} + \tilde{\mathbf{y}}^{(1)} = (1.0000, 1.0000, 0.99999)^t,$$

and the actual error in this approximation is

$$\|\mathbf{x} - \tilde{\mathbf{x}}^{(2)}\|_{\infty} = 1 \times 10^{-5}.$$

Using the suggested stopping technique for the algorithm, we compute $\mathbf{r}^{(2)} = \mathbf{b} - A\tilde{\mathbf{x}}^{(2)}$ and solve the system $A\mathbf{y}^{(2)} = \mathbf{r}^{(2)}$, which gives

$$\tilde{\mathbf{y}}^{(2)} = (1.5002 \times 10^{-9}, 2.0951 \times 10^{-10}, 1.0000 \times 10^{-5})^t.$$

Since $\|\tilde{\mathbf{y}}^{(2)}\|_{\infty} \leq 10^{-5}$, we conclude that

$$\tilde{\mathbf{x}}^{(3)} = \tilde{\mathbf{x}}^{(2)} + \tilde{\mathbf{v}}^{(2)} = (1.0000, 1.0000, 1.0000)^t$$

is sufficiently accurate, which is certainly correct.

Throughout this section it has been assumed that in the linear system $A\mathbf{x} = \mathbf{b}$, A and \mathbf{b} can be represented exactly. Realistically, the entries a_{ij} and b_j will be altered or perturbed by an amount δa_{ij} and δb_j , causing the linear system

$$(A + \delta A)\mathbf{x} = \mathbf{b} + \delta \mathbf{b}$$

to be solved in place of $A\mathbf{x} = \mathbf{b}$. Normally, if $\|\delta A\|$ and $\|\delta \mathbf{b}\|$ are small (on the order of 10^{-t}), the *t*-digit arithmetic should yield a solution $\tilde{\mathbf{x}}$ for which $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ is correspondingly small. However, in the case of ill-conditioned systems, we have seen that even if A and \mathbf{b} are represented exactly, rounding errors can cause $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ to be large. The following theorem

relates the perturbations of linear systems to the condition number of a matrix. The proof of this result can be found in [Or2], p. 33.

Theorem 7.29 Suppose A is nonsingular and

$$\|\delta A\|<\frac{1}{\|A^{-1}\|}.$$

The solution $\tilde{\mathbf{x}}$ to $(A + \delta A)\tilde{\mathbf{x}} = \mathbf{b} + \delta \mathbf{b}$ approximates the solution \mathbf{x} of $A\mathbf{x} = \mathbf{b}$ with the error estimate

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \frac{K(A)\|A\|}{\|A\| - K(A)\|\delta A\|} \left(\frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta A\|}{\|A\|}\right). \tag{7.25}$$

The estimate in inequality (7.25) states that if the matrix A is well-conditioned (that is, K(A) is not too large), then small changes in A and \mathbf{b} produce correspondingly small changes in the solution \mathbf{x} . If, on the other hand, A is ill-conditioned, then small changes in A and \mathbf{b} may produce large changes in \mathbf{x} .

The theorem is independent of the particular numerical procedure used to solve $A\mathbf{x} = \mathbf{b}$. It can be shown, by means of a backward error analysis (see [Wil1] or [Wil2]), that if Gaussian elimination with pivoting is used to solve $A\mathbf{x} = \mathbf{b}$ in *t*-digit arithmetic, the numerical solution $\tilde{\mathbf{x}}$ is the actual solution of a linear system:

$$(A + \delta A)\tilde{\mathbf{x}} = \mathbf{b}$$
, where $\|\delta A\|_{\infty} \le f(n)10^{1-t} \max_{i,j,k} |a_{ij}^{(k)}|$.

for some function f(n). Wilkinson found that in practice $f(n) \approx n$ and, at worst, $f(n) \le 1.01(n^3 + 3n^2)$.

James Hardy Wilkinson (1919–1986) is best known for his extensive work in numerical methods for solving linear equations and eigenvalues problems. He also developed the technique of backward error analysis.

EXERCISE SET 7.5

1. Compute the condition numbers of the following matrices relative to $\|\cdot\|_{\infty}$.

a.
$$\begin{bmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{4} \end{bmatrix}$$
 b. $\begin{bmatrix} 3.9 & 1.6 \\ 6.8 & 2.9 \end{bmatrix}$ **c.** $\begin{bmatrix} 1 & 2 \\ 1.00001 & 2 \end{bmatrix}$ **d.** $\begin{bmatrix} 1.003 & 58.09 \\ 5.550 & 321.8 \end{bmatrix}$

2. Compute the condition numbers of the following matrices relative to $\|\cdot\|_{\infty}$.

a.
$$\begin{bmatrix} 0.03 & 58.9 \\ 5.31 & -6.10 \end{bmatrix}$$
b. $\begin{bmatrix} 58.9 & 0.03 \\ -6.10 & 5.31 \end{bmatrix}$ **c.** $\begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix}$ **d.** $\begin{bmatrix} 0.04 & 0.01 & -0.01 \\ 0.2 & 0.5 & -0.2 \\ 1 & 2 & 4 \end{bmatrix}$

3. The following linear systems $A\mathbf{x} = \mathbf{b}$ have \mathbf{x} as the actual solution and $\tilde{\mathbf{x}}$ as an approximate solution. Using the results of Exercise 1, compute

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty}$$
 and $K_{\infty}(A) \frac{\|\mathbf{b} - A\tilde{\mathbf{x}}\|_{\infty}}{\|A\|_{\infty}}$.

a.
$$\frac{1}{2}x_1 + \frac{1}{3}x_2 = \frac{1}{63},$$
$$\frac{1}{3}x_1 + \frac{1}{4}x_2 = \frac{1}{168},$$
$$\mathbf{x} = \left(\frac{1}{7}, -\frac{1}{6}\right)^t,$$
$$\tilde{\mathbf{x}} = (0.142, -0.166)^t.$$

$$\mathbf{x} = (1, 1)^t,$$
 $\tilde{\mathbf{x}} = (0.98, 1.1)^t.$

b. $3.9x_1 + 1.6x_2 = 5.5$, $6.8x_1 + 2.9x_2 = 9.7$,

c.
$$x_1 + 2x_2 = 3$$
,
 $1.0001x_1 + 2x_2 = 3.0001$,
 $\mathbf{x} = (1, 1)^t$,
 $\tilde{\mathbf{x}} = (0.96, 1.02)^t$.

d.
$$1.003x_1 + 58.09x_2 = 68.12,$$

 $5.550x_1 + 321.8x_2 = 377.3,$
 $\mathbf{x} = (10, 1)^t,$
 $\tilde{\mathbf{x}} = (-10, 1)^t.$

4. The following linear systems $A\mathbf{x} = \mathbf{b}$ have \mathbf{x} as the actual solution and $\tilde{\mathbf{x}}$ as an approximate solution. Using the results of Exercise 2, compute

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_{\infty}$$
 and $K_{\infty}(A) \frac{\|\mathbf{b} - A\tilde{\mathbf{x}}\|_{\infty}}{\|A\|_{\infty}}$.

a.
$$0.03x_1 + 58.9x_2 = 59.2$$
, $5.31x_1 - 6.10x_2 = 47.0$, $\mathbf{x} = (10, 1)^t$, $\tilde{\mathbf{x}} = (30.0, 0.990)^t$.
c. $x_1 - x_2 - x_3 = 2\pi$,

b.
$$58.9x_1 + 0.03x_2 = 59.2,$$

 $-6.10x_1 + 5.31x_2 = 47.0,$
 $\mathbf{x} = (1, 10)^t,$
 $\tilde{\mathbf{x}} = (1.02, 9.98)^t.$

$$x_{2} - x_{3} = 0,$$

$$-x_{3} = \pi.$$

$$\mathbf{x} = (0, -\pi, -\pi)^{t},$$

$$\tilde{\mathbf{x}} = (-0.1, -3.15, -3.14)^{t}.$$

d.
$$0.04x_1 + 0.01x_2 - 0.01x_3 = 0.06,$$

 $0.2x_1 + 0.5x_2 - 0.2x_3 = 0.3,$
 $x_1 + 2x_2 + 4x_3 = 11,$
 $\mathbf{x} = (1.827586, 0.6551724, 1.965517)^t,$
 $\tilde{\mathbf{x}} = (1.8, 0.64, 1.9)^t.$

