# SYSTEM OF LINEAR ALGEBRAIC EQUATIONS

## Learning Objectives

After reading this unit you should be able to:

1. State when LU Decomposition is numerically more efficient than Gaussian Elimination,
2. Decompose a non-singular matrix into LU,
3. Show how LU Decomposition is used to find matrix inverse

## Unit content

**SESSION 1-2: Matrices**

1-2.1 Matrix Inversion and Cramer's Rule

1-2.2 Triangular System of Equations

1-2.3 Back Substitution

1-2.4 Division by zero:

1-2.5 Finding the determinant of a square matrix using Naïve Gauss Elimination methods

**SESSION 2-2: Techniques for improving Naïve Gauss Elimination Method**

2-2.1 Gaussian Elimination with partial pivoting

2-2.2 LU Decomposition

2-2.3 Decomposing a non-singular matrix A into the form A=LU.

2-2.4 Finding the inverse of a square matrix using LU Decomposition

**SESSION 3-1: Iterative Method**

3-1.1 Jacobi Method

3-1.2 Gauss-Seidel Method

3-1.3 Successive over–relaxation method

3-1.4 A necessary and sufficient condition for the convergence of iterative schemes

# SESSION 1-2: MATRICES

## 1-2.1 Matrix Inversion and Cramer's Rule

A system of linear Algebraic equations is nothing but a system of n algebraic linear equations satisfied by a set of n unknown quantities.  The aim is to find these *n* unknown quantities satisfying the *n* equations.

It is a very common practice to write the system of *n* equations in matrix form as $Ax = b$ , where *A* is an $n \times n$, non-singular matrix and *x* and *b* are $n \times 1$ matrices out of which *b* is known.  i.e.,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

For small *n* the elementary methods like matrix inversion or Cramers rule given below are very convenient to get the unknown vector *x* from the system $Ax = b$.

Let $A$ be a nonsingular matrix of order *n* and *b* be an *n*-vector. The solution $x$ of the system $Ax = b$ is given by:

a) Matrix Inversion as $x = A^{-1}.b$

b) Cramer's rule as $x_i = \frac{\det(A_i)}{\det(A)}, \qquad i = 1, \ldots, n$

where $A_i$ is a matrix obtained by replacing the $i^{th}$ column of $A$ by the vector $b$ and $x = (x_1, x_2, \ldots, x_n)^T$, i.e., Cramer's Rule or by matrix inversion formula given as $x = A^{-1}b$

However, for large ' $n$ ' these methods will become computationally very expensive because of the evaluation of matrix determinants involved in these methods. Hence to make the solution methods computationally less expensive one has to find alternate means which doesn't require the evaluation of any determinants or inverses to find *x* form $Ax = b$.

## 1-2.2 Triangular System of Equations

The system of equations below is an upper triangular system of equations

$$
\begin{array}{rcl}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \quad \cdots \quad + a_{1n}x_n & = & b_1 \\
a_{22}x_2 + a_{23}x_3 + \quad \cdots \quad + a_{2n}x_n & = & b_2 \\
a_{33}x_3 + \quad \cdots \quad + a_{3n}x_n & = & b_3 \\
\vdots & & \\
a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n & = & b_{n-1} \\
a_{nn}x_n & = & b_n
\end{array}
$$

Now the equations are solved starting from the last equation as it has only one unknown.

$$x_n^* = \frac{b_n}{a_{nn}}$$

Then the second last equation, that is the ($n$-1)$^{th}$ equation, has two unknowns ( $x_n$ and $x_{n-1}$), but $x_n^*$ is already known. This reduces the ($n$-1)$^{th}$ equation also to one unknown and we have:

$$x_{n-1} = \frac{1}{a_{n-1,n-1}}\left(b_{n-1} - a_{n-1,n}x_n^*\right)$$

Back substitution hence can be represented for all equations by the formula

$$x_n^* = \frac{b_n}{a_{nn}} \qquad \text{and} \quad x_i = \frac{1}{a_{i,i}}\left(b_i - \sum_{j=i+1}^{n} a_{i,j}x_j^*\right) \quad \text{for } i = n-1, n-2, \ldots, 1$$

**Note** $\sum\limits_{j=n+1}^{n} \quad = 0$.

## 1-2.3 Back Substitution

Now the equations are solved starting from the last equation as it has only one unknown.

