3

Numerical methods for linear system

CONTENTS

In this chapter we present the solution of n linear simultaneous algebraic equations in n unknowns. Linear systems of equations are associated with many problems in engineering and science, as well as with applications of mathematics to the social sciences and quantitative study of business and economic problems. A system of algebraic equations has the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{3.1}$$

where the coefficients a_{ij} and the constants b_j are known and x_i represents the unknowns. In matrix notation, the equations are written as

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

or simply

$$Ax = b.$$

where A is *coefficient matrix.*

A system of linear equations in n unknowns has a unique solution, provided that the determinant of the coefficient matrix is non-singular i.e., if $|A| \neq 0$. The rows and columns of a non-singular matrix are linearly independent in the sense that no row (or column) is a linear combination of the other rows (or columns). In this chapter, we assume A is non-singular. The Cramer's Rule implies

$$x_i = \frac{D_i}{D}, \quad (i = 1, 2, \cdots, n), \tag{3.2}$$

where $x^* = [x_1, x_2, \cdots, x_n]^T$, $D = \det A$ and D_i is the determinant of a new matrix formed by replacing the ith column of A with the b vector. Cramer's rule is slow because we have to evaluate a determinant for each x_i. If we compute the $n + 1$ determinants directly, $(n + 1)!(n - 1)$ multiplications are required. Taking $n = 20$ as an example, we need $21! \times 19 = 9.707 \times 10^{20}$ multiplications which means the computation time would be more than 11 days if it is carried on High Performance Computer with a performance of petaflops. Numerical method for solving the linear system will be discussed in this chapter.

There are two classes of methods for solving system of linear, algebraic equations: direct and iterative methods.

The common characteristics of **direct methods** are that they transform the original equation into equivalent equations (equations that have the same solution) that can be solved more easily. The transformation is carried out by applying certain operations. The solution are not with any truncation errors but the round off errors is introduced due to floating point operations.

Iterative or **indirect methods**, start with a guess of the solution x, and then repeatedly refine the solution until a certain convergence criterion is reached. Iterative methods are generally less efficient than direct methods due to the large number of operations or iterations required. Iterative procedures are self-correcting, meaning that round off errors (or even arithmetic mistakes) in one iteration cycle are corrected in subsequent cycles. The solution contains truncation error. A serious drawback of iterative methods is that they do not always converge to the solution. The initial guess affects only the number of iterations that are required for convergence. The indirect solution technique (iterative) is more useful to solve sparse matrix with large n.

3.1 Direct Method

In this section, we will discuss Gaussian elimination method, Gaussian elimination method with partial pivoting and Thomas algorithm for tridiagonal system.

3.1.1 Gaussian Elimination Method

Gaussian elimination is a popular technique for solving simultaneous linear algebraic equations. It reduces the coefficient matrix into an upper triangular matrix through a sequence of operations carried out on the matrix. The vector b is also modified in the process. We call this stage **forward elimination**. Then the solution x is obtained from a **backward substitution** procedure.

Firstly, we write the augmented matrix in the form

$$\overline{\boldsymbol{A}}^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{i1}^{(1)} & a_{i2}^{(1)} & \cdots & a_{in}^{(1)} & a_{i,n+1}^{(1)} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & a_{n,n+1}^{(1)} \end{bmatrix} \tag{3.3}$$

where

$$a_{ij}^{(1)} = a_{ij}, \quad a_{i,n+1}^{(1)} = b_i \quad (1 \leqslant i, j \leqslant n)$$

Step 1 Suppose $a_{11}^{(1)} \neq 0$. Eliminate x_1 from the second equation to the nth equation by addition of multipling $-l_{i1}$ $\left(l_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \right)$ on the first column to others.The resulting matrix has the form

$$\bar{A}^{(1)} \xrightarrow[\substack{\vdots \\ r_n+(-l_{n1})r_1}]{\substack{r_2+(-l_{21})r_1 \\ r_3+(-l_{31})r_1}} \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ 0 & a_{32}^{(2)} & \cdots & a_{3n}^{(2)} & a_{3,n+1}^{(2)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & a_{n,n+1}^{(2)} \end{bmatrix} \equiv \bar{A}^{(2)} \tag{3.4}$$

where

$$a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i1} a_{1j}^{(1)} \quad (2 \leqslant i \leqslant n, 2 \leqslant j \leqslant n+1).$$

Step k:

Suppose that $(k-1)$ eliminations have been finished, and the augmented matrix of the equivalent linear equations of (3.3) is as follows:

$$\overline{\boldsymbol{A}}^{(k)}=\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} & a_{k-1}^{(k-1)} \\ 0 & 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & a_{k,n+1}^{(k)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{ik}^{(k)} & \cdots & a_{in}^{(k)} & a_{i,n+1}^{(k)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{mi}^{(k)} & a_{nn+1}^{(k)} \end{bmatrix}$$

If $a_{kk}^{(k)} \neq 0$, we multiply $-l_{ik}\left(l_{ik}=\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right)$ on the kth column of $\overline{\mathbf{A}}^{(k)}$ and add it to ith column $(i=k+1,k+2,\cdots n)$. Then the augmented matrix of the equivalent linear equations of (3.3) is

$$\bar{A}^{(k)} \xrightarrow[\substack{\vdots \\ r_n+(-l_{nk})r_k}]{\substack{r_{k+1}+(-l_{k+1,k})r_k \\ r_{k+2}+(-l_{k+2,k})r_k}} \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & a_{2,k+1}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{kn}^{(k)} & a_{k,n+1}^{(k)} \\ 0 & 0 & \cdots & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} & a_{k+1,n+1}^{(k+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} & a_{n,n+1}^{(k+1)} \end{bmatrix} \equiv \bar{A}^{(k+1)} \tag{3.5}$$

where

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}-l_{ik}a_{kj}^{(k)} \quad (k+1 \leqslant i \leqslant n, k+1 \leqslant j \leqslant n+1). \tag{3.6}$$

In the procedure, l_{ik} is said **multiplier**, and the coefficient $a_{kk}^{(k)}$ is **pivot element**.

Reapply the above step after $n-1$ eliminations, and the augmented matrix of the equivalent linear equations is

$$\overline{\boldsymbol{A}}^{(n)}=\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,n-1}^{(1)} & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2,n-1}^{(2)} & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ & & \ddots & \vdots & & \vdots \\ & & & a_{n-1,n-1}^{(n-1)} & a_{n-1,n}^{(n-1)} & a_{n-1,n+1}^{(n-1)} \\ & & & & a_{nn}^{(n)} & a_{n,n+1}^{(n)} \end{bmatrix}. \tag{3.7}$$

Let

$$\boldsymbol{U} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,n-1}^{(1)} & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2,n-1}^{(2)} & a_{2n}^{(2)} \\ & & \ddots & \vdots & \vdots \\ & & & a_{n-1,n-1}^{(n-1)} & a_{n-1,n}^{(n-1)} \\ & & & & a_{nn}^{(n)} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} a_{1,n+1}^{(1)} \\ a_{2,n+1}^{(2)} \\ \vdots \\ a_{n-1,n+1}^{(n-1)} \\ a_{n,n+1}^{(n)} \end{bmatrix}$$

and the equivalent equation becomes

$$\boldsymbol{U}\boldsymbol{x} = \boldsymbol{y},$$

where $\boldsymbol{U}$ is upper triangular matrix.

Finally, x_n can be solved

$$x_n = a_{n,n+1}^{(n)} / a_{n,n}^{(n)}. \tag{3.8}$$

In general, x_i can be solved by $x_n, x_{n-1}, \cdots, x_{i+1}$, i.e.

$$x_i = \left(a_{i,n+1}^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j \right) / a_{ii}^{(i)} \quad (i = n-1, n-2, \cdots, 1). \tag{3.9}$$

The procedure is called **backward substitution**.

Operation Counts

Both the amount of time required to complete the calculations and the subsequent round-off error depend on the number of floating-point arithmetic operations needed to solve a routine problem. In general, the amount of time required to perform a multiplication or division on a computer is approximately the same and is considerably greater than that required to perform an addition or subtraction. The actual differences in execution time, however, depend on the particular computing system. To demonstrate the counting operations for a given method, we will count the operations required to solve a typical linear system of n equations in n unknowns. We will keep the count of the additions/subtractions separate from the count of the multiplications/divisions because of the time differential.

In forward elimination stage, Eq. (3.6) requires $(n-k)$ divisions being performed. The replacement of the equation r_k by $(r_k - l_{ik} r_i)$ requires l_{ik} being multiplied by each term in r_i, resulting in a total of $(n-i)(n-i+1)$ multiplications. After this is completed, each term of the resulting equation is subtracted from the corresponding term in r_j. This requires $(n-k)(n-k+1)$ subtractions. For each $k = 1, 2, \ldots, n-1$, the operations required in Step k as follows.

(1) Multiplications/divisions

$$(n-k) + (n-k)(n-k+1) = (n-k)(n-k+2);$$

(2) Additions/subtractions

$$(n-k)(n-k+1).$$

The total number of operations required by $n-1$ eliminations is obtained by summing the operation counts for each k. Recalling from calculus that

$$\sum_{j=1}^{m} 1 = m, \quad \sum_{j=1}^{m} j = \frac{m(m+1)}{2}, \quad \text{and} \quad \sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6},$$

we have the following operation counts.

Multiplications/divisions

$$\begin{aligned}\sum_{k=1}^{n-1}(n-k)(n-k+2) &= \sum_{k=1}^{n-1}\left(n^2-2nk+k^2+2n-2k\right)\\ &= \sum_{k=1}^{n-1}(n-k)^2+2\sum_{k=1}^{n-1}(n-k)=\sum_{k=1}^{n-1}k^2+2\sum_{k=1}^{n-1}k\\ &= \frac{(n-1)n(2n-1)}{6}+2\frac{(n-1)n}{2}=\frac{2n^3+3n^2-5n}{6};\end{aligned}$$

Additions/subtractions

$$\begin{aligned}\sum_{k=1}^{n-1}(n-k)(n-k+1) &= \sum_{k=1}^{n-1}\left(n^2-2nk+k^2+n-k\right)\\ &= \sum_{k=1}^{n-1}(n-k)^2+\sum_{k=1}^{n-1}(n-k)=\sum_{k=1}^{n-1}k^2+\sum_{k=1}^{n-1}k\\ &= \frac{(n-1)n(2n-1)}{6}+\frac{(n-1)n}{2}=\frac{n^3-n}{3}.\end{aligned}$$

In the backward substitution, (3.8) requires one division and (3.9) requires $(n-i)$ multiplications and $(n-i-1)$ additions for each summation term and then one subtraction and one division. The total number of operations in (3.8) and (3.9) is as follows.

Multiplications/divisions

$$\begin{aligned}1+\sum_{i=1}^{n-1}((n-i)+1) &= 1+\left(\sum_{i=1}^{n-1}(n-i)\right)+n-1\\ 1 &= n+\sum_{i=1}^{n-1}(n-i)=n+\sum_{i=1}^{n-1}i=\frac{n^2+n}{2};\end{aligned}$$

Additions/subtractions

$$\sum_{i=1}^{n-1}((n-i-1)+1)=\sum_{i=1}^{n-1}(n-i)=\sum_{i=1}^{n-1}i=\frac{n^2-n}{2}.$$

The total number of arithmetic operations in Gaussian elimination method is, therefore:

Multiplications/divisions

$$\frac{2n^3+3n^2-5n}{6}+\frac{n^2+n}{2}=\frac{n^3}{3}+n^2-\frac{n}{3};$$

Additions/subtractions

$$\frac{n^3-n}{3}+\frac{n^2-n}{2}=\frac{n^3}{3}+\frac{n^2}{2}-\frac{5n}{6}.$$

For large n, the total number of multiplications and divisions is approximately $n^3/3$, as is the total number of additions and subtractions. Thus the amount of computation and the time required increases with n in proportion to n^3. The computation amount is far less than that of Cramer's method.

Exercise 3.1. *Solving the linear system by Gaussian elimination*

$$\begin{bmatrix} 2 & -4 & 6 \\ 4 & -9 & 2 \\ 1 & -1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 3 \\ 5 \\ 4 \end{bmatrix}.$$

Solution

$$\overline{\boldsymbol{A}}^{(1)}=\begin{bmatrix} 2 & -4 & 6 & 3 \\ 4 & -9 & 2 & 5 \\ 1 & -1 & 3 & 4 \end{bmatrix}\xrightarrow[r_3+\left(-\frac{1}{2}\right)r_1]{r_2+(-2)r_1}\begin{bmatrix} 2 & -4 & 6 & 3 \\ 0 & -1 & -10 & -1 \\ 0 & 1 & 0 & \frac{5}{2} \end{bmatrix}.$$

$$\xrightarrow{r_3+r_2}\begin{bmatrix} 2 & -4 & 6 & 3 \\ 0 & -1 & -10 & -1 \\ 0 & 0 & -10 & \frac{3}{2} \end{bmatrix}$$

The equivalent linear equations are

$$\begin{cases} 2x_1-4x_2+6x_3=3, \\ \quad -x_2-10x_3=-1, \\ \quad\quad -10x_3=\dfrac{3}{2}. \end{cases}$$

Then by backward substitution we obtain

$$\begin{aligned} x_3&=\tfrac{3}{2}/(-10)=-\tfrac{3}{20}, \\ x_2&=(-1+10x_3)/(-1)=\tfrac{5}{2}, \\ x_1&=(3+4x_2-6x_3)/2=\tfrac{139}{20}. \end{aligned}$$

In Gaussian elimination, we suppose $a_{kk}^{(k)}\neq 0(k=1,2,\cdots,n-1)$. The next theorem implies what properties the coefficient matrix A has to satisfy this requirement.

Theorem 7. *Given the linear equation $Ax = b$. If all the determinants of leading principal submatrices are nonzero , i.e.*

$$a_{11} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \neq 0, \quad \cdots, \quad \det(\boldsymbol{A}) \neq 0$$

the pivot element in every elimination step $a_{kk}^{(k)}(k = 1, 2, \cdots, n) \neq 0$.