5. (i) Use Gaussian elimination and three-digit rounding arithmetic to approximate the solutions to the following linear systems. **(ii)** Then use one iteration of iterative refinement to improve the approximation, and compare the approximations to the actual solutions.

a.
$$0.03x_1 + 58.9x_2 = 59.2$$
, $5.31x_1 - 6.10x_2 = 47.0$. Actual solution $(10, 1)^t$.

b.
$$3.3330x_1 + 15920x_2 + 10.333x_3 = 7953,$$
 $2.2220x_1 + 16.710x_2 + 9.6120x_3 = 0.965,$ $-1.5611x_1 + 5.1792x_2 - 1.6855x_3 = 2.714.$ Actual solution $(1, 0.5, -1)^t$.

c.
$$1.19x_1 + 2.11x_2 - 100x_3 + x_4 = 1.12,$$
 $14.2x_1 - 0.122x_2 + 12.2x_3 - x_4 = 3.44,$ $100x_2 - 99.9x_3 + x_4 = 2.15,$ $15.3x_1 + 0.110x_2 - 13.1x_3 - x_4 = 4.16.$ Actual solution $(0.17682530, 0.01269269, -0.02065405, -1.18260870)^t$.

d.
$$\pi x_1 - ex_2 + \sqrt{2}x_3 - \sqrt{3}x_4 = \sqrt{11},$$

 $\pi^2 x_1 + ex_2 - e^2 x_3 + \frac{3}{7}x_4 = 0,$
 $\sqrt{5}x_1 - \sqrt{6}x_2 + x_3 - \sqrt{2}x_4 = \pi,$
 $\pi^3 x_1 + e^2 x_2 - \sqrt{7}x_3 + \frac{1}{9}x_4 = \sqrt{2}.$
Actual solution $(0.78839378, -3.12541367, 0.16759660, 4.55700252)^t.$

6. Repeat Exercise 5 using four-digit rounding arithmetic.

7. The linear system

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

has solution $(1,1)^t$. Change A slightly to

$$\left[\begin{array}{cc} 1 & 2 \\ 0.9999 & 2 \end{array}\right],$$

and consider the linear system

$$\left[\begin{array}{cc} 1 & 2 \\ 0.9999 & 2 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \left[\begin{array}{c} 3 \\ 3.0001 \end{array}\right].$$

Compute the new solution using five-digit rounding arithmetic, and compare the actual error to the estimate (7.25). Is A ill-conditioned?

8. The linear system $A\mathbf{x} = \mathbf{b}$ given by

$$\left[\begin{array}{cc} 1 & 2 \\ 1.00001 & 2 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \left[\begin{array}{c} 3 \\ 3.00001 \end{array}\right]$$

has solution $(1,1)^t$. Use seven-digit rounding arithmetic to find the solution of the perturbed system

$$\begin{bmatrix} 1 & 2 \\ 1.000011 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3.00001 \\ 3.00003 \end{bmatrix},$$

and compare the actual error to the estimate (7.25). Is A ill-conditioned?

9. Show that if B is singular, then

$$\frac{1}{K(A)} \le \frac{||A - B||}{||A||}.$$

[Hint: There exists a vector with $||\mathbf{x}|| = 1$, such that $B\mathbf{x} = \mathbf{0}$. Derive the estimate using $||A\mathbf{x}|| \ge ||\mathbf{x}|| / ||A^{-1}||$.]

10. Using Exercise 9, estimate the condition numbers for the following matrices:

$$\mathbf{a.} \quad \left[\begin{array}{cc} 1 & 2 \\ 1.0001 & 2 \end{array} \right]$$

b.
$$\begin{bmatrix} 3.9 & 1.6 \\ 6.8 & 2.9 \end{bmatrix}$$

11. The $n \times n$ Hilbert matrix $H^{(n)}$ (see page 512) defined by

$$H_{ij}^{(n)} = \frac{1}{i+j-1}, \quad 1 \le i, j \le n,$$

is an ill-conditioned matrix that arises in solving the normal equations for the coefficients of the least-squares polynomial (see Example 1 of Section 8.2).

a. Show that

$$[H^{(4)}]^{-1} = \begin{bmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{bmatrix},$$

and compute $K_{\infty}(H^{(4)})$.

b. Show that

$$[H^{(5)}]^{-1} = \begin{bmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{bmatrix},$$

and compute $K_{\infty}(H^{(5)})$.

c. Solve the linear system

$$H^{(4)} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

using five-digit rounding arithmetic, and compare the actual error to that estimated in (7.25).

12. Use four-digit rounding arithmetic to compute the inverse H^{-1} of the 3×3 Hilbert matrix H, and then compute $\hat{H} = (H^{-1})^{-1}$. Determine $||H - \hat{H}||_{\infty}$.

7.6 The Conjugate Gradient Method

The conjugate gradient method of Hestenes and Stiefel [HS] was originally developed as a direct method designed to solve an $n \times n$ positive definite linear system. As a direct method it is generally inferior to Gaussian elimination with pivoting. Both methods require n steps to determine a solution, and the steps of the conjugate gradient method are more computationally expensive than those of Gaussian elimination.

However, the conjugate gradient method is useful when employed as an iterative approximation method for solving large sparse systems with nonzero entries occurring in predictable patterns. These problems frequently arise in the solution of boundary-value problems. When the matrix has been preconditioned to make the calculations more effective, good results are obtained in only about \sqrt{n} iterations. Employed in this way, the method is preferred over Gaussian elimination and the previously-discussed iterative methods.

Throughout this section we assume that the matrix *A* is positive definite. We will use the *inner product* notation

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^t \mathbf{y},\tag{7.26}$$

where **x** and **y** are *n*-dimensional vectors. We will also need some additional standard results from linear algebra. A review of this material is found in Section 9.1.

The next result follows easily from the properties of transposes (see Exercise 12).

Magnus Hestenes (1906-1991)

and Eduard Steifel (1907-

1998) published the original

the Institute for Numerical

Analysis on the campus of

UCLA.

paper on the conjugate gradient

method in 1952 while working at

Theorem 7.30 For any vectors \mathbf{x} , \mathbf{y} , and \mathbf{z} and any real number α , we have

(a)
$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$$
;

(b)
$$\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \alpha \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle;$$

(c)
$$\langle \mathbf{x} + \mathbf{z}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{z}, \mathbf{y} \rangle;$$

(d)
$$\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$$
;

(e)
$$\langle \mathbf{x}, \mathbf{x} \rangle = 0$$
 if and only if $\mathbf{x} = \mathbf{0}$.

When A is positive definite, $\langle \mathbf{x}, A\mathbf{x} \rangle = \mathbf{x}^t A\mathbf{x} > 0$ unless $\mathbf{x} = \mathbf{0}$. Also, since A is symmetric, we have $\mathbf{x}^t A\mathbf{y} = \mathbf{x}^t A^t \mathbf{y} = (A\mathbf{x})^t \mathbf{y}$, so in addition to the results in Theorem 7.30, we have for each \mathbf{x} and \mathbf{y} ,

$$\langle \mathbf{x}, A\mathbf{y} \rangle = (A\mathbf{x})^t \mathbf{y} = \mathbf{x}^t A^t \mathbf{y} = \mathbf{x}^t A \mathbf{y} = \langle A\mathbf{x}, \mathbf{y} \rangle. \tag{7.27}$$

The following result is a basic tool in the development of the conjugate gradient method.