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

Then the second last equation, that is the ($n$-1)$^{th}$ equation, has two unknowns - $x_n$ and $x_{n-1}$, but $x_n$ is already known. This reduces the ($n$-1)$^{th}$ equation also to one unknown. Back substitution hence can be represented for all equations by the formula $x_i = \frac{1}{a_{ii}^{(i-1)}}\left( b_i^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)} x_j \right)$ for $i = n, n-1, n-2, \ldots, 1$

**Note**: $\sum_{j=n+1}^{n} \quad = 0$.

The procedure just described above is called the Naïve Gaussian Elimination method or the Gaussian Elimination without pivoting.

**Example 1**

Use Naïve Gaussian Elimination to solve $Ax = b$, where

$$A = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}, \; b = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

**Solution**

Forward Elimination of Unknowns: Since there are three equations, there will be two steps of forward elimination of unknowns.

First step: Divide *Row 1* by 25 and then multiply it by 64, i.e.,

$\left[\frac{Row\ 1}{25}\right] \times (64) = Row\ 1 \times 2.56\,6$ gives *Row 1* as $[64 \quad 12.8 \quad 2.56] \qquad [273.408]$

Subtract the result from Row 2

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.81 \\ -96.21 \\ 279.2 \end{bmatrix}$$

Divide Row 1 by 25 and then multiply it by 144

$\left[\frac{Row\ 1}{25}\right] \times (144) = Row\ 1 \times 5.76$ gives *Row 1* as $[144 \quad 28.8 \quad 5.76] \qquad [615.2256]$

Subtract the result from Row 3

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}\begin{bmatrix} \mathrm{a}_1 \\ \mathrm{a}_2 \\ \mathrm{a}_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ -336.0 \end{bmatrix}$$

<u>Second step</u>: We now divide Row 2 by –4.8 and then multiply by –16.8

$$\left[\frac{Row\ 2}{-4.8}\right]\times(-16.8) = Row\ 2\times 3.5 \text{ gives } Row\ 2 \text{ as } \begin{bmatrix}0 & -16.8 & -5.46\end{bmatrix} \qquad \begin{bmatrix}-336.735\end{bmatrix}$$

Subtract the result from Row 3

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}\begin{bmatrix} \mathrm{a}_1 \\ \mathrm{a}_2 \\ \mathrm{a}_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ 0.735 \end{bmatrix}$$

<u>Back substitution</u>: From the third equation

$$0.7x_3 = 0.735 \quad\Rightarrow\quad x_3 = \frac{0.735}{0.7} = 1.050$$

Substituting the value of $x_3$ in the second equation,

$$-4.8x_2 - 1.56x_3 = -96.21 \quad\Rightarrow\quad x_2 = \frac{-96.21 + 1.56x_3}{-4.8} = \frac{-96.21 + 1.56(1.050)}{-4.8} = 19.70$$

Substituting the value of $x_2$ and $x_3$ in the first equation,

$$25x_1 + 5x_2 + x_3 = 106.8 \quad\Rightarrow$$

$$x_1 = \frac{106.8 - 5x_2 - x_3}{25} = \frac{106.8 - 5(19.70) - 1.050}{25} = 0.2900$$

Hence the solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.2900 \\ 19.70 \\ 1.050 \end{bmatrix}$$

**Example 2**

Use Naïve Gauss Elimination to solve

$$\begin{aligned} 10x_1 - \quad 7x_2 \qquad &= 7 \\ -3x_1 + 2.099x_2 + 6x_3 &= 3.901 \\ 5x_1 - \qquad x_2 + 5x_3 &= 6 \end{aligned}$$

Use six significant digits with chopping in your calculations.