Proof

Let $\Delta_1 = a_{11}, \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{z2} \end{vmatrix}, \cdots, \Delta_n = \det(\boldsymbol{A})$. We adopt the induction to prove the theorem.

If $k = 1$, $\Delta_1 = a_{11} \neq 0$, and $a_{11}^{(1)} = a_{11} \neq 0$. The theorem holds.

Suppose $\Delta_1 \neq 0, \Delta_2 \neq 0, \cdots, \Delta_{k-1} \neq 0$. We have $a_{ii}^{(i)} \neq 0(i = 1, 2, \cdots, k-1)$. We will prove $a_{kk}^{(k)} \neq 0$ if $\Delta_k \neq 0$.

Since $a_{ii}^{(i)} \neq 0(i = 1, 2, \cdots, k-1)$, $A^{(1)}$ can be reduced to $A^{(k)}$ using Gauss elimination:

$$\boldsymbol{A}^{(1)} \longrightarrow \boldsymbol{A}^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots & & \vdots \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix}$$

Thus

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \rightarrow \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{kk}^{(k)} \end{bmatrix}$$

$$\Delta_k = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix} = \begin{vmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{kk}^{(k)} \end{vmatrix} = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{kk}^{(k)}$$

From $\Delta_k \neq 0$ and $a_{ii}^{(i)} \neq 0(i = 1, 2, \cdots, k-1)$, we obtain $a_{kk}^{(k)} \neq 0$. The theorem is proved.

Remark It easy to get $\Delta_i \neq 0(i = 1, 2, \cdots, k)$ from $a_{ii}^{(i)} \neq 0(i = 1, 2, \cdots, k)$. Hence, in Theorem 7 $a_{kk}^{(k)} \neq 0(k = 1, 2, \cdots, n)$ if and only if $\Delta_k \neq 0(k = 1, 2, \cdots, n)$.

3.1.2 Gaussian Elimination Method with Partial Pivoting

The Gaussian algorithm, in the simple form just described, can be carried under the condition $a_{kk}^{(k)} \neq 0(k = 1, 2, \cdots, n-1)$. But it is noted that if $\left|a_{kk}^{(k)}\right|$ is small relative to the entries $\left|a_{ik}^{(k)}\right| (k+1 \leqslant i \leqslant n)$, the Gaussian elimination is not adaptable. In the kth step elimination, the kth column of $\bar{A}^{(k)}$ is multiplied by $(-l_{ik})$ and added to ith column. If the coefficient $\left(a_{k,k+1}^{(k)}, a_{k,k+2}^{(k)}, \cdots, a_{k,n+1}^{(k)}\right)$ of the kth column have some errors $(\varepsilon_{k+1}, \varepsilon_{k+2}, \cdots, \varepsilon_{n+1})$, the errors are enlarged $(-l_{ik})$ times and added to the i column of $\overline{\boldsymbol{A}}^{(k+1)}$. Since $|l_{ik}| \left(= |\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}|\right)$ may be very large, the round-off error may be enlarged much more.

Suppose $(k-1)$ eliminations have been finished, and

$$\overline{\boldsymbol{A}}^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ & & \ddots & \vdots & & \vdots & \vdots \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & a_{k,n+1}^{(k)} \\ & & & \vdots & & \vdots & \vdots \\ & & & a_{ik}^{(k)} & \cdots & a_{in}^{(k)} & a_{i,n+1}^{(k)} \\ & & & \vdots & & \vdots & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & a_{n,n+1}^{(k)} \end{bmatrix}$$

To avoid this problem, pivoting is performed by selecting an element $a_{tk}^{(k)}$ with a larger magnitude as the pivot, and interchanging the kth and tth rows. We determine the $t \geq k$ such that

$$\left|a_{tk}^{(k)}\right| = \max_{k \leqslant i \leqslant n} \left|a_{ik}^{(k)}\right|$$

and perform $r_t \leftrightarrow r_k$:

$$\overline{\boldsymbol{A}}^{(k)} \xrightarrow{r_t \leftrightarrow r_k} \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ & & \ddots & \vdots & & \vdots & \vdots \\ & & & a_{tk}^{(k)} & \cdots & a_{kn}^{(k)} & a_{k,n+1}^{(k)} \\ & & & \vdots & & \vdots & \vdots \\ & & & a_{ik}^{(k)} & \cdots & a_{in}^{(k)} & a_{i,n+1}^{(k)} \\ & & & \vdots & & \vdots & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & a_{n,n+1}^{(k)} \end{bmatrix}$$

Next the step k-th elimination is used and the procedure is said **Gaussian elimination with partial pivoting**. The multiplier satisfies

$$|l_{ik}| = \left| \frac{a_{ik}^{(k)}}{a_{kk}} \right| \leqslant 1 \quad (i = k+1, k+2, \cdots, n) \tag{3.10}$$

which can ensure the round-off error is not enlarged generally. The algorithm is numerical stable.

Exercise 3.2. *Use Gaussian elimination with partial pivoting to solve the following linear equations*

$$\begin{bmatrix} 2 & -4 & 6 \\ 4 & -9 & 2 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 4 \end{bmatrix}$$

Solution

$$\overline{\boldsymbol{A}}^{(1)} = \begin{bmatrix} 2 & -4 & 6 & 3 \\ \boxed{4} & -9 & 2 & 5 \\ 1 & -1 & 3 & 4 \end{bmatrix} \xrightarrow{r_2 \leftrightarrow r_1} \begin{bmatrix} 4 & -9 & 2 & 5 \\ 2 & -4 & 6 & 3 \\ 1 & -1 & 3 & 4 \end{bmatrix}$$

$$\xrightarrow[r_3+(-\frac{1}{4})r_1]{r_2+(-\frac{1}{2})r_1} \begin{bmatrix} 4 & -9 & 2 & 5 \\ 0 & \frac{1}{2} & 5 & \frac{1}{2} \\ 0 & \boxed{\dfrac{5}{4}} & \frac{5}{2} & \frac{11}{4} \end{bmatrix} \xrightarrow{r_3 \leftrightarrow r_2} \begin{bmatrix} 4 & -9 & 2 & 5 \\ 0 & \frac{5}{4} & \frac{5}{2} & \frac{11}{4} \\ 0 & \frac{1}{2} & 5 & \frac{1}{2} \end{bmatrix}$$

$$\xrightarrow{r_3+(-\frac{2}{5})r_2} \begin{bmatrix} 4 & -9 & 2 & 5 \\ 0 & \frac{5}{4} & \frac{5}{2} & \frac{11}{4} \\ 0 & 0 & 4 & -\frac{3}{5} \end{bmatrix}$$

The equivalent upper triangular equations are

$$\begin{cases} 4x_1 - 9x_2 + 2x_3 = 5 \\ \dfrac{5}{4}x_2 + \dfrac{5}{2}x_3 = \dfrac{11}{4} \\ 4x_3 = -\dfrac{3}{5} \end{cases}$$

By backward substitution procedure, we have

$$x_3 = -\frac{3}{20}, \quad x_2 = \frac{5}{2}, \quad x_1 = \frac{139}{20}.$$

Definition 8. *If* $n \times n$ *matrix* A *is said to be* **strictly diagonally dominant** *when*

$$|a_{ii}| > \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}|$$

holds for each $i = 1, 2, \cdots, n$.

Theorem 8. *If A is strictly diagonally dominant, then $|A| \neq 0$.*

Remark:

(1) If the coefficient matrix is symmetric and positive or strictly diagonally dominant, the Gaussian eliminations is stable without partial pivoting.

(2) The computation amount of Gaussian elimination of partial pivoting is almost the same as that of Gaussian algorithm except choosing the pivot element and exchange the two columns.

3.1.3 Thomas Algorithm for Tridiagonal System

Consider the system of linear simultaneous algebraic equations given by

$$\begin{bmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix} \tag{3.11}$$

where coefficient matrix is a tridiagonal matrix, and the coefficients satisfy

(1) $|b_1| > |c_1 |> 0$;

(2) $|b_i| \geqslant |a_i| + |c_i|, a_i c_i \neq 0 \quad (i = 2, 3, \cdots, n-1)$;

(3) $|b_n| > |a_n| > 0$;

After elimination, the augmented matrix of the equivalent linear equations is

$$\begin{bmatrix} \beta_1 & c_1 & & & & y_1 \\ & \beta_2 & c_2 & & & y_2 \\ & & \ddots & \ddots & & \vdots \\ & & & \beta_{n-1} & c_{n-1} & y_{n-1} \\ & & & & \beta_n & y_n \end{bmatrix},$$

where

$$\begin{cases} \beta_1 = b_1, \quad y_1 = d_1 \\ l_i = \frac{a_i}{\beta_{i-1}}, \quad \beta_i = b_i - l_i c_{i-1}, \quad y_i = d_i - l_i y_{i-1} \quad (i = 2, 3, \cdots, n) \end{cases} \tag{3.12}$$

Then we can get x by backward substitution

$$\begin{cases} x_n = y_n / \beta_n, \\ x_i = (y_i - c_i x_{i+1}) / \beta_i \quad (i = n-1, n-2, \cdots, 1). \end{cases} \tag{3.13}$$

The above procedure is said **Thomas Algorithm**. $(5n - 4)$ Multiplications/divisions and $(3n-3)$ Additions/subtractions are needed in the Thomas algorithm.

3.2 Norms and Error Analysis

The direct methods for solving linear equations required a large number of arithmetic operations, and using finite-digit arithmetic leads only to an approximation to an actual solution of the system. We first need to determine a way to measure the difference between the approximations and the exact solution which are n-dimensional column vectors.

In actuality, to discuss iterative methods for solving linear systems in Sec. 3.3 , 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.

3.2.1 Norms of Vectors

Let $\mathbf{R}^n$ denote the set of all n-dimensional column vectors $x = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ with real-number components.

Definition 9. *A* **vector norm** *on* $\mathbf{R}^n$ *is a function* $f(x) = \|x\|$, *from* $\mathbf{R}^n$ *to* $\mathbf{R}$ *with the following properties:*

$1°$ $\|x\| \geqslant 0$ *for all* $x \in \mathbf{R}^n$, *and* $\|x\| = 0$*if and only if* $x = 0$*;*

$2°$ $\|\lambda x\| = |\lambda| \cdot \|x\|$ *for all* $\lambda \in \mathbf{R}$ *and* $x \in \mathbf{R}^n$ *;*

$3°$ *(triangular inequality)* $\|x + y\| \leqslant \|x\| + \|y\|$ *for all* $x, y \in \mathbf{R}^n$ *.*

From the triangular inequality, there is

$$| \, \|x\| - \|y\|| \leqslant \|x - y\| \tag{3.14}$$

for all $x \in \mathbf{R}^n, y \in \mathbf{R}^n$.

There are three common vector norms defined by:

$\|x\|_1 = \sum\limits_{i=1}^{n} |x_i|\, , \|\boldsymbol{x}\|_\infty = \max\limits_{1 \leqslant i \leqslant n} |x_i|\, , \|\boldsymbol{x}\|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}.$

It is easy to prove they are vector norms. The l_2 norm is called the **Euclidean norm** of the vector x because it represents the usual notion of distance from the origin in case x is in $\mathbf{R}^1, \mathbf{R}^2$ or $\mathbf{R}^3$. l_∞ is called **infinity** or **maximum norm**.

Exercise 3.3. *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 $\mathbf{R}^3$ has norms

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

and

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

Theorem 9. *Suppose $f(x) = \|x\|$ is a vector norm on $\mathbf{R}^n$. $f(x)$ is continuous on the components of the vector x.*

Proof

Suppose$\boldsymbol{x} = (x_1, \cdots, x_n)^{\mathrm{T}} \in \mathbf{R}^n, \boldsymbol{y} = (y_1, \cdots, y_n)^{\mathrm{T}} \in \mathbf{R}^n$.
Let$e_1 = (1, 0, 0, \cdots, 0)^{\mathrm{T}}, \boldsymbol{e}_2 = (0, 1, 0, \cdots, 0)^{\mathrm{T}}, \cdots, e_n = (0, 0, \cdots, 0, 1)^{\mathrm{T}}$.
Then there are

$$x = x_1e_1 + x_2e_2 + \cdots + x_ne_n, \quad y = y_1e_1 + y_2e_2 + \cdots + y_ne_n.$$

Since (3.14) and triangular inequality, we have

$$\begin{aligned} |f(\boldsymbol{y}) - f(\boldsymbol{x})| &= |\ \|\boldsymbol{y}\| - \|\boldsymbol{x}\|\ | \leqslant \|\boldsymbol{x} - \boldsymbol{y}\| \\ &= \left\| \sum_{i=1}^{n} (x_i - y_i)\, \boldsymbol{e}_i \right\| \leqslant \max_{1 \leqslant i \leqslant n} |x_i - y_i| \sum_{i=1}^{n} \|\boldsymbol{e}_i\| \end{aligned}$$

For any $\varepsilon > 0$, we take $\delta = \varepsilon / \sum\limits_{i=1}^{n} \|e_i\|$ .When $|x_i - y_i| \leqslant \delta (i = 1, 2, \cdots, n)$, we have

$$|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leqslant \varepsilon$$

i.e. $f(x)$ is continuous on $x_1, x_2, \cdots, x_n$.

Definition 10. *Suppose $\| \cdot \|_p$ and $\| \cdot \|_q$ are vector norms on $\mathbf{R}^n$. If there exist two positive constants c_1 and c_2, such that*

$$c_1\|\boldsymbol{x}\|_p \leqslant \|\boldsymbol{x}\|_q \leqslant c_2\|\boldsymbol{x}\|_p \tag{3.15}$$

for all $x \in \mathbf{R}^n$. We call the vector norms $\| \cdot \|_p$ and $\| \cdot \|_q$ are **equivalent**.

