Sect.1 Solving systems of linear equations

The principal objective of this chapter is to discuss the numerical aspects of solving systems of linear equations having the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n & = & b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{1n}x_n & = & b_3 \\ \vdots & & \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n & = & b_n \end{cases} \tag{1}$$

This is a system of n equations in the n unknowns x_1, x_2, ..., x_n.
The system (1) can be written in matrix form as

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} \cdots\cdots & a_{n1} \\ a_{12} & a_{22} & a_{32} \cdots\cdots & a_{n2} \\ a_{13} & a_{23} & a_{33} \cdots\cdots & a_{n3} \\ \vdots & \vdots & \vdots \quad \ddots & \vdots \\ a_{1n} & a_{2n} & a_{3n} \cdots\cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

Then we can denote these matrices by A, x and b, so that the equation becomes simply

$$Ax = b$$

In dealing with systems of linear equations there is a concept of equivalence that is important. Let two systems be given, each consisting of n equations with n unknowns

$$Ax = b \qquad\qquad Cx = d$$

If the two systems have precisely the same solutions, we call them equivalent systems. Thus, to solve a system of equations, we can instead solve any equivalent system; no solutions are lost and no new ones appear. This simple idea is at the heart of our numerical procedures. Given a system of equations to be solved, we transform it by certain elementary operations into a simpler equivalent system which we then solve instead.

The elementary operations alluded to in the previous paragraph are of the following three types. (Here $\mathcal{E}_i$ denotes the i^{th} equation in the system.)

1 Interchanging two equations in the system: $\mathcal{E}_i \longleftrightarrow \mathcal{E}_j$

2 Multiplying an equation by a nonzero number: $\lambda\mathcal{E}_i \longrightarrow \mathcal{E}_i$

3 Adding to an equation a multiple of some other equation: $\mathcal{E}_i + \lambda\mathcal{E}_j \longrightarrow \mathcal{E}_i$

Theorem 1.

If one system of equations is obtained from another by a finite sequence of elementary operations, then the two systems are equivalent.

Proof. It suffices to consider the effect of a single application of each elementary operation. Suppose that an elementary operation transforms the system $Ax = b$ into the system $Bx = d$. If the operation is of type (1), then the two systems consist of precisely the same equations although written in a different order. Clearly, if x solves the first system then it solves the second and vice versa. If the operation is of type (2), then suppose that the i^{th} equation has been multiplied by a scalar $\lambda \neq 0$. The i^{th} and the j^{th} equations in $Ax = b$ are

$$a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \cdots + a_{in}x_n = b_i \tag{2}$$

and

$$a_{j1}x_1 + a_{j2}x_2 + a_{j3}x_3 + \cdots + a_{jn}x_n = b_j \tag{3}$$

and the i^{th} equation in $Bx = d$ is

$$\lambda a_{i1}x_1 + \lambda a_{i2}x_2 + \lambda a_{i3}x_3 + \cdots + \lambda a_{in}x_n = \lambda b_i \tag{4}$$

Any vector x that satisfies Equation (2) satisfies Equation (3) and vice versa, because $\lambda \neq 0$. Finally, suppose that the operation is of type (3). Assume that λ times the j^{th} equation has been added to the i^{th}. Then the i^{th} equation in $Bx = d$ is

$$\left(a_{i1} + \lambda a_{j1}\right)x_1 + \left(a_{i2} + \lambda a_{j2}\right)x_2 + \cdots + \left(a_{in} + \lambda a_{jn}\right)x_n = \left(b_1 + \lambda b_j\right) \tag{5}$$

Observe particularly that the j^{th} equation in the system $Bx = d$ has not been changed. If $Ax = b$, then Equations (2) and (3) are true. Hence, (5) is true. Thus, $Bx = d$. On the other hand, if we suppose that x solves $Bx = d$, then Equations (5) and (3) are true. If λ times Equation (3) is subtracted from Equation (5) the result is Equation (2). Hence, $Ax = b$. □

The $n \times n$ matrix

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

is called an identity matrix. It has the property that $IA = AI = A$ for any matrix A of size $n \times n$.

If A and B are two matrices such that $AB = I$, then we say that B is a right inverse of A and that A is a left inverse of B.

For example

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

We see from this example that if a matrix has a right inverse, the latter is not necessarily unique. For square matrices the situation is better.

Theorem 2.
A square matrix can possess at most one right inverse. If A and B are square matrices such that $AB = I$, then $BA = I$.

It follows from Theorems 2 that if a square matrix A has a right inverse B, then B is unique and $BA = AB = I$. We then call B the inverse of A and say that A is invertible or nonsingular. (Of course, B is therefore invertible and A is its inverse.)
Therefore, if A is invertible, then the system of equations $Ax = b$ has the solution $x = A^{-1}b$.

The elementary operations discussed earlier can be carried out with matrix multiplications, as we now indicate. An elementary matrix is defined to be an $n \times n$ matrix that arises when an elementary operation is applied to the $n \times n$ identity matrix. The elementary operations, expressed in terms of the rows of a matrix A are:

1 Interchanging two rows of A: $A_i \longleftrightarrow A_j$

2 Multiplying one row by a nonzero constant: $\lambda A_i \longrightarrow A_i$

3 Adding to one row a multiple of another: $A_i + \lambda A_j \longrightarrow A_i$

If we wish to apply a succession of elementary row operations to A, we introduce elementary matrices $E_1, E_2, ..., E_m$ and then write the transformed matrix as

$$E_1 E_2 ... E_m A$$

If a matrix is invertible, such a sequence of elementary row operations can be applied to A, reducing it to I.Thus

$$E_1 E_2 ... E_m A = I$$

From this it follows that $A^{-1} = E_1 E_2 ... E_m$. Consequently, A^{-1} can be obtained by subjecting I to the same sequence of elementary row operations.

Example 1.