Theorem 7.31 The vector \mathbf{x}^* is a solution to the positive definite linear system $A\mathbf{x} = \mathbf{b}$ if and only if \mathbf{x}^* produces the minimal value of

$$g(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle.$$

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it

Proof Let x and $\mathbf{v} \neq \mathbf{0}$ be fixed vectors and t a real number variable. We have

$$g(\mathbf{x} + t\mathbf{v}) = \langle \mathbf{x} + t\mathbf{v}, A\mathbf{x} + tA\mathbf{v} \rangle - 2\langle \mathbf{x} + t\mathbf{v}, \mathbf{b} \rangle$$

$$= \langle \mathbf{x}, A\mathbf{x} \rangle + t\langle \mathbf{v}, A\mathbf{x} \rangle + t\langle \mathbf{x}, A\mathbf{v} \rangle + t^2 \langle \mathbf{v}, A\mathbf{v} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle - 2t\langle \mathbf{v}, \mathbf{b} \rangle$$

$$= \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle + 2t\langle \mathbf{v}, A\mathbf{x} \rangle - 2t\langle \mathbf{v}, \mathbf{b} \rangle + t^2 \langle \mathbf{v}, A\mathbf{v} \rangle,$$

so

$$g(\mathbf{x} + t\mathbf{v}) = g(\mathbf{x}) - 2t\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle + t^2 \langle \mathbf{v}, A\mathbf{v} \rangle. \tag{7.28}$$

With \mathbf{x} and \mathbf{v} fixed we can define the quadratic function h in t by

$$h(t) = g(\mathbf{x} + t\mathbf{v}).$$

Then h assumes a minimal value when h'(t) = 0, because its t^2 coefficient, $\langle \mathbf{v}, A\mathbf{v} \rangle$, is positive. Because

$$h'(t) = -2\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle + 2t\langle \mathbf{v}, A\mathbf{v} \rangle,$$

the minimum occurs when

$$\hat{t} = \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle},$$

and, from Equation (7.28),

$$h(\hat{t}) = g(\mathbf{x} + \hat{t}\mathbf{v})$$

$$= g(\mathbf{x}) - 2\hat{t}\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle + \hat{t}^2\langle \mathbf{v}, A\mathbf{v}\rangle$$

$$= g(\mathbf{x}) - 2\frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle} \langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle + \left(\frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle}\right)^2 \langle \mathbf{v}, A\mathbf{v}\rangle$$

$$= g(\mathbf{x}) - \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle^2}{\langle \mathbf{v}, A\mathbf{v}\rangle}.$$

So for any vector $\mathbf{v} \neq \mathbf{0}$, we have $g(\mathbf{x} + \hat{t}\mathbf{v}) < g(\mathbf{x})$ unless $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle = 0$, in which case $g(\mathbf{x}) = g(\mathbf{x} + \hat{t}\mathbf{v})$. This is the basic result we need to prove Theorem 7.31.

Suppose \mathbf{x}^* satisfies $A\mathbf{x}^* = \mathbf{b}$. Then $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}^* \rangle = 0$ for any vector \mathbf{v} , and $g(\mathbf{x})$ cannot be made any smaller than $g(\mathbf{x}^*)$. Thus, \mathbf{x}^* minimizes g.

On the other hand, suppose that \mathbf{x}^* is a vector that minimizes g. Then for any vector \mathbf{v} , we have $g(\mathbf{x}^* + \hat{t}\mathbf{v}) \ge g(\mathbf{x}^*)$. Thus, $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}^* \rangle = 0$. This implies that $\mathbf{b} - A\mathbf{x}^* = \mathbf{0}$ and, consequently, that $A\mathbf{x}^* = \mathbf{b}$.

To begin the conjugate gradient method, we choose \mathbf{x} , an approximate solution to $A\mathbf{x}^* = \mathbf{b}$, and $\mathbf{v} \neq \mathbf{0}$, which gives a *search direction* in which to move away from \mathbf{x} to improve the approximation. Let $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ be the residual vector associated with \mathbf{x} and

$$t = \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle} = \frac{\langle \mathbf{v}, \mathbf{r} \rangle}{\langle \mathbf{v}, A\mathbf{v} \rangle}.$$

If $\mathbf{r} \neq \mathbf{0}$ and if \mathbf{v} and \mathbf{r} are not orthogonal, then $\mathbf{x} + t\mathbf{v}$ gives a smaller value for g than $g(\mathbf{x})$ and is presumably closer to \mathbf{x}^* than is \mathbf{x} . This suggests the following method.

Let $\mathbf{x}^{(0)}$ be an initial approximation to \mathbf{x}^* , and let $\mathbf{v}^{(1)} \neq \mathbf{0}$ be an initial search direction. For $k = 1, 2, 3, \ldots$, we compute

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle},$$
$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$$

and choose a new search direction $\mathbf{v}^{(k+1)}$. The object is to make this selection so that the sequence of approximations $\{\mathbf{x}^{(k)}\}$ converges rapidly to \mathbf{x}^* .

To choose the search directions, we view g as a function of the components of $\mathbf{x} = (x_1, x_2, \dots, x_n)^t$. Thus,

$$g(x_1, x_2, \dots, x_n) = \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j - 2 \sum_{i=1}^n x_i b_i.$$

Taking partial derivatives with respect to the component variables x_k gives

$$\frac{\partial g}{\partial x_k}(\mathbf{x}) = 2\sum_{i=1}^n a_{ki}x_i - 2b_k,$$

which is the kth component of the vector $2(A\mathbf{x} - \mathbf{b})$. Therefore, the gradient of g is

$$\nabla g(\mathbf{x}) = \left(\frac{\partial g}{\partial x_1}(\mathbf{x}), \frac{\partial g}{\partial x_2}(\mathbf{x}), \dots, \frac{\partial g}{\partial x_n}(\mathbf{x})\right)^t = 2(A\mathbf{x} - \mathbf{b}) = -2\mathbf{r},$$

where the vector \mathbf{r} is the residual vector for \mathbf{x} .

From multivariable calculus, we know that the direction of greatest decrease in the value of $g(\mathbf{x})$ is the direction given by $-\nabla g(\mathbf{x})$; that is, in the direction of the residual \mathbf{r} . The method that chooses

$$\mathbf{v}^{(k+1)} = \mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$$

is called the *method of steepest descent*. Although we will see in Section 10.4 that this method has merit for nonlinear systems and optimization problems, it is not used for linear systems because of slow convergence.

An alternative approach uses a set of nonzero direction vectors $\{\mathbf{v}^{(1)},\dots,\mathbf{v}^{(n)}\}$ that satisfy

$$\langle \mathbf{v}^{(i)}, A\mathbf{v}^{(j)} \rangle = 0$$
, if $i \neq j$.

This is called an *A*-orthogonality condition, and the set of vectors $\{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$ is said to be *A*-orthogonal. It is not difficult to show that a set of *A*-orthogonal vectors associated with the positive definite matrix *A* is linearly independent. (See Exercise 13(a).) This set of search directions gives

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{v}^{(k)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$

and $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$.

The following theorem shows that this choice of search directions gives convergence in at most *n*-steps, so as a direct method it produces the exact solution, assuming that the arithmetic is exact.

Theorem 7.32 Let $\{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$ be an *A*-orthogonal set of nonzero vectors associated with the positive definite matrix *A*, and let $\mathbf{x}^{(0)}$ be arbitrary. Define

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$
 and $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$,

for k = 1, 2, ..., n. Then, assuming exact arithmetic, $A\mathbf{x}^{(n)} = \mathbf{b}$.

Proof Since, for each k = 1, 2, ..., n, $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$, we have

$$A\mathbf{x}^{(n)} = A\mathbf{x}^{(n-1)} + t_n A\mathbf{v}^{(n)}$$

$$= (A\mathbf{x}^{(n-2)} + t_{n-1}A\mathbf{v}^{(n-1)}) + t_n A\mathbf{v}^{(n)}$$

$$\vdots$$

$$= A\mathbf{x}^{(0)} + t_1 A\mathbf{v}^{(1)} + t_2 A\mathbf{v}^{(2)} + \dots + t_n A\mathbf{v}^{(n)}.$$

Subtracting **b** from this result yields

$$A\mathbf{x}^{(n)} - \mathbf{b} = A\mathbf{x}^{(0)} - \mathbf{b} + t_1 A\mathbf{v}^{(1)} + t_2 A\mathbf{v}^{(2)} + \dots + t_n A\mathbf{v}^{(n)}.$$

We now take the inner product of both sides with the vector $\mathbf{v}^{(k)}$ and use the properties of inner products and the fact that A is symmetric to obtain

$$\langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle = \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_1 \langle A\mathbf{v}^{(1)}, \mathbf{v}^{(k)} \rangle + \dots + t_n \langle A\mathbf{v}^{(n)}, \mathbf{v}^{(k)} \rangle$$
$$= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_1 \langle \mathbf{v}^{(1)}, A\mathbf{v}^{(k)} \rangle + \dots + t_n \langle \mathbf{v}^{(n)}, A\mathbf{v}^{(k)} \rangle.$$

The A-orthogonality property gives, for each k,

$$\langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle = \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle. \tag{7.29}$$

However $t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle$ so

$$t_{k}\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)}\rangle = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} + A\mathbf{x}^{(0)} - A\mathbf{x}^{(1)} + \dots - A\mathbf{x}^{(k-2)} + A\mathbf{x}^{(k-2)} - A\mathbf{x}^{(k-1)}\rangle$$

$$= \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)}\rangle + \langle \mathbf{v}^{(k)}, A\mathbf{x}^{(0)} - A\mathbf{x}^{(1)}\rangle + \dots + \langle \mathbf{v}^{(k)}, A\mathbf{x}^{(k-2)} - A\mathbf{x}^{(k-1)}\rangle.$$