Theorem 10. *Suppose $\| \cdot \|_p$ and $\| \cdot \|_q$ are vector norms on $\mathbf{R}^n$, then $\| \cdot \|_p$ and $\| \cdot \|_q$ are equivalent.*

Proof Consider the set

$$S = \{\boldsymbol{y} \mid \boldsymbol{y} \in \mathbf{R}^n, \|\boldsymbol{y}\|_2 = 1\},$$

which is closed bounded in $\mathbf{R}^n$. Since the function $f_1(\boldsymbol{x}) = \|x\|_p$ is continuous on S, the minimum a_1 and maximum a_2 of the function exist, i.e. there exists $\boldsymbol{y}_1 \in S, \boldsymbol{y}_2 \in S$ such that for any $\boldsymbol{y} \in S$

$$a_1 \leqslant \|\boldsymbol{y}\|_p \leqslant a_2 \tag{3.16}$$

where $a_1 = \|\boldsymbol{y}_1\|_p and\ a_2 = \|\boldsymbol{y}_2\|_p$. Because $\boldsymbol{y}_1 \neq \mathbf{0}$, we have $a_1 > 0$. For all $\boldsymbol{x} \in \mathbf{R}^n, \boldsymbol{x} \neq \mathbf{0}$, there is $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \in S$. From (3.16),

$$a_1 \leqslant \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \right\|_p \leqslant a_2,$$

i.e.

$$a_1\|\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{x}\|_p \leqslant a_2\|\boldsymbol{x}\|_2, \tag{3.17}$$

which holds for $x = 0$.

Similarly, $f_2(\boldsymbol{x}) = \|\boldsymbol{x}\|_q$ is also continuous on S. Thus there exist positive constant b_1 and b_2 such that

$$b_1\|\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{x}\|_q \leqslant b_2\|\boldsymbol{x}\|_2 \tag{3.18}$$

for all $\boldsymbol{x} \in \mathbf{R}^n$.

From (3.17) and (3.18),

$$b_1 \cdot \frac{1}{a_2}\|\boldsymbol{x}\|_p \leqslant \|\boldsymbol{x}\|_q \leqslant b_2 \cdot \frac{1}{a_1}\|\boldsymbol{x}\|_p,$$

$foranyx \in \mathbf{R}^n$.

Let

$$c_1 = \frac{b_1}{a_2}, \quad c_2 = \frac{b_2}{a_1}$$

(3.15) can be derived.

Definition 11. *Suppose* $\|\cdot\|$ *is a vector norm on* R^n. $\|x - y\|$ *is the called* **distance** *between* x *and* y*.*

The distance can indicate the difference between the approximation x of $Ax = b$ and the actual solution x^*. If the value $\|x^* - \tilde{x}\|$ is small, x is close to x^*. When consider the value of x^*, we can use the relative error $\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}\| / \|\boldsymbol{x}^*\|$ or $\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}\| / \|\tilde{\boldsymbol{x}}\|$ to measure the approximation is better or not.

Definition 12. *A sequence of vectors* $\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty}$ *in* $\mathbf{R}^n$ *is said to* **converge** *to* $\mathbf{c}$ *with respect to the norm* $\|\cdot\|$ *if*

$$\lim_{k\to\infty} \left\|\boldsymbol{x}^{(k)} - \boldsymbol{c}\right\| = 0,$$

and denote

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = c.$$

Since the norms are equivalent, the convergency of sequence of vectors is irrelevant to the definition of vector norm $\|\cdot\|$.

3.2.2 Norms of Matrices

Let $\mathbf{R}^{n\times n}$ denote the set of $n \times n$ real matrices

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

Definition 13. *A* **matrix norm** *on $\mathbf{R}^{n\times n}$ is a real-valued function, $\|\cdot\|$, defined on this set, satisfying for all $n \times n$ matrices A and B and all real number λ:*

(1) $\|\boldsymbol{A}\| \geqslant 0$, and $\|\boldsymbol{A}\| = 0$ if and only if $\boldsymbol{A} = \mathbf{0}$.

(2) $\|\lambda\mathbf{A}\| = |\lambda|\|\boldsymbol{A}\|$;

(3) $\|\boldsymbol{A}+\boldsymbol{B}\| \leqslant \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$.

Theorem 11. *If $\|\cdot\|$ is a vector norm on $\mathbf{R}^n$, then*

$$\|\boldsymbol{A}\| = \max_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} \tag{3.19}$$

is a matrix norm.

Proof Suppose $\mathbf{A} \in \mathbf{R}^{n\times n}, \boldsymbol{x} \in \mathbf{R}^n$, $\|\cdot\|$ is a vector norm on $\mathbf{R}^n$. $\|\boldsymbol{Ax}\|$ is continuous on $\mathbf{R}^n$. Then $\|\boldsymbol{Ax}\|$ has maximum M on the set

$$\tilde{S} = \{\boldsymbol{y} \mid \boldsymbol{y} \in \mathbf{R}^n, \|\boldsymbol{y}\| = 1\},$$

i.e. there exists $\tilde{\boldsymbol{x}} \in \tilde{S}$ such that

$$\max_{y\in\tilde{S}} \|Ay\| = M, \quad M = \|A\tilde{\boldsymbol{x}}\|.$$

For all $\boldsymbol{x} \in \mathbf{R}^n, \boldsymbol{x} \neq 0$, we have

$$\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in \tilde{S}, \quad \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \left\|A\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right\|$$

Thus

$$\max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\neq 0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{y}\in\tilde{S}} \|A\boldsymbol{y}\| = M.$$

Furthermore, it is easy to prove (3.19) satisfy (1)-(3) in matrix norm definition.

Remark

(1) 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.

(2) From (3.19), $\|\mathbf{Ax}\| \leqslant \|\boldsymbol{A}\|\|\boldsymbol{x}\|$ for all $x \in \mathbf{R}^n$, and $\boldsymbol{A} \in \mathbf{R}^{n\times n}$.

(3) We also have $\|\boldsymbol{AB}\| \leqslant \|\boldsymbol{A}\|\|\boldsymbol{B}\|$.

Definition 14. *The* **spectral radius** $\rho(\boldsymbol{B})$ *of a matrix B is defined by*

$$\rho(\boldsymbol{B}) = \max_{1\leq i\leqslant n} \{|\lambda_i|\},$$

where $\lambda_1, \lambda_2, \cdots, \lambda_n$ *are the* n *eigenvalues of* $\boldsymbol{B} \in \mathbf{R}^{n\times n}$.

Theorem 12. *Suppose* $\boldsymbol{A} \in \mathbf{R}^{n\times n}$, *then*

$$(1)\|\boldsymbol{A}\|_1 = \max_{\substack{\boldsymbol{x}\in\mathbf{R}^n\\ \boldsymbol{x}\neq 0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|; \tag{3.20}$$

$$(2)\|\boldsymbol{A}\|_\infty = \max_{\substack{\boldsymbol{x}\in\mathbf{R}^n\\ \boldsymbol{x}\neq 0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|; \tag{3.21}$$

$$(3)\|\boldsymbol{A}\|_2 = \max_{\substack{\boldsymbol{x}\in\mathbf{R}^n\\ \boldsymbol{x}\neq 0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sqrt{\rho(\boldsymbol{A}^T\boldsymbol{A})}. \tag{3.22}$$

Proof (1) Let

$$\mu = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|$$

For any $x \in \mathbf{R}^n$, we have

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{x}\|_1 &= \sum_{i=1}^{n} \left|\sum_{j=1}^{n} a_{ij}x_j\right| \leqslant \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|\,|x_j| \\ &= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} |a_{ij}|\right) |x_j| \leqslant \mu \sum_{j=1}^{n} |x_j| = \mu\|\boldsymbol{x}\|_1.\end{aligned}$$

Thus

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} \leqslant \mu \tag{3.23}$$

for $\boldsymbol{x} \neq 0$. On the other hand, let

$$\mu = \sum_{i=1}^{n} |a_{ik}|.$$

Take

$$\tilde{\boldsymbol{x}} = (0, \cdots, 0, \underset{\substack{\uparrow\\ \text{the } k\text{th componet}}}{1}, 0, \cdots, 0)^{\mathrm{T}}$$

then

$$\|\tilde{\boldsymbol{x}}\|_1 = 1, \quad \|\boldsymbol{A}\tilde{\boldsymbol{x}}\|_1 = \left\| \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{bmatrix} \right\|_1 = \mu$$

Thus

$$\frac{\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|_1}{\|\tilde{\boldsymbol{x}}\|_1} = \mu. \tag{3.24}$$

Combining (3.23) and (3.24)

$$\max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\neq 0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} = \mu. \tag{3.25}$$

(2) Denote

$$v = \max_{1<i\leqslant n} \sum_{j=1}^{n} |a_{ij}|\,.$$

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{x}\|_\infty &= \max_{1\leqslant i\leqslant n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \leqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|\,|x_j| \\ &\leqslant \max_{1\leqslant i\leqslant n} \left(\sum_{j=1}^{n} |a_{ij}| \right) \max_{1\leqslant j<n} |x_j| \leqslant v\|\boldsymbol{x}\|_\infty \end{aligned}$$

for any $\boldsymbol{x} \neq 0$.

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leqslant v \tag{3.26}$$

On the other hand, let

$$v = \sum_{j=1}^{n} |a_{kj}|\,.$$

Let

$$\tilde{\boldsymbol{x}} = (\xi_1, \xi_2, \cdots, \xi_n)^{\mathrm{T}}$$

where

$$\xi_j = \begin{cases} |a_{kj}|/a_{kj}, & a_{kj} \neq 0 \\ 1, & a_{kj} = 0 \end{cases} \quad (j = 1, 2, \cdots, n)$$

Clearly, $\|\tilde{\boldsymbol{x}}\|_\infty = 1$ and the kth component of $\boldsymbol{A}\tilde{\boldsymbol{x}}$ is $\sum\limits_{j=1}^{n} a_{kj}\boldsymbol{\xi}_j = \sum\limits_{j=1}^{n} |a_{kj}| = v$. Thus

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|_\infty = v.$$

We obtain

$$\frac{\|\boldsymbol{A}\tilde{\boldsymbol{x}}|_\infty}{\|\tilde{\boldsymbol{x}}\|_\infty} = v \tag{3.27}$$

By (3.26) and (3.27), there is

$$\max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\neq 0}} \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = v \tag{3.28}$$

(3) For any $\boldsymbol{x} \in \mathbf{R}^n$,

$$\|\boldsymbol{Ax}\|_2^2 = (\boldsymbol{Ax})^{\mathrm{T}}(\boldsymbol{Ax}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Ax} \geqslant 0$$

Thus $A^{\mathrm{T}}A$ is symmetric positive semidefinite. So $A^{\mathrm{T}}A$ has n nonnegative real eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0$ and standard orthogonal eigenvector system $\boldsymbol{\mu}_1, \boldsymbol{\mu}_{\not\asymp}, \cdots, \boldsymbol{\mu}_n$ corresponding to eigenvalues satisfies

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}_i = \lambda_i\boldsymbol{\mu}_i \quad (i = 1, 2, \cdots, n).$$

Any vector $\boldsymbol{x} \in \mathbf{R}^n$ can be written as

$$\boldsymbol{x} = c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2 + \cdots + c_n\boldsymbol{\mu}_n.$$

If $\|\boldsymbol{x}\|_2 = \sqrt{\sum\limits_{i=1}^{n} c_i^2} = 1$, then

$$\begin{aligned}
\|\boldsymbol{Ax}\|_2^2 = (\boldsymbol{Ax})^{\mathrm{T}}\boldsymbol{Ax} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ax} &= \left(\sum_{i=1}^{n} c_i\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}\sum_{i=1}^{n} c_i\boldsymbol{\mu}_i \\
&= \left(\sum_{i=1}^{n} c_i\boldsymbol{\mu}_i\right)^{\mathrm{T}} \sum_{i=1}^{n} c_i\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}_i = \left(\sum_{i=1}^{n} c_i\boldsymbol{\mu}_i\right)^{\mathrm{T}} \sum_{i=1}^{n} c_i\lambda_i\boldsymbol{\mu}_i \\
&= \sum_{i=1}^{n} \lambda_i c_i^2 \leqslant \lambda_1 \sum_{i=1}^{n} c_i^2 = \lambda_1\|\boldsymbol{x}\|_2^2
\end{aligned}$$

When $\|\boldsymbol{x}\|_2 \neq 0$,

$$\frac{\|\boldsymbol{Ax}\|_2}{\|\boldsymbol{x}\|_2} \leqslant \sqrt{\lambda_1}. \tag{3.29}$$

Let $\tilde{\boldsymbol{x}} = \boldsymbol{\mu}_1$, then $\|\tilde{\boldsymbol{x}}\|_2 = 1$.

$$\|A\boldsymbol{\mu}_1\|_2^2 = \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}_1 = \lambda_1\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1 = \lambda_1$$

Thus

$$\frac{\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|_2}{\|\tilde{\boldsymbol{x}}\|_2} = \sqrt{\lambda_1} \tag{3.30}$$

Since (3.29) and (3.30)

$$\max_{\boldsymbol{x}\in\mathbb{R}^n} \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sqrt{\lambda_1}. \tag{3.31}$$

Exercise 3.4. *Suppose* $\boldsymbol{A}=\begin{bmatrix}1 & -1\\ 2 & 3\end{bmatrix}$. *Compute* $\|\boldsymbol{A}\|_\infty, \|\boldsymbol{A}\|_1$ *and* $\|\boldsymbol{A}\|_2$

Solution

$$\begin{aligned}&\|\boldsymbol{A}\|_\infty=\max\{1+|-1|,2+3\}=5\\&\|\boldsymbol{A}\|_1=\max\{1+2,|-1|+3\}=4\end{aligned}$$

By $\|\boldsymbol{A}\|_2=\sqrt{\rho\left(\boldsymbol{A}^\top\boldsymbol{A}\right)}$, and

$$\begin{aligned}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}&=\begin{bmatrix}1 & 2\\ -1 & 3\end{bmatrix}\begin{bmatrix}1 & -1\\ 2 & 3\end{bmatrix}=\begin{bmatrix}5 & 5\\ 5 & 10\end{bmatrix}\\ |\lambda\boldsymbol{I}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}|&=\begin{vmatrix}\lambda-5 & -5\\ -5 & \lambda-10\end{vmatrix}=\lambda^2-15\lambda+25=0\end{aligned}$$

The two roots of the characteristic polynomial are solved

$$\lambda_{1.2}=\frac{1}{2}(15\pm\sqrt{125})=\frac{5}{2}(3\pm\sqrt{5}).$$

So

$$\|\boldsymbol{A}\|_2=\sqrt{\lambda_1}=\sqrt{\frac{5}{2}(3+\sqrt{5})}=3.6180340.$$

Theorem 13. *If* $A\in\mathbf{R}^{n\times n}$ *is symmetric matrix, then* $\rho(\boldsymbol{A})=\|\boldsymbol{A}\|_2$.

