Numerical Methods

King Saud University

Aims

In this lecture, we will . . .

- Consider the System of Linear Equations
- Introduce the Gaussian elimination method
- Introduce the Jacobi Iterative Method
- Introduce the Gauss-Seidel Iterative Method

Introduction

When engineering systems are modeled, the mathematical description is frequently developed in terms of set of algebraic simultaneous equations. Sometimes these equations are non-linear and sometimes linear. Here we discuss systems of simultaneous **linear equations** and describe the numerical methods for the approximate solutions of such systems.

Important Points

I. We will look for the solutions of systems of linear equations.

II. Linear systems may be simultaneous (number of linear equations and unknowns variables are equal) or underdetermined (number of linear equations less than unknowns variables) or overdetermined (number of linear equations more than unknowns variables). Here, we shall discuss only simultaneous systems.

III. Matrix form of linear system is $A\mathbf{x} = \mathbf{b}$, where A called coefficient matrix, column matrix $\mathbf{b}$ is right hand constant and column matrix $\mathbf{x}$ be the unknowns.

IV. Linear systems may be nonhomogeneous (right hand vector $\mathbf{b} \neq 0$) or homogeneous ($\mathbf{b} = 0$).

V. Linear systems may have unique solution or no solution or infinitely many solutions.

VI. Linear systems may be nonsingular (determinant of coefficients matrix A not equal to zero) or singular (determinant of coefficients matrix A equal to zero). Nonsingular systems have unique solution while singular systems have either no solution or infinitely many solutions.

VII. Solutions of linear systems can be obtained by both direct and indirect (iterative) methods.

VII. Linear systems may be well-conditioned (small condition number) or ill-conditioned (large condition number) .

Definition 1

(Linear equation)

It is an equation in which the highest exponent in a variable term is no more than one. The graph of such equation is a straight line. •

A**linear equation** in two variables x_1 and x_2 is an equation that can be written in the form

$$a_1x_1 + a_2x_2 = b,$$

where a_1, a_2 and b are real numbers. Note that this is the equation of a straight line in the plane. For example, the equations

$$5x_1 + 2x_2 = 2, \qquad \frac{4}{5}x_1 + 2x_2 = 1, \qquad 2x_1 - 4x_2 = \pi,$$

are all linear equations in two variables.

A linear equation in n variables $x_1, x_2, \ldots, x_n$ is an equation that can be written as

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b,$$

where $a_1, a_2, \ldots, a_n$ are real numbers and called the *coefficients* of unknown variables $x_1, x_2, \ldots, x_n$ and the real number b, the right-hand side of equation, is called the *constant term* of the equation.

Definition 2

(System of Linear Equations)

A **system of linear equations** (or linear system) is simply a finite set of linear equations.

For example,

$$\begin{array}{rcrcl} 4x_1 & - & 2x_2 & = & 5 \\ 3x_1 & + & 2x_2 & = & 4 \end{array}$$

is the system of two equations in two variables x_1 and x_2, while

$$\begin{array}{rcrcrcrcl} 2x_1 & + & x_2 & - & 5x_3 & + & 2x_4 & = & 9 \\ 4x_1 & + & 3x_2 & + & 2x_3 & + & 4x_4 & = & 3 \\ x_1 & + & 2x_2 & + & 3x_3 & + & 2x_4 & = & 11 \end{array}$$

is the system of three equations in the four variables x_1, x_2, x_3 and x_4.

In order to write a general system of m linear equations in the n variables $x_1, \ldots, x_n$, we have

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \cdots & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array} \tag{1}$$

or, in compact form the system (1) can be written

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \qquad i = 1, 2, \ldots, m. \tag{2}$$

For such system we seek all possible ordered sets of numbers $c_1, \ldots, c_n$ which satisfies all m equations when they are substituted for the variables $x_1, x_2, \ldots, x_n$. Any such set $\{c_1, c_2, \ldots, c_n\}$, is called a *solution* of the system of linear equations (1) or (2).

Theorem 3

(Solution of a Linear System)

Every system of linear equations has either no solution, exactly one solution, or infinitely many solutions. •

Linear System in Matrix Notation

To write the general simultaneous system of n linear equations in the n unknown variables $x_1, x_2, \ldots, x_n$, is

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n \end{array} \tag{3}$$

The system of linear equations (3) can be written as the single matrix equation

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \tag{4}$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix},$$

the **coefficient matrix**, the **column matrix of unknowns**, and the **column matrix of constants**, respectively, then the system (3) can be written very compactly as

$$A\mathbf{x} = \mathbf{b}, \tag{5}$$

which is called the *matrix form* of the system of linear equations (3). The column matrices $\mathbf{x}$ and $\mathbf{b}$ are called *vectors*.

Solutions of Linear Systems of Equations

Now we shall discuss numerical methods for solving system of linear equations. We shall discuss both direct and indirect (iterative) methods for the solution of given linear systems. In direct method we shall discuss the familiar technique called the method of elimination to find the solution of linear systems. This method starts with the augmented matrix of the given linear system and obtain a matrix of a certain form. This new matrix represents a linear system that has exactly the same solutions as the given origin system. In indirect methods we shall discuss Jacobi and Gauss-Seidel methods.

Gaussian Elimination Method

Simple Gaussian Elimination Method

The Gaussian elimination procedure start with *forward elimination*, in which the first equation in the linear system is used to eliminate the first variable from the rest of $(n-1)$ equations. Then the new second equation is used to elimination second variable from the rest of $(n-2)$ equations, and so on. If $(n-1)$ such elimination is performed then the resulting system will be the triangular form. Once this forward elimination is completed, we can determine whether the system is overdetermined or underdetermined or has a unique solution. If it has a unique solution, then the *backward substitution* is used to solve the triangular system easily and one can find the unknown variables involve in the system.

Now we shall describe the method in detail for a system of n linear equations. Consider the following a system of n linear equations:

$$\begin{array}{ccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \cdots & + & a_{2n}x_n & = & b_2 \\
a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & + & \cdots & + & a_{3n}x_n & = & b_3 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
a_{n1}x_1 & + & a_{n2}x_2 & & a_{n3}x_3 & + & \cdots & + & a_{nn}x_n & = & b_n
\end{array} \tag{6}$$

Forward Elimination

Consider first equation of the given system (6)

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n = b_1, \tag{7}$$

as first pivotal equation with first pivot element a_{11}. Then the first equation times multiples $m_{i1} = (a_{i1}/a_{11}),\ i = 2, 3, \ldots, n$, is subtracted from the *ith* equation to eliminate first variable x_1, producing an equivalent system

$$\begin{array}{ccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
 & & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & + & \cdots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\
 & & a_{32}^{(1)}x_2 & + & a_{33}^{(1)}x_3 & + & \cdots & + & a_{3n}^{(1)}x_n & = & b_3^{(1)} \\
 & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
 & & a_{n2}^{(1)}x_2 & + & a_{n3}^{(1)}x_3 & + & \cdots & + & a_{nn}^{(1)}x_n & = & b_n^{(1)}
\end{array} \tag{8}$$

Now consider a second equation of the system (8), which is

$$a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \cdots + a_{2n}^{(1)}x_n = b_2^{(1)}, \tag{9}$$

as second pivotal equation with second pivot element $a_{22}^{(1)}$. Then the second equation times multiples $m_{i2} = (a_{i2}^{(1)}/a_{22}^{(1)}),\ i = 3, \ldots, n$, is subtracted from the *ith* equation to eliminate second variable x_2, producing an equivalent system

$$\begin{array}{ccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
 & & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & + & \cdots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\
 & & & & a_{33}^{(2)}x_3 & + & \cdots & + & a_{3n}^{(2)}x_n & = & b_3^{(2)} \\
 & & & & \vdots & & \vdots & & \vdots & & \vdots \\
 & & & & a_{n3}^{(2)}x_3 & + & \cdots & + & a_{nn}^{(2)}x_n & = & b_n^{(2)}
\end{array} \tag{10}$$

Now consider a third equation of the system (10), which is

$$a_{33}^{(2)} x_3 + \cdots + a_{3n}^{(2)} x_n = b_3^{(2)}, \tag{11}$$

as the third pivotal equation with third pivot element $a_{33}^{(2)}$. Then the third equation times multiples $m_{i3} = (a_{i3}^{(2)}/a_{33}^{(2)})$, $i = 4, \ldots, n$, is subtracted from the *ith* equation to eliminate third variable x_3. Similarly, after *(n-1)th* steps, we have the *nth* pivotal equation which have only one unknown variable x_n, that is

$$\begin{array}{ccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & + & \cdots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\
 & & & + & a_{33}^{(2)}x_3 & + & \cdots & + & a_{3n}^{(2)}x_n & = & b_3^{(2)} \\
 & & & & & & & & \vdots & & \vdots \\
 & & & & & & & & a_{nn}^{(n-1)}x_n & = & b_n^{(n-1)}
\end{array} \tag{12}$$

with *nth* pivotal element $a_{nn}^{(n-1)}$. After getting the upper-triangular system which is equivalent to the original system, the forward elimination is completed. After the triangular set of equations has been obtained, the last equation of the system (12) yields the value of x_n directly. The value is then substituted into the equation next to the last one of the system (12) to obtain a value of x_{n-1}, which is, in turn, used along with the value of x_n in the second to the last equation to obtain a value of x_{n-2}, and so on.