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

2 $A = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 5 \\ 1 & 1 & 4 \end{bmatrix}$,

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

4 $\begin{bmatrix} 1 & 2 & 3 & -1 \\ 1 & 1 & -3 & 1 \\ 2 & 3 & 1 & 0 \\ 1 & 2 & 1 & -2 \end{bmatrix}$,

5 $\begin{bmatrix} 1 & 2 & 3 & 0 \\ 2 & 3 & 0 & 2 \\ 3 & 0 & 2 & 1 \\ 0 & 2 & -1 & 1 \end{bmatrix}$,

6 $\begin{bmatrix} -2 & 1 & 5 & 1 \\ -5 & 1 & 2 & 1 \\ 1 & 1 & 2 & 7 \\ -7 & 1 & 1 & -1 \end{bmatrix}$

Sect.2 Easy to solve systems

We begin by looking for special types of systems that can be easily solved.

1 Suppose that the $n \times n$ matrix A has a **diagonal structure**. This means that all the nonzero elements of A are on the main diagonal and System (1)

$$a_{ii}x_i = b_i \quad \Longrightarrow \quad x_i = \frac{b_i}{a_{ii}}$$

If $a_{ii} = 0$ for some index i, and if $b_i = 0$ also, then x_i can be any real number.
If $a_{ii} = 0$ and $b_i \neq 0$, no solution of the system exists.

2 Assume a **lower triangular** structure for A. This means that all the nonzero elements of A are situated on or below the main diagonal and System (1) is

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{6}$$

To solve this, assume that $a_{ii} \neq 0$ for all i, then obtain x_1 from the first equation. With the known value of x_1 substituted in the second equation, solve the second equation for x_2. We proceed in the same way, obtaining x_1, x_2, ..., x_n one at a time, and in this order. A formal algorithm for the solution in this case is called forward substitution:

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} \\ x_i &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j\right), \quad i = 2, \dots, n. \end{aligned} \tag{7}$$

3 Assume a **upper triangular** structure for A. This means that all the nonzero elements of A are situated on or below the main diagonal and System (1) is

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{8}$$

The case of a linear system whose matrix is invertible and upper triangular is treated in a similar manner, by the method called back substitution

$$
\begin{aligned}
x_n &= \frac{b_n}{a_{nn}} \\
x_i &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right), \quad i = n-1, \dots, 1.
\end{aligned} \tag{9}
$$

Sect.3 Triangularization Method (LU Factorization)

In this method, the coefficient matrix A is decomposed into the product of a lower triangular matrix L and an upper triangular matrix U. We write

$$A = LU, \quad \text{where} \quad l_{ij} = 0,\ j > i, \quad \text{and} \quad u_{ij} = 0,\ i > j,\ u_{ii} = 1$$

This method is also called the **Crout's method**. Instead of $u_{ii} = 1$, if we take $l_{ii} = 1$, then the method is also called the **Doolittle's method**.
Comparing the elements of the matrices on both sides, we obtain n^2 equations in n^2 unknowns, which uniquely determines L and U. We get

$$
\begin{aligned}
l_{ij} &= a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}, \qquad i \geqslant j \\
u_{ij} &= \frac{1}{l_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right), \qquad i < j. \\
u_{ii} &= 1.
\end{aligned} \tag{10}
$$

When using LU decomposition to solve equations, first solve the lower triangular system $Ly = b$. This finds $y = Ux$. Now solve the upper triangular system $Ux = y$ to find x. Each of these solution steps is simple.

Example 2.
Let the decomposition $A = LU$

$$\begin{bmatrix} 1 & -1 & 3 \\ 2 & -3 & 1 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -5 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 3 \\ 0 & -1 & -5 \\ 0 & 0 & -33 \end{bmatrix}$$

Execute forward substitution to solve $Ly = b = [1 \ \ 3 \ \ 1]^t$

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -5 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \quad \Longrightarrow \quad \begin{cases} y_1 = 1 \\ y_2 = 1 \\ y_3 = 3 \end{cases}$$

Execute back substitution to solve $Ux = y$

$$\begin{bmatrix} 1 & -1 & 3 \\ 0 & -1 & -5 \\ 0 & 0 & -33 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix} \quad \Longrightarrow \quad \begin{cases} x_1 = -1/11 \\ x_2 = -6/11 \\ x_3 = 8/11 \end{cases}$$

Sect.4 Gaussian elimination

To see which matrices have an LU factorization and to find how it is determined, first suppose that Gaussian elimination can be performed on the system $Ax = b$ **without row interchanges**.

Let's denote by Eq_i the i^{th} equation in the system (1). The first step in the Gaussian elimination process consists of performing, for each $j = 2, 3, ..., n$, the operations

$$Eq_j - m_{j,1} Eq_1 \to Eq_j \qquad \text{where} \quad m_{j,1} = \frac{a_{j1}^{(1)}}{a_{11}^{(1)}} \tag{11}$$

These operations transform the system into one in which all the entries in the first column below the diagonal are zero. The system of operations in (11) can be viewed in another way. It is simultaneously accomplished by multiplying the original matrix A on the left by the matrix

$$M^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -m_{21} & 1 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -m_{n1} & 0 & \cdots & 0 & 1 \end{bmatrix}$$

This is called the first Gaussian transformation matrix. We denote the product of this matrix with $A^{(1)} = A$ by $M^{(1)}$ and with b by $b^{(2)}$, so

$$A^{(2)} x = M^{(1)} A x = M^{(1)} b = b^{(2)}.$$

In general, with $A^{(k)} x = b^{(k)}$ already formed, multiply by the k^{th} Gaussian transformation matrix $M^{(k)}$ to obtain

$$A^{(k+1)} x = M^{(k)} A^{(k)} x = M^{(k)} ... M^{(1)} A x = M^{(k)} b^{(k)} = b^{(k+1)} M^{(k)} ... M^{(1)} b.$$

The process ends with the formation of $A^{(n)} x = b^{(n)}$, where $A^{(n)}$ is an upper triangular matrix

$$A^{(n)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nn}^{(n)} \end{bmatrix}, \qquad \text{with} \quad A^{(n)} = M^{(n-1)}...M^{(1)}A.$$

This process forms the factorization $A = LU$. To determine the complementary lower triangular matrix L, fist recall that Gaussian transformation $M^{(k)}$ generates the row operations

$$(Eq_j - m_{j,k}Eq_k) \longrightarrow (E_j)$$

Theorem 3.
If Gaussian elimination can be performed on the linear system $Ax = b$ without row interchanges, then the matrix A can be factored into the product of a lower-triangular matrix L and an upper-triangular matrix U that is, $A = LU$ where $m_{ji} = a_{ji}^{(i)}/a_{ii}^{(i)}$

$$U = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nn}^{(n)} \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & \dots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \dots & m_{n,n-1} & 1 \end{bmatrix}$$

LU decomposition is not always guaranteed for arbitrary matrices. Decomposition is guaranteed when the matrix A is positive definite (see Definition 2).
Here is a sufficient condition for a square matrix A to have an LU-decomposition.