Proof Since $\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}$, $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{A}^2$. Thus

$$\|\boldsymbol{A}\|_2=\sqrt{\rho\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)}=\sqrt{\rho\left(\boldsymbol{A}^2\right)}=\rho(\boldsymbol{A}).$$

Theorem 14. *Suppose* $\|\cdot\|$ *is a matrix norm in* $\mathbf{R}^{n\times n}$, $\boldsymbol{A}\in\mathbf{R}^{n\times n}$, *we have*

$$\rho(\boldsymbol{A})\leqslant\|\boldsymbol{A}\|$$

Proof Suppose λ is the dominant eigenvalue of A –that is, the eigenvalue with the largest value, and $\boldsymbol{x}$ is eigenvalue vector corresponding to λ. Then

$$A\boldsymbol{x}=\lambda\boldsymbol{x}\tag{3.32}$$

and

$$\rho(\boldsymbol{A})=|\lambda|$$

1° If $\lambda\in\mathbf{R}$, $\boldsymbol{x}\in\mathbf{R}^n$, From (3.32) , we have

$$\|\lambda\boldsymbol{x}\|=\|\boldsymbol{A}\boldsymbol{x}\|$$

For

$$\|\lambda\boldsymbol{x}\|=|\lambda|\|\boldsymbol{x}\|,\quad \|\boldsymbol{A}\boldsymbol{x}\|\leqslant\|\boldsymbol{A}\|\|\boldsymbol{x}\|$$

so

$$|\lambda|\|\boldsymbol{x}\| \leqslant \|\boldsymbol{A}\|\|\boldsymbol{x}\|.$$

Because $\|\boldsymbol{x}\| \neq 0$, we can have

$$|\lambda| \leqslant \|\boldsymbol{A}\|.$$

Thus

$$\rho(\boldsymbol{A}) \leqslant \|\boldsymbol{A}\|$$

2° If λ is complex number, x would be a complex vector. Similarly as 1°, the theorem can be prooved.

From Theorem 14, a matrix norm is a bound of eigenvalues of a matrix.

Theorem 15. *Suppose $\|\cdot\|_p$ and $\|\cdot\|_q$ are matrices norms on $\mathbf{R}^{n\times n}$. There exist two positive constants d_1 and d_2 such that*

$$d_1\|\boldsymbol{A}\|_p \leqslant \|\boldsymbol{A}\|_q \leqslant d_2\|\boldsymbol{A}\|_p$$

for all $\boldsymbol{A} \in \mathbf{R}^{n\times n}$.

Definition 15. *Suppose $\|\cdot\|$ is a matrix norm on $\mathbf{R}^{n\times n}$.*

$$\|\boldsymbol{A}-\boldsymbol{B}\|$$

is said the **distance** *between A and B.*

Definition 16. *Suppose $\|\cdot\|$ is a matrix norm in $\mathbf{R}^{n\times n}$. A sequence of matrices $\boldsymbol{A}^{(0)}, \boldsymbol{A}^{(1)}, \cdots, \boldsymbol{A}^{(k)}, \cdots$ $\mathbf{R}^{n\times n}$ is said* **converge** *to A if*

$$\lim_{k\to\infty}\left\|\boldsymbol{A}^{(k)}-\boldsymbol{A}\right\| = 0.$$

And denote

$$\lim_{k\to\infty}\mathbf{A}^{(k)} = \boldsymbol{A}.$$

Theorem 16. *Suppose $B \in \mathbf{R}^{n\times n}$. The sequence $\boldsymbol{B}^k(k = 0, 1, 2, \cdots)$ converge to $O_{n\times n}$ $\left(\lim\limits_{k\to\infty}\boldsymbol{B}^{(k)} = \boldsymbol{O}\right)$ if and only if $\rho(\boldsymbol{B}) < 1$.*

3.2.3 Error analysis

Exercise 3.5.

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

has solutions $x_1 = x_2 = 1$. If there is small perturbations in coefficient matrix, it becomes

$$\begin{bmatrix} 1 & -1 \\ 1 & 1.0005 \end{bmatrix}\begin{bmatrix} \tilde{x}_1 \\ \widetilde{x}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

And the solutions are

$$\tilde{x}_1 = \tilde{x}_2 = \frac{2}{2.0005} = 0.99975006.$$

It can be seen the small perturbations in A lead to small perturbations in x.

Exercise 3.6. $x_1 = x_2 = 1$ *are solutions of*

$$\begin{bmatrix} 10 & -10 \\ -1 & 1.001 \end{bmatrix} \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.001 \end{bmatrix}$$

If there is same perturbations in coefficient matrix, i.e.

$$\begin{bmatrix} 10 & -10 \\ -1 & 1.0015 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.001 \end{bmatrix}$$

And the solutions are

$$\tilde{x}_1 = \tilde{x}_2 = \frac{2}{3} = 0.66666667.$$

We can see that small relative perturbation induces a greater perturbation in solution.

The solution $\boldsymbol{x}^*$ of linear system $\boldsymbol{Ax} = \boldsymbol{b}$ is determined by the coefficient matrix $\boldsymbol{A}$ and $\boldsymbol{b}$. We will analyze how a small perturbation on A or $\boldsymbol{b}$ influence the exact solution $\boldsymbol{x}^*$.

In the following discussion, suppose $\boldsymbol{A}$ is non-singular and $\boldsymbol{b} \neq \boldsymbol{0}$. Then $\boldsymbol{x}^* \neq \boldsymbol{0}$.

(1) If $\boldsymbol{b}$ has a smaller perturbation $\boldsymbol{\delta b}$, then the solutions become $\boldsymbol{x}^* + \boldsymbol{\delta x}^*$, i.e.

$$\boldsymbol{A}\left(\boldsymbol{x}^* + \boldsymbol{\delta x}^*\right) = \boldsymbol{b} + \boldsymbol{\delta b}.$$

Noticing that

$$\boldsymbol{A x}^* = \boldsymbol{b}, \tag{3.33}$$

we have

$$\boldsymbol{A \delta x}^* = \boldsymbol{\delta b}$$

i.e.

$$\boldsymbol{\delta x}^* = \boldsymbol{A}^{-1} \boldsymbol{\delta b}.$$

Thus

$$\|\boldsymbol{\delta x}^*\| \leqslant \left\|\boldsymbol{A}^{-1}\right\| \cdot \|\boldsymbol{\delta b}\|.$$

From (3.33), we can obtain

$$\|\boldsymbol{b}\| \leqslant \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}^*\| .$$

From the above equations, there is

$$\frac{\|\boldsymbol{\delta x}^*\|}{\|\boldsymbol{x}^*\|} \leq \frac{\left\|\boldsymbol{A}^{-1}\right\| \cdot \|\boldsymbol{\delta b}\|}{\frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}\|}} = \left\|\boldsymbol{A}^{-1}\right\| \cdot \|\boldsymbol{A}\| \cdot \frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}. \tag{3.34}$$

(2) If there is a smaller perturbation on $\boldsymbol{A}$, i.e. $\boldsymbol{\delta A}$. Then

$$(\boldsymbol{A}+\boldsymbol{\delta A})\left(\boldsymbol{x}^*+\boldsymbol{\delta x}^*\right)=\boldsymbol{b}$$

From (3.33) , we get

$$\boldsymbol{\delta}\mathbf{A}\left(\boldsymbol{x}^*+\boldsymbol{\delta x}^*\right)+\boldsymbol{A\delta x}^*=\mathbf{0} \tag{3.35}$$

or

$$\boldsymbol{\delta x}^*=-\boldsymbol{A}^{-1}\cdot\boldsymbol{\delta}\mathbf{A}\left(\boldsymbol{x}^*+\boldsymbol{\delta x}^*\right)$$

Then

$$\|\boldsymbol{\delta x}^*\| \leqslant \|\boldsymbol{A}^{-1}\| \cdot \|\boldsymbol{\delta}\mathbf{A}\| \cdot \|\boldsymbol{x}^*+\boldsymbol{\delta x}^*\|. \tag{3.36}$$

It is easy to know $\boldsymbol{x}^*+\boldsymbol{\delta x}^* \neq 0$. If not, from $\boldsymbol{x}^*+\boldsymbol{\delta x}^*=0$ and (3.35) we have $\boldsymbol{\delta x}^*=0$. Then $\boldsymbol{x}^*=\mathbf{0}$, which contract with $\boldsymbol{x}^* \neq \mathbf{0}$. Thus

$$\frac{\|\boldsymbol{\delta x}^*\|}{\|\boldsymbol{x}^*+\boldsymbol{\delta x}^*\|} \leqslant \left\|\boldsymbol{A}^{-1}\right\| \cdot \|\boldsymbol{\delta A}\| = \left\|\boldsymbol{A}^{-1}\right\| \cdot \|\boldsymbol{A}\| \cdot \frac{\|\boldsymbol{\delta}\mathbf{A}\|}{\|\boldsymbol{A}\|}. \tag{3.37}$$

It can be seen the relative error of solutions is less than $\left\|\boldsymbol{A}^{-1}\right\|\|\boldsymbol{A}\|$ times the relative error of $\boldsymbol{A}$ or $\boldsymbol{b}$ from (3.34) and (3.37) .If $\left\|\boldsymbol{A}^{-1}\right\|\|\boldsymbol{A}\|$ is greater, the relative error will be much greater although smaller δb or δA.

Definition 17. *Suppose A is non-singular. $\left\|A^{-1}\right\|\|A\|$ is said* **condition number** *and denote*

$$\operatorname{cond}(\boldsymbol{A})=\left\|\boldsymbol{A}^{-1}\right\|\|\boldsymbol{A}\|.$$

It is clear the condition number is related to what the matrix norm is. In general,

(1) $\operatorname{cond}(\boldsymbol{A})_\infty=\left\|\boldsymbol{A}^{-1}\right\|_\infty\|\boldsymbol{A}\|_\infty$;

(2) The spectral condition number $\boldsymbol{A}$ is

$$\operatorname{cond}(\boldsymbol{A})_2=\left\|\boldsymbol{A}^{-1}\right\|_2\|\boldsymbol{A}\|_2=\sqrt{\frac{\lambda_{\max}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)}{\lambda_{\min}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)}}.$$

If A is symmetric and positive definite

$$\operatorname{cond}(\boldsymbol{A})_2=\frac{\lambda_1}{\lambda_n}$$

where $\lambda_{\max}\left(\mathbf{A}^{\mathrm{T}}\boldsymbol{A}\right)$ and $\lambda_{\min}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)$ are the maximum and minimum eigenvalues of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, and λ_1 and λ_n are the maximum and minimum eigenvalues of A.

Definition 18. *For* $Ax = b$*,* A *is non-sigular. If* A *is with large condition number i.e.* $\operatorname{cond}(\boldsymbol{A}) \gg 1$*,* $\boldsymbol{Ax} = \boldsymbol{b}$ *is* **ill conditioned***. If the condition number of* $\boldsymbol{A}$ *is moderate,* $Ax = b$ *is said to be* **well conditioned**.

Exercise 3.7. *Compute the condition numbers of the coefficient matrices in Example 3.5 and Example 3.6.*

Solution

In Example 3.5 ,

$$\boldsymbol{A} = \left[\begin{array}{cc} 1 & -1 \\ 1 & 1 \end{array}\right], \quad \boldsymbol{\delta}\mathbf{A} = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0.0005 \end{array}\right], \quad \boldsymbol{\delta b} = \left[\begin{array}{c} 0 \\ 0 \end{array}\right], \quad \boldsymbol{A}^{-1} = \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{array}\right]$$

$$\|\boldsymbol{A}\|_\infty = 2, \quad \left\|\boldsymbol{A}^{-1}\right\|_\infty = 1, \quad \operatorname{cond}(\boldsymbol{A})_\infty = 2$$

Thus it is well conditioned.

From (3.37) ,

$$\frac{\|\boldsymbol{\delta x}^*\|_\infty}{\|\boldsymbol{x}^* + \boldsymbol{\delta x}^*\|_\infty} \leqslant \operatorname{cond}(\boldsymbol{A})_\infty \frac{\|\boldsymbol{\delta}\mathbf{A}\|_\infty}{\|\boldsymbol{A}\|_\infty} = 2 \times \frac{0.0005}{2} = 0.05\%$$

we can see that the approximation is accurate to the exact solutions.

In Example 3.6,

$$\boldsymbol{A} = \left[\begin{array}{cc} 10 & -10 \\ -1 & 1.001 \end{array}\right], \quad \boldsymbol{\delta}\mathbf{A} = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0.0005 \end{array}\right], \quad \boldsymbol{\delta b} = \left[\begin{array}{c} 0 \\ 0 \end{array}\right]$$

$$\boldsymbol{A}^{-1} = \left[\begin{array}{cc} 100.1 & 1000 \\ 100 & 1000 \end{array}\right]$$

$$\|\boldsymbol{A}\|_\infty = 20, \left\|\boldsymbol{A}^{-1}\right\|_\infty = 1100.1, \quad \text{cond } (\boldsymbol{A})_\infty = 22002$$

Thus, the linear system is ill conditioned.