But for any i,

$$\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + t_i \mathbf{v}^{(i)}$$
 and $A\mathbf{x}^{(i)} = A\mathbf{x}^{(i-1)} + t_i A\mathbf{v}^{(i)}$,

so

$$A\mathbf{x}^{(i-1)} - A\mathbf{x}^{(i)} = -tA\mathbf{y}^{(i)}$$

Thus

$$t_k\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)}\rangle = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)}\rangle - t_1\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(1)}\rangle - \cdots - t_{k-1}\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k-1)}\rangle.$$

Because of the A-orthogonality, $\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(i)} \rangle = 0$, for $i \neq k$, so

$$\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)}\rangle t_k = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)}\rangle.$$

From Eq.(7.29),

$$\begin{split} \langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} \rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + \langle \mathbf{b} - A\mathbf{x}^{(0)}, \mathbf{v}^{(k)} \rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle - \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle = 0. \end{split}$$

Hence the vector $A\mathbf{x}^{(n)} - \mathbf{b}$ is orthogonal to the *A*-orthogonal set of vectors $\{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$. From this, it follows (see Exercise 13(b)) that $A\mathbf{x}^{(n)} - \mathbf{b} = \mathbf{0}$, so $A\mathbf{x}^{(n)} = \mathbf{b}$.

Example 1 The linear system

$$4x_1 + 3x_2 = 24,$$

 $3x_1 + 4x_2 - x_3 = 30,$
 $-x_2 + 4x_3 = -24$

has the exact solution $\mathbf{x}^* = (3, 4, -5)^t$. Show that the procedure described in Theorem 7.32 with $\mathbf{x}^{(0)} = (0, 0, 0)^t$ produces this exact solution after three iterations.

Solution We established in Example 2 of Section 7.4 that the coefficient matrix

$$A = \left[\begin{array}{rrr} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{array} \right].$$

of this system is positive definite. Let $\mathbf{v}^{(1)} = (1,0,0)^t$, $\mathbf{v}^{(2)} = (-3/4,1,0)^t$, and $\mathbf{v}^{(3)} = (-3/7,4/7,1)^t$. Then

$$\langle \mathbf{v}^{(1)}, A\mathbf{v}^{(2)} \rangle = \mathbf{v}^{(1)t} A\mathbf{v}^{(2)} = (1, 0, 0) \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} -\frac{3}{4} \\ 1 \\ 0 \end{bmatrix} = 0,$$

$$\langle \mathbf{v}^{(1)}, A\mathbf{v}^{(3)} \rangle = (1, 0, 0) \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} -\frac{3}{7} \\ \frac{4}{7} \\ 1 \end{bmatrix} = 0,$$

and

$$\langle \mathbf{v}^{(2)}, A\mathbf{v}^{(3)} \rangle = \begin{pmatrix} -\frac{3}{4}, 1, 0 \end{pmatrix} \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} -\frac{3}{7} \\ \frac{4}{7} \\ 1 \end{bmatrix} = 0.$$

Hence $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\}\$ is an A-orthogonal set.

Applying the iterations described in Theorem 7.22 for A with $\mathbf{x}^{(0)} = (0,0,0)^t$ and $\mathbf{b} = (24,30,-24)^t$ gives

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \mathbf{b} = (24, 30, -24)^t$$

so

$$\langle \mathbf{v}^{(1)}, \mathbf{r}^{(0)} \rangle = \mathbf{v}^{(1)t} \mathbf{r}^{(0)} = 24, \quad \langle \mathbf{v}^{(1)}, A \mathbf{v}^{(1)} \rangle = 4, \quad \text{and} \quad t_0 = \frac{24}{4} = 6.$$

Hence

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + t_0 \mathbf{v}^{(1)} = (0, 0, 0)^t + 6(1, 0, 0)^t = (6, 0, 0)^t.$$

Continuing, we have

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)} = (0, 12, -24)^{t}; \quad t_{1} = \frac{\langle \mathbf{v}^{(2)}, \mathbf{r}^{(1)} \rangle}{\langle \mathbf{v}^{(2)}, A\mathbf{v}^{(2)} \rangle} = \frac{12}{7/4} = \frac{48}{7};$$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + t_{1}\mathbf{v}^{(2)} = (6, 0, 0)^{t} + \frac{48}{7} \left(-\frac{3}{4}, 1, 0 \right)^{t} = \left(\frac{6}{7}, \frac{48}{7}, 0 \right)^{t};$$

$$\mathbf{r}^{(2)} = \mathbf{b} - A\mathbf{x}^{(2)} = \left(0, 0, -\frac{120}{7} \right); \quad t_{2} = \frac{\langle \mathbf{v}^{(3)}, \mathbf{r}^{(2)} \rangle}{\langle \mathbf{v}^{(3)}, A\mathbf{v}^{(3)} \rangle} = \frac{-120/7}{24/7} = -5;$$

and

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} + t_2 \mathbf{v}^{(3)} = \left(\frac{6}{7}, \frac{48}{7}, 0\right)^t + (-5)\left(-\frac{3}{7}, \frac{4}{7}, 1\right)^t = (3, 4, -5)^t.$$

Since we applied the technique n = 3 times, this must be the actual solution.

Before discussing how to determine the A-orthogonal set, we will continue the development. The use of an A-orthogonal set $\{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$ of direction vectors gives what is called a *conjugate direction* method. The following theorem shows the orthogonality of the residual vectors $\mathbf{r}^{(k)}$ and the direction vectors $\mathbf{v}^{(j)}$. A proof of this result using mathematical induction is considered in Exercise 14.

Theorem 7.33 The residual vectors $\mathbf{r}^{(k)}$, where k = 1, 2, ..., n, for a conjugate direction method, satisfy the equations

$$\langle \mathbf{r}^{(k)}, \mathbf{v}^{(j)} \rangle = 0$$
, for each $j = 1, 2, \dots, k$.

The conjugate gradient method of Hestenes and Stiefel chooses the search directions $\{\mathbf{v}^{(k)}\}$ during the iterative process so that the residual vectors $\{\mathbf{r}^{(k)}\}$ are mutually orthogonal. To construct the direction vectors $\{\mathbf{v}^{(1)},\mathbf{v}^{(2)},\ldots\}$ and the approximations $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots\}$, we start with an initial approximation $\mathbf{x}^{(0)}$ and use the steepest descent direction $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ as the first search direction $\mathbf{v}^{(1)}$.

Assume that the conjugate directions $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(k-1)}$ and the approximations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k-1)}$ have been computed with

$$\mathbf{x}^{(k-1)} = \mathbf{x}^{(k-2)} + t_{k-1} \mathbf{v}^{(k-1)},$$

where

$$\langle \mathbf{v}^{(i)}, A\mathbf{v}^{(j)} \rangle = 0$$
 and $\langle \mathbf{r}^{(i)}, \mathbf{r}^{(j)} \rangle = 0$, for $i \neq j$.

If $\mathbf{x}^{(k-1)}$ is the solution to $A\mathbf{x} = \mathbf{b}$, we are done. Otherwise, $\mathbf{r}^{(k-1)} = \mathbf{b} - A\mathbf{x}^{(k-1)} \neq \mathbf{0}$ and Theorem 7.33 implies that $\langle \mathbf{r}^{(k-1)}, \mathbf{v}^{(i)} \rangle = 0$, for each $i = 1, 2, \dots, k-1$.

We use $\mathbf{r}^{(k-1)}$ to generate $\mathbf{v}^{(k)}$ by setting

$$\mathbf{v}^{(k)} = \mathbf{r}^{(k-1)} + s_{k-1} \mathbf{v}^{(k-1)}.$$

We want to choose s_{k-1} so that

$$\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = 0.$$

Since

$$A\mathbf{v}^{(k)} = A\mathbf{r}^{(k-1)} + s_{k-1}A\mathbf{v}^{(k-1)}$$

and

$$\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = \langle \mathbf{v}^{(k-1)}, A\mathbf{r}^{(k-1)} \rangle + s_{k-1} \langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k-1)} \rangle,$$

we will have $\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = 0$ when

$$s_{k-1} = -\frac{\langle \mathbf{v}^{(k-1)}, A\mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k-1)} \rangle}.$$

It can also be shown that with this choice of s_{k-1} we have $\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(i)} \rangle = 0$, for each $i = 1, 2, \dots, k-2$ (see [Lu], p. 245). Thus $\{\mathbf{v}^{(1)}, \dots \mathbf{v}^{(k)}\}$ is an A-orthogonal set.