Theorem 4.
If all n leading principal minors of the $n \times n$ matrix A are nonsingular, then A has an LU-decomposition.

In the previous discussion we assumed that A is such that a linear system of the form $Ax = b$ can be solved using Gaussian elimination that does not require row interchanges. Although many systems we encounter, the row interchanges are required.

Sect.5 PA=LU Factorization

As its name implies, the $PA = LU$ factorization is simply the LU factorization of a row-exchanged version of A. Under partial pivoting, the rows that need exchanging are not known at the outset, so we must be careful about fitting the row exchange information into the factorization.

An $n \times n$ permutation matrix P is a matrix with precisely one entry whose value is 1 in each column and each row and all of whose other entries are 0. If P is obtained from the identity by interchanging the i^{th} and j^{th} rows, then for any $n \times n$ matrix A, multiplying on the left by P has the effect of interchanging the i^{th} and j^{th} rows of A. Similarly, multiplying on the right by P interchanges the i^{th} and j^{th} columns of A. Furthermore, the inverse P^{-1} exists and satisfy $P^{-1} = P^t$.

Example 3.

Let us consider the following matrix $A = \begin{bmatrix} 0 & 1 & -1 & 1 \\ 1 & 1 & -1 & 2 \\ -1 & -1 & 1 & 0 \\ 1 & 2 & 0 & 2 \end{bmatrix}$.

Since $a_{11} = 0$*, the matrix* A *does not have an LU factorization. However, using the row interchange* $A_1 \leftrightarrow A_2$*, followed by* $(A_3 + A_1) \to A_3$ *and* $(A_4 - A_1) \to A_4$*, produces*

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

Then the row interchange $A_3 \leftrightarrow A_4$*, followed by* $(A_3 - A_2) \to A_3$*, gives the matrix*

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

The permutation matrix associated with the row interchanges $A_1 \leftrightarrow A_2$ *and* $A_3 \leftrightarrow A_4$ *is*

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = P^{-1}$$

Gaussian elimination can be performed on PA *without row interchanges to give the LU factorization*

$$PA = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix} = LU \implies A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

Exercise 1:

Obtain factorizations of the form $A = P^{-1}LU$ for the following matrices.

$$\begin{bmatrix} 1 & -2 & 3 & 0 \\ 3 & -6 & 9 & 3 \\ 2 & 1 & 4 & 1 \\ 1 & -2 & 2 & -2 \end{bmatrix}, \quad \begin{bmatrix} 1 & -2 & 3 & 0 \\ 3 & -6 & 9 & 3 \\ 2 & 1 & 4 & 1 \\ 1 & -2 & 2 & -2 \end{bmatrix}, \quad \begin{bmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 2 & 0 \\ 2 & 0 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 1 & 3 \\ 2 & 3 & -1 & 4 \\ 1 & 0 & 2 & 1 \\ 3 & 1 & -2 & 2 \end{bmatrix}$$

Sect.6 Cholesky Method (Square Root Method)

If the coefficient matrix in (1) is symmetric and positive definite (see the next section), then A can be decomposed as

$$A = LL^t, \qquad \text{with} \quad l_{ij} = 0,\ j > i.$$

The elements of L are given by

$$l_{ii} = \left(a_{ii} - \sum_{j=1}^{i-1} l_{ij}^2 \right)^{1/2}, \qquad i = 1, \dots n, \tag{12}$$

$$l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{jk} l_{ik} \right), \qquad i = j+1, \dots, n, \quad j = 1, \dots n$$

Theorem 5.
If A is a real, symmetric, and positive definite matrix, then it has a unique factorization $A = LL^t$, in which L is lower triangular with a positive diagonal.

Example 4.

$$\begin{bmatrix} 1 & 2 & 3 & 2 \\ 2 & 5 & 9 & 9 \\ 3 & 9 & 34 & 25 \\ 2 & 9 & 25 & 79 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 3 & 4 & 0 \\ 2 & 5 & 1 & 7 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 2 \\ 0 & 1 & 3 & 5 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 7 \end{bmatrix}$$

Exercise 2:
Give the Cholesky (LL^t) decomposition of the following matrices.

$$\begin{bmatrix} 1 & 0 & 7 & 5 \\ 0 & 1 & 2 & 3 \\ 7 & 2 & 54 & 44 \\ 5 & 3 & 44 & 44 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & -2 & -5 \\ 0 & 1 & -2 & 3 \\ -2 & -2 & 9 & 1 \\ -5 & 3 & 1 & 44 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 4 & 3 \\ 2 & 4 & 21 & 13 \\ 1 & 3 & 13 & 12 \end{bmatrix}, \quad \begin{bmatrix} 49 & 21 & 14 & 7 \\ 21 & 10 & 10 & 6 \\ 14 & 10 & 21 & 19 \\ 7 & 6 & 19 & 36 \end{bmatrix}$$

Sect.7 Techniques for Special Matrices

Definition 1.
The $n \times n$ matrix A is said to be diagonally dominant when

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

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

Example 5.
E $\begin{bmatrix} 7 & 3 & 0 \\ 2 & -5 & 1 \\ -1 & 2 & 4 \end{bmatrix}$, F $\begin{bmatrix} 6 & 3 & -4 \\ 3 & -2 & 2 \\ -4 & 2 & 1 \end{bmatrix}$.
E is diagonally dominant, while F isn't.

Theorem 6.
Every diagonally dominant matrix is nonsingular and has an LU-factorization.

Moreover, in this case, Gaussian elimination can be performed on any linear system of the form $Ax = b$ to obtain its unique solution without row or column interchanges. Moreover, we can prove by induction that

Theorem 7.
Gaussian elimination without pivoting preserves the diagonal dominance of a matrix.