Example 0.1

Solve the following linear system using the simple Gaussian elimination method

$$\begin{array}{rcrcrcl} x_1 & + & 2x_2 & + & x_3 & = & 2 \\ 2x_1 & + & 5x_2 & + & 3x_3 & = & 1 \\ x_1 & + & 3x_2 & + & 4x_3 & = & 5 \end{array}$$

solution. The process begins with the augmented matrix form

$$\left(\begin{array}{cccc:c} 1 & 2 & 1 & \vdots & 2 \\ 2 & 5 & 3 & \vdots & 1 \\ 1 & 3 & 4 & \vdots & 5 \end{array}\right).$$

Since $a_{11} = 1 \neq 0$, so we wish to eliminate the elements a_{21} and a_{31} by subtracting from the second and third rows the appropriate multiples of the first row. In this case the multiples are $m_{21} = \dfrac{2}{1} = 2$ and $m_{31} = \dfrac{1}{1} = 1$. Hence

$$\left(\begin{array}{cccc:c} 1 & 2 & 1 & \vdots & 2 \\ 0 & 1 & 1 & \vdots & -3 \\ 0 & 1 & 3 & \vdots & 3 \end{array}\right).$$

As $a_{22}^{(1)} = 1 \neq 0$, therefore, we eliminate entry in $a_{32}^{(1)}$ position by subtracting the multiple $m_{32} = \dfrac{1}{1} = 1$ of the second row from the third row, to get

$$\left(\begin{array}{cccc:c} 1 & 2 & 1 & & 2 \\ 0 & 1 & 1 & & -3 \\ 0 & 0 & 2 & & 6 \end{array}\right).$$

Obviously, the original set of equations has been transformed to an upper-triangular form. Since all the diagonal elements of the obtaining upper-triangular matrix are nonzero, which means that the coefficient matrix of the given system is nonsingular and therefore, the given system has a unique solution. Now expressing the set in algebraic form yields

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & x_3 & = & 2 \\ & & x_2 & + & x_3 & = & -3 \\ & & & & 2x_3 & = & 6 \end{array}$$

Now using backward substitution, we get

$$\begin{array}{lcll} 2x_3 = 6, & & \text{gives} & x_3 = 3, \\ x_2 = -x_3 - 3 = -(3) - 3 = -6, & & \text{gives} & x_2 = -6, \\ x_1 = 2 - 2x_2 - x_3 = 2 - 2(-6) - 3, & & \text{gives} & x_1 = 11, \end{array}$$

which is the required solution of the given system. •

Example 0.2

Solve the following linear system using the simple Gaussian elimination method

$$\begin{array}{rcrcrcl} & & x_2 & + & x_3 & = & 1 \\ x_1 & + & 2x_2 & + & 2x_3 & = & 1 \\ 2x_1 & + & x_2 & + & 2x_3 & = & 3 \end{array}$$

Solution. Writing the given system in the augmented matrix form

$$\left(\begin{array}{ccc:c} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 1 \\ 2 & 1 & 2 & 3 \end{array}\right).$$

To solve this system, the simple Gaussian elimination method will fail immediately because the element in the first row on the leading diagonal, the pivot, is zero. Thus it is impossible to divide that row by the pivot value. Clearly, this difficulty can be overcome by rearranging the order of the rows; for example by making the first row the second, gives

$$\left(\begin{array}{ccc:c} 1 & 2 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 2 & 1 & 2 & 3 \end{array}\right).$$

Now we use the usual elimination process. The first elimination step is to eliminate the element $a_{31} = 2$ from the third row by subtracting a multiple $m_{31} = \dfrac{2}{1} = 2$ of row 1 from row 3, gives

$$\left(\begin{array}{cccc:c} 1 & 2 & 2 & & 1 \\ 0 & 1 & 1 & & 1 \\ 0 & -3 & -2 & & 1 \end{array}\right).$$

We finished with the first elimination step since the element a_{21} is already eliminated from second row. The second elimination step is to eliminate the element $a_{32}^{(1)} = -3$ from the third row by subtracting a multiple $m_{32} = \dfrac{-3}{1}$ of row 2 from row 3, gives

$$\left(\begin{array}{cccc:c} 1 & 2 & 2 & & 1 \\ 0 & 1 & 1 & & 1 \\ 0 & 0 & 1 & & 4 \end{array}\right).$$

Obviously, the original set of equations has been transformed to an upper-triangular form. Now expressing the set in algebraic form yields

$$\begin{array}{rcrcrcl} x_1 & + & 2x_2 & + & 2x_3 & = & 1 \\ & & x_2 & + & x_3 & = & 1 \\ & & & & x_3 & = & 4 \end{array}$$

Using backward substitution, we get, $x_1 = -1,\ x_2 = -3,\quad x_3 = 4$, the solution of the system.

Example 0.3

For what values of α the following linear system has (i) Unique solution, (ii) No solution, (iii) Infinitely many solutions, by using the simple Gaussian elimination method. Use smallest positive integer value of α to get the unique solution of the system.

$$\begin{array}{rcrcrcl} x_1 & + & 3x_2 & + & \alpha x_3 & = & 4 \\ 2x_1 & - & x_2 & + & 2\alpha x_3 & = & 1 \\ \alpha x_1 & + & 5x_2 & + & x_3 & = & 6 \end{array}$$

Solution. Writing the given system in the augmented matrix form

$$\begin{pmatrix} 1 & 3 & \alpha & 4 \\ 2 & -1 & 2\alpha & 1 \\ \alpha & 5 & 1 & 6 \end{pmatrix},$$

and by using the following multiples $m_{21} = 2$ and $m_{31} = \alpha$, we get

$$\begin{pmatrix} 1 & 3 & \alpha & 4 \\ 0 & -7 & 0 & -7 \\ 0 & 5-3\alpha & 1-\alpha^2 & 6-4\alpha \end{pmatrix}.$$

Now using the multiple $m_{32} = \dfrac{5-3\alpha}{-7}$, gives

$$\begin{pmatrix} 1 & 3 & \alpha & 4 \\ 0 & -7 & 0 & -7 \\ 0 & 0 & 1-\alpha^2 & 1-\alpha \end{pmatrix}.$$

So if $1-\alpha^2 \neq 0$, then we have the unique solution of the given system while for $\alpha = \pm 1$, we have no unique solution. If $\alpha = 1$, then we have infinitely many solution because third row of above matrix gives

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

and when $\alpha = -1$, we have

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

which is not possible, so no solution.
Since we can not take $\alpha = 1$ for the unique solution, so can take next positive integer $\alpha = 2$, which gives us upper-triangular system of the form

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & + & 2x_3 & = & 4 \\ & - & 7x_2 & & & = & -7 \\ & & & - & 3x_3 & = & -1 \end{array}$$

Solving this system using backward substitution, we get,
$x_1 = 1/3,\ x_2 = 1,\ x_3 = 1/3$, the required unique solution of the given system using smallest positive integer value of α. •

Theorem 4

An upper-triangular matrix A is nonsingular if and only if all its diagonal elements are different from zero. •

Example 0.4

Use the simple Gaussian elimination method to find all the values of α which make the following matrix singular.

$$A = \begin{pmatrix} 1 & -1 & \alpha \\ 2 & 2 & 1 \\ 0 & \alpha & -1.5 \end{pmatrix}.$$

Solution. Applying the forward elimination step of the simple Gaussian elimination on the given matrix A and eliminate the element a_{21} by subtracting from the second row the appropriate multiple of the first row. In this case the multiple is given as

$$\begin{pmatrix} 1 & -1 & \alpha \\ 0 & 4 & 1-2\alpha \\ 0 & \alpha & -1.5 \end{pmatrix}.$$

We finished with the first elimination step. The second elimination step is to eliminate element $a_{32}^{(1)} = \alpha$ by subtracting a multiple $m_{32} = \dfrac{\alpha}{4}$ of row 2 from row 3, gives

$$\begin{pmatrix} 1 & -1 & \alpha \\ 0 & 4 & 1-2\alpha \\ 0 & 0 & -1.5 - \dfrac{\alpha(1-2\alpha)}{4} \end{pmatrix}.$$

To show that the given matrix is singular, we have to set the third diagonal element equal to zero (by Theorem 4), that is

$$-1.5 - \frac{\alpha(1-2\alpha)}{4} = 0, \quad \text{or} \quad 2\alpha^2 - \alpha - 6 = 0.$$

Solving the above quadratic equation, we get, $\alpha = -\frac{3}{2}$ and $\alpha = 2$, the possible values of α which make the given matrix singular. •

Procedure

[Gaussian Elimination Method]

1. Form the augmented matrix, $B = [A|\mathbf{b}]$.
2. Check first pivot element $a_{11} \neq 0$, then move to the next step; otherwise, interchange rows so that $a_{11} \neq 0$.
3. Multiply row one by multiplier $m_{i1} = \dfrac{a_{i1}}{a_{11}}$ and subtract to the *ith* row for $i = 2, 3, \ldots, n$.
4. Repeat the steps *2* and *3* for the remaining pivots elements unless coefficient matrix A becomes upper-triangular matrix U.
5. Use backward substitution to solve x_n from the *nth* equation $x_n = \dfrac{b_n^{n-1}}{a_{nn}}$ and solve the other (n-1) unknowns variables by using

$$\left.\begin{aligned} x_n &= \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}} \\ x_{n-1} &= \frac{1}{a_{n-1n-1}^{(n-2)}}\left(b_{n-1}^{(n-2)} - a_{n-1n}^{(n-2)} x_n\right) \\ &\vdots \\ x_1 &= \frac{1}{a_{11}}\left(b_1 - \sum_{j=2}^{n} a_{1j} x_j\right) \end{aligned}\right\} \tag{13}$$

Partial Pivoting

In using the Gaussian elimination by partial pivoting(or row pivoting), the basic approach is to use the largest (in absolute value) element on or below the diagonal in the column of current interest as the pivotal element for elimination in the rest of that column.
One immediate effect of this will be to force all the multiples used to be not greater than 1 in absolute value. This will inhibit the growth of error in the rest of elimination phase and in subsequent backward substitution.
At stage k of forward elimination, it is necessary, therefore, to be able to identify the largest element from $|a_{kk}|, |a_{k+1,k}|, \ldots, |a_{nk}|$, where these a_{ik}'s are the elements in the current partially triangularized coefficient matrix. If this maximum occurs in row p, then pth and kth rows of the augmented matrix are interchange and the elimination proceed as usual. In solving n linear equations, a total of $N = \dfrac{n(n+1)}{2}$ coefficients must be examined.