Having chosen $\mathbf{v}^{(k)}$, we compute

$$t_{k} = \frac{\langle \mathbf{v}^{(k)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{r}^{(k-1)} + s_{k-1}\mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$
$$= \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} + s_{k-1} \frac{\langle \mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}.$$

By Theorem 7.33, $\langle \mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle = 0$, so

$$t_k = \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}.$$
 (7.30)

Thus

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}.$$

To compute $\mathbf{r}^{(k)}$, we multiply by A and subtract **b** to obtain

$$A\mathbf{x}^{(k)} - \mathbf{b} = A\mathbf{x}^{(k-1)} - \mathbf{b} + t_k A\mathbf{v}^{(k)}$$

or

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - t_k A \mathbf{v}^{(k)}.$$

This gives

$$\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle = \langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k)} \rangle - t_k \langle A\mathbf{v}^{(k)}, \mathbf{r}^{(k)} \rangle = -t_k \langle \mathbf{r}^{(k)}, A\mathbf{v}^{(k)} \rangle.$$

Further, from Eq. (7.30),

$$\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle = t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle,$$

so

$$s_k = -\frac{\langle \mathbf{v}^{(k)}, A\mathbf{r}^{(k)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = -\frac{\langle \mathbf{r}^{(k)}, A\mathbf{v}^{(k)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{(1/t_k)\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{(1/t_k)\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle} = \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}.$$

In summary, we have

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$$
: $\mathbf{v}^{(1)} = \mathbf{r}^{(0)}$:

and, for k = 1, 2, ..., n,

$$t_k = \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A \mathbf{v}^{(k)} \rangle}, \quad \mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}, \quad \mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - t_k A \mathbf{v}^{(k)}, \quad s_k = \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle},$$

and

$$\mathbf{v}^{(k+1)} = \mathbf{r}^{(k)} + s_k \mathbf{v}^{(k)}. \tag{7.31}$$

Preconditioning

Rather than presenting an algorithm for the conjugate gradient method using these formulas, we extend the method to include *preconditioning*. If the matrix A is ill-conditioned, the conjugate gradient method is highly susceptible to rounding errors. So, although the exact answer should be obtained in n steps, this is not usually the case. As a direct method the conjugate gradient method is not as good as Gaussian elimination with pivoting. The main use of the conjugate gradient method is as an iterative method applied to a better-conditioned system. In this case an acceptable approximate solution is often obtained in about \sqrt{n} steps.

Preconditioning replaces a given system with one having the same solutions but with better convergence characteristics. When preconditioning is used, the conjugate gradient method is not applied directly to the matrix A but to another positive definite matrix that a smaller condition number. We need to do this in such a way that once the solution to this new system is found it will be easy to obtain the solution to the original system. The expectation is that this will reduce the rounding error when the method is applied. To maintain the positive definiteness of the resulting matrix, we need to multiply on each side by a nonsingular matrix. We will denote this matrix by C^{-1} , and consider

$$\tilde{A} = C^{-1}A(C^{-1})^t,$$

with the hope that \tilde{A} has a lower condition number than A. To simplify the notation, we use the matrix notation $C^{-t} \equiv (C^{-1})^t$. Later in the section we will see a reasonable way to select C, but first we will consider the conjugate applied to \tilde{A} .

Consider the linear system

$$\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$$
.

where $\tilde{\mathbf{x}} = C^t \mathbf{x}$ and $\tilde{\mathbf{b}} = C^{-1} \mathbf{b}$. Then

$$\tilde{A}\tilde{\mathbf{x}} = (C^{-1}AC^{-t})(C^t\mathbf{x}) = C^{-1}A\mathbf{x}.$$

Thus, we could solve $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ for $\tilde{\mathbf{x}}$ and then obtain \mathbf{x} by multiplying by C^{-t} . However, instead of rewriting equations (7.31) using $\tilde{\mathbf{r}}^{(k)}$, $\tilde{\mathbf{v}}^{(k)}$, \tilde{t}_k , $\tilde{\mathbf{x}}^{(k)}$, and \tilde{s}_k , we incorporate the preconditioning implicitly.

Since

$$\tilde{\mathbf{x}}^{(k)} = C^t \mathbf{x}^{(k)}.$$

we have

$$\tilde{\mathbf{r}}^{(k)} = \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}^{(k)} = C^{-1}\mathbf{b} - (C^{-1}AC^{-t})C^{t}\mathbf{x}^{(k)} = C^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}) = C^{-1}\mathbf{r}^{(k)}.$$

Let $\tilde{\mathbf{v}}^{(k)} = C^t \mathbf{v}^{(k)}$ and $\mathbf{w}^{(k)} = C^{-1} \mathbf{r}^{(k)}$. Then

$$\tilde{s}_k = \frac{\langle \tilde{\mathbf{r}}^{(k)}, \tilde{\mathbf{r}}^{(k)} \rangle}{\langle \tilde{\mathbf{r}}^{(k-1)}, \tilde{\mathbf{r}}^{(k-1)} \rangle} = \frac{\langle C^{-1} \mathbf{r}^{(k)}, C^{-1} \mathbf{r}^{(k)} \rangle}{\langle C^{-1} \mathbf{r}^{(k-1)}, C^{-1} \mathbf{r}^{(k-1)} \rangle},$$

so

$$\tilde{s}_k = \frac{\langle \mathbf{w}^{(k)}, \mathbf{w}^{(k)} \rangle}{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}.$$
(7.32)

Thus

$$\tilde{t}_k = \frac{\langle \tilde{\mathbf{r}}^{(k-1)}, \tilde{\mathbf{r}}^{(k-1)} \rangle}{\langle \tilde{\mathbf{v}}^{(k)}, \tilde{A}\tilde{\mathbf{v}}^{(k)} \rangle} = \frac{\langle C^{-1}\mathbf{r}^{(k-1)}, C^{-1}\mathbf{r}^{(k-1)} \rangle}{\langle C^t\mathbf{v}^{(k)}, C^{-1}AC^{-t}C^t\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}{\langle C^t\mathbf{v}^{(k)}, C^{-1}A\mathbf{v}^{(k)} \rangle}$$

and, since

$$\langle C^t \mathbf{v}^{(k)}, C^{-1} A \mathbf{v}^{(k)} \rangle = [C^t \mathbf{v}^{(k)}]^t C^{-1} A \mathbf{v}^{(k)}$$
$$= [\mathbf{v}^{(k)}]^t C C^{-1} A \mathbf{v}^{(k)} = [\mathbf{v}^{(k)}]^t A \mathbf{v}^{(k)} = \langle \mathbf{v}^{(k)}, A \mathbf{v}^{(k)} \rangle,$$

we have

$$\tilde{t}_k = \frac{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A \mathbf{v}^{(k)} \rangle}.$$
(7.33)

Further,

$$\tilde{\mathbf{x}}^{(k)} = \tilde{\mathbf{x}}^{(k-1)} + \tilde{t}_k \tilde{\mathbf{v}}^{(k)}, \quad \text{so} \quad C^t \mathbf{x}^{(k)} = C^t \mathbf{x}^{(k-1)} + \tilde{t}_k C^t \mathbf{v}^{(k)}$$

and

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \tilde{t}_k \mathbf{v}^{(k)}. \tag{7.34}$$

Continuing,

$$\tilde{\mathbf{r}}^{(k)} = \tilde{\mathbf{r}}^{(k-1)} - \tilde{t}_k \tilde{A} \tilde{\mathbf{v}}^{(k)},$$

so

$$C^{-1}\mathbf{r}^{(k)} = C^{-1}\mathbf{r}^{(k-1)} - \tilde{t}_k C^{-1}AC^{-t}\tilde{v}^{(k)}, \qquad \mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \tilde{t}_k AC^{-t}C^t\mathbf{v}^{(k)},$$

and

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \tilde{t}_k A \mathbf{v}^{(k)}. \tag{7.35}$$

Finally,

$$\tilde{\mathbf{v}}^{(k+1)} = \tilde{\mathbf{r}}^{(k)} + \tilde{s}_k \tilde{\mathbf{v}}^{(k)}$$
 and $C^t \mathbf{v}^{(k+1)} = C^{-1} \mathbf{r}^{(k)} + \tilde{s}_k C^t \mathbf{v}^{(k)}$,

so

$$\mathbf{v}^{(k+1)} = C^{-t}C^{-1}\mathbf{r}^{(k)} + \tilde{s}_k\mathbf{v}^{(k)} = C^{-t}\mathbf{w}^{(k)} + \tilde{s}_k\mathbf{v}^{(k)}.$$
 (7.36)

The preconditioned conjugate gradient method is based on using equations (7.32)–(7.36) in the order (7.33), (7.34), (7.35), (7.32), and (7.36). Algorithm 7.5 implements this procedure.