Definition 2.
A matrix A is positive definite if it is symmetric and if $x^t Ax > 0$ for every n-dimensional column vector $x \neq 0$.

Example 6.
A $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$: *we have* $\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (x_1 + x_2)^2 + x_1^2 + x_2^2 > 0, \quad \forall x_1, x_2 \neq 0.$

Using the definition to determine whether a matrix is positive definite can be difficult. Fortunately, there are more easily verified criteria for identifying members that are and are not of this important class.

Sect.8 Vector norms

We shall require some measure of the 'magnitude' of a vector for satisfactory error analyses of methods of solving linear equations.

Definition 3.
For any vector $\boldsymbol{x} \in \mathbb{C}^n$, we define the norm of $\boldsymbol{x}$, written as $\|\boldsymbol{x}\|$, to be a rear number satisfying the following conditions

(i) $\|\boldsymbol{x}\| > 0$, *unless* $\boldsymbol{x} = 0$ *and* $\|0\| = 0$.

(ii) For any scalar λ, *and any vector* $\boldsymbol{x}$, $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$.

(iii) for any vectors $\boldsymbol{x}$ *and* $\boldsymbol{y}$

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|.$$

There are many ways in which one may choose norms. One of the most common is the Euclidean norm (or length) $\|\boldsymbol{x}\|_2$, defined by

$$\|\boldsymbol{x}\|_2 = \left(|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2\right)^{1/2} = \left(\boldsymbol{x}^t \boldsymbol{x}\right)^{1/2}. \tag{13}$$

Other possible norms include

$$\|\boldsymbol{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n| = \sum_{i=1}^{n} |x_i|. \tag{14}$$

$$\|\boldsymbol{x}\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i| \tag{15}$$

These are particularly useful for numerical methods as the corresponding matrix norms are easily computed. All of these norms may be considered special cases of

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \qquad p \geqslant 1. \tag{16}$$

The norms $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_2$ correspond to $p = 1$ and $p = 2$, respectively. The maximum norm $\|\boldsymbol{x}\|_\infty$ may be considered the result of allowing p to increase indefinitely.

Example 7.
The norms of the vectors $\boldsymbol{x} = (1 \;\; -3 \;\; 0 \;\; 2)^t$, $\boldsymbol{y} = (1 \;\; 3 \;\; 2 \;\; -2)^t$ *are*

$$\begin{aligned} &\|\boldsymbol{x}\|_1 = 6, && \|\boldsymbol{y}\|_1 = 8 \\ &\|\boldsymbol{x}\|_2 = \sqrt{14}, && \|\boldsymbol{y}\|_2 = \sqrt{18} \\ &\|\boldsymbol{x}\|_\infty = \|\boldsymbol{y}\|_\infty = 3 \end{aligned}$$

You will notice that two different vectors may have the same norm, so that

$$\|\boldsymbol{x}\| = \|\boldsymbol{y}\| \quad \not\Rightarrow \quad \boldsymbol{x} = \boldsymbol{y}$$

However, from condition (i)

$$\|\boldsymbol{x} - \boldsymbol{y}\| = 0 \quad \Longrightarrow \quad \boldsymbol{x} - \boldsymbol{y} = 0 \quad \Longrightarrow \quad \boldsymbol{x} = \boldsymbol{y}.$$

We say that a sequence of vectors $\{\boldsymbol{x}_n\}_{n=0}^{\infty}$ converges to a vector $\boldsymbol{\alpha}$ if the sequence formed from the i^{th} elements of the $\boldsymbol{x}_n$ converges to the i^{th} element of $\boldsymbol{\alpha}$, for each component i. We write

$$\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{\alpha}.$$

For any type of vector norm, we have

$$\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{\alpha} \quad \Longleftrightarrow \quad \lim_{n\to\infty} \|\boldsymbol{x}_n - \boldsymbol{\alpha}\| = 0.$$

Notice that

$$\lim_{n\to\infty} \|\boldsymbol{x}_n\| = \|\boldsymbol{\alpha}\|$$

is not sufficient to ensure that $\boldsymbol{x}_n$ converges to $\boldsymbol{\alpha}$.

Sect.9 Matrix norms

Definition 4.
For an $n \times n$ matrix A, we define the matrix norm $\|A\|$ subordinate to a given vector norm to be

$$\|A\| = \sup_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|$$

The supremum is taken over all n-dimensional vectors $\boldsymbol{x}$ with unit norm. The supremum is attained and, therefore, we write

$$\|A\| = \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\| \tag{17}$$

Proposition 1.
A subordinate matrix norm satisfies the following properties.

(i) $\|A\| > 0$, *unless* $A = 0$ *and* $\|0\| = 0$.

(ii) For any λ, and any A, $\|\lambda A\| = |\lambda|\|A\|$.

(iii) for any two matrices A and B we have

$$\|A + B\| \leqslant \|A\| + \|B\|.$$

(iv) For any n-dimensional vector $\boldsymbol{x}$ and any matrix A

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

(v) For any two matrices A and B

$$\|AB\| \leqslant \|A\|.\|B\|$$

Proof. (ii) From condition (ii) for vector norms, it follows that

$$\|\lambda A\| = \max_{\|\boldsymbol{x}\|=1} \|\lambda A\boldsymbol{x}\| = \max_{\|\boldsymbol{x}\|=1} |\lambda|.\|A\boldsymbol{x}\| = |\lambda|. \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\| = |\lambda|.\|A\|$$

(iii) From condition (iii) for vector norms, we have

$$\begin{aligned}\|A+B\| &= \max_{\|\boldsymbol{x}\|=1} \|(A+B)\boldsymbol{x}\| \leqslant \max_{\|\boldsymbol{x}\|=1} \left(\|A\boldsymbol{x}\| + \|B\boldsymbol{x}\|\right)\\ &\leqslant \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\| + \leqslant \max_{\|\boldsymbol{x}\|=1} \|B\boldsymbol{x}\| = \|A\| + \|B\|\end{aligned}$$

(iv) For nay $\boldsymbol{x} \neq 0$,

$$\|A\boldsymbol{x}\| = \|\boldsymbol{x}\| \left\| A\left(\frac{1}{\|\boldsymbol{x}\|}\boldsymbol{x}\right)\right\| \leqslant \|\boldsymbol{x}\|.\|A\|, \qquad \text{as} \qquad \left\|\frac{1}{\|\boldsymbol{x}\|}\boldsymbol{x}\right\| = 1$$

(v) By (iv) we have

$$\|AB\| = \max_{\|\boldsymbol{x}\|=1} \|AB\boldsymbol{x}\| \leqslant \|A\| \max_{\|\boldsymbol{x}\|=1} \|Bx\| = \|A\|.\|B\|$$

□

Matrix norms other than those subordinate to vector norms m a y also be defined. Such matrix norms are real numbers which satisfy conditions of Proposition 1.