Example 0.5

Solve the following linear system using the Gaussian elimination with **partial pivoting**

$$\begin{array}{rcrcrcr} x_1 & + & x_2 & + & x_3 & = & 1 \\ 2x_1 & + & 3x_2 & + & 4x_3 & = & 3 \\ 4x_1 & + & 9x_2 & + & 16x_3 & = & 11 \end{array}$$

Solution. For the first elimination step, since 4 is the largest absolute coefficient of first variable x_1, therefore, the first row and the third row are interchange, giving us

$$\begin{array}{rcrcrcr} 4x_1 & + & 9x_2 & + & 16x_3 & = & 11 \\ 2x_1 & + & 3x_2 & + & 4x_3 & = & 3 \\ x_1 & + & x_2 & + & x_3 & = & 1 \end{array}$$

Eliminate first variable x_1 from the second and third rows by subtracting the multiples $m_{21} = \dfrac{2}{4}$ and $m_{31} = \dfrac{1}{4}$ of row 1 from row 2 and row 3 respectively, gives

$$\begin{array}{rcrcrcr} 4x_1 & + & 9x_2 & + & 16x_3 & = & 11 \\ & - & 3/2x_2 & - & 4x_3 & = & -5/2 \\ & - & 5/4x_2 & - & x_3 & = & -7/5 \end{array}$$

For the second elimination step, $-3/2$ is the largest absolute coefficient of second variable x_2, so eliminate second variable x_2 from the third row by subtracting the multiple $m_{32} = \dfrac{5}{6}$ of row 2 from row 3, gives

$$\begin{array}{rcrcrcr} 4x_1 & + & 9x_2 & + & 16x_3 & = & 11 \\ & - & 3/2x_2 & - & 4x_3 & = & -5/2 \\ & & & & 1/3x_3 & = & 1/3 \end{array}$$

Obviously, the original set of equations has been transformed to an equivalent upper-triangular form. Now using backward substitution, gives,
$x_1 = 1,\ x_2 = -1,\ x_3 = 1$, the required solution. •

Strictly Diagonally Dominant Matrix

Definition 5

A square matrix is said to be strictly diagonally dominant (SDD) if the absolute value of each element on the main diagonal is greater than the sum of the absolute values of all the other elements in that row. Thus, **strictly diagonally dominant matrix** is defined as

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \qquad \text{for} \quad i = 1, 2, \ldots, n. \tag{14}$$

Example 0.6

The matrix

$$A = \begin{pmatrix} 7 & 3 & 1 \\ 1 & 6 & 3 \\ -2 & 4 & 8 \end{pmatrix},$$

is strictly diagonally dominant since

$$\begin{array}{lcll} |7| & > & |3| + |1|, & \text{that is,} \quad 7 > 4, \\ |6| & > & |1| + |3|, & \text{that is,} \quad 6 > 4, \\ |8| & > & |-2| + |4|, & \text{that is,} \quad 8 > 6, \end{array}$$

but the following matrix

$$B = \begin{pmatrix} 6 & -3 & 4 \\ 3 & 7 & 3 \\ 5 & -4 & 10 \end{pmatrix},$$

is not strictly diagonally dominant since

$$|6| > |-3| + |4|, \quad \text{that is,} \quad 6 > 7,$$

Theorem 6

If a matrix A is strictly diagonally dominant, then:

1. *Matrix A is nonsingular.*
2. *Gaussian elimination without row interchange can be performed on the linear system $A\mathbf{x} = \mathbf{b}$.* •

Example 0.7

Solve the following linear system using the simple Gaussian elimination method.

$$\begin{array}{rcrcrcr} 5x_1 & + & x_2 & + & x_3 & = & 7 \\ 2x_1 & + & 6x_2 & + & x_3 & = & 9 \\ x_1 & + & 2x_2 & + & 9x_3 & = & 12 \end{array}$$

solution. Start with the augmented matrix form

$$\left(\begin{array}{cccc} 5 & 1 & 1 & 7 \\ 2 & 6 & 1 & 9 \\ 1 & 2 & 9 & 12 \end{array}\right),$$

and since $a_{11} = 5 \neq 0$, so we can eliminate the elements a_{21} and a_{31} by subtracting from the second and third rows the appropriate multiples of the first row. In this case the multiples are given

$$m_{21} = \frac{2}{5} \quad \text{and} \quad m_{31} = \frac{1}{5}.$$

Hence

$$\left(\begin{array}{ccc:c} 5 & 1 & 1 & 7 \\ 0 & 28/5 & 3/5 & 31/5 \\ 0 & 9/5 & 44/5 & 53/5 \end{array}\right).$$

As $a_{22}^{(1)} = 28/5 \neq 0$, therefore, we eliminate entry in $a_{32}^{(1)}$ position by subtracting the multiple $m_{32} = \dfrac{1.8}{5.6} = 9/28$ of the second row from the third row, to get

$$\left(\begin{array}{ccc:c} 5 & 1 & 1 & 7 \\ 0 & 28/5 & 3/5 & 31/5 \\ 0 & 0 & 43/5 & 43/5 \end{array}\right).$$

Obviously, the original set of equations has been transformed to an upper-triangular form. Since all the diagonal elements of the obtaining upper-triangular matrix are nonzero, which means that the coefficient matrix of the given system is *nonsingular* and therefore, the given system has a unique solution. Now expressing the set in algebraic form yields

$$\begin{array}{rcrcrcr} 5x_1 & + & x_2 & + & x_3 & = & 7 \\ & & (28/5)x_2 & + & (3/5)x_3 & = & 31/5 \\ & & & & (43/5)x_3 & = & 43/5 \end{array}$$

Now using backward substitution to get the solution of the system as

$$\begin{array}{lcll} (43/5)x_3 & = & 43/5, & \text{gives} \quad x_3 = 1, \\ (28/5)x_2 & = & -(3/5)x_3 + 31/5, & \text{gives} \quad x_2 = 1, \\ 5x_1 & = & 7 - x_2 - x_3, & \text{gives} \quad x_1 = 1. \end{array}$$

.

Norms of Vectors and Matrices

For solving linear systems, we discuss a method for quantitatively measuring the distance between vectors in $\mathbf{R}^n$, the set of all column vectors with real components, to determine whether the sequence of vectors that results from using an direct method converges to a solution of the system. To define a distance in $\mathbf{R}^n$, we use the notation of the *norm* of a vector.

Vector Norms

It is sometimes useful to have a scalar measure of the magnitude of a vector. Such a measure is called a *vector norm* and for a vector $\mathbf{x}$ is written as $\|\mathbf{x}\|$.
A **vector norm** on $\mathbf{R}^n$ is a function, from $\mathbf{R}^n$ to $\mathbf{R}$ satisfying:

1. $\|\mathbf{x}\| > 0 \quad$ for all $\quad \mathbf{x} \in \mathbf{R}^n$.
2. $\|\mathbf{x}\| = 0 \quad$ if and only if $\quad \mathbf{x} = \mathbf{0}$.
3. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|, \quad$ for all $\quad \alpha \in \mathbf{R}, \quad \mathbf{x} \in \mathbf{R}^n$.
4. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|, \quad$ for all $\quad \mathbf{x}, \mathbf{y} \in \mathbf{R}^n$.

There are three norms in $\mathbf{R}^n$ that are most commonly used in applications, called l_1-norm, l_2-norm, and l_∞-norm, and are defined for the given **vectors**

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$$

as

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}, \qquad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

The l_1-norm is called the *absolute norm*, the l_2-norm is frequently called the *Euclidean norm* as it is just the formula for distance in ordinary three-dimensional Euclidean space extended to dimension n. Finally, the l_∞-norm is called the *maximum norm* or occasionally the *uniform norm*. All these three norms are also called the *natural norms*.
Here we will consider the vector l_∞-norm only.

Example 0.8

Compute l_p-norms $(p = 1, 2, \infty)$ of the vector $\mathbf{x} = [-5, 3, -2]^T$ in $\mathbf{R}^3$.