Preconditioned Conjugate Gradient Method

To solve $A\mathbf{x} = \mathbf{b}$ given the preconditioning matrix C^{-1} and the initial approximation $\mathbf{x}^{(0)}$:

INPUT the number of equations and unknowns n; the entries a_{ij} , $1 \le i, j \le n$ of the matrix A; the entries b_j , $1 \le j \le n$ of the vector \mathbf{b} ; the entries γ_{ij} , $1 \le i, j \le n$ of the preconditioning matrix C^{-1} , the entries x_i , $1 \le i \le n$ of the initial approximation $\mathbf{x} = \mathbf{x}^{(0)}$, the maximum number of iterations N; tolerance TOL.

OUTPUT the approximate solution $x_1, \ldots x_n$ and the residual $r_1, \ldots r_n$ or a message that the number of iterations was exceeded.



```
Step 1 Set \mathbf{r} = \mathbf{b} - A\mathbf{x}; (Compute \mathbf{r}^{(0)}.)
                      \mathbf{w} = C^{-1}\mathbf{r}; (Note: \mathbf{w} = \mathbf{w}^{(0)})
                     \mathbf{v} = C^{-t}\mathbf{w}; (Note: \mathbf{v} = \mathbf{v}^{(1)})

\alpha = \sum_{j=1}^{n} w_j^2.
Step 2 Set k = 1.
Step 3 While (k \le N) do Steps 4–7.
          Step 4 If \|\mathbf{v}\| < TOL, then
                              OUTPUT ('Solution vector'; x_1, \ldots, x_n);
                              OUTPUT ('with residual'; r_1, \ldots, r_n);
                              (The procedure was successful.)
                              STOP
          Step 5 Set \mathbf{u} = A\mathbf{v}; (Note: \mathbf{u} = A\mathbf{v}^{(k)})
                                t = \frac{\alpha}{\sum_{j=1}^{n} v_j u_j}; (Note: t = t_k)
                                 \mathbf{x} = \mathbf{x} + t\mathbf{v}; (Note: \mathbf{x} = \mathbf{x}^{(k)})
                                \mathbf{r} = \mathbf{r} - t\mathbf{u}; (Note: \mathbf{r} = \mathbf{r}^{(k)})
\mathbf{w} = C^{-1}\mathbf{r}; (Note: \mathbf{w} = \mathbf{w}^{(k)})
\beta = \sum_{j=1}^{n} w_j^2. (Note: \beta = \langle \mathbf{w}^{(k)}, \mathbf{w}^{(k)} \rangle)
          Step 6 If |\beta| < TOL then
                              if \|\mathbf{r}\| < TOL then
                                 OUTPUT('Solution vector'; x_1, \ldots, x_n);
                                 OUTPUT('with residual'; r_1, \ldots, r_n);
                                 (The procedure was successful.)
                                 STOP
          Step 7 Set s = \beta/\alpha; (s = s_k)
                                 \mathbf{v} = C^{-t}\mathbf{w} + s\mathbf{v}; (Note: \mathbf{v} = \mathbf{v}^{(k+1)})
                                 \alpha = \beta; (Update \alpha.)
                                 k = k + 1.
Step 8 If (k > n) then
                OUTPUT ('The maximum number of iterations was exceeded.');
                (The procedure was unsuccessful.)
                STOP.
```

The next example illustrates the calculations for an elementary problem.

Example 2 The linear system $A\mathbf{x} = \mathbf{b}$ given by

$$4x_1 + 3x_2 = 24,$$

$$3x_1 + 4x_2 - x_3 = 30,$$

$$-x_2 + 4x_3 = -24$$

has solution $(3, 4, -5)^t$. Use the conjugate gradient method with $\mathbf{x}^{(0)} = (0, 0, 0)^t$ and no preconditioning, that is, with $C = C^{-1} = I$, to approximate the solution.

Solution The solution was considered in Example 2 of Section 7.4 where the SOR method were used with a nearly optimal value of $\omega = 1.25$.

For the conjugate gradient method we start with

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \mathbf{b} = (24, 30, -24)^{t};$$

$$\mathbf{w} = C^{-1}\mathbf{r}^{(0)} = (24, 30, -24)^{t};$$

$$\mathbf{v}^{(1)} = C^{-t}\mathbf{w} = (24, 30, -24)^{t};$$

$$\alpha = \langle \mathbf{w}, \mathbf{w} \rangle = 2052.$$

We start the first iteration with k = 1. Then

$$\mathbf{u} = A\mathbf{v}^{(1)} = (186.0, 216.0, -126.0)^{t};$$

$$t_{1} = \frac{\alpha}{\langle \mathbf{v}^{(1)}, \mathbf{u} \rangle} = 0.1469072165;$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + t_{1}\mathbf{v}^{(1)} = (3.525773196, 4.407216495, -3.525773196)^{t};$$

$$\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - t_{1}\mathbf{u} = (-3.32474227, -1.73195876, -5.48969072)^{t};$$

$$\mathbf{w} = C^{-1}\mathbf{r}^{(1)} = \mathbf{r}^{(1)};$$

$$\beta = \langle \mathbf{w}, \mathbf{w} \rangle = 44.19029651;$$

$$s_{1} = \frac{\beta}{\alpha} = 0.02153523222;$$

$$\mathbf{v}^{(2)} = C^{-t}\mathbf{w} + s_{1}\mathbf{v}^{(1)} = (-2.807896697, -1.085901793, -6.006536293)^{t}.$$

Set

$$\alpha = \beta = 44.19029651.$$

For the second iteration we have

$$\mathbf{u} = A\mathbf{v}^{(2)} = (-14.48929217, -6.760760967, -22.94024338)^{t};$$

$$t_{2} = 0.2378157558;$$

$$\mathbf{x}^{(2)} = (2.858011121, 4.148971939, -4.954222164)^{t};$$

$$\mathbf{r}^{(2)} = (0.121039698, -0.124143281, -0.034139402)^{t};$$

$$\mathbf{w} = C^{-1}\mathbf{r}^{(2)} = \mathbf{r}^{(2)};$$

$$\beta = 0.03122766148;$$

$$s_{2} = 0.0007066633163;$$

$$\mathbf{v}^{(3)} = (0.1190554504, -0.1249106480, -0.03838400086)^{t}.$$

Set $\alpha = \beta = 0.03122766148$.

The third iteration gives

$$\mathbf{u} = A\mathbf{v}^{(3)} = (0.1014898976, -0.1040922099, -0.0286253554)^{t};$$

$$t_{3} = 1.192628008;$$

$$\mathbf{x}^{(3)} = (2.9999999998, 4.000000002, -4.9999999998)^{t};$$

$$\mathbf{r}^{(3)} = (0.36 \times 10^{-8}, 0.39 \times 10^{-8}, -0.141 \times 10^{-8})^{t}.$$

Since $\mathbf{x}^{(3)}$ is nearly the exact solution, rounding error did not significantly effect the result. In Example 2 of Section 7.4, the SOR method with $\omega = 1.25$ required 14 iterations

for an accuracy of 10^{-7} . It should be noted, however, that in this example, we are really comparing a direct method to iterative methods.

The next example illustrates the effect of preconditioning on a poorly conditioned matrix. In this example, we use $D^{-1/2}$ to represent the diagonal matrix whose entries are the reciprocals of the square roots of the diagonal entries of the coefficient matrix A. This is used as the preconditioner. Because the matrix A is positive definite we expect the eigenvalues of $D^{-1/2}AD^{-1/2}$ to be close to 1, with the result that the condition number of this matrix will be small relative to the condition number of A.

Example 3 Use Maple to find the eigenvalues and condition number of the matrix

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix}$$

and compare these with the eigenvalues and condition number of the preconditioned matrix $D^{-1/2}AD^{-t/2}$.