We now obtain more explicit expressions for the matrix norms subordinate to the vector norms of (14)-(16).

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

The second factor of the last expression is unity if $\|\boldsymbol{x}\|_1 = 1$ and thus

$$\|A\|_1 = \max_{\|\boldsymbol{x}\|_1=1} \|A\boldsymbol{x}\|_1 \leqslant \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}| \tag{18}$$

Suppose the maximum in the last expression is attained for $j = p$. We now choose x such that

$$x_p = j, \quad \text{and} \quad x_i = 0, \quad j \neq p$$

when $\|\boldsymbol{x}\| = 1$ and

$$\|A\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |a_{ip}|$$

Equality is obtained in (18) for this choice of $\boldsymbol{x}$ and, therefore,

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} \left|a_{ij}\right|$$

Thus $\|A\|_1$, is the maximum column sum of the absolute values of elements. Similarly, (15)

$$\begin{aligned} \|A\|_\infty = \max_{\|\boldsymbol{x}\|_\infty=1} \|A\boldsymbol{x}\|_\infty &\leqslant \max_{\|\boldsymbol{x}\|_\infty=1} \left(\max_i \left|\sum_{j=1}^{n} a_{ij}x_j\right|\right) \\ &\leqslant \max_i \sum_{j=1}^{n} \left|a_{ij}\right| \qquad (19) \end{aligned}$$

as $|x_j| \leqslant 1$ for $\|\boldsymbol{x}\|_\infty = 1$.
Suppose the maximum in (19) occurs for $i = p$. Assuming A is real, we now choose $\boldsymbol{x}$ such that

$$x_j = \operatorname{sign} a_{pj}$$

We obtain equality in (19) and, therefore,

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} \left|a_{ij}\right|$$

To consider the Euclidean norm (13) we start with (10.18) but assume $\boldsymbol{x}$ is real

$$\|A\boldsymbol{x}\|_2^2 = (A\boldsymbol{x})^t A\boldsymbol{x} = \boldsymbol{x}^t A^t A\boldsymbol{x} = \boldsymbol{x}^t B\boldsymbol{x}$$

where $B = A^t A$ is a symmetric matrix. Thus there exists an orthogonal matrix Q such that

$$\boldsymbol{x}^t B\boldsymbol{x} = \boldsymbol{x}^t QDQ^t\boldsymbol{x} = \boldsymbol{y}^t D\boldsymbol{y}$$

where D is the diagonal matrix formed from the eigenvalues of B and $\boldsymbol{y}Q^t\boldsymbol{x}$. Since Q is orthogonal, Q^t is orthogonal and $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$. Thus $\|\boldsymbol{y}\|_2 = 1$, so that

$$\sum_{i=1}^{n} y_i^2 = 1$$

Now,

$$\|A\boldsymbol{x}\|_2^2 = \boldsymbol{y}^t D\boldsymbol{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2 \qquad (20)$$

As $B = A^t A$ is semi-positive definite all its eigenvalues are non-negative. We assume that they are arranged in the order $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$. Hence from (20)

$$\|A\boldsymbol{x}\|_2^2 \leqslant \lambda_1\left(y_1^2 + \cdots + y_n^2\right) = \lambda_1 \qquad (21)$$

Now choose $\boldsymbol{x}$ such that $\boldsymbol{y} = (1\ 0\ \dots\ 0)^t$, when $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$, and from (20) we have equality in (21). We deduce that

$$\|A\|_2 = \lambda_1^{1/2} = \sqrt{\rho\left(A^t A\right)}$$

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

$$\rho(A) = \max|\lambda|, \qquad \lambda \quad \text{is an eigenvalue of} \quad A.$$

And in general, we have $\rho(A) \leqslant \|A\|$, for any natural norm. More precisely, we have

Theorem 8.
spectral radius function satisfies the equation

$$\rho(A) = \inf_{\|.\|}\|A\| \tag{22}$$

in which the infimum is taken over all subordinate matrix norms.

Proof. It is easy to prove that $\rho(A) \leqslant \|A\|$. To do so, let λ be any eigenvalue of A. Select a nonzero eigenvector x corresponding to λ. Then for any vector norm and its subordinate matrix norm, we have

$$|\lambda|\|x\| = |\lambda x| = \|Ax\| \leqslant \|A\|\|x\| \qquad \Longrightarrow \qquad |\lambda| \leqslant \|A\|.$$

It follows that $\rho(A) \leqslant \|A\|$. By taking an infimum we get $\rho(A) \leqslant \inf\limits_{\|.\|}\|A\|$.
For the reverse inequality, we use the fact that every square matrix is similar to a triangular matrix. The latter asserts that for any $\epsilon > 0$ there exists a nonsingular matrix S such that $S^{-1}AS = D + T$, where D is diagonal and T is strictly upper triangular, with $\|T\|_\infty \leqslant \epsilon$. Then we have

$$\left\|S^{-1}AS\right\|_\infty = \left\|D+T\right\|_\infty \leqslant \left\|D\right\|_\infty + \left\|T\right\|_\infty$$

Since D has the eigenvalues of A on its diagonal, it follows that

$$\left\|D\right\|_\infty = \max_{1\leqslant i\leqslant n}|\lambda_i| = \rho(A)$$

Hence, we have

$$\left\|S^{-1}AS\right\|_\infty \leqslant \rho(A) + \epsilon$$

Since ϵ was arbitrary, $\inf\limits_{\|.\|}\|A\| \leqslant \rho(A)$. □

Note that if A is symmetric $A^t A = A^2$, whence $\rho\left(A^2\right) = \left(\rho(A)\right)^2$ and thus

$$\|A\|_2 = \rho(A).$$

Exercise 3:

1 Show that $\|A\|_\infty \leqslant \|A\|_2 \leqslant \|A\|_1$ for all $x \in \mathbb{R}^n$, and that equalities can occur, even for nonzero vectors.