**Solution**

Working in the matrix form

$$\begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 3.901 \\ 6 \end{bmatrix}$$

Forward Elimination of Unknowns

Dividing Row 1 by 10 and multiplying by –3, that is, multiplying Row 1 by -0.3, and subtract it from Row 2 would eliminate $a_{21}$,

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 6 \end{bmatrix}$$

Again dividing Row 1 by 10 and multiplying by 5, that is, multiplying Row 1 by 0.5, and subtract it from Row 3 would eliminate $a_{31,}$

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 2.5 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 2.5 \end{bmatrix}$$

This is the end of the first step of forward elimination.

Now for the second step of forward elimination, we would use Row 2 as the pivot equation and eliminate Row 3 – Column 2. Dividing Row 2 by –0.001 and multiplying by 2.5, that is multiplying Row 2 by –2500, and subtracting from Row 3 gives

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 0 & 15005 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 15005 \end{bmatrix}$$

This is the end of the forward elimination steps.

Back substitution

We can now solve the above equations by back substitution. From the third equation,

$$15005x_3 = 15005 \quad \Rightarrow \quad x_3 = \frac{15005}{15005} = 1$$

Substituting the value of $x_3$ in the second equation

$$-0.001x_2 + 6x_3 = 6.001 \quad \Rightarrow$$

$$x_2 = \frac{6.001 - 6x_3}{-0.001} = \frac{6.001\text{-}6(1)}{-0.001} = \frac{0.001}{-0.001} = -1$$

Substituting the value of $x_3$ and $x_2$ in the first equation,

$$10x_1 - 7x_2 + 0x_3 = 7 \quad \Rightarrow \quad x_1 = \frac{7 + 7\,x_2 - 0x_3}{10} = \frac{7 + 7(-1) - 0(1)}{10} = 0$$

Hence the solution is

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

**Note**:

There are two pitfalls of Naïve Gauss Elimination method:

1. Division by zero,
2. Round-off error.

## 1-2.4 Division by zero

It is possible that division by zero may occur during forward elimination steps. For example for the set of equations

$$\begin{aligned} 10x_2 - 7x_3 &= 7 \\ 6x_1 + 2.099x_2 - 3x_3 &= 3.901 \\ 5x_1 - x_2 + 5x_3 &= 6 \end{aligned}$$

during the first forward elimination step, the coefficient of $x_1$ is zero and hence normalization would require division by zero.

**Round-off error:**

Naïve Gauss Elimination Method is prone to round-off errors. This is true when there are large numbers of equations as errors propagate. Also, if there is subtraction of numbers from each other, it may create large errors. See the example below.

**Example 3**

Remember the previous example where we used Naïve Gauss Elimination to solve

$$\begin{aligned} 10x_1 - 7x_2 &= 7 \\ -3x_1 + 2.099x_2 + 6x_3 &= 3.901 \\ 5x_1 - x_2 + 5x_3 &= 6 \end{aligned}$$

using <u>six</u> significant digits with chopping in your calculations. Repeat the problem, but now use <u>five</u> significant digits with chopping in your calculations.

**Solution**

Writing in the matrix form

$$\begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 3.901 \\ 6 \end{bmatrix}$$

<u>Forward Elimination of Unknowns</u>

Dividing Row 1 by 10 and multiplying by –3, that is, multiplying Row 1 by -0.3, and subtract it from Row 2 would eliminate $a_{21}$,

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 6 \end{bmatrix}$$

Again dividing Row 1 by 10 and multiplying by 5, that is, multiplying the Row 1 by 0.5, and subtract it from Row 3 would eliminate $a_{31,}$

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 2.5 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 2.5 \end{bmatrix}$$

This is the end of the first step of forward elimination.