Solution We first need to load the *LinearAlgebra* package and then enter the matrix.

with(LinearAlgebra):

$$A := Matrix([[0.2, 0.1, 1, 1, 0], [0.1, 4, -1, 1, -1], [1, -1, 60, 0, -2], [1, 1, 0, 8, 4], [0, -1, -2, 4, 700]])$$

To determine the preconditioned matrix we first need the diagonal matrix, which being symmetric is also its transpose. its diagonal entries are specified by

$$a1 := \frac{1}{\sqrt{0.2}}; \ a2 := \frac{1}{\sqrt{4.0}}; \ a3 := \frac{1}{\sqrt{60.0}}; \ a4 := \frac{1}{\sqrt{8.0}}; \ a5 := \frac{1}{\sqrt{700.0}}$$

and the preconditioning matrix is

CI := Matrix([[a1, 0, 0, 0, 0], [0, a2, 0, 0, 0], [0, 0, a3, 0, 0], [0, 0, 0, a4, 0], [0, 0, 0, 0, a5]]) which Maple returns as

$$\begin{bmatrix} 2.23607 & 0 & 0 & 0 & 0 \\ 0 & .500000 & 0 & 0 & 0 \\ 0 & 0 & .129099 & 0 & 0 \\ 0 & 0 & 0 & .353553 & 0 \\ 0 & 0 & 0 & 0 & 0.0377965 \end{bmatrix}$$

The preconditioned matrix is

AH := CI.A.Transpose(CI)

The eigenvalues of A and AH are found with

Eigenvalues(A); Eigenvalues(AH)

Maple gives these as

Eigenvalues of A: 700.031, 60.0284, 0.0570747, 8.33845, 3.74533

Eigenvalues of AH: 1.88052, 0.156370, 0.852686, 1.10159, 1.00884

The condition numbers of A and AH in the l_{∞} norm are found with

ConditionNumber(A); ConditionNumber(AH)

which Maple gives as 13961.7 for A and 16.1155 for AH. It is certainly true in this case that AH is better conditioned that the original matrix A.

Illustration The linear system $A\mathbf{x} = \mathbf{b}$ with

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

has the solution

 $\mathbf{x}^* = (7.859713071, 0.4229264082, -0.07359223906, -0.5406430164, 0.01062616286)^t$

Table 7.5 lists the results obtained by using the Jacobi, Gauss-Seidel, and SOR (with $\omega=1.25$) iterative methods applied to the system with A with a tolerance of 0.01, as well as those when the Conjugate Gradient method is applied both in its unpreconditioned form and using the preconditioning matrix described in Example 3. The preconditioned conjugate gradient method not only gives the most accurate approximations, it also uses the smallest number of iterations.

Table 7.5

| Method | Number of Iterations | $\mathbf{x}^{(k)}$ | $\ \mathbf{x}^* - \mathbf{x}^{(k)}\ _{\infty}$ |
|-------------------------------------|----------------------|--|--|
| Jacobi | 49 | $(7.86277141, 0.42320802, -0.07348669, -0.53975964, 0.01062847)^{t}$ | 0.00305834 |
| Gauss-Seidel | 15 | $(7.83525748, 0.42257868, -0.07319124, -0.53753055, 0.01060903)^t$ | 0.02445559 |
| SOR ($\omega = 1.25$) | 7 | $(7.85152706, 0.42277371, -0.07348303, -0.53978369, 0.01062286)^t$ | 0.00818607 |
| Conjugate Gradient | 5 | $(7.85341523, 0.42298677, -0.07347963, -0.53987920, 0.008628916)^t$ | 0.00629785 |
| Conjugate Gradient (Preconditioned) | 4 | $(7.85968827, 0.42288329, -0.07359878, -0.54063200, 0.01064344)^t$ | 0.00009312 |

The preconditioned conjugate gradient method is often used in the solution of large linear systems in which the matrix is sparse and positive definite. These systems must be solved to approximate solutions to boundary-value problems in ordinary-differential equations (Sections 11.3, 11.4, 11.5). The larger the system, the more impressive the conjugate gradient method becomes because it significantly reduces the number of iterations required. In these systems, the preconditioning matrix C is approximately equal to L in the Cholesky

factorization LL^t of A. Generally, small entries in A are ignored and Cholesky's method is applied to obtain what is called an incomplete LL^t factorization of A. Thus, $C^{-t}C^{-1} \approx A^{-1}$ and a good approximation is obtained. More information about the conjugate gradient method can be found in [Kelley].

EXERCISE SET 7.6

1. The linear system

$$x_1 + \frac{1}{2}x_2 = \frac{5}{21},$$
$$\frac{1}{2}x_1 + \frac{1}{3}x_2 = \frac{11}{84}$$

has solution $(x_1, x_2)^t = (1/6, 1/7)^t$.

- **a.** Solve the linear system using Gaussian elimination with two-digit rounding arithmetic.
- **b.** Solve the linear system using the conjugate gradient method $(C = C^{-1} = I)$ with two-digit rounding arithmetic.
- c. Which method gives the better answer?
- **d.** Choose $C^{-1} = D^{-1/2}$. Does this choice improve the conjugate gradient method?
- 2. The linear system

$$0.1x_1 + 0.2x_2 = 0.3,$$

 $0.2x_1 + 113x_2 = 113.2$

has solution $(x_1, x_2)^t = (1, 1)^t$. Repeat the directions for Exercise 1 on this linear system.

3. The linear system

$$x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = \frac{5}{6},$$
$$\frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 = \frac{5}{12},$$
$$\frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = \frac{17}{60}$$

has solution $(1, -1, 1)^t$.

- a. Solve the linear system using Gaussian elimination with three-digit rounding arithmetic.
- b. Solve the linear system using the conjugate gradient method with three-digit rounding arithmetic.
- **c.** Does pivoting improve the answer in (a)?
- **d.** Repeat part (b) using $C^{-1} = D^{-1/2}$. Does this improve the answer in (b)?
- **4.** Repeat Exercise 3 using single-precision arithmetic on a computer.
- **5.** Perform only two steps of the conjugate gradient method with $C = C^{-1} = I$ on each of the following linear systems. Compare the results in parts (b) and (c) to the results obtained in parts (b) and (c) of Exercise 1 of Section 7.3 and Exercise 1 of Section 7.4.

a.
$$3x_1 - x_2 + x_3 = 1$$
, **b.** $10x_1 - x_2 = 9$, $-x_1 + 6x_2 + 2x_3 = 0$, $-x_1 + 10x_2 - 2x_3 = 7$, $x_1 + 2x_2 + 7x_3 = 4$. $-2x_2 + 10x_3 = 6$.

c.
$$10x_1 + 5x_2 = 6$$
, **d.** $5x_1 + 10x_2 - 4x_3 = 25$, $-4x_2 + 8x_3 - x_4 = -11$, $-x_3 + 5x_4 = -11$.

d.
$$4x_1 + x_2 - x_3 + x_4 = -2,$$

 $x_1 + 4x_2 - x_3 - x_4 = -1,$
 $-x_1 - x_2 + 5x_3 + x_4 = 0,$
 $x_1 - x_2 + x_3 + 3x_4 = 1.$

$$x_1 + 3x_2 + x_3 + x_4 = 6,$$

$$x_1 + x_2 + 5x_3 - x_4 - x_5 = 6,$$

$$x_2 - x_3 + 4x_4 = 6,$$

$$x_1 - x_3 + 4x_5 = 6.$$

e.
$$4x_1 + x_2 + x_3 + x_4 = 6$$
,
 $x_1 + 3x_2 + x_3 + x_4 = 6$,
 $x_1 + x_2 + 5x_3 - x_4 - x_5 = 6$,
 $x_2 - x_3 + 4x_4 = 6$,
 $x_1 - x_2 + x_3 + x_4 = 6$,
 $x_2 - x_3 + 4x_4 = 6$,
 $x_1 - x_2 + x_3 + x_4 = 1$.

f. $4x_1 - x_2 - x_4 = 0$,
 $-x_1 + 4x_2 - x_3 - x_5 = 5$,
 $-x_2 + 4x_3 - x_6 = 0$,
 $-x_1 + 4x_4 - x_5 = 6$,
 $-x_2 - x_4 + 4x_5 - x_6 = -2$,