Solution. These l_p-norms $(p = 1, 2, \infty)$ of the given vector are:

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + |x_3| = |-5| + |3| + |-2| = 10.$$

$$\|\mathbf{x}\|_2 = (x_1^2 + x_2^2 + x_3^2)^{1/2} = \left[(-5)^2 + (3)^2 + (-2)^2\right]^{1/2} \approx 6.1644.$$

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, |x_3|\} = \max\{|-5|, |3|, |-2|\} = 5.$$

Matrix Norms

A **matrix norm** is a measure of how well one matrix approximates another, or, more accurately, of how well their difference approximates the zero matrix. An iterative procedure for inverting a matrix produces a sequence of approximate inverses. Since in practices such a process must be terminated, it is desirable to have some measure of the error of approximate inverse.
So a matrix norm on the set of all $n \times n$ matrices is a real-valued function, $\|.\|$, defined on this set, satisfying for all $n \times n$ matrices A and B and all real number α as follows:

1. $\|A\| > 0, \qquad A \neq \mathbf{0}.$
2. $\|A\| = 0, \qquad A = \mathbf{0}.$
3. $\|I\| = 1, \qquad I \textit{ is the identity matrix.}$
4. $\|\alpha A\| = |\alpha|\|A\|, \quad \text{for some scalar} \quad \alpha \in \mathbf{R}.$
5. $\|A + B\| \leq \|A\| + \|B\|.$
6. $\|AB\| \leq \|A\|\|B\|.$
7. $\|A - B\| \geq \Big|\|A\| - \|B\|\Big|.$

Several norms for matrices have been defined, we shall use the following three natural norms l_1, l_2, and l_∞ for a square matrix of order n:

$$\|A\|_1 = \max_j \left(\sum_{i=1}^{n} |a_{ij}| \right) = \text{maximum column-sum.}$$

$$\|A\|_2 = \max_{\|x\|_2=1} \|A\mathbf{x}\|_2 = \text{spectral norm.}$$

$$\|A\|_\infty = \max_i \left(\sum_{j=1}^{n} |a_{ij}| \right) = \text{row-sum norm.}$$

For $m \times n$ matrix, we can paraphrase the *Frobenius norm* (or *Euclidean norm*), which is not a natural norm and is define as

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

It can be shown that

$$\|A\|_F = \sqrt{tr(A^T A)},$$

where $tr(A^T A)$ is the *trace* of a matrix $A^T A$, that is, the sum of the diagonal entries of $A^T A$. The Frobenius norm of a matrix is a good measure of the magnitude of a matrix. It is to be noted that $\|A\|_F \neq \|A\|_2$. For a diagonal matrix, all norms have the same values.

Also, here we will consider the matrix l_∞-norm only.

Example 0.9

Compute l_p-norms $(p = \infty, F)$ of the following matrix

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

l_∞-norm is defined as

$$\sum_{j=1}^{3} |a_{1j}| = |4| + |2| + |-1| = 7,$$
$$\sum_{j=1}^{3} |a_{2j}| = |3| + |5| + |-2| = 10,$$
$$\sum_{j=1}^{3} |a_{3j}| = |1| + |-2| + |7| = 10,$$

so

$$\|A\|_\infty = max\{7, 10, 10\} = 10.$$

In addition, we have the l_F-norm of the matrix as

$$\|A\|_F = (16 + 4 + 1 + 9 + 25 + 4 + 1 + 4 + 49)^{1/2} \approx 10.6301,$$

the Frobenius norm of the given matrix. •

Iterative Methods for Solving Linear Systems

The methods discussed in the previous section for the solution of the system of linear equations have been direct, which required a finite number of arithmetic operations. The elimination methods of solving such systems usually yield sufficiently accurate solutions for approximately 20 to 25 simultaneous equations, where most of the unknowns are present in all of the equations. When the coefficients matrix is sparse (has many zeros), a considerably large number of equations can be handled by the **elimination methods**. But these methods are generally impractical when many hundreds or thousands of equations must be solved simultaneously.

There are, however, several methods which can be used to solve large numbers of simultaneous equations. These methods are, called *iterative methods* by which an approximation to the solution of a system of linear equations may be obtained. Here, we consider just two of these iterative methods. These two forms the basis of a family of methods which are designed either to accelerate the convergence or to suit some particular computer architecture.

Jacobi Iterative Method

This is one of the easiest iterative method to find the approximate solution of the system of linear equations

$$A\mathbf{x} = \mathbf{b}, \tag{15}$$

To explain its procedure, consider a system of three linear equations as follows:

$$\begin{array}{rcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & b_3 \end{array}$$

The solution process starts by solving for the first variable x_1 from first equation, second variable x_2 from second equation and third variable x_3 from third equation, gives

$$\begin{array}{rcrcrcl} a_{11}x_1 & = & b_1 & - & a_{12}x_2 & - & a_{13}x_3 \\ a_{22}x_2 & = & b_2 & - & a_{21}x_1 & - & a_{23}x_3 \\ a_{33}x_3 & = & b_3 & - & a_{31}x_1 & - & a_{32}x_2 \end{array}$$

Divide both sides of the above three equations by their diagonal elements, a_{11}, a_{22} and a_{33} respectively, to have

$$\begin{array}{rcrcrcl} x_1 & = & \frac{1}{a_{11}}\Big[b_1 & - & a_{12}x_2 & - & a_{13}x_3\Big] \\ x_2 & = & \frac{1}{a_{22}}\Big[b_2 & - & a_{21}x_1 & - & a_{23}x_3\Big] \\ x_3 & = & \frac{1}{a_{33}}\Big[b_3 & - & a_{31}x_1 & - & a_{32}x_2\Big] \end{array}$$

Let $\mathbf{x}^{(k)} = \left[x_1^{(k)}, x_2^{(k)}, x_3^{(k)}\right]^T$ be an initial solution of the exact solution $\mathbf{x}$ of the linear system (22), then we define an iterative sequence

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{11}}\left[b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right] \\ x_2^{(k+1)} &= \frac{1}{a_{22}}\left[b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)}\right] \\ x_3^{(k+1)} &= \frac{1}{a_{33}}\left[b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)}\right] \end{aligned} \tag{16}$$

where k is the number of iterative steps. Then the form (16) is called the Jacobi formula for system of three equations. For a general system of n linear equations, the **Jacobi method** is defined by

$$\begin{aligned} x_i^{(k+1)} &= \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right] \\ i &= 1, 2, \ldots, n, \quad k = 0, 1, 2, \ldots, \end{aligned} \tag{17}$$

provided that the diagonal elements $a_{ii} \neq 0$ for each $i = 1, 2, \ldots, n$. If the diagonal elements equal to zero, then reordering of the equations can be performed so that no element in the diagonal position equal to zero. As usual with iterative methods, an initial approximation $x_i^{(0)}$ must be supplied. If we don't have knowledge about the exact solution, it is conventional to start with $x_i^{(0)} = \mathbf{0}$ for all i. The iterations defined by (17) are stopped when

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \epsilon, \tag{18}$$

or by using other possible stopping criteria

$$\frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|}{\|\mathbf{x}^{(k+1)}\|} < \epsilon, \tag{19}$$

where ϵ is a preassigned small positive number. For this purpose, any convenient norm can be used, the most usual being the l_∞-norm.

Example 0.10

Solve the following system of equations using the Jacobi iterative method, using $\epsilon = 10^{-6}$ in the l_∞-norm.

$$\begin{array}{rcrcrcr} 5x_1 & - & x_2 & + & x_3 & = & 10 \\ 2x_1 & + & 8x_2 & - & x_3 & = & 11 \\ -x_1 & + & x_2 & + & 4x_3 & = & 3 \end{array}$$

Start with the initial solution $\mathbf{x}^{(0)} = [0, 0, 0]^T$.

Solution. The Jacobi iterative method for the given system has the form

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{5}\left[10 + x_2^{(k)} - x_3^{(k)}\right] \\ x_2^{(k+1)} &= \frac{1}{8}\left[11 - 2x_1^{(k)} + x_3^{(k)}\right] \\ x_3^{(k+1)} &= \frac{1}{4}\left[3 + x_1^{(k)} - x_2^{(k)}\right] \end{aligned}$$

and starting with initial approximation $x_1^{(0)} = 0, x_2^{(0)} = 0, x_3^{(0)} = 0$, then for $k = 0$, we obtain

$$\begin{aligned}
x_1^{(1)} &= \frac{1}{5}\Big[10 + x_2^{(0)} - x_3^{(0)}\Big] = \frac{1}{5}\Big[10 + 0 - 0\Big] = 2,\\
x_2^{(1)} &= \frac{1}{8}\Big[11 - 2x_1^{(0)} + x_3^{(0)}\Big] = \frac{1}{8}\Big[11 - 0 + 0\Big] = 1.375,\\
x_3^{(1)} &= \frac{1}{4}\Big[3 + x_1^{(0)} - x_2^{(0)}\Big] = \frac{1}{4}\Big[3 + 0 - 0\Big] = 0.75.
\end{aligned}$$

The first and subsequent iterations are listed in Table 1.

Table: Solution of the Example 0.10

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000	0.000000	0.000000
1	2.000000	1.375000	0.750000
2	2.125000	0.968750	0.906250
⋮	⋮	⋮	⋮
15	2.000000	0.999999	1.000000
16	2.000000	1.000000	1.000000

Note that the Jacobi method converges and after 16 iterations we obtained what is obviously the exact solution. Ideally the iteration should stop automatically when we obtained the required accuracy using one of the stopping criteria mentioned by (18) or (19).
To get the above results using MATLAB command, we do the following:

$$>> Ab = [A|b] = [5\ \ -1\ 1\ 10; 2\ 8\ \ -1\ 11; -1\ 1\ 4\ 3];$$
$$>> x = [0\ 0\ 0];\ \ acc = 0.5e-6;\ \ JacobiM(Ab, x, acc);$$

Example 0.11

Solve the following system of equations using the Jacobi iterative method.

$$\begin{array}{rcrcrcr} 2x_1 & + & 8x_2 & - & x_3 & = & 11 \\ 5x_1 & - & x_2 & + & x_3 & = & 10 \\ -x_1 & + & x_2 & + & 4x_3 & = & 3 \end{array}$$

Start with the initial solution $\mathbf{x}^{(0)} = [0, 0, 0]^T$.