2 Show that $\|A\|_1 \leqslant n\|A\|_\infty$ and $\leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$ for all $x \in \mathbb{R}^n$.

3 Deduce the following $\frac{1}{n}\|A\|_2 \leqslant \frac{1}{\sqrt{n}}\|A\|_\infty \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1 \leqslant n\|A\|_2$

4 Prove this inequality for conditioning numbers $k(AB) \leqslant k(A)k(B)$.

5 The conditioning number $k(A)$ can be expressed by $k(A) = \sup_{\|x\|=\|y\|} \|Ax\|/\|Ay\|$

The following result will be useful in subsequent error analyses.

Proposition 2.
If $\|A\| < 1$, then $I + A$ and $I - A$ are non-singular and

$$\frac{1}{I + \|A\|} \leqslant \|I \pm A\|^{-1} \leqslant \frac{1}{I - \|A\|} \tag{23}$$

Proof. From (17) we remark that

$$\|I\| = \max_{\|\boldsymbol{x}\|=1} \|I\boldsymbol{x}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{x}\| = 1$$

If $I + A$ is singular, there exists a vector $\boldsymbol{x} \neq 0$ such that

$$(I + A)\, x = 0.$$

We will assume that $\boldsymbol{x}$ is scaled so that $\|\boldsymbol{x}\| = 1$, when

$$A\boldsymbol{x} = -I\boldsymbol{x} = -\boldsymbol{x}, \qquad \text{and} \qquad \|A\boldsymbol{x}\| = |-1|.\|\boldsymbol{x}\| = 1.$$

It follows from (17) that $\|A\| \geqslant 1$. Thus $I + A$ is non-singular for $\|A\| < 1$.
From

$$I = (I + A)^{-1}(I + A) \tag{24}$$

and Proposition 1, we have

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

Dividing by the last factor gives the left side of (23). We do the same thing of the right side by considering $I - A$ instead. We rearrange (24) as

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

and, on using Proposition 1

$$\|\,(I + A)^{-1}\,\| \leqslant \|I\| + \|\,(I + A)^{-1}\,\|\|A\|$$

Thus

$$(1-\|A\|)\,\|(I+A)^{-1}\|\leqslant 1$$

and, as $\|A\|<1$ the fight side of (23) follows. □

The convergence of a sequence of matrices $\{A_m\}_{m=0}^{\infty}$ is defined in an identical manner to that for vectors and we write

$$\lim_{m\to\infty} A_m = B$$

if, for all i and j, the sequence of $(i,j)^{\text{th}}$ elements of the A converges to the $(i,j)^{\text{th}}$ element of B. We have the following.

Proposition 3.
For any type of matrix norm

$$\lim_{m\to\infty} A_m = B \qquad \Longleftrightarrow \qquad \lim_{m\to\infty} \|A_m - B\| = 0$$

One important sequence is

$$I,\ A,\ A^2,\ A^3,\ \ldots \tag{25}$$

It follows from Proposition 1 that

$$\|A^m\|\leqslant \|A\|\,\|A^{m-1}\|\leqslant\cdots\leqslant\|A\|^m.$$

Hence, if for some norm $\|A\|<1$, then the sequence (25) converges to the zero matrix. The infinite series

$$I+A+A^2+A^3+\ \ldots \tag{26}$$

is said to be convergent if the sequence of partial sums is convergent. The partial sums are

$$S_m = I+A+A^2+\cdots+\ A^m$$

and if $\lim\limits_{m\to\infty} S_m = S$, we say that (26) is convergent with sum S. Now

$$\begin{aligned}(I-A)\,S_m &= (I-A)\left(I+A+A^2+\cdots+\ A^m\right)\\ &= \left(I+A+A^2+\cdots+\ A^m\right)-\left(A+A^2+\cdots+\ A^{m+1}\right)\\ &= I-A^{m+1}.\end{aligned}$$

and, if $\|A\|<1$, from Proposition 2, $I-A$ is non-singular, so that

$$S_m = (I-A)^{-1}\left(I-A^{m+1}\right).$$

Thus

$$S_m - (I - A)^{-1} = -(I - A)^{-1} A^{m+1}$$

and

$$\left\| S_m - (I - A)^{-1} \right\| \leqslant \left\| (I - A)^{-1} \right\| \|A\|^{m+1}$$

For $\|A\| < 1$ the fight side tends to zero as m increases and by Proposition 3

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

Thus, we have proved

Corollary 1.
If A is an $n \times n$ matrix such that $\|A\| < 1$, then the series (26) converges with sum $(I - A)^{-1}$.

Here is a variant of the latter Corollary

Theorem 9.
If A and B are $n \times n$ matrices such that $\|I - AB\| < 1$, then A and B are invertible. Furthermore,

$$A^{-1} = B \sum_{k=0}^{\infty} (I - AB)^k, \qquad B^{-1} = \sum_{k=0}^{\infty} (I - AB)^k B$$

Proof. By the preceding Proposition 2, AB is invertible and its inverse is

$$(AB)^{-1} = \sum_{k=0}^{\infty} (I - AB)^k$$

Hence

$$A^{-1} = BB^{-1}A^{-1} = B(AB)^{-1} = B \sum_{k=0}^{\infty} (I - AB)^k,$$

$$B^{-1} = B^{-1}A^{-1}A = (AB)^{-1} A = \sum_{k=0}^{\infty} (I - AB)^k A$$ □

Example 8.
If $A = \begin{bmatrix} 0.3 & 0.2 \\ 0.5 & 0.7 \end{bmatrix}$, *then* $\|A\|_1 = 0.9$, $\quad \|A\|_\infty = 1.2$
Using Proposition 2 with $\|A\|_1$ we deduce that $I + A$ is non-singular and

$$\frac{10}{19} \leqslant \left\| (I - A)^{-1} \right\| \leqslant 10.$$

The sequence $\{A_m\}_{m=0}^{\infty}$ converges to the zero matrix and the series $\sum A^m$ is convergent with sum $(I - A)^{-1}$. Notice that we cannot make these deductions using $\|A\|_\infty$.