- Repeat Exercise 5 using $C^{-1} = D^{-1/2}$.
- 7. Repeat Exercise 5 with $TOL=10^{-3}$ in the l_{∞} norm. Compare the results in parts (b) and (c) to those obtained in Exercises 5 and 7 of Section 7.3 and Exercise 5 of Section 7.4.
- **8.** Repeat Exercise 7 using $C^{-1} = D^{-1/2}$.
- **9.** Approximate solutions to the following linear systems $A\mathbf{x} = \mathbf{b}$ to within 10^{-5} in the l_{∞} norm. (i)

$$a_{i,j} = \begin{cases} 4, & \text{when } j = i \text{ and } i = 1, 2, \dots, 16, \\ j = i + 1 \text{ and } i = 1, 2, 3, 5, 6, 7, 9, 10, 11, 13, 14, 15, \\ j = i - 1 \text{ and } i = 2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15, 16, \\ j = i + 4 \text{ and } i = 1, 2, \dots, 12, \\ j = i - 4 \text{ and } i = 5, 6, \dots, 16, \end{cases}$$

and

 $\mathbf{b} = (1.902207, 1.051143, 1.175689, 3.480083, 0.819600, -0.264419,$ -0.412789, 1.175689, 0.913337, -0.150209, -0.264419, 1.051143, $1.966694, 0.913337, 0.819600, 1.902207)^t$

$$a_{i,j} = \begin{cases} 4, & \text{when } j = i \text{ and } i = 1, 2, \dots, 25, \\ j = i + 1 \text{ and } i = \begin{cases} 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, \\ 16, 17, 18, 19, 21, 22, 23, 24, \end{cases} \\ -1, & \text{when } \begin{cases} j = i + 1 \text{ and } i = \begin{cases} 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, \\ 17, 18, 19, 20, 22, 23, 24, 25, \end{cases} \\ j = i + 5 \text{ and } i = 1, 2, \dots, 20, \\ j = i - 5 \text{ and } i = 6, 7, \dots, 25, \end{cases} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\mathbf{b} = (1, 0, -1, 0, 2, 1, 0, -1, 0, 2, 1, 0, -1, 0, 2, 1, 0, -1, 0, 2, 1, 0, -1, 0, 2)^{t}$$

(iii)

$$a_{i,j} = \begin{cases} 2i, & \text{when } j = i \text{ and } i = 1, 2, \dots, 40, \\ -1, & \text{when } \begin{cases} j = i + 1 \text{ and } i = 1, 2, \dots, 39, \\ j = i - 1 \text{ and } i = 2, 3, \dots, 40, \end{cases}$$

$$0, & \text{otherwise}$$

and $b_i = 1.5i - 6$, for each i = 1, 2, ..., 40

- **a.** Use the Jacobi method.
- **b.** Use the Gauss-Seidel method.
- **c.** Use the SOR method with $\omega = 1.3$ in (i), $\omega = 1.2$ in (ii), and $\omega = 1.1$ in (iii).
- **d.** Use the conjugate gradient method and preconditioning with $C^{-1} = D^{-1/2}$.
- **10.** Solve the linear system in Exercise 16(b) of Exercise Set 7.3 using the conjugate gradient method with $C^{-1} = I$.
- **11.** Let

$$A_1 = \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix}, \quad -I = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \quad \text{and}$$

Form the 16×16 matrix A in partitioned form,

$$A = \left[\begin{array}{cccc} A_1 & -I & O & O \\ -I & A_1 & -I & O \\ O & -I & A_1 & -I \\ O & O & -I & A_1 \end{array} \right].$$

Let $\mathbf{b} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 0, 1, 2, 3, 4, 5, 6)^t$.

- **a.** Solve $A\mathbf{x} = \mathbf{b}$ using the conjugate gradient method with tolerance 0.05.
- **b.** Solve $A\mathbf{x} = \mathbf{b}$ using the preconditioned conjugate gradient method with $C^{-1} = D^{-1/2}$ and tolerance 0.05.
- c. Is there any tolerance for which the methods of part (a) and part (b) require a different number of iterations?
- 12. Use the transpose properties given in Theorem 6.14 on page 390 to prove Theorem 7.30.
- **13. a.** Show that an *A*-orthogonal set of nonzero vectors associated with a positive definite matrix is linearly independent.
 - **b.** Show that if $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(n)}\}$ is a set of *A*-orthogonal nonzero vectors in \mathbb{R} and $\mathbf{z}^t \mathbf{v}^{(i)} = \mathbf{0}$, for each $i = 1, 2, \dots, n$, then $\mathbf{z} = \mathbf{0}$.
- **14.** Prove Theorem 7.33 using mathematical induction as follows:
 - **a.** Show that $\langle \mathbf{r}^{(1)}, \mathbf{v}^{(1)} \rangle = 0$.
 - **b.** Assume that $\langle \mathbf{r}^{(k)}, \mathbf{v}^{(j)} \rangle = 0$, for each $k \leq l$ and j = 1, 2, ..., k, and show that this implies that $\langle \mathbf{r}^{(l+1)}, \mathbf{v}^{(j)} \rangle = 0$, for each j = 1, 2, ..., l.
 - **c.** Show that $\langle \mathbf{r}^{(l+1)}, \mathbf{v}^{(l+1)} \rangle = 0$.
- 15. In Example 3 the eigenvalues were found for the matrix A and the conditioned matrix AH. Use these to determine the condition numbers of A and AH in the l_2 norm, and compare your results to those given with the Maple commands ConditionNumber(A,2) and ConditionNumber(AH,2).

7.7 Survey of Methods and Software

In this chapter we studied iterative techniques to approximate the solution of linear systems. We began with the Jacobi method and the Gauss-Seidel method to introduce the iterative methods. Both methods require an arbitrary initial approximation $\mathbf{x}^{(0)}$ and generate a sequence of vectors $\mathbf{x}^{(i+1)}$ using an equation of the form

$$\mathbf{x}^{(i+1)} = T\mathbf{x}^{(i)} + \mathbf{c}.$$

It was noted that the method will converge if and only if the spectral radius of the iteration matrix $\rho(T) < 1$, and the smaller the spectral radius, the faster the convergence. Analysis of the residual vectors of the Gauss-Seidel technique led to the SOR iterative method, which involves a parameter ω to speed convergence.

These iterative methods and modifications are used extensively in the solution of linear systems that arise in the numerical solution of boundary value problems and partial differential equations (see Chapters 11 and 12). These systems are often very large, on the order of 10,000 equations in 10,000 unknowns, and are sparse with their nonzero entries in predictable positions. The iterative methods are also useful for other large sparse systems and are easily adapted for efficient use on parallel computers.

Almost all commercial and public domain packages that contain iterative methods for the solution of a linear system of equations require a preconditioner to be used with the method. Faster convergence of iterative solvers is often achieved by using a preconditioner. A preconditioner produces an equivalent system of equations that hopefully exhibits better convergence characteristics than the original system. The IMSL Library has a preconditioned conjugate gradient method, and the NAG Library has several subroutines for the iterative solution of linear systems.

All of the subroutines are based on Krylov subspaces. Saad [Sa2] has a detailed description of Krylov subspace methods. The packages LINPACK and LAPACK contain only direct methods for the solution of linear systems; however, the packages do contain many subroutines that are used by the iterative solvers. The public domain packages IML++, ITPACK, SLAP, and Templates, contain iterative methods. MATLAB contains several iterative methods that are also based on Krylov subspaces.

The concepts of condition number and poorly conditioned matrices were introduced in Section 7.5. Many of the subroutines for solving a linear system or for factoring a matrix into an LU factorization include checks for ill-conditioned matrices and also give an estimate of the condition number. LAPACK has numerous routines that include the estimate of a condition number, as do the ISML and NAG libraries.

LAPACK, LINPACK, the IMSL Library, and the NAG Library have subroutines that improve on a solution to a linear system that is poorly conditioned. The subroutines test the condition number and then use iterative refinement to obtain the most accurate solution possible given the precision of the computer.

More information on the use of iterative methods for solving linear systems can be found in Varga [Var1], Young [Y], Hageman and Young [HY], and Axelsson [Ax]. Iterative methods for large sparse systems are discussed in Barrett et al [Barr], Hackbusch [Hac], Kelley [Kelley], and Saad [Sa2].

Aleksei Nikolaevich Krylov (1863–1945) worked in applied mathematics, primarily in the areas of boundary value problems, the acceleration of convergence of Fourier series, and various classical problems involving mechanical systems. During the early 1930s he was the Director of the Physics-Mathematics Institute of the Soviet Academy of Sciences.