From (3.37) , there is

$$\tfrac{\|\boldsymbol{\delta x}^*\|}{\|\boldsymbol{x}^* + \boldsymbol{\delta x}^*\|} \leqslant \operatorname{cond}(A)_\infty \tfrac{\|\boldsymbol{\delta A}\|_\infty}{\|\boldsymbol{A}\|_\infty} = 22002 \times \tfrac{0.0005}{20} = 55.005\%$$

The relative error of the solution can reach 55%.

If the linear equations are ill conditioned, pretreatment algorithm is a good way to solve the system, i.e. non-singular diagonal matrices $\boldsymbol{D}$ and $\boldsymbol{C}$, such that condition number of $\boldsymbol{DAC}$ in the equivalent linear equations $\mathbf{DAC}\left[\boldsymbol{C}^{-1}\boldsymbol{x}\right] = \boldsymbol{Db}$, can be improved.

Exercise 3.8. *Suppose*

$$\left[\begin{array}{cc} 1 & 10^4 \\ 1 & 1 \end{array}\right]\left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \left[\begin{array}{c} 10^4 \\ 2 \end{array}\right]$$

Let $\boldsymbol{Ax} = \boldsymbol{b}$*, and compute* $\operatorname{cond}(\boldsymbol{Ax})_\infty$*. Try to improve the condition number* $\operatorname{cond}(\widetilde{A})_\infty$ *in the equivalent system* $\widetilde{A}x = \tilde{b}$*.*

Solution

$$\boldsymbol{A}^{-1}=\frac{1}{10^4-1}\begin{bmatrix}-1 & 10^4\\ 1 & -1\end{bmatrix}$$

So

$$\operatorname{cond}(\boldsymbol{A})_\infty=\frac{\left(1+10^4\right)^2}{10^4-1}\approx 10^4.$$

Take $\boldsymbol{D}=\begin{bmatrix}10^{-4} & 0\\ 0 & 1\end{bmatrix}$, and $\tilde{\boldsymbol{A}}=\boldsymbol{D}\boldsymbol{A}$ is

$$\tilde{\boldsymbol{A}}=\begin{bmatrix}10^{-4} & 1\\ 1 & 1\end{bmatrix},\quad \tilde{\boldsymbol{A}}^{-1}=\frac{1}{1-10^{-4}}\begin{bmatrix}-1 & 1\\ 1 & -10^{-4}\end{bmatrix}$$

Then

$$\operatorname{cond}(\tilde{\boldsymbol{A}})_\infty=\frac{4}{1-10^{-4}}\approx 4.$$

Solve the following equations by Gaussian elimination with partial pivoting

$$\widetilde{A}x=\tilde{b}$$

and get the solutions $\boldsymbol{x}=\begin{bmatrix}1\\ 1\end{bmatrix}$, which are better approiximations.

In (3.34) and (3.37), the relative error of the approximation can be bouned which is theoretical and always be enlarged in general. Moreover A^{-1} is difficult to be obtained, and $\boldsymbol{\delta}\mathbf{A}$ and $\boldsymbol{\delta b}$ is a hypothesis in theory. In fact, they are random, so the above estimation is not adaptable. A better way is to estimate the **residual** r :

$$\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}. \tag{3.38}$$

If $\boldsymbol{r}=0$, $\tilde{\boldsymbol{x}}$ is exact solution. Generally speaking $\boldsymbol{r}\neq 0$.

Theorem 17. *Let*
tilde$\boldsymbol{x}$ is an approximation of solutions of $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ and the exact solutions $\boldsymbol{x}^*,\boldsymbol{r}$ is the residual of $\boldsymbol{x}$, we have

$$\frac{\|\boldsymbol{x}^*-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}^*\|}\leqslant\operatorname{cond}(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|} \tag{3.39}$$

Proof Since

$$\boldsymbol{A}\boldsymbol{x}^*=\boldsymbol{b},\quad \boldsymbol{A}\left(\boldsymbol{x}^*-\tilde{\boldsymbol{x}}\right)=\boldsymbol{r},$$

we have

$$\|\boldsymbol{b}\|=\|\boldsymbol{A}\boldsymbol{x}^*\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}^*\|$$
$$\|\boldsymbol{x}^*-\tilde{\boldsymbol{x}}\|=\left\|\boldsymbol{A}^{-1}\boldsymbol{r}\right\|\leqslant\left\|\boldsymbol{A}^{-1}\right\|\,\|\boldsymbol{r}\|$$

Hence

$$\frac{\|\boldsymbol{x}^*-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}^*\|}\leqslant\frac{\left\|\boldsymbol{A}^{-1}\right\|\,\|\boldsymbol{r}\|\|\boldsymbol{A}\|}{\|\boldsymbol{b}\|}=\operatorname{cond}(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}.$$

From Theorem 17, if cond($\boldsymbol{A}$) is bigger, the relative error would be larger although smaller $\|r\|$.

For example, the matrix $A = \begin{bmatrix} 1.000 & 1.001 \\ 1.000 & 1.000 \end{bmatrix}$, and

$$\boldsymbol{A}^{-1} = \begin{bmatrix} -1000 & 1001 \\ 1000 & -1000 \end{bmatrix}$$

The condition number is

$$\text{cond}(\boldsymbol{A})_\infty = \left\|\boldsymbol{A}^{-1}\right\|_\infty \|\boldsymbol{A}\|_\infty = 4004.001.$$

$\boldsymbol{A}$ is ill conditioned.

For the linear system,

$$\begin{bmatrix} 1.000 & 1.001 \\ 1.000 & 1.000 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 2.001 \\ 2.000 \end{bmatrix}$$

its exact solutions are $\boldsymbol{x}^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Suppose the approximated solutions are $\tilde{\boldsymbol{x}} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, which is far away from $\boldsymbol{x}$ *. But the residual $\boldsymbol{r}(\tilde{\boldsymbol{x}}) = \begin{bmatrix} 0.001 \\ 0 \end{bmatrix}$ is very small.

3.3 Iterative Methods

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.

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

Consider

$$\boldsymbol{Ax} = \boldsymbol{b}, \tag{3.40}$$

where $\boldsymbol{A}$ is assumed to be non-singular, so it has unique solution $\boldsymbol{x}^*$.

Iterative techniques involve a process that converts the system (3.40) into an equivalent linear systems of the form

$$\boldsymbol{x} = \boldsymbol{Bx} + \boldsymbol{f} \tag{3.41}$$

for some fixed matrix $\boldsymbol{B} \in \mathbf{R}^{n\times n}$ and vector $\boldsymbol{f} \in \mathbf{R}^n$.

After the initial approximation $x^{(0)} \in \mathbf{R}^n$ is selected, the sequence of approximate solution vectors $\left\{\boldsymbol{x}^{(k)}\right\}_{k-0}$ is generated by computing

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{Bx}^{(k)} + \boldsymbol{f} \quad (k = 0, 1, 2, \cdots). \tag{3.42}$$

If

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*,$$

then

$$\boldsymbol{x}^* = \boldsymbol{Bx}^* + \boldsymbol{f},$$

which is the solution of (3.40). Eq. (3.42) is said **iterative scheme** and $\boldsymbol{B}$ is **iterative matrix**. Let $\boldsymbol{e}^{(k)} = \boldsymbol{x}^* - \boldsymbol{x}^{(k)}$ denote the error vector of the kth iteration. An iterative scheme (3.42) is said to be **convergent** if the sequence of vectors $\left\{\boldsymbol{x}^{(k)}\right\}_{k=0}^{\infty}$ generated by (3.42) is convergent for any initial guess $\boldsymbol{x}^{(0)}$.

Next we will discuss the following topics:

(1) How to construct the iterative scheme (3.42)?

(2) What is the convergent condition of (3.42)?

3.3.1 Jacobi iterative method

The equations of (3.40) can be rewritten in the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}$$

Solve the ith equation for x_i, (provided $a_{ii} \neq 0(i = 1, 2, \cdots, n)$)

$$\begin{cases} x_1 = (b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n)/a_{11} \\ x_2 = (b_2 - a_{21}x_1 - a_{23}x_3 - \cdots - a_{2n}x_n)/a_{22} \\ \qquad\qquad\vdots \\ x_n = (b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n,n-1}x_{n-1})/a_{nn} \end{cases}$$

and generate each x_i^{k+1} from x_i^k for $k \geq 0$ by

$$\begin{cases} x_1^{(k+1)} = \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)}\right)/a_{11} \\ x_2^{(k+1)} = \left(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}\right)/a_{22} \\ \qquad\qquad\vdots \\ x_n^{(k+1)} = \left(b_n - a_{n1}x_1^{(k)} - a_{n2}x_2^{(k)} - \cdots - a_{n,n-1}x_{n-1}^{(k)}\right)/a_{nn} \end{cases} \tag{3.43}$$

The method (3.43) is called **Jacobi iterative method**.

Exercise 3.9. *Use the Jacobi method to solve the following linear system*

$$\begin{bmatrix} 8 & -1 & 1 \\ 2 & 10 & -1 \\ 1 & 1 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}$$

and the approximation solutions are with 3 significant figures.

Solution Jacobi iteration is

$$\begin{cases} x_1^{(k+1)} = \left(1 + x_2^{(k)} - x_3^{(k)}\right)/8 \\ x_2^{(k+1)} = \left(4 - 2x_1^{(k)} + x_3^{(k)}\right)/10 \\ x_3^{(k+1)} = \left(3 - x_1^{(k)} - x_2^{(k)}\right)/(-5) \end{cases}$$

Let $x^{(0)} = (0, 0, 0)^{\mathrm{T}}$, and the computed results are listed as followed:

The approximation solutions $x^* = (0.225, 0.306, -0.494)^{\mathrm{T}}$ with 3 significant figures.

k	0	1	2	3	4	5	6
$x_1^{(k)}$	0.0000	0.1250	0.2500	0.2263	0.2235	0.2251	0.2250
$x_2^{(k)}$	0.0000	0.4000	0.3150	0.3005	0.3060	0.3058	0.3056
$x_3^{(k)}$	0.0000	-0.6000	-0.4950	-0.4870	-0.4946	-0.4941	-0.4938

Split the matrix $\boldsymbol{A}$ to $\boldsymbol{L}+\boldsymbol{D}+\boldsymbol{U}$, where

$$\boldsymbol{L}=\begin{bmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{bmatrix},$$

$$\boldsymbol{D}=\begin{bmatrix} a_{11} & & & & \\ & a_{22} & & & \\ & & a_{33} & & \\ & & & \ddots & \\ & & & & a_{nn} \end{bmatrix}, \boldsymbol{U}=\begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ & 0 & a_{23} & \cdots & a_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & 0 & a_{n-1,n} \\ & & & & 0 \end{bmatrix}$$

Thus, we have

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

or

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

$\boldsymbol{D}$ is non-singular since $a_{ii}\neq 0(i=1,2,\cdots,n)$. Hence

$$\boldsymbol{x}=\boldsymbol{D}^{-1}[\boldsymbol{b}-(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{x}],$$

and the iterative scheme is

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{D}^{-1}\left[\boldsymbol{b}-(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{x}^{(k)}\right]$$

It is **Jacobi iteration** in matrix form.Let

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{J}\boldsymbol{x}^{(k)}+\boldsymbol{f}_J,$$

where

$$\boldsymbol{J}=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U}),\quad \boldsymbol{f}_J=\boldsymbol{D}^{-1}\boldsymbol{b},$$

where $\boldsymbol{J}$ is Jacobi iterative matrix.

3.3.2 Gauss-Seidel Iterative method

In Jacobi iteration, the components $x_1^{(k)}$, $x_2^{(k)}, \cdots, x_n^{(k)}$ are used to compute $\boldsymbol{x}^{k+1}$. For $i > 1$, the components $x_1^{(k+1)}, x_2^{(k+1)}, \cdots, x_{i-1}^{(k+1)}$ have been computed and are expected to be better approximations to the actual solutions $x_1, \cdots, x_i$ than are $x_1^{(k-1)}, \cdots, x_{i-1}^{(k-1)}$. It seems reasonable, then, to compute $x_i^{(k)}$ using these most recently calculated values. That is

$$\begin{cases} x_1^{(k+1)} = \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)}\right)/a_{11} \\ x_2^{(k+1)} = \left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}\right)/a_{22} \\ \qquad\vdots \\ x_n^{(k+1)} = \left(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \cdots - a_{n,n-1}x_{n-1}^{(k+1)}\right)/a_{nn} \end{cases} \tag{3.44}$$

This modification is called the **Gauss-Seidel iteration** and it can be represented in the matrix form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{D}^{-1}\left(\boldsymbol{b} - \boldsymbol{L}\boldsymbol{x}^{(k+1)} - \boldsymbol{U}\boldsymbol{x}^{(k)}\right) \tag{3.45}$$

Then

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

Since $\boldsymbol{D}+\boldsymbol{L}$ is non-singular, we have

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{f}_G \tag{3.46}$$

where

$$\boldsymbol{G} = -(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{U}, \quad f_G = (\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{b}, \tag{3.47}$$

and G is Gauss-Seidel iterative matrix.

Exercise 3.10. *Use Gauss-Seidel iteration to solve the following linear equations*

$$\begin{bmatrix} 8 & -1 & 1 \\ 2 & 10 & -1 \\ 1 & 1 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}$$

and the approximation solutions are with 3 significant figures.