Sect.10 Conditioning

Definition 5.
We define the condition number of an $n \times n$ non-singular matrix A for the norm $\|A\|_p$ to be

$$k_p(A) = \|A\|_p \|A^{-1}\|_p \tag{27}$$

The condition number of a matrix A gives a measure of how sensitive systems of equations, with coefficient matrix A, are to small perturbations such as those caused by rounding. We shall see that, for large $k_p(A)$ perturbations may have a large effect on the solution. Suppose that

$$Ax = b \tag{28}$$

where A is non-singular.
(i) If A^{-1} is perturbed to obtain a new matrix B, the solution $x = A^{-1}b$ is perturbed to become a new vector $\tilde{x} = Bb$. How large is this latter perturbation in absolute and relative terms?
We have, using any vector norm and its subordinate matrix norm:

$$\|x - \tilde{x}\| = \|x - Bb\| = \|x - BAx\| = \|(I - BA)x\| \leqslant \|I - BA\| . \|x\|$$

This gives the magnitude of the perturbation in x. If the relative perturbation is being measured, we can write

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leqslant \|I - BA\| \tag{29}$$

Inequality (29) gives an upper bound of the relative error between x and $\tilde{x}$.
(ii) Suppose that the vector b is perturbed to obtain a vector $\tilde{b}$. If x and $\tilde{x}$ satisfy $Ax = b$ and $A\tilde{x} - \tilde{b}$ by how much do x and $\tilde{x}$ differ, in absolute and relative terms?
Assuming that A is invertible, we have

$$\|x - \tilde{x}\| = \|A^{-1}b - A^{-1}\tilde{b}\| = \|A^{-1}\left(b - \tilde{b}\right)\| \leqslant \left\|A^{-1}\right\| . \|b - \tilde{b}\|$$

This gives a measure of the perturbation in x. To estimate the relative perturbation, remark first that

$$\|b\| \leqslant \|A\| \|x\| \qquad \Longrightarrow 1 \leqslant \frac{\|A\| \|x\|}{\|b\|}$$

then we can write

$$\|x - \tilde{x}\| \leqslant \left\|A^{-1}\right\| . \|b - \tilde{b}\| \leqslant \left\|A^{-1}\right\| \|A\| \|x\| \frac{\|b - \tilde{b}\|}{\|b\|} \tag{30}$$

therefore,

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leqslant \left\|A^{-1}\right\| \|A\| \frac{\|b-\tilde{b}\|}{\|b\|} = k(A)\frac{\|b-\tilde{b}\|}{\|b\|},$$

If we solve a system of equations $Ax = b$ numerically, we obtain not the exact solution x but an approximate solution $\tilde{x}$. One can test $\tilde{x}$ by forming $A\tilde{x}$ to see whether it is close to b. Thus, we have the residual vector

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

The difference between the exact solution x and the approximate solution $\tilde{x}$ is called the error vector

$$e = x - \tilde{x}$$

The following relationship

$$Ae = r$$

between the error vector and the residual vector is of fundamental importance.
The following theorem shows that the condition number $k(A)$ plays an important role.

Theorem 10.
We have

$$\frac{1}{k(A)}\frac{\|r\|}{\|b\|} \leqslant \frac{\|e\|}{\|x\|} \leqslant k(A)\frac{\|r\|}{\|b\|}$$

Proof. The inequality on the right can be written as

$$\|e\|\|b\| \leqslant \|A\| \left\|A^{-1}\right\| \|r\|\|x\|$$

and this is true since

$$\|e\|\|b\| = \left\|A^{-1}r\right\| \|Ax\| \leqslant \left\|A^{-1}\right\| \|A\|\|r\|\|x\|$$

which is the inequality (30). The inequality on the left can be written as

$$\|r\|\|x\| \leqslant \|A\| \left\|A^{-1}\right\| \|b\|\|e\|$$

and this follows at once from

$$\|r\|\|x\| = \|Ae\| \left\|A^{-1}b\right\| \leqslant \|A\| \left\|A^{-1}\right\| \|b\|\|e\| \qquad \square$$

It is easily seen that for any non-zero scalar λ

$$k(\lambda A) = k(A)$$

Increasing (or decreasing) λ will increase the elements of λA (or $(\lambda A)^{-1}$) but the condition number will not change. If $k(A) >> 1$ we say that A is ill-conditioned.

Sect.11 Iterative Methods

We again consider the problem of solving n non-singular equations in n unknowns

$$Ax = b \tag{31}$$

If E and F are $n \times n$ matrices such that

$$A = E - F \tag{32}$$

then we call (31) an additive splitting of the matrix A. For such a splitting, (32) may be written as

$$Ex = Fx + b$$

This form of the equations suggests an iterative procedure

$$Ex_{n+1} = Fx_n + b, \qquad n = 0, 1, 2, \ldots$$

for arbitrary x_0. If the sequence is to be uniquely defined for a given x_0, we require E to be non-singular, when

$$x_{n+1} = E^{-1}Fx_n + E^{-1}b. \tag{33}$$

This is of the form (34) and Theorem 11 below which states that the sequence $\{x_n\}_{n=0}^{\infty}$ converges if

$$\|E^{-1}F\| < 1$$

It can also be seen that, in this case, $\{x_n\}_{n\geqslant 0}$ converges to the solution vector x.

Theorem 11.
For the iteration formula

$$x_{k+1} = Ax_k + d \tag{34}$$

to produce a sequence converging to the unique solution $(I - A)^{-1}\, d$*, for any starting vector* x_0 *it is necessary and sufficient that the spectral radius of* A *be less than* 1.