Solution. Results for this linear system are listed in Table 2. •

Table: Solution of the Example 0.11

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000	0.000000	0.000000
1	5.500000	-10.0000	0.750000
2	45.87500	18.25000	4.625000
3	-65.1875	224.0000	7.656250

Notice that Jacobi method diverges rapidly. Although the given linear system is same as the linear system of the previous Example 0.10 except the first and second equations are interchanged. From this example we concluded that Jacobi iterative method is not always convergent.

Program 3.10

MATLAB m-file for the Jacobi Iterative Method for Linear System
function x=JacobiM(Ab,x,acc)
[n,t]=size(Ab); b=Ab(1:n,t); R=1; k=1; d(1,1:n+1)=[0 x]; while $R > acc$
for i=1:n; sum=0; for j=1:n; if j ~ =i
$sum = sum + Ab(i,j) * d(k,j+1)$; end; $x(1,i) = (1/Ab(i,i)) * (b(i,1) - sum)$;end;end
k=k+1; d(k,1:n+1)=[k-1 x]; R=max(abs((d(k,2:n+1)-d(k-1,2:n+1))));
if $k > 10$ & $R > 100$ ('Jacobi Method is diverges') break; end; end; x=d;

Procedure [Jacobi Method]

1. Check the coefficient matrix A is strictly diagonally dominant (for guaranteed convergence).
2. Initialize the first approximation $\mathbf{x}^{(0)}$ and pre-assigned accuracy ϵ.
3. Compute the constant $\mathbf{c} = D^{-1}\mathbf{b} = \dfrac{b_i}{a_{ii}}$, for $i = 1, 2, \ldots, n$.
4. Compute the Jacobi iteration matrix $T_J = -D^{-1}(L+U)$.
5. Solve for the approximate solutions $\mathbf{x}_i^{(k+1)} = T_J\mathbf{x}_i^{(k)} + \mathbf{c}$, $i = 1, 2, \ldots, n$ and $k = 0, 1, \ldots$
6. Repeat step *5* until $\|\mathbf{x}_i^{(k+1)} - \mathbf{x}_i^{(k)}\| < \epsilon$.

Gauss-Seidel Iterative Method

This is one of the most popular and widely used iterative method to find the approximate solution of the system of linear equations. This iterative method is a modification of the Jacobi iterative method and give us good accuracy by using the most recently calculated values.
From the Jacobi iterative formula (17), it is seen that the new estimates for solution $\mathbf{x}$ are computed from the old estimates and only when all the new estimates have been determined are then used in the right-hand side of the equation to perform the next iteration. But the Gauss-Seidel method is to make use of the new estimates in the right-hand side of the equation as soon as they become available. For example, the Gauss-Seidel formula for the system of three equations can be define an iterative sequence

$$
\begin{aligned}
x_1^{(k+1)} &= \frac{1}{a_{11}}\left[b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right] \\
x_2^{(k+1)} &= \frac{1}{a_{22}}\left[b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}\right] \qquad (20) \\
x_3^{(k+1)} &= \frac{1}{a_{33}}\left[b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)}\right]
\end{aligned}
$$

For a general system of n linear equations, the **Gauss-Seidel iterative method** defined as

$$
\begin{aligned}
x_i^{(k+1)} &= \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right] \qquad (21) \\
i &= 1, 2, \ldots, n, \quad k = 0, 1, 2, \ldots
\end{aligned}
$$

The Gauss-Seidel iterative method is sometimes called the method of *successive iteration*, because the most recent values of all $\mathbf{x}_i$ are used in the calculation.

Example 0.12

Solve the following system of equations using the Gauss-Seidel iterative method, with $\epsilon = 10^{-6}$ in l_∞-norm.

$$\begin{array}{rcrcrcr} 5x_1 & - & x_2 & + & x_3 & = & 10 \\ 2x_1 & + & 8x_2 & - & x_3 & = & 11 \\ -x_1 & + & x_2 & + & 4x_3 & = & 3 \end{array}$$

Start with the initial solution $\mathbf{x}^{(0)} = [0, 0, 0]^T$.

Solution. The Gauss-Seidel iteration for the given system is

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{5}\left[10 + x_2^{(k)} - x_3^{(k)}\right] \\ x_2^{(k+1)} &= \frac{1}{8}\left[11 - 2x_1^{(k+1)} + x_3^{(k)}\right] \\ x_3^{(k+1)} &= \frac{1}{4}\left[3 + x_1^{(k+1)} - x_2^{(k+1)}\right] \end{aligned}$$

and starting with initial approximation $x_1^{(0)} = 0, x_2^{(0)} = 0, x_3^{(0)} = 0$, then for $k = 0$, we obtain

$$\begin{aligned}
x_1^{(1)} &= \frac{1}{5}\Big[10 + x_2^{(0)} - x_3^{(0)}\Big] = \frac{1}{5}\Big[10 + 0 - 0\Big] = 2,\\
x_2^{(1)} &= \frac{1}{8}\Big[11 - 2x_1^{(1)} + x_3^{(0)}\Big] = \frac{1}{8}\Big[11 - 4 + 0\Big] = 0.875,\\
x_3^{(1)} &= \frac{1}{4}\Big[3 + x_1^{(1)} - x_2^{(1)}\Big] = \frac{1}{4}\Big[3 + 2 - 0.875\Big] = 1.03125.
\end{aligned}$$

The first and subsequent iterations are listed in Table 3.

Table: Solution of the Example 0.12

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000	0.000000	0.000000
1	2.000000	0.875000	1.031250
2	1.968750	1.011719	0.989258
3	2.004492	0.997534	1.001740
⋮	⋮	⋮	⋮
9	2.000000	0.999999	1.000000
10	2.000000	1.000000	1.000000

Note that the Gauss-Seidel method converged and required 10 iterations to obtain the correct solution for the given system, which is 6 iterations less than required by the Jacobi method for the same Example 0.10.

The above results can be obtained using MATLAB command as follows:

$$>> Ab = [A|b] = [5\ \ -1\ 1\ 10; 2\ 8\ \ -1\ 11; -1\ 1\ 4\ 3];$$
$$>> x = [0\ 0\ 0];\ \ acc = 0.5e-6;\ \ GaussSM(Ab, x, acc);$$

Example 0.13

Solve the following system of equations using the Gauss-Seidel iterative method.

$$\begin{array}{rcrcrcr} 2x_1 & + & 8x_2 & - & x_3 & = & 11 \\ 5x_1 & - & x_2 & + & x_3 & = & 10 \\ -x_1 & + & x_2 & + & 4x_3 & = & 3 \end{array}$$

Start with the initial solution $\mathbf{x}^{(0)} = [0, 0, 0]^T$.

Solution. Results for this linear system are listed in Table 4. Note that in this case the Gauss-Seidel method diverges rapidly. Although the given linear system is same as the linear system of the previous Example 0.12 except the first and second equations are interchanged. From this example we concluded that the Gauss-Seidel iterative method is not always convergent. •

Table: Solution of the Example 0.13

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000	0.000000	0.000000
1	5.500000	17.5000	-2.25000
2	-65.6250	-340.375	69.43750
3	1401.719	7068.031	-1415.83

Procedure

[Gauss-Seidel Method]

1. Check the coefficient matrix A is strictly diagonally dominant (for guaranteed convergence).
2. Initialize the first approximation $\mathbf{x}^{(0)} \in \mathbf{R}$ and pre-assigned accuracy ϵ.
3. Compute the constant $\mathbf{c} = (D + L)^{-1}\mathbf{b}$.
4. Compute the Gauss-Seidel iteration matrix $T_G = -(D + L)^{-1}U$.
5. Solve for the approximate solutions $x_i^{(k+1)} = T_G x_i^{(k)} + \mathbf{c}, \quad i = 1, 2, \ldots, n$ and $k = 0, 1, \ldots$
6. Repeat step *5* until $\|\mathbf{x}_i^{(k+1)} - \mathbf{x}_i^{(k)}\| < \epsilon$.

Note that from the Examples (0.10) and (0.12), we noted that the solution by the Gauss-Seidel method converges more quickly than the Jacobi method. In general, we may state that **if both the Jacobi method and the Gauss-Seidel method are converge, then the Gauss-Seidel method converges more quickly**. This is generally the case but not always true. In fact, there are some linear systems for which the Jacobi method converges but the Gauss-Seidel method does not, and others for which the Gauss-Seidel method converges but the Jacobi method does not.