Solution The Gauss-Seidel iterative technique is

$$\begin{cases} x_1^{(k+1)} = \left(1 + x_2^{(k)} - x_3^{(k)}\right)/8 \\ x_2^{(k+1)} = \left(4 - 2x_1^{(k+1)} + x_3^{(k)}\right)/10 \\ x_3^{(k+1)} = \left(3 - x_1^{(k+1)} - x_2^{(k+1)}\right)/(-5) \end{cases}$$

Take $\boldsymbol{x}^{(0)} = (0,0,0)^{\mathrm{T}}$ as staring vector, the computed results are listed Thus, the approximation solutions $x^* = (0.225, 0.306, -0.494)^{\mathrm{T}}$ are with 3 significant figures.

k	0	1	2	3	4
$x_1^{(k)}$	0.0000	0.1250	0.2344	0.2245	0.2250
$x_2^{(k)}$	0.0000	0.3750	0.3031	0.3059	0.3056
$x_3^{(6)}$	0.0000	-0.5000	-0.4925	-0.4939	-0.4936

3.3.3 Successive Over-relaxation method

The method called **Successive Over-Relaxation (SOR)** takes the Gauss–Seidel direction toward the solution and “overshoots” to try to speed convergence. Let ω be a real number, and define each component of the new guess $\boldsymbol{x}^{k+1}$ as a weighted average of ω times the Gauss–Seidel formula and $1-\omega$ times the current guess $\boldsymbol{x}^k$, i.e.

$$\begin{cases} x_1^{(k+1)} = (1-\omega)x_1^{(k)} + \omega\left(b_1 - a_{12}x_2^{(k)} - a_{13}a_3^{(k)} - \cdots - a_{1n}x_n^{(k)}\right)/a_{11} \\ x_2^{(k+1)} = (1-\omega)x_2^{(k)} + \omega\left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}\right)/a_{22} \\ \qquad\vdots \\ x_n^{(k+1)} = (1-\omega)x_n^{(k)} + \omega\left(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \cdots - a_{n,n-1}x_{n-1}^{(k+1)}\right)/a_{nn} \end{cases} \tag{3.48}$$

where ω is called **relaxation parameter**. If $\omega > 1$, it is referred to over-relaxation; If $0 < \omega < 1$,it is referred to under-relaxation; In special, $\omega = 1$, it is exactly Gauss-Seidel iteration. If the appropriate value of ω is chosen, SOR method is converging to solution vector faster than Gauss-Seidel iteration.

We rewrite (3.48) in the matrix form:

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega\boldsymbol{D}^{-1}\left(\boldsymbol{b} - \boldsymbol{L}\boldsymbol{x}^{(k+1)} - \boldsymbol{U}\boldsymbol{x}^{(k)}\right).$$

Multiply $\boldsymbol{D}$ on the above equation, there is

$$\boldsymbol{D}\boldsymbol{x}^{(k+1)} = (1-\omega)\boldsymbol{D}\boldsymbol{x}^{(k)} + \omega\left(\boldsymbol{b} - \boldsymbol{L}\boldsymbol{x}^{(k+1)} - \boldsymbol{U}\boldsymbol{x}^{(k)}\right).$$

Then

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

Multiplying $(\boldsymbol{D} + \omega\boldsymbol{L})^{-1}$, we have

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{S}_\omega\boldsymbol{x}^{(k)} + \boldsymbol{f}_\omega \tag{3.49}$$

where

$$\boldsymbol{S}_\omega = (\boldsymbol{D} + \omega\boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D} - \omega\boldsymbol{U}], \quad \boldsymbol{f}_\omega = \omega(\boldsymbol{D} + \boldsymbol{\omega}\boldsymbol{L})^{-1}\boldsymbol{b} \tag{3.50}$$

3.3.4 Convergency of Iterative Method

Theorem 18. *The iterative scheme* (3.42) *is convergent if and only if* $\rho(\boldsymbol{B}) < 1$.

Proof

1° (Necessity) Suppose $\left\{\boldsymbol{x}^{(k)}\right\}$ generated by the iteration (3.42) converging to $\boldsymbol{x}^*$. Taking limit on (3.42) ($k \to \infty$), we have

$$\boldsymbol{x}^* = \boldsymbol{B}\boldsymbol{x}^* + \boldsymbol{f}.$$

Let

$$\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{x}^* - \boldsymbol{x}^{(k)},$$

then

$$\begin{aligned}\boldsymbol{\varepsilon}^{(k)} &= \boldsymbol{x}^* - \boldsymbol{x}^{(k)} = (\boldsymbol{B}\boldsymbol{x}^* + \boldsymbol{f}) - \left(\boldsymbol{B}\boldsymbol{x}^{(k-1)} + \boldsymbol{f}\right) \\ &= \boldsymbol{B}\left(\boldsymbol{x}^* - \boldsymbol{x}^{(k-1)}\right) \quad (k = 0, 1, 2, \cdots)\end{aligned}$$

By recursion, there is

$$\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{B}\boldsymbol{\varepsilon}^{(k-1)} = \cdots = \boldsymbol{B}^k \boldsymbol{\varepsilon}^{(0)} \quad (k = 0, 1, 2, \cdots).$$

Since for any $x^{(0)}$ and $\boldsymbol{\varepsilon}^{(0)}$, $\lim\limits_{k\to\infty} \boldsymbol{B}^k \boldsymbol{\varepsilon}^{(0)} = \boldsymbol{O}$, we have $\lim\limits_{k\to\infty} \boldsymbol{B}^k = \boldsymbol{O}$.

From Theroem16, $\rho(\boldsymbol{B}) < 1$

2° (Sufficient) $If \rho(\boldsymbol{B}) < 1$, then $|\boldsymbol{I} - \boldsymbol{B}| \neq 0$ i.e. $(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{x} = f$ has unique solution x^* satisfy

$$\boldsymbol{x}^* = \boldsymbol{B}\boldsymbol{x}^* + \boldsymbol{f}.$$

Hence

$$\begin{aligned}\boldsymbol{\varepsilon}^{(k)} &= \boldsymbol{x}^* - \boldsymbol{x}^{(k)} = (\boldsymbol{B}\boldsymbol{x}^* + \boldsymbol{f}) - \left(\boldsymbol{B}\boldsymbol{x}^{(k-1)} + \boldsymbol{f}\right) \\ &= B\left(\boldsymbol{x}^* - \boldsymbol{x}^{(k-1)}\right) = B\boldsymbol{\varepsilon}^{(k-1)} \quad (k = 0, 1, 2, \cdots)\end{aligned}$$

By rerecursion, there is

$$\boldsymbol{\varepsilon}^{(k)} = B^k \boldsymbol{\varepsilon}^{(0)} \quad (k = 0, 1, 2, \cdots)$$

From Theorem 16 , we obtain

$$\lim_{k\to\infty} \boldsymbol{B}^k = \boldsymbol{O}.$$

So

$$\lim_{k\to\infty} \boldsymbol{\varepsilon}^{(k)} = \boldsymbol{O}.$$

i.e.

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*.$$

From Theorem 14, $\rho(\boldsymbol{B}) \leqslant \|\boldsymbol{B}\|$. If $\|\boldsymbol{B}\| < 1$, then $\rho(\boldsymbol{B}) < 1$.

Theorem 19. *If* $\|B\| < 1$, *the iteration* (3.5.3) *is convergent.*

According to the proof of Theorem 2, it is easy to get the following corollary.

Corollary 1. *If* $\|\boldsymbol{B}\| < 1$, *then*

$$\left\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\right\| \leqslant \frac{\|\boldsymbol{B}\|}{1-\|\boldsymbol{B}\|}\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\| \quad (k = 1, 2, 3, \cdots)$$

$$\left\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\right\| \leqslant \frac{\|\boldsymbol{B}\|^k}{1-\|\boldsymbol{B}\|}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\right\| \quad (k = 1, 2, 3, \cdots)$$

$$\left\|\boldsymbol{x}^* - \boldsymbol{x}^{(k+1)}\right\| \leqslant \|\boldsymbol{B}\|\left\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\right\| \quad (k = 0, 1, 2, \cdots)$$

(1) Convergency of Jacobi iterative method

The Jacobi iterative matrix is

$$\boldsymbol{J} = -\boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U}),$$

and its characteristic equation is

$$|\lambda \boldsymbol{I} - \boldsymbol{J}| = 0,$$

i.e.

$$\left|\lambda \boldsymbol{I} + \boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U})\right| = 0.$$

We rewrite the above equation in the form

$$\left|\boldsymbol{D}^{-1}\right| \cdot |\lambda \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}| = 0.$$

Since $\left|\boldsymbol{D}^{-1}\right| \neq 0$, then

$$|\lambda \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}| = 0, \tag{3.51}$$

which is the characteristic equation of Jacobi iterative matrix.

From Theorem 18, Jacobi iterative method is convergent if and only if $\rho(\boldsymbol{J}) < 1$.

Exercise 3.11. *Show Jacobi iteration is convergent or not to solve the linear system:*

$$\begin{bmatrix} 8 & -1 & 1 \\ 2 & 10 & -1 \\ 1 & 1 & -5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}.$$

Solution The characteristic equation of Jacobi iterative matrix $\boldsymbol{J}$ is

$$\begin{vmatrix} 8\lambda & -1 & 1 \\ 2 & 10\lambda & -1 \\ 1 & 1 & -5\lambda \end{vmatrix} = 0.$$

Then

$$400\lambda^3 + 12\lambda - 3 = 0.$$

The only real root $\lambda_1 = 0.146084$ obtained by Newton's method. By Vieta theorem, the other two roots λ_2 and λ_3 satisfy $\lambda_3 = \bar{\lambda}_2$, and $\lambda_1\lambda_2\lambda_3 = \frac{3}{400}$.

Then

$$|\lambda_2| = |\lambda_3| = \sqrt{\frac{3}{400\lambda_1}} = 0.226584,$$
$$\rho(\boldsymbol{J}) = \max\{|\lambda_1|, |\lambda_2|, |\lambda_3|\} = 0.226584 < 1.$$

Hence the Jacobi iteration is convergent.

Theorem 20. *Jacobi iteration is convergent for the linear system $\boldsymbol{Ax} = \boldsymbol{b}$, which coefficient matrix $\boldsymbol{A}$ is strictly diagonally dominant.*

Proof Let

$$\boldsymbol{B}(\lambda) = \begin{bmatrix} \lambda a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \lambda a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & \lambda a_{nn} \end{bmatrix},$$

and the characteristic equation of Jacobi iterative matrix $\boldsymbol{J}$ is

$$|\boldsymbol{B}(\lambda)| = 0.$$

Since $\boldsymbol{A}$ is strictly diagonally dominant, there is

$$|\lambda a_{ii}| \geqslant |a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad (i = 1, 2, \cdots, n)$$

for $|\lambda| \geqslant 1$. Hence $\boldsymbol{B}(\lambda) = 0$ is also strictly diagonally dominant for $|\lambda| \geqslant 1$ result in $\boldsymbol{B}(\lambda) \neq 0$. The n roots $\lambda_1, \lambda_2, \cdots, \lambda_n$ of $|\boldsymbol{B}(\lambda)| = 0$ should satisfy $|\lambda_i| < 1 (i = 1, 2, \cdots, n)$. Then $\rho(\boldsymbol{J}) < 1$ and the Jacobi method is convergent.

(2) Convergency of Gauss-Seidel Method

The characteristic equation of Gauss-Seidel iterative matrix is

$$|\lambda \boldsymbol{I} - \boldsymbol{G}| = 0,$$

where

$$\boldsymbol{G} = -(\boldsymbol{D} + \boldsymbol{L})^{-1}\boldsymbol{U},$$

Then we rewrite the above equation in the form

$$\left|(\boldsymbol{D} + \boldsymbol{L})^{-1}\right| \cdot |\lambda(\boldsymbol{D} + \boldsymbol{L}) + \boldsymbol{U}| = 0.$$

Since $\left|(\boldsymbol{D} + \boldsymbol{L})^{-1}\right| \neq 0$,

$$|\lambda(\boldsymbol{D} + \boldsymbol{L}) + \boldsymbol{U}| = 0. \tag{3.52}$$

From Theorem 18, Gauss-Seidel iterative method converges if and only if $\rho(\boldsymbol{G}) < 1$, i.e. the module of all the roots of (3.52) less than 1.

Exercise 3.12. *Show the Gauss-Seidel iteration is convergent or not to solve the following linear system*

$$\begin{bmatrix} 8 & -1 & 1 \\ 2 & 10 & -1 \\ 1 & 1 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}.$$

Solution The characteristic equation of Gauss-Seidel iterative matrix $\boldsymbol{G}$ is

$$\begin{vmatrix} 8\lambda & -1 & 1 \\ 2\lambda & 10\lambda & -1 \\ \lambda & \lambda & -5\lambda \end{vmatrix} = 0.$$

Then we have

$$\lambda\left(400\lambda^2 + 10\lambda - 1\right) = 0.$$

The roots of the above equation are

$$\lambda_1 = 0, \quad \lambda_2 = \frac{-1+\sqrt{17}}{80}, \quad \lambda_3 = \frac{-1-\sqrt{17}}{80}.$$

Thus

$$\rho(\boldsymbol{G}) = \max\left\{|\lambda_1|, |\lambda_2|, |\lambda_3|\right\} = \frac{1+\sqrt{17}}{80} = 0.0640388 < 1,$$

result that Gauss-Seidel iteration is convergent.