Proof. Suppose that $\rho(A) < 1$. By Corollary 1, there is a subordinate matrix norm such that $\|A\| < 1$. We write

$$\begin{aligned} x_1 &= Ax_0 + d \\ x_2 &= Ax_1 + d = A(Ax_0 + d) + d = A^2x_0 + Ad + d \end{aligned}$$

The general formula is

$$x_{k+1} = A^k x_0 + \sum_{i=0}^{k-1} A^i d \tag{35}$$

Using the vector norm that engendered our matrix norm, we have

$$\left\| A^k x_0 \right\| \leqslant \left\| A^k \right\| \| x_0 \| \leqslant \left\| A \right\|^k \| x_0 \| \xrightarrow[k\to\infty]{} 0$$

By the Corollary 1, we have

$$\sum_{i=0}^{\infty} A^i d = (I - A)^{-1} d$$

Thus by letting $k \to \infty$ in Equation (35), we obtain

$$\lim_{k\to\infty} x_k = (I - A)^{-1} d.$$

For the converse, suppose that $\rho(A) \geqslant 1$. Select u and λ so that

$$Au = \lambda u, \qquad |\lambda| \geqslant 1, \qquad u \neq 0.$$

If $|\lambda| = 1$, let $d = u$ and $x_0 = 0$. By equation (35)

$$x_{k+1} = \sum_{i=0}^{k-1} A^i d = \pm k u$$

which converges as $k \to \infty$. If $|\lambda| > 1$, let $d = 0$ and $x_0 = u$. Similarlym by Equation (35) we have $x_k = A^k u = \lambda^k u$, which converges as $k \to \infty$. □

The simplest choice of E is a diagonal matrix, usually the diagonal of A provided all the diagonal elements are non-zero. We obtain the splitting $A = D - B$, where D is diagonal with non-zero diagonal elements and B has zeros on its diagonal. The relation (10.50) becomes

$$x_{n+1} = D^{-1} B x_n + D^{-1} b, \qquad n = 0, 1, 2, \ldots$$

This is known as **the Jacobi iterative method**. D^{-1} is simply the diagonal matrix whose diagonal elements are the inverses of those of D. The method is convergent if

$$\| D^{-1} B \| < 1$$

This condition is certainly satisfied if the matrix A is a strictly diagonally dominant matrix, that is, a matrix such that

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

$D^{-1}B$ has off-diagonal elements $-a_{ji}/a_{ii}$, and zeros on the diagonal, so that for a strictly diagonally dominant matrix A.

$$\|D^{-1}B\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \max_{1\leqslant i\leqslant n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$$

Another suitable choice of E is a triangular matrix. We split A into

$$A = (D-L) - U$$

where D is diagonal with non-zero diagonal elements, L is lower triangular with zeros on the diagonal and U is upper triangular with zeros on the diagonal. We obtain in place of (33),

$$(D-L)\,x_{n+1} = Ux_n + b. \tag{36}$$

This is known as **the Gauss-Seidel iterative method**. The coefficient matrix $(D-L)$ is lower triangular and x_{n+1} is easily found by forward substitution. There is no need to compute $(D-L)^{-1}U$ explicitly. The Gauss-Seidel method converges if

$$\|\,(D-L)^{-1}U\| < 1$$

Again it can be shown that the Gauss-Seidel method is convergent if the original matrix is diagonally dominant. You will notice that both the Jacobi and Gauss-Seidel methods require the elements on the diagonal of A to be non-zero.

Theorem 12.
If A is diagonally dominant, then the Gauss-Seidel method converges for any starting vector.

Proof. It suffices to prove that

$$\rho\left(I - Q^{-1}A\right) < 1$$

To this end, let A be any eigenvalue of $I-Q^{-1}A$. Let x be a corresponding eigenvector. We assume, with no loss of generality, that $\|x\|_\infty = 1$. We have now

$$\left(I - Q^{-1}A\right)x = \lambda x, \qquad Qx - Ax = \lambda Qx$$

Since Q is the lower triangular part of A, including its diagonal,

$$-\sum_{j=i+1}^{n} a_{ij}x_j = \lambda \sum_{j=1}^{i} a_{ij}x_j \qquad (1\leqslant i\leqslant n)$$

By transposing terms in this equation, we obtain

$$\lambda a_{ii}x_i = -\lambda \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j, \qquad (1\leqslant i\leqslant n)$$

Select an index i such that $|x_i| = 1 \geqslant |x_j|$ for all j. Then

$$|\lambda||a_{ij}| \leqslant |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

Solving for $|A|$ and using the diagonal dominance of A, we get

$$|\lambda| \leqslant \left\{ \sum_{j=i+1}^{n} |a_{ij}| \right\} \left\{ |a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| \right\}^{-1} < 1. \qquad \square$$

Example 9.

Let us apply the Jacobi, then the Gauss-Seidel iterative method on the following matrix

$$\begin{bmatrix} -5 & 1 & -2 & 1 \\ -1 & 5 & 2 & -1 \\ 1 & 2 & -7 & -3 \\ -1 & 1 & 1 & 7 \end{bmatrix} = \begin{bmatrix} -5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & -7 & 0 \\ 0 & 0 & 0 & 7 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & -2 & 0 & 0 \\ 1 & -1 & -1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & -1 & 2 & -1 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

In case of Jacobi iteration method, we have

$$X_{k+1} = D^{-1}(L+U)X_k + D^{-1}b,$$

with

$$M := D^{-1}(L+U) = \begin{bmatrix} 0 & \frac{1}{5} & -\frac{2}{5} & \frac{1}{5} \\ -\frac{1}{5} & 0 & -\frac{2}{5} & \frac{1}{5} \\ -\frac{1}{7} & -\frac{2}{7} & 0 & -\frac{3}{7} \\ -\frac{1}{7} & \frac{1}{7} & \frac{1}{7} & 0 \end{bmatrix}, \qquad \|M\|_\infty = \frac{6}{7}, \quad \|M\|_1 = \frac{33}{35}$$

In case of Gauss-Seidel iteration method, we have

$$X_{k+1} = (D-L)^{-1}UX_k + (D-L)^{-1}b,$$

with

$$N := (D-L)^{-1}U = \begin{bmatrix} 0 & \frac{1}{5} & -\frac{2}{5} & \frac{1}{5} \\ 0 & \frac{1}{25} & -\frac{12}{25} & \frac{6}{25} \\ 0 & \frac{1}{25} & -\frac{34}{175} & -\frac{58}{175} \\ 0 & \frac{3}{175} & \frac{48}{1225} & \frac{51}{1225} \end{bmatrix}, \qquad \|N\|_\infty = \frac{4}{5}, \qquad \|N\|_1 = \frac{1364}{1225}.$$

Therefore, both of the two methods (Jacobi and Gauss-Seidel) converges.