Matrix Forms of Iterative Methods for Linear System

The iterative methods to solve the system of linear equations

$$A\mathbf{x} = \mathbf{b}, \tag{22}$$

start with an initial approximation $\mathbf{x}^{(0)} \in \mathbf{R}$ to the solution $\mathbf{x}$ of the linear system (22), and generates a sequence of vectors $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ that converges to $\mathbf{x}$. Most of these iterative methods involve a process that converts the system (22) into an equivalent system of the form

$$\mathbf{x} = T\mathbf{x} + \mathbf{c}, \tag{23}$$

for some square matrix T and vector $\mathbf{c}$. After the initial vector $\mathbf{x}^{(0)}$ is selected, the sequence of approximate solutions vector is generated by computing

$$\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + \mathbf{c}, \qquad \text{for} \quad k = 0, 1, 2, \ldots \tag{24}$$

The sequence is terminated when the error is sufficiently small, that is

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \epsilon, \quad \text{for small positive} \quad \epsilon. \tag{25}$$

Note that a **matrix T is called iteration matrix** and a **vector c is a column matrix**. We can find the forms of these matrices easily for both iterative methods as follows. Let a matrix A can be written as

$$A = L + D + U, \tag{26}$$

where L is strictly lower-triangular, U is strictly upper-triangular, and D is the diagonal parts of the coefficients matrix A, that is

$$L = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ 0 & 0 & 0 & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

and

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix}.$$

Then the linear system (22) can be written as

$$(L + D + U)\mathbf{x} = \mathbf{b}. \tag{27}$$

Now we find forms of both matrices T and $\mathbf{c}$ which help us to solve the linear system.

$$T_J = -D^{-1}(L+U) \quad \text{and} \quad \mathbf{c}_j = D^{-1}\mathbf{b}, \tag{28}$$

are called **Jacobi iteration matrix** and **Jacobi constant column matrix**, respectively, and their elements are defined by

$$\begin{aligned} t_{ij} &= \frac{a_{ij}}{a_{ii}}, \quad i,j = 1,2,\ldots,n, \quad i \neq j, \\ t_{ij} &= 0, \qquad i = j, \\ c_i &= \frac{b_i}{a_{ii}}, \quad i = 1,2,\ldots,n. \end{aligned}$$

Note that the diagonal elements of Jacobi iteration matrix T_J are always zero.

$$T_G = -(L+D)^{-1}U \quad \text{and} \quad \mathbf{c}_G = (L+D)^{-1}\mathbf{b}, \tag{29}$$

are called **Gauss-Seidel iteration matrix** and **Gauss-seidel constant column matrix**, respectively.

Example 0.14

Consider the following system

$$\begin{array}{rcrcrcr} 6x_1 & + & 2x_2 & & & = & 1 \\ x_1 & + & 7x_2 & - & 2x_3 & = & 2 \\ 3x_1 & - & 2x_2 & + & 9x_3 & = & -1 \end{array}$$

Find the matrix form of iterative (Jacobi and Gauss-Seidel) methods.

Solution. Since the given matrix A is

$$A = \begin{pmatrix} 6 & 2 & 0 \\ 1 & 7 & -2 \\ 3 & -2 & 9 \end{pmatrix},$$

and so

$$A = L + U + D = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 3 & -2 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 6 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 9 \end{pmatrix}.$$

Jacobi Iterative Method

Since the matrix form of the Jacobi iterative method can be written as

$$\mathbf{x}^{(k+1)} = T_J \mathbf{x}^{(k)} + \mathbf{c}_J, \qquad k = 0, 1, 2, \ldots,$$

where

$$T_J = -D^{-1}(L + U) \qquad \text{and} \qquad \mathbf{c}_J = D^{-1}\mathbf{b}.$$

One can easily compute the Jacobi iteration matrix T_J and the vector $\mathbf{c}_J$ as follows:

$$T_J = -\begin{pmatrix} \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{7} & 0 \\ 0 & 0 & \frac{1}{9} \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 \\ 1 & 0 & -2 \\ 3 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{2}{6} & 0 \\ -\frac{1}{7} & 0 & \frac{2}{7} \\ -\frac{3}{9} & \frac{2}{9} & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} \frac{1}{6} \\ \frac{2}{7} \\ -\frac{1}{9} \end{pmatrix}$$

Thus the matrix form of Jacobi iterative method is

$$\mathbf{x}^{(k+1)} = \begin{pmatrix} 0 & -\frac{2}{6} & 0 \\ -\frac{1}{7} & 0 & \frac{2}{7} \\ -\frac{3}{9} & \frac{2}{9} & 0 \end{pmatrix} \mathbf{x}^{(k)} + \begin{pmatrix} \frac{1}{6} \\ \frac{2}{7} \\ -\frac{1}{9} \end{pmatrix}, \quad k = 0, 1, 2.$$

Gauss-Seidel Iterative Method

Now by using Gauss-Seidel method, first we compute the Gauss-Seidel iteration matrix T_G and the vector $\mathbf{c}_G$ as follows:

$$T_G = \begin{pmatrix} 0 & -\frac{1}{3} & 0 \\ 0 & \frac{1}{21} & \frac{2}{7} \\ 0 & \frac{23}{189} & \frac{4}{63} \end{pmatrix} \quad \text{and} \quad \mathbf{c}_G = \begin{pmatrix} \frac{1}{6} \\ \frac{11}{42} \\ -\frac{41}{378} \end{pmatrix}.$$

Thus the matrix form of Gauss-Seidel iterative method is

$$\mathbf{x}^{(k+1)} = \begin{pmatrix} 0 & -\frac{1}{3} & 0 \\ 0 & \frac{1}{21} & \frac{2}{7} \\ 0 & \frac{23}{189} & \frac{4}{63} \end{pmatrix} \mathbf{x}^{(k)} + \begin{pmatrix} \frac{1}{6} \\ \frac{11}{42} \\ -\frac{41}{378} \end{pmatrix}, \quad k = 0, 1, 2.$$

Theorem 7

(Second Sufficient Condition for Convergence)
For any initial approximation $\mathbf{x}^{(0)} \in \mathbf{R}$*, the sequence* $\{x^{(k)}\}_{k=0}^{\infty}$ *of approximations defined by*

$$\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + \mathbf{c}, \quad \textit{for each} \quad k \geq 0, \quad \textit{and} \quad \mathbf{c} \neq 0, \tag{30}$$

converges to the unique solution of $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ *if* $\|T\| < 1$ *for any natural matrix norm, and the following* **error bounds** *hold:*

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \|T\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\|,$$

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|T\|^k}{1 - \|T\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|.$$

Note that smaller the value of the $\|T\|$, faster the convergence of the iterative methods.

Example 0.15

Consider the following nonhomogeneous linear system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{pmatrix} 5 & 0 & -1 \\ -1 & 3 & 0 \\ 0 & -1 & 4 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}.$$

Find the matrix form of iterative (Jacobi and Gauss-Seidel) methods and show that Gauss-Seidel iterative method converges faster than Jacobi iterative method for the given system.

Solution. Here we will show that the l_∞-norm of the Gauss-Seidel iteration matrix T_G is less than the l_∞-norm of the Jacobi iteration matrix T_J, that is

$$\|T_G\|_\infty < \|T_J\|_\infty.$$

The Jacobi iteration matrix T_J can be obtained from the given matrix A as follows

$$T_J = -D^{-1}(L+U) = -\begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \frac{1}{5} \\ \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{pmatrix}.$$

Thus the matrix form of Jacobi iterative method is

$$\mathbf{x}^{(k+1)} = \begin{pmatrix} 0 & 0 & \frac{1}{5} \\ \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{pmatrix} \mathbf{x}^{(k)} + \begin{pmatrix} \frac{1}{5} \\ \frac{2}{3} \\ 1 \end{pmatrix}, \quad k \geq 0.$$

Similarly, Gauss-Seidel iteration matrix T_G is defined as

$$T_G = -(D+L)^{-1}U = -\begin{pmatrix} 5 & 0 & 0 \\ -1 & 3 & 0 \\ 0 & -1 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and it gives

$$T_G = -\begin{pmatrix} \frac{1}{5} & 0 & 0 \\ \frac{1}{15} & \frac{1}{3} & 0 \\ \frac{1}{60} & \frac{1}{15} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{15} \\ 0 & 0 & \frac{1}{60} \end{pmatrix}.$$

So the matrix form of Gauss-Seidel iterative method is

$$\mathbf{x}^{(k+1)} = \begin{pmatrix} 0 & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{15} \\ 0 & 0 & \frac{1}{60} \end{pmatrix} \mathbf{x}^{(k)} + \begin{pmatrix} \frac{1}{5} \\ \frac{11}{15} \\ \frac{71}{60} \end{pmatrix}, \quad k \geq 0.$$

Since the l_∞-norm of the matrix T_J is

$$\|T_J\|_\infty = \max\left\{\frac{1}{5}, \frac{1}{3}, \frac{1}{4}\right\} = \frac{1}{3} = 0.3333 < 1,$$

and the l_∞-norm of the matrix T_G is

$$\|T_G\|_\infty = \max\left\{\frac{1}{5}, \frac{1}{15}, \frac{1}{60}\right\} = \frac{1}{5} = 0.2000 < 1.$$

Since $\|T_G\|_\infty < \|T_J\|_\infty$, which shows that Gauss-Seidel method will converge faster than Jacobi method for the given linear system. •

Example 0.16

Consider the following linear system of equations

$$\begin{array}{rcrcrcr} 4x_1 & - & x_2 & + & x_3 & = & 12 \\ -x_1 & + & 3x_2 & + & x_3 & = & 1 \\ x_1 & + & x_2 & + & 5x_3 & = & -14 \end{array}$$

(a) Show that both iterative methods (Jacobi and Gauss-Seidel) will converge by using $\|T\|_\infty < 1$.
(b) Find second approximation $\mathbf{x}^{(2)}$ when the initial solution is $\mathbf{x}^{(0)} = [4, 3, -3]^T$.
(c) Compute the error bounds for your approximations.
(d) How many iterations needed to get an accuracy within 10^{-4}.

Solution. From (26), we have

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

$$= L + U + D.$$

Jacobi Method:
(a) Since the Jacobi iteration matrix is defined as

$$T_J = -D^{-1}(L+U),$$

and by using the given information, we have

$$T_J = -\begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{3} & 0 & -\frac{1}{3} \\ -\frac{1}{5} & -\frac{1}{5} & 0 \end{pmatrix}.$$

Then the l_∞ norm of the matrix T_J is

$$\|T_J\|_\infty = \max\left\{\frac{2}{4}, \frac{2}{3}, \frac{2}{5}\right\} = \frac{2}{3} < 1.$$

Thus the Jacobi method will converge for the given linear system.