Theorem 21. *Gauss-Seidel iteration is convergent for the linear system* $\boldsymbol{Ax} = \boldsymbol{b}$, *which coefficient matrix* $\boldsymbol{A}$ *is strictly diagonally dominant.*

Proof Let

$$\boldsymbol{C}(\lambda) = \begin{bmatrix} \lambda a_{11} & a_{12} & \cdots & a_{1n} \\ \lambda a_{31} & \lambda a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \lambda a_{n1} & \lambda a_{n2} & \cdots & \lambda a_{nn} \end{bmatrix},$$

and the characteristic equation of Gauss-Seidel iterative matrix $\boldsymbol{G}$ is

$$|\boldsymbol{C}(\lambda)| = 0.$$

Since $\boldsymbol{A}$ is strictly diagonally dominant,

$$\begin{aligned} |\lambda a_{ii}| = |\lambda|\,|a_{ii}| &> |\lambda| \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \\ &\geqslant |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \quad (i = 1, 2, \cdots, n) \end{aligned}$$

for $|\lambda| \geqslant 1$. Hence $\boldsymbol{C}(\lambda)$ is also is strictly diagonally dominant. That is $|\boldsymbol{C}(\lambda)|$ for $|\lambda| \geqslant 1. \neq 0$. So the n roots $\lambda_1, \lambda_2, \cdots, \lambda_n$ of the linear equations $|\boldsymbol{C}(\lambda)| = 0$ satisfy $|\lambda_i| < 1 (i = 1, 2, \cdots, n)$. We have $\rho(\boldsymbol{G}) < 1$ and Gauss-Seidel method is convergent.

(3) Convergency of SOR method

The iteration matrix of SOR is

$$\boldsymbol{S}_w = (\boldsymbol{D} + \omega \boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D} - \omega \boldsymbol{U}].$$

It is clear that SOR method is convergent if and only if

$$\rho(\boldsymbol{S}_\omega) < 1.$$

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 22. *The SOR method converge only if* $0 < \omega < 2$.

Proof Suppose $\lambda_1, \lambda_2, \cdots, \lambda_n$ are eigenvalues of $\boldsymbol{S}_\omega$. Then

$$|\det(\boldsymbol{S}_\omega)| = |\lambda_1 \lambda_2 \cdots \lambda_n| \leqslant \rho(\boldsymbol{S}_\omega)^n.$$

SOR method converge if and only if $\rho(\boldsymbol{S}_\omega) < 1$. This implies

$$|\det(\boldsymbol{S}_\omega)| < 1$$

and

$$\begin{aligned}\det(\boldsymbol{S}_w) &= \det\left((\boldsymbol{D} + \omega \boldsymbol{L})^{-1}\right) \cdot \det((1-\omega)\boldsymbol{D} - \omega \boldsymbol{U}) \\ &= \left(\prod_{i=1}^{n} a_{ii}\right)^{-1} \prod_{i=1}^{n} [(1-\omega)a_{ii}] = (1-\omega)^n.\end{aligned}$$

Thus

$$|(1-\omega)^n| < 1,$$

i.e.

$$0 < \omega < 2.$$

In the theorem, $\omega \in (0, 2)$ is necessary that SOR method converges. But if $\omega \in (0, 2)$, SOR method may not converge for any linear system.

Theorem 23. *If $\boldsymbol{A}$ is a positive definite matrix and $\omega \in (0, 2)$, then the SOR method converges for any choice of initial approximate vector $\boldsymbol{x}^{(0)}$.*

The question of choosing ω for the most rapid convergence of the SOR iteration addressed by Young (1950) . If $\boldsymbol{A}$ is positive definite and tridiagonal , then the optimal choice of ω for the SOR method is

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho^2(\boldsymbol{J})}},$$

where $\rho(\boldsymbol{J})$ is the spectral radius of Jacobi iterative matrix. But it is difficult to compute $\rho(\boldsymbol{J})$, the approximation ω_{opt} can be determined by trial method in fact. The trial method is to compare the residuals of $r = b - Ax^{(k)}$ (or $x^{(k)} - x^{(k-1)}\big)$ where $x^{(k)}$ generated by the different ω and same initial guess $x^{(0)}$. If the residual $\|r\|$ is minimum, the ω which is used in the SOR method is the approximation ω_{opt}.

3.3.5 Power Method

The **Power method** is an iterative technique used to determine the dominant eigenvalue of a matrix—that is, the eigenvalue with the largest magnitude. One useful feature of the Power method is that it produces not only an eigenvalue, but also an associated eigenvector. In fact, the Power method is often applied to find an eigenvector for an eigenvalue that is determined by some other means.

Assume that the $n \times n$ matrix $A = (a_{ij})_{n\times n}$ has n eigenvalues $\lambda_j (j = 1, 2, \cdots, n)$, with an associated collection of linearly independent eigenvectors $x_1, x_2, \cdots, x_n$.Moreover, we assume

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant | \lambda_n |$$

where λ_1 is the dominant eigenvalue.

Choose $v_0 \neq 0$, and we can get the iteration

$$\boldsymbol{v}_k = \boldsymbol{A}\boldsymbol{v}_{k-1} \quad (k = 1, 2, \cdots).$$

A sequence of vectors $\{\boldsymbol{v}_k\}$ is generated:

$$\begin{cases} \boldsymbol{v}_1 = \boldsymbol{A}\boldsymbol{v}_0 \\ \boldsymbol{v}_2 = \boldsymbol{A}\boldsymbol{v}_1 = \boldsymbol{A}^2\boldsymbol{v}_0 \\ \qquad \vdots \\ \boldsymbol{v}_k = \boldsymbol{A}\boldsymbol{v}_{k-1} = \cdots = A^k\boldsymbol{v}_0 \end{cases} \tag{3.53}$$

The fact that $v_k (k = 1, 2, \cdots, n)$ is linear independent implies that constants exists with

$$\boldsymbol{v}_0 = \alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_n\boldsymbol{x}_n = \sum_{i=1}^{n} \alpha_i\boldsymbol{x}_i, \tag{3.54}$$

and suppose $\alpha_1 \neq 0$. Taking v_0 into (3.53) , we get

$$\begin{aligned} \boldsymbol{v}_1 = \boldsymbol{A}\boldsymbol{v}_0 &= \boldsymbol{A}(\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_n\boldsymbol{x}_n) \\ &= \alpha_1\lambda_1\boldsymbol{x}_1 + \alpha_2\lambda_2\boldsymbol{x}_2 + \cdots + \alpha_n\lambda_n\boldsymbol{x}_n \\ &= \sum_{i=1}^{n} \alpha_i\lambda_i\boldsymbol{x}_i. \end{aligned}$$

Similarly,

$$\begin{aligned}\boldsymbol{v}_k = \boldsymbol{A}\boldsymbol{v}_{k-1} &= \alpha_1\lambda_1^k\boldsymbol{x}_1 + \alpha_2\lambda_2^k\boldsymbol{x}_2 + \cdots + \alpha_n\lambda_n^k\boldsymbol{x}_n \\ &= \sum_{i=1}^{n}\alpha_i\lambda_i^k\boldsymbol{x}_i \quad (k = 1, 2, \cdots) \qquad (3.55)\end{aligned}$$

(1) $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$

Rewrite (3.55) in the following form:

$$\boldsymbol{v}_k = \lambda_1^k\left[\alpha_1\boldsymbol{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\boldsymbol{x}_2 + \cdots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\boldsymbol{x}_n\right],$$

and

$$\boldsymbol{v}_{k+1} = \lambda_1^{k+1}\left[\alpha_1\boldsymbol{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}\boldsymbol{x}_2 + \cdots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}\boldsymbol{x}_n\right].$$

Since

$$|\lambda_1| > |\lambda_i| \quad (i = 2, 3, \cdots, n),$$

there is

$$\lim_{k\to\infty}\left(\frac{\lambda_i}{\lambda_1}\right)^k = 0 \quad (i = 2, 3, \cdots, n).$$

Thus

$$\boldsymbol{v}_k \approx \alpha_1\lambda_1^k\boldsymbol{x}_1 \qquad (3.56)$$

and

$$\boldsymbol{v}_{k+1} \approx \alpha_1\lambda_1^{k+1}\boldsymbol{x}_1 \approx \lambda_1\boldsymbol{v}_k. \qquad (3.57)$$

The sequence in (3.57) converges to zero if $|\lambda| < 1$ and diverges if $|\lambda| > 1$. As a consequence, the entries in the $A^k\mathbf{x}$ will grow with k if $|\lambda_1| > 1$ and will go to 0 if $|\lambda_1| < 1$, perhaps resulting in overflow or underflow. To take care of that possibility, we scale the powers of $A^k\mathbf{x}$ in an appropriate manner to ensure that the limit of the sequence is finite and nonzero. The scaling begins by choosing $\mathbf{x}$ to be a unit vector $\mathbf{u}^{(0)}$.

$$\begin{cases}\boldsymbol{u}_0 = \boldsymbol{v}_0 \\ \boldsymbol{v}_k = \boldsymbol{A}\boldsymbol{u}_{k-1} \\ m_k = \max(\boldsymbol{v}_k) \\ \boldsymbol{u}_k = \boldsymbol{v}_k/m_k\end{cases} \qquad (3.58)$$

where $m_k = \max(\boldsymbol{v}_k)$ means the maximum absolute value of the component of $\boldsymbol{v}_k$. For example, $\boldsymbol{v}_k = (3, -7, 7)^{\mathrm{T}}$, then $\max(\boldsymbol{v}_k) = m_k = -7$, the scaling vector is

$$\boldsymbol{u}_k = \left(-\frac{3}{7}, 1, -1\right)^{\mathrm{T}}$$

Theorem 24. *Suppose* $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$, *then the sequences of vectors* $\{\boldsymbol{u}_k\}$ *and* $\{m_k\}$, *generated by* (3.58) *are convergent and*

$$\lim_{k\to\infty} \boldsymbol{u}_k = \frac{\boldsymbol{x}_1}{\max(\boldsymbol{x}_1)} \tag{3.59}$$

and

$$\lim_{k\to\infty} m_k = \lambda_1 \tag{3.60}$$

Proof From (3.58), we have

$$\begin{aligned}\boldsymbol{u}_k &= \frac{1}{m_k}\boldsymbol{v}_k = \frac{1}{m_k}\boldsymbol{A}\boldsymbol{u}_{k-1} = \frac{1}{m_k}\left(\boldsymbol{A}\frac{1}{m_{k-1}}\boldsymbol{A}\boldsymbol{u}_{k-2}\right) = \frac{1}{m_k m_{k-1}}\boldsymbol{A}^2\boldsymbol{u}_{k-2}\\ &= \cdots = \frac{1}{m_k m_{k-1}\cdots m_1}\boldsymbol{A}^k\boldsymbol{u}_0.\end{aligned}$$

Since $\boldsymbol{u}_k$ are scaling vector, $m_k m_{k-1}\cdots m_1 = \max\left(\boldsymbol{A}^k\boldsymbol{u}_0\right)$. Thus

$$\begin{aligned}\boldsymbol{u}_k =& \frac{\boldsymbol{A}^k\boldsymbol{u}_0}{\max\left(\boldsymbol{A}^k\boldsymbol{u}_0\right)} = \frac{\lambda_1^k\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)}{\max\left(\lambda_1^k\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)\right)}\\ =& \frac{\lambda_1^k\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)}{\lambda_1^k\max\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n a_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)} = \frac{\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i}{\max\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)}\end{aligned}$$

Let $k \to \infty$, and we get

$$\lim_{k\to\infty} \boldsymbol{u}_k = \frac{\boldsymbol{x}_1}{\max(\boldsymbol{x}_1)}.$$

Similarly,

$$\begin{aligned}\boldsymbol{v}_k =& \boldsymbol{A}\boldsymbol{u}_{k-1} = \boldsymbol{A}\frac{\boldsymbol{A}^{k-1}\boldsymbol{u}_0}{\max\left(\boldsymbol{A}^{k-1}\boldsymbol{u}_0\right)} = \frac{\boldsymbol{A}^k\boldsymbol{u}_0}{\max\left(\boldsymbol{A}^{k-1}\boldsymbol{u}_0\right)}\\ =& \frac{\lambda_1^k\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)}{\max\left(\lambda_1^{k-1}\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^{k-1}\boldsymbol{x}_i\right)\right)}\end{aligned}$$

$$m_k = \max(\boldsymbol{v}_k) = \frac{\lambda_1\max\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)}{\max\left(\alpha_1\boldsymbol{x}_1 + \sum\limits_{i=2}^n \alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^{k-1}\boldsymbol{x}_i\right)}.$$

TABLE 3.1
Computing results

k	$\boldsymbol{u}_k^{\mathrm{T}}$	$m_k = \max(\boldsymbol{v}_k)$
0	(0.0000,0.0000,1.0000)	1.0000
1	(0.5000,1.0000,0.2500)	4.0000
2	(0.5000,1.0000,0.8611)	9.0000
3	(0.5000,1.0000,0.7306)	11.4400
4	(0.5000,1.0000,0.7535)	10.9224
5	(0.5000,1.0000,0.7493)	11.0140
6	(0.5000,1.0000,0.7501)	10.9927
7	(0.5000,1.0000,0.7500)	11.0004
8	(0.5000,1.0000,0.7500)	11.0000

Thus

$$\lim_{k\to\infty} \max(\boldsymbol{v}_k) = \lambda_1.$$

Remark The converging speed of the power method is related to $\left|\frac{\lambda_2}{\lambda_1}\right|$.

Exercise 3.13. *Use power method to compute the dominant eigenvalue and eigenvector for the following matrix*

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 & 2 \\ 10 & 3 & 4 \\ 3 & 6 & 1 \end{bmatrix}.$$

Solution Choose $\boldsymbol{u}_0 = (0,0,1)^{\mathrm{T}}$, and we can get by (3.58)

$$\boldsymbol{v}_1 = \boldsymbol{A}\boldsymbol{u}_0 = \begin{bmatrix} 2 & 3 & 2 \\ 10 & 3 & 4 \\ 3 & 6 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix}.$$

Then

$$m_1 = 4, \quad \boldsymbol{u}_1 = \left(\frac{1}{2}, 1, \frac{1}{4}\right)^{\mathrm{T}} = (0.5, 1.0, 0.25)^{\mathrm{T}}$$

Repeat this iteration, the computing results are listed in the table.

From the table, the dominant eigenvalue is $\lambda_1 = 11.0000$ and the corresponding eigenvector is $\boldsymbol{x}_1 = (0.5000, 1.0000, 0.7500)$. The exact eigenvalues are $11, -3, -2$, and its converging speed is $\frac{3}{11}$.