Now for the second step of forward elimination, we would use Row 2 as the pivoting equation and eliminate Row 3 – Column 2. Dividing Row 2 by –0.001 and multiplying by 2.5, that is, multiplying Row 2 by –2500, and subtract from Row 3 gives

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 0 & 15005 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 15004 \end{bmatrix}$$

This is the end of the forward elimination steps.

Back substitution

We can now solve the above equations by back substitution. From the third equation,

$$15005x_3 = 15004 \quad \Rightarrow \quad x_3 = \frac{15004}{15005} = 0.99993$$

Substituting the value of $x_3$ in the second equation

$$-0.001x_2 + 6x_3 = 6.001 \quad \Rightarrow$$

$$x_2 = \frac{6.001 - 6x_3}{-0.001} = \frac{6.001\text{-}6(0.99993)}{\text{-}0.001} = \frac{6.001\text{-}5.9995}{\text{-}0.001} = \frac{0.0015}{-0.001} = -1.5$$

Substituting the value of x3 and x2 in the first equation, $10x_1 - 7x_2 + 0x_3 = 7 \quad \Rightarrow$

$$x_1 = \frac{7 + 7x_2 - 0x_3}{10} = \frac{7 + 7(-1.5) - 0(1)}{10} = \frac{7 + 7(-1.5) - 0(1)}{10}$$
$$= \frac{7 - 10.5 - 0}{10} = \frac{-3.5}{10} = -0.3500$$

Hence the solution is

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -0.35 \\ -1.5 \\ 0.99993 \end{bmatrix}$$

Compare this with the exact solution of

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

### 1-2.5 Finding the determinant of a square matrix using Naïve Gaussian Elimination methods

One of the more efficient ways to find the determinant of a square matrix is by taking advantage of the following two theorems on a determinant of matrices coupled with Naïve Gauss Elimination.

**Theorem 1:**

Let *A* be a *n*x*n* matrix. Then, if *B* is a matrix that results from adding or subtracting a multiple of one row to another row, then det(*B*) = det(*A*). (The same is true for column operations also).

**Theorem 2:**

Let *A* be a *n*x*n* matrix that is upper triangular, lower triangular or diagonal, then $\det(A) = a_{11} * a_{22} * \cdots * a_{nn} = \prod_{i=1}^{n} a_{ii}$

This implies that if we apply the forward elimination steps of Naive Gauss Elimination method, the determinant of the matrix stays the same according the Theorem 1. Then since at the end of the forward elimination steps, the resulting matrix is upper triangular, the determinant will be given by Theorem 2.

**Example 5**
Find the determinant of

$$A = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

**Solution**
Remember earlier in this chapter, we conducted the steps of forward elimination of unknowns using Naïve Gauss Elimination method on $A$ to give

$$B = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

According to Theorem 2

$$\det(A) = \det(B) = \left(25\right)(-4.8)(0.7) = -84.00$$

**Note:**

If you cannot find the determinant of the matrix using Naive Gauss Elimination method due to a division by zero problems during Naïve Gauss Elimination method, you can apply Gaussian Elimination with Partial Pivoting. However, the determinant of the resulting upper triangular matrix may differ by a sign. The following theorem applies in addition to the previous two to find determinant of a square matrix.

**Theorem 3:**

Let $A$ be a *nxn* matrix. Then, if $B$ is a matrix that results from switching one row with another row, then $\det(A) = -\det(B)$.

**Example 6**

Find the determinant of

$$A=\begin{bmatrix}10 & -7 & 0\\ -3 & 2.099 & 6\\ 5 & -1 & 5\end{bmatrix}$$

**Solution**

Remember from that at the end of the forward elimination steps of Gaussian elimination with partial pivoting, we obtained

$$B=\begin{bmatrix}10 & -7 & 0\\ 0 & 2.5 & 5\\ 0 & 0 & 6.002\end{bmatrix}$$

$\det\left(B\right)=(10)(2.5)(6.002)=150.05$

Since rows were switched once during the forward elimination steps of Gaussian elimination with partial pivoting, $\det\left(A\right)=-\det(B)=-150.05$

Prove that $\det\left(A\right)=\dfrac{1}{\det\left(A^{-1}\right)}$

***Proof:***

$$AA^{-1}=I\ \ \det\left(A\,A^{-1}\right)=\det\left(I\right)\ \ \Rightarrow\ \ \det\left(A\right)\det\left(A^{-1}\right)=1\ \ \Rightarrow\ \ \det\left(A\right)=\frac{1}{\det\left(A^{-1}\right)}.$$

If *A* is a *n*x*n* matrix and det (*A*) ≠ 0, what other statements are equivalent to it?