(b) The Jacobi method for the given system is

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{4}\left[12 + x_2^{(k)} - x_3^{(k)}\right] \\ x_2^{(k+1)} &= \frac{1}{3}\left[1 + x_1^{(k)} - x_3^{(k)}\right] \\ x_3^{(k+1)} &= \frac{1}{5}\left[-14 - x_1^{(k)} - x_2^{(k)}\right] \end{aligned}$$

Starting with initial approximation $x_1^{(0)} = 4, x_2^{(0)} = 3, x_3^{(0)} = -3$, and for $k = 0, 1$, we obtain the first and the second approximations as

$$\mathbf{x}^{(1)} = [4.5, 2.6667, -4.2]^T \quad \text{and} \quad \mathbf{x}^{(2)} = [4.7167, 3.2333, -4.2333]^T.$$

(c) Using the error bound formula (31), we obtain

$$\|\mathbf{x} - \mathbf{x}^{(2)}\| \leq \frac{(2/3)^2}{1 - 2/3} \left\| \begin{pmatrix} 4.5 \\ 2.6667 \\ -4.2 \end{pmatrix} - \begin{pmatrix} 4 \\ 3 \\ -3 \end{pmatrix} \right\| \leq \frac{4}{3}(1.2) = 1.6.$$

(d) To find the number of iterations, we use the formula (31) as

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|T_J\|^k}{1 - \|T_J\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \leq 10^{-4}.$$

It gives

$$\frac{(2/3)^k}{1/3}(1.2) \leq 10^{-4}, \quad \text{or} \quad (2/3)^k \leq \frac{10^{-4}}{3.6}.$$

Taking ln on both sides, we obtain

$$k \ln(2/3) \leq \ln\left(\frac{10^{-4}}{3.6}\right), \quad \text{gives} \quad k \geq 25.8789, \quad \text{or} \quad k = 26,$$

which is the required number of iterations.

Gauss-Seidel Method:
(a) Since the Gauss-Seidel iteration matrix is defined as

$$T_G = -(D + L)^{-1}U,$$

and by using the given information, we have

$$T_G = -\begin{pmatrix} \frac{1}{4} & 0 & 0 \\ \frac{1}{12} & \frac{1}{3} & 0 \\ -\frac{4}{60} & -\frac{1}{15} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 0 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{1}{12} & -\frac{5}{12} \\ 0 & -\frac{4}{60} & \frac{8}{60} \end{pmatrix}.$$

Then the l_∞ norm of the matrix T_G is

$$\|T_G\|_\infty = \max\left\{\frac{2}{4}, \frac{6}{12}, \frac{12}{60}\right\} = \frac{1}{2} < 1.$$

Thus the Gauss-Seidel method will converge for the given linear system.

(b) The Gauss-Seidel method for the given system is

$$\begin{aligned}
x_1^{(k+1)} &= \frac{1}{4}\left[12 + x_2^{(k)} - x_3^{(k)}\right] \\
x_2^{(k+1)} &= \frac{1}{3}\left[1 + x_1^{(k+1)} - x_3^{(k)}\right] \\
x_3^{(k+1)} &= \frac{1}{5}\left[-12 - x_1^{(k+1)} - x_2^{(k+1)}\right]
\end{aligned}$$

Starting with initial approximation $x_1^{(0)} = 4, x_2^{(0)} = 3, x_3^{(0)} = -3$, and for $k = 0, 1$, we obtain the first and the second approximations as

$$\mathbf{x}^{(1)} = [4.5, 2.8333, -4.2667]^T \quad \text{and} \quad \mathbf{x}^{(2)} = [4.775, 3.3472, -4.4244]^T.$$

(c) Using the error bound formula (31), we obtain

$$\|\mathbf{x} - \mathbf{x}^{(2)}\| \leq \frac{(1/2)^2}{1 - 1/2} \left\| \begin{pmatrix} 4.5 \\ 2.8333 \\ -4.2667 \end{pmatrix} - \begin{pmatrix} 4 \\ 3 \\ -3 \end{pmatrix} \right\| \leq \frac{1}{2}(1.2667) = 0.6334.$$

(d) To find the number of iterations, we use the formula (31) as

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|T_J\|^k}{1 - \|T_J\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \leq 10^{-4}.$$

It gives

$$\frac{(1/2)^k}{1/2}(1.2667) \leq 10^{-4}, \quad \text{or} \quad (1/2)^k \leq \frac{10^{-4}}{2.5334}.$$

Taking ln on both sides, we obtain

$$k \ln(1/2) \leq \ln\left(\frac{10^{-4}}{2.5334}\right), \quad \text{gives} \quad k \geq 14.6084 \quad \text{or} \quad k = 15,$$

which is the required number of iterations. •

Errors in Solving Linear Systems

Any computed solution of a linear system must, because of round-off and other errors, be considered an approximate solution. Here we shall consider the most natural method for determining the accuracy of a solution of the linear system. One obvious way of estimating the accuracy of the computed solution $\mathbf{x}^*$ is to compute $A\mathbf{x}^*$ and to see how close $A\mathbf{x}^*$ comes to $\mathbf{b}$. Thus if $\mathbf{x}^*$ is an approximate solution of the given system $A\mathbf{x} = \mathbf{b}$, we compute a vector

$$\mathbf{r} = \mathbf{b} - A\mathbf{x}^*, \tag{31}$$

which is called the *residual vector* and can be easily calculated. The quantity

$$\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} = \frac{\|\mathbf{b} - A\mathbf{x}^*\|}{\|\mathbf{b}\|},$$

is called the *relative residual*.

The smallness of the residual then provides a measure of the goodness of the approximate solution $\mathbf{x}^*$. If every component of vector $\mathbf{r}$ vanishes, then $\mathbf{x}^*$ is the exact solution. If $\mathbf{x}^*$ is a good approximation then we would expect each component of $\mathbf{r}$ to be small, at least in a relative sense. For example, the following linear system

$$\begin{array}{rcrcr} x_1 & + & 2x_2 & = & 3 \\ 1.0001x_1 & + & 2x_2 & = & 3.0001 \end{array}$$

has the exact solution $\mathbf{x} = [1, 1]^T$ but has a poor approximate solution $\mathbf{x}^* = [3, 0]^T$. To see how good this solution is, we compute the residual, $\mathbf{r} = [0, -0.0002]^T$, and so $\|\mathbf{r}\|_\infty = 0.0002$. Although the norm of the residual vector is small, the approximate solution $\mathbf{x}^* = [3, 0]^T$ is obviously quite poor; in fact $\|\mathbf{x} - \mathbf{x}^*\|_\infty = 2$.

Program 3.12

MATLAB m-file for finding Residual Vector

```
function r=RES(A,b,x0)
[n,n]=size(A);
for i=1:n; R(i) = b(i); for j=1:n
R(i)=R(i)-A(i,j)*x0(j);end; RES(i)=R(i); end; r=RES'
```

To get above results using MATLAB command, we do the following:

$>> A = [1\ 2; 1.0001\ 2];\ b = [3\ 3.0001];\ x0 = [3\ 0];$
$>> RESID(A, b, x0);\ x = [1\ 1];\ Error = norm((x - x0), inf);$

We can conclude from the residual that the approximate solution is correct to at most three decimal places. Also, the following linear system

$$\begin{array}{rcrcrcrcl} 1.0000x_1 & + & 0.9600x_2 & + & 0.8400x_3 & + & 0.6400x_4 & = & 3.4400 \\ 0.9600x_1 & + & 0.9214x_2 & + & 0.4406x_3 & + & 0.2222x_4 & = & 2.5442 \\ 0.8400x_1 & + & 0.4406x_2 & + & 1.0000x_3 & + & 0.3444x_4 & = & 2.6250 \\ 0.6400x_1 & + & 0.2222x_2 & + & 0.3444x_3 & + & 1.0000x_4 & = & 2.2066 \end{array}$$

has exact solution $\mathbf{x} = [1, 1, 1, 1]^T$ and having the approximate solution due to the Gaussian elimination without pivoting is

$$\mathbf{x}^* = [1.0000322, 0.99996948, 0.99998748, 1.0000113]^T,$$

and the residual is

$$\mathbf{r} = [0.6 \times 10^{-7}, 0.6 \times 10^{-7}, -0.53 \times 10^{-5}, -0.21 \times 10^{-4}]^T$$

The approximate solution due to the Gaussian elimination with partial pivoting is

$$\mathbf{x}^* = [0.9999997, 0.99999997, 0.99999996, 1.0000000]^T,$$

and the residual is

$$\mathbf{r} = [0.3 \times 10^{-7}, 0.3 \times 10^{-7}, 0.6 \times 10^{-7}, 0.1 \times 10^{-8}]^T.$$

We found that all the elements of the residual for second case (with pivoting) are less than 0.6×10^{-7}, whereas for first case (without pivoting) they are as large as 0.2×10^{-4}. Even without knowing the exact solution, it is clear that the solution obtained in second case is much better than that of first case. The residual provides a reasonable measure of the accuracy of a solution in those cases where the error is primarily due to the accumulation of round-off errors.
Intuitively it would seem reasonable to assume that when $\|\mathbf{r}\|$ is small for a given vector norm, then the error $\|\mathbf{x} - \mathbf{x}^*\|$ would be small as well. In fact this is true for some systems. However, there are systems of equations which do not satisfy this property. Such systems are said to be *ill-conditioned*.