(2) If $|\lambda_1| = |\lambda_2|$, and $|\lambda_2| > |\lambda_3|$, there are three cases:

(2.1) $\lambda_1 = \lambda_2$. From (3.58), there is

$$\boldsymbol{u}_k = \frac{\boldsymbol{A}^k\boldsymbol{v}_0}{\max\left(\boldsymbol{A}^k\boldsymbol{v}_0\right)} = \frac{\lambda_1^k\left(\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \sum\limits_{i=3}^{n}\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)}{\lambda_1^k\max\left(\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \sum\limits_{i=3}^{n}\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i\right)}.$$

If $k \to \infty$, we have

$$\lim_{k\to\infty} \boldsymbol{u}_k = \frac{\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2}{\max\left(\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2\right)},$$

where $|\alpha_1| + |\alpha_2| \neq 0$. Similarly

$$\lim_{k\to\infty} m_k = \lim_{k\to\infty} \max\left(\boldsymbol{v}_k\right) = \lambda_1 = \lambda_2.$$

It is clear that converging speed is $\left|\frac{\lambda_3}{\lambda_1}\right|$.

(2.2) $\lambda_1 = -\lambda_2$

From (3.58), there is

$$\begin{aligned}\boldsymbol{u}_k =& \frac{\boldsymbol{A}^k \boldsymbol{v}_0}{\max\left(\boldsymbol{A}^k \boldsymbol{v}_0\right)} = \frac{\lambda_1^k \left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i\right)}{\lambda_1^k \max\left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i\right)}\\ =& \frac{\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i}{\max\left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i\right)}\end{aligned}$$

where $|\alpha_1| + |\alpha_2| \neq 0$. If $\alpha_2 \neq 0$, $\boldsymbol{u}_k$ is divergent .

Multiplying $\boldsymbol{A}$ twice on $\boldsymbol{u}_k$, then

$$\boldsymbol{A}^2 \boldsymbol{u}_k = \lambda_1^2 \frac{\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^{k+2} \boldsymbol{x}_i}{\max\left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i\right)}.$$

Thus

$$\lim_{k\to\infty} \max\left(\boldsymbol{A}^2 \boldsymbol{u}_k\right) = \lim_{k\to\infty} \lambda_1^2 \frac{\max\left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^{k+2} \boldsymbol{x}_i\right)}{\max\left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i\right)} = \lambda_1^2$$

Similarly, the converging speed is $\left|\frac{\lambda_3}{\lambda_1}\right|$. Moveover

$$\begin{aligned}A\boldsymbol{u}_k + \lambda_1 \boldsymbol{u}_k =& \frac{2\lambda_1 \alpha_1 \boldsymbol{x}_1 + \sum\limits_{i=3}^n \alpha_i \left(\lambda_i + \lambda_1\right) \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i}{\max\left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i\right)}\\ \boldsymbol{A}\boldsymbol{u}_k - \lambda_1 \boldsymbol{u}_k =& \frac{(-1)^{k+1} 2\lambda_1 \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \alpha_i \left(\lambda_i - \lambda_1\right) \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i}{\max\left(\alpha_1 \boldsymbol{x}_1 + (-1)^k \alpha_2 \boldsymbol{x}_2 + \sum\limits_{i=3}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i\right)}.\end{aligned}$$

From the aboving equations, we get $\boldsymbol{A}\boldsymbol{u}_k + \lambda_1 \boldsymbol{u}_k$ and $\boldsymbol{A}\boldsymbol{u}_k - \lambda_1 \boldsymbol{u}_k$ which can be seen as the eigenvectors corresponding to λ_1 and $\lambda_2\,(= -\lambda_1)$ respectivly.

If $\lambda_1 = \lambda_2 = \cdots = \lambda_r, \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_{r+l} = -\lambda_1$, and $|\lambda_1| > |\lambda_{r+l+1}|$, then we have similar conclusions.

$(2.3)\lambda_1 = \bar{\lambda}_2$

Since $\mathbf{A}$ is real matrix, the complex roots of $\boldsymbol{A}$ are conjugate. From (3.55), there is

$$\begin{aligned} \boldsymbol{v}_k &= \alpha_1 \lambda_1^k \boldsymbol{x}_1 + \alpha_2 \lambda_2^k \boldsymbol{x}_2 + \alpha_3 \lambda_3^k \boldsymbol{x}_3 + \cdots + \alpha_n \lambda_n^k \boldsymbol{x}_n \\ &= \alpha_1 \lambda_1^k \boldsymbol{x}_1 + a_2 \bar{\lambda}_1^k \boldsymbol{x}_1 + \alpha_3 \lambda_3^k \boldsymbol{x}_3 + \cdots + \alpha_n \lambda_n^k \boldsymbol{x}_n \\ &= \lambda_1^k \left[\alpha_1 \boldsymbol{x}_1 + \alpha_2 \left(\frac{\bar{\lambda}_1}{\lambda_1} \right)^k \bar{\boldsymbol{x}}_1 + \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \boldsymbol{x}_i \right]. \end{aligned}$$

Since $\left| \frac{\lambda_i}{\lambda_1} \right| < 1 (i = 3, 4, \cdots, n)$, we can get with $k \to \infty$

$$\begin{gathered} \boldsymbol{v}_k \approx \alpha_1 \lambda_1^k \boldsymbol{x}_1 + \alpha_2 \bar{\lambda}_1^k \overline{\boldsymbol{x}}_1 \\ \boldsymbol{v}_{k+1} \approx \alpha_1 \lambda_1^{k+1} \boldsymbol{x}_1 + \boldsymbol{\alpha}_2 \bar{\lambda}_1^{k+1} \overline{\boldsymbol{x}}_1 \\ \boldsymbol{v}_{k+2} \approx \alpha_1 \lambda_1^{k+2} \boldsymbol{x}_1 + \boldsymbol{\alpha}_2 \bar{\lambda}_1^{k+2} \overline{\boldsymbol{x}}_1. \end{gathered}$$

Multiplying $\lambda_1 \bar{\lambda}_1, -\left(\lambda_1 + \bar{\lambda}_1\right)$ and 1, on the above three equations respectively and added together, we obtain

$$\boldsymbol{v}_{k+2} - \left(\lambda_1 + \bar{\lambda}_1\right) \boldsymbol{v}_{k+1} + \lambda_1 \bar{\lambda}_1 \boldsymbol{v}_k \approx 0. \tag{3.61}$$

It is clear $\boldsymbol{v}_{k+2}, \boldsymbol{v}_{k+1}$ and $\boldsymbol{v}_k$ almost linearly dependent from (3.61).Let

$$\begin{gathered} p = -\left(\lambda_1 + \bar{\lambda}_1\right), \quad q = \lambda_1 \bar{\lambda}_1 \\ \boldsymbol{v}_{k+2} + p \boldsymbol{v}_{k+1} + q \boldsymbol{v}_k \approx 0, \end{gathered}$$

which is the linear system with p, q unknowns. We can use least-square method to solve the equations. Then the eigenvalues $\lambda_1, \bar{\lambda}_1$ can be obtained by

$$\left\{ \begin{array}{l} \lambda_1 = -\frac{p}{2} + \sqrt{\frac{p^2}{4} - q} \\ \bar{\lambda}_1 = -\frac{p}{2} - \sqrt{\frac{p}{4} - q}. \end{array} \right.$$

The eigenvectors are

$$\left\{ \begin{array}{l} \boldsymbol{v}_{k+1} - \bar{\lambda}_1 \boldsymbol{v}_k = \lambda_1^k \left(\lambda_1 - \bar{\lambda}_1\right) \alpha_1 \boldsymbol{x}_1 \\ \boldsymbol{v}_{k+1} - \lambda_1 \boldsymbol{v}_k = \bar{\lambda}_1^k \left(\bar{\lambda}_1 - \lambda_1\right) \alpha_1 \boldsymbol{x}_1. \end{array} \right.$$

So the eigenvectors are x_1 and x_2.

The Power method has the disadvantage that it is unknown at the outset whether or not the matrix has a single dominant eigenvalue. Nor is it known how $x^{(0)}$ should be chosen so as to ensure that its representation in terms of the eigenvectors of the matrix will contain a nonzero contribution from the eigenvector associated with the dominant eigenvalue, should it exist.

3.4 Exercise

1. Use Gaussian elimination with partial pivoting to solve

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 2 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 8 \\ 13 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 1 \\ 3 & 4 & 0 \\ 2 & 10 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 10 \end{bmatrix}$$

2. Determine $\|\boldsymbol{x}\|_\infty, \|\boldsymbol{x}\|_1, \|\boldsymbol{x}\|_2$, where $\boldsymbol{x} = [0, -1, 2]^{\mathrm{T}}$.

3. Suppose $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in \mathbf{R}^n, \omega_i > 0 (i = 1, 2, \cdots, n)$. Show

$$\|\boldsymbol{x}\| = \sum_{i=1}^{n} \omega_i \, |x_i|$$

is a vector norm in $\mathbf{R}^n$.

4. Show
(1) $\|\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{x}\|_1 \leqslant \sqrt{n}\|\boldsymbol{x}\|_2$;
(2) $\|x\|_\infty \leqslant \|x\|_1 \leqslant n\|x\|_\infty$;
(3) $\|\boldsymbol{x}\|_\infty \leqslant \|\boldsymbol{x}\|_2 \leqslant \sqrt{n}\|\boldsymbol{x}\|$,
where $x \in \mathbf{R}^n$.

5. Compute $\|\boldsymbol{A}\|_\infty, \|\boldsymbol{A}\|_1, \|\boldsymbol{A}\|_2$ where

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

6. Supose $\|\boldsymbol{A}\|_p, \|\boldsymbol{A}\|_q$ are matrix norms on $\mathbf{R}^{n\times n}$. Show there exist constants $d_1, d_2 > 0$, such that

$$d_1\|\boldsymbol{A}\|_p \leqslant \|\boldsymbol{A}\|_q \leqslant d_2\|\boldsymbol{A}\|_p$$

for $\boldsymbol{A} \in \mathbf{R}^{n\times n}$.

7. If $\boldsymbol{A} \in \mathbf{R}^{n\times n}, \|\boldsymbol{A}\| < 1$, show $\boldsymbol{I} + \boldsymbol{A}$ is invertible and

$$\left\|(\boldsymbol{I} + \boldsymbol{A})^{-1}\right\| \leqslant \frac{1}{1 - \|\boldsymbol{A}\|}.$$

8. Show the sequence

$$S_k = I + A + \cdots + A^k \quad (k = 0, 1, 2, \cdots)$$

is convergent if $\rho(\boldsymbol{A}) < 1$, $\boldsymbol{A} \in \mathbf{R}^{n\times n}$,and compute the limit.

9. Let $\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 4 & -2 \end{bmatrix}$. Compute $\mathrm{cond}(\boldsymbol{A})_\infty, \mathrm{cond}(\boldsymbol{A})_1$ and $\mathrm{cond}(\boldsymbol{A})_2$.

10. Let

$$\boldsymbol{A} = \begin{bmatrix} 2.000 & 1 & -1 \\ -2 & 1 & \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 7.0003 \\ -7 \end{bmatrix}$$

The linear system $Ax = b$ has exact solutions

$$x = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

(1) Compute cond($\boldsymbol{A}$) $_\infty$; (2) Taking

$$\boldsymbol{y} = \begin{bmatrix} 2,91 \\ -1.01 \end{bmatrix}$$

compute $r_y = b - Ay$;

(3) If $z = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$, compute $r_z = b - Az$;

(5) What do you understand?

11. Use Jacobi method and Gauss-Seidel method to solve the following equations with initial guess $x^{(0)} = (0, 0, 0)^{\mathrm{T}}$,

$$\begin{bmatrix} 20 & 2 & 3 \\ 1 & 8 & 1 \\ 2 & -3 & 15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 24 \\ 12 \\ 30 \end{bmatrix}$$

and the approximation solutions are with 2 significant figures.

12. Show Jacobi iterative method converge if and only if

$$\left| \frac{a_{12}a_{21}}{a_{11}a_{22}} \right| < 1$$

for

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases} \qquad (a_{11}, a_{22} \neq 0)$$

13. Illustrate at least eigenvalues of the Gauss-Seidel iterative matrix $\boldsymbol{G} = -(\boldsymbol{D} + \boldsymbol{L})^{-1}\boldsymbol{U}$ is zero.

14. Discuss the convergency by Jacobi method and Gauss-Seidel method for the following linear equaitons:

(1) $\begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$;

(2) $\begin{bmatrix} 5 & 2 & 1 \\ -1 & 4 & 2 \\ 2 & -3 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -12 \\ 30 \\ 3 \end{bmatrix}$.

15. Use power method to compute the dominant eigenvalue and eigenvector for

$$A = \begin{bmatrix} 7 & 3 & -2 \\ 3 & 4 & -1 \\ -2 & -1 & 3 \end{bmatrix}$$

with 2 significant figures.

16. Computer Problem

Program the Gauss elimination with partial pivoting to solve the following linear equations

$$Rx = v$$

where

$$\boldsymbol{R} = \begin{bmatrix} 31 & -13 & 0 & 0 & 0 & -10 & 0 & 0 & 0 \\ -13 & 35 & -9 & 0 & -11 & 0 & 0 & 0 & 0 \\ 0 & -9 & 31 & -10 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -10 & 79 & -30 & 0 & 0 & 0 & -9 \\ 0 & 0 & 0 & -30 & 57 & -7 & 0 & -5 & 0 \\ 0 & 0 & 0 & 0 & -7 & 47 & -30 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -30 & 41 & 0 & 0 \\ 0 & 0 & 0 & 0 & -5 & 0 & 0 & 27 & -2 \\ 0 & 0 & 0 & -9 & 0 & 0 & 0 & -2 & 29 \end{bmatrix}.$$