Conditioning of Linear Systems

In solving the linear system numerically we have to see the problem conditioning, algorithm stability, and cost. Above we discussed efficient elimination schemes to solve a linear system and these schemes are stable when pivoting is employed. But there are some ill-conditioned systems which are tough to solve by any method.

Definition 8

(Condition Number of a Matrix)
The number $\|A\|\|A^{-1}\|$ is called the *condition number* of a nonsingular matrix A and is denoted by $K(A)$, that is

$$cond(A) = K(A) = \|A\|\|A^{-1}\|. \tag{32}$$

Note that the **condition number** $K(A)$ for A depends on the matrix norm used and can, for some matrices, vary considerably as the matrix norm is changed. Since

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = K(A),$$

therefore, the condition number is always in the range $1 \leq K(A) \leq \infty$ regardless of any natural norm. The lower limit is attained for identity matrices and $K(A) = \infty$ if A is singular. So the matrix A is *well-behaved* (well-conditioned) if $K(A)$ is close to 1 and is increasingly *ill-conditioned* when $K(A)$ is significantly greater than 1, that is, $K(A) \to \infty$.

Example 0.17

Compute the condition number of the following matrix using the l_∞-norm

$$A = \begin{pmatrix} 2 & -1 & 0 \\ 2 & -4 & -1 \\ -1 & 0 & 2 \end{pmatrix}.$$

Solution. Since the condition number of a matrix is defined as

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

First we calculate the inverse of the given matrix which is

$$A^{-1} = \begin{pmatrix} \frac{8}{13} & -\frac{2}{13} & -\frac{1}{13} \\ \frac{3}{13} & -\frac{4}{13} & -\frac{2}{13} \\ \frac{4}{13} & -\frac{1}{13} & \frac{6}{13} \end{pmatrix}.$$

Now we calculate the l_∞-norm of both the matrices A and A^{-1}. Since the l_∞-norm of a matrix is the maximum of the absolute row sums, we have

$$\|A\|_\infty = max\{|2| + |-1| + |0|, |2| + |-4| + |-1|, |-1| + |0| + |2|\} = 7,$$

and

$$\|A^{-1}\|_\infty = max\left\{\left|\frac{8}{13}\right| + \left|\frac{-2}{13}\right| + \left|\frac{-1}{13}\right|, \left|\frac{3}{13}\right| + \left|\frac{-4}{13}\right| + \left|\frac{-2}{13}\right|, \left|\frac{4}{13}\right| + \left|\frac{-1}{13}\right| + \left|\frac{6}{13}\right|\right\},$$

which gives

$$\|A^{-1}\|_\infty = \frac{11}{13}.$$

Therefore,

$$K(A) = \|A\|_\infty \|A^{-1}\|_\infty = (7)\left(\frac{11}{13}\right) \approx 5.9231.$$

Depending on the application, we might consider this number to be reasonably small and conclude that the given matrix A is reasonably well-conditioned. •

Example 0.18

If the condition number of following matrix A is 8.8671, then find the l_∞-norm of its inverse matrix, that is, $\|A^{-1}\|_\infty$

$$A = \begin{pmatrix} 10.2 & 2.4 & 4.5 \\ -2.3 & 7.7 & 11.1 \\ -5.5 & -3.2 & 0.9 \end{pmatrix}.$$

Solution. Since the condition number of a matrix is defined as

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

First we calculate the l_∞-norm of the given matrix A which is the maximum of the absolute row sums, we have

$$\|A\|_\infty = max\{17.1000, 21.1000, 9.6\} = 21.1000,$$

and as it is given $K(A) = 8.8671$, so we have

$$8.8671 = (21.1000)\|A^{-1}\|_\infty.$$

Simplifying this, we get $\|A^{-1}\|_\infty = 0.4202$. •

Theorem 9

(Error in Linear Systems)

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

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \|\mathbf{r}\| \|A^{-1}\|, \tag{33}$$

and the relative error is

$$\frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}\|} \leq K(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}, \quad \textit{provided that} \quad \mathbf{x} \neq 0, \ \mathbf{b} \neq 0. \tag{34}$$

Example 0.19

Find the condition number of the following matrix (for $n = 2, 3, \ldots$)

$$A_n = \begin{bmatrix} 1 & 1 \\ 1 & 1 - 1/n \end{bmatrix}.$$

If $n = 2$ and $x^* = [-1.99, 2.99]^T$ be the approximate solution of the linear system $A\mathbf{x} = [1, -0.5]^T$, then find the relative error.

Solution. We can easily find the inverse of the given matrix as

$$A_n^{-1} = \frac{1}{(1 - 1/n) - 1} \begin{bmatrix} 1 - 1/n & -1 \\ -1 & 1 \end{bmatrix} = -n \begin{bmatrix} 1 - 1/n & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 - n & n \\ n & -n \end{bmatrix}.$$

Then the l_∞-norm of both matrices A_n and A_n^{-1} are

$$\|A_n\|_\infty = 2 \quad \text{and} \quad \|A_n^{-1}\|_\infty = 2n,$$

and so the condition number of the matrix can be computed as follows:

$$K(A) = \|A_n\|_\infty \|A_n^{-1}\|_\infty = (2)(2n) = 4n \qquad \text{and} \qquad \lim_{n \to \infty} K(A) = \infty,$$

which shows that the matrix A_n is obviously ill-conditioned.

Here we expect that the relative error in the calculated solution to a linear system of the form $A_n\mathbf{x} = \mathbf{b}$ could be as much as $4n$ times the relative residual.
The residual vector (by taking $n = 2$) can be calculated as

$$\mathbf{r} = \mathbf{b} - A_2\mathbf{x}^* = \begin{pmatrix} 1 \\ -0.5 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 0.5 \end{pmatrix} \begin{pmatrix} -1.99 \\ 2.99 \end{pmatrix} = \begin{pmatrix} 0.000 \\ -0.005 \end{pmatrix},$$

and it gives $\|\mathbf{r}\|_\infty = 0.005$. Now using (35), we obtain

$$\frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}\|} \leq K(A)\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} = (8)\frac{0.005}{1} = 0.0400,$$

which is the required relative error. •

Example 0.20

Consider a following linear system

$$\begin{array}{rcrcrcr} x_1 & + & x_2 & - & x_3 & = & 1 \\ x_1 & + & 2x_2 & - & 2x_3 & = & 0 \\ -2x_1 & + & x_2 & + & x_3 & = & -1 \end{array}$$

(a) Discuss the ill-conditioning of the given linear system.
(b) If $\mathbf{x}^* = [2.01, 1.01, 1.98]^T$ be an approximate solution of the given system, then find the residual vector $\mathbf{r}$ and its norm $\|\mathbf{r}\|_\infty$.
(c) Estimate the relative error using

$$\frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}\|} \leq K(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}, \quad \text{provided that} \quad \mathbf{x} \neq 0, \ \mathbf{b} \neq 0. \tag{35}$$

.
(d) Use the simple Gaussian elimination method to find approximate error using

$$A(\mathbf{x} - \mathbf{x}^*) = \mathbf{r}, \quad \text{or} \quad \mathbf{x} - \mathbf{x}^* = A^{-1}\mathbf{r}. \tag{36}$$

.

Solution. (a) Given the matrix

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & -2 \\ -2 & 1 & 1 \end{pmatrix},$$

and whose inverse can be computed as

$$A^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ 1.5 & -0.5 & 0.5 \end{pmatrix}.$$

(b) The residual vector can be calculated as

$$\mathbf{r} = \mathbf{b} - A\mathbf{x}^* = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & -2 \\ -2 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2.01 \\ 1.01 \\ 1.98 \end{pmatrix} = \begin{pmatrix} -0.04 \\ -0.07 \\ 0.03 \end{pmatrix},$$

and it gives

$$\|\mathbf{r}\|_\infty = 0.07.$$

```
>> A = [1 1 -1; 1 2 -2; -2 1 1]; b = [1 0 -1]';
>> x0 = [2.01 1.01 1.98]'; r = RES(A, b, x0); rnorm = norm(r, inf);
```

(c) From (35), we have

$$\frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}\|} \leq K(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

By using above parts (a) and (b) and the value $\|\mathbf{b}\|_\infty = 1$, we obtain

$$\frac{\|\mathbf{x} - \mathbf{x}^*\|}{\|\mathbf{x}\|} \leq (22.5) \frac{(0.07)}{1} = 1.575.$$

```
>> RelErr = (K(A) * rnorm)/norm(b, inf);
```

(d) To solve the linear system $A\mathbf{e} = \mathbf{r}$, where

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & -2 \\ -2 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{r} = \begin{pmatrix} -0.04 \\ -0.07 \\ 0.03 \end{pmatrix},$$

and $\mathbf{e} = \mathbf{x} - \mathbf{x}^*$. Writing the above system in the augmented matrix form

$$\left(\begin{array}{ccc:c} 1 & 1 & -1 & -0.04 \\ 1 & 2 & -2 & -0.07 \\ -2 & 1 & 1 & 0.03 \end{array}\right).$$

After applying forward elimination step of the simple Gauss elimination method, we obtain

$$\left(\begin{array}{ccc:c} 1 & 1 & -1 & -0.04 \\ 0 & 1 & -1 & -0.03 \\ 0 & 0 & 2 & 0.04 \end{array}\right).$$

Now by using the backward substitution, we obtain the solution

$$\mathbf{e}^* = [-0.01, -0.01, 0.02]^T,$$

which is the required approximation of the exact error. •

Summary

In this lecture, we ...

- Considered the System of Linear Equations
- Introduced the Gaussian elimination method
- Introduced the Jacobi Iterative Method
- Introduced the Gauss-Seidel Iterative Method