1. $A$ is invertible.
2. $A^{-1}$ exists.
3. $Ax=b$ has a unique solution.
4. $Ax=0$ solution is $x=0$
5. $AA^{-1}=I=A^{-1}A$.

## SESSION 2-2: Techniques for improving Naïve Gauss Elimination Method

As seen in the example, round off errors were large when five significant digits were used as opposed to six significant digits. So, one way of decreasing round off error would be to use more significant digits, that is, use double or quad precision. However, this would not avoid division by zero errors in Naïve Gauss Elimination. To avoid division by zero as well as reduce (not eliminate) round off error, Gaussian Elimination with partial pivoting is the method of choice.

### 2-2.1 Gaussian Elimination with partial pivoting

The Gaussian elimination with partial pivoting and the Naïve Gauss elimination methods are the same, except in the beginning of each step of forward elimination; a row switching is done based on the following criterion. If there are $n$ equations, then there are $(n-1)$ forward elimination steps. At the beginning of the $k^{th}$ step of forward elimination, one finds the maximum of

$|a_{kk}|, |a_{k+1,k}|, \ldots, |a_{nk}|$

Then if the maximum of these values is $|a_{pk}|$ in the $p^{th}$ row, $k \le p \le n$, then switch rows $p$ and $k$. The other steps of forward elimination are the same as Naïve Gauss elimination method. The back substitution steps stay exactly the same as Naïve Gauss Elimination method.

**Example 4**

In the previous two examples, we used Naïve Gauss Elimination to solve

$$\begin{aligned} 10x_1 - 7x_2 \quad &= 7 \\ -3x_1 + 2.099x_2 + 6x_3 &= 3.901 \\ 5x_1 - x_2 + 5x_3 &= 6 \end{aligned}$$

using five and six significant digits with chopping in the calculations. Using five significant digits with chopping, the solution found was

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -0.35 \\ -1.5 \\ 0.99993 \end{bmatrix}$$

This is different from the exact solution

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

Find the solution using Gaussian elimination with partial pivoting using five significant digits with chopping in your calculations.

**Solution**

$$\begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2.901 \\ 6 \end{bmatrix}$$

Forward Elimination of Unknowns

Now for the first step of forward elimination, the absolute values of first column elements are $|10|, |-3|, |5|$ or 10, 3, 5.

So the largest absolute value is in the Row 1. So as per Gaussian Elimination with partial pivoting, the switch is between Row 1 and Row 1 to give

$$\begin{bmatrix} 10 & 7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 3.901 \\ 6 \end{bmatrix}$$

Dividing Row 1 by 10 and multiplying by –3, that is, multiplying the Row 1 by -0.3, and subtract it from Row 2 would eliminate $a_{21}$,

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 5 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 6 \end{bmatrix}$$

Again dividing Row 1 by 10 and multiplying by 5, that is, multiplying the Row 1 by 0.5, and subtract it from Row 3 would eliminate $a_{31}$,

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 2.5 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 6.001 \\ 2.5 \end{bmatrix}$$

This is the end of the first step of forward elimination.

Now for the second step of forward elimination, the absolute value of the second column elements below the Row 2 is $|-0.001|$, $|2.5|$ or 0.001, 2.5

So the largest absolute value is in Row 3. So the Row 2 is switched with the Row 3 to give

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & -0.001 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2.5 \\ 6.001 \end{bmatrix}$$

Dividing row 2 by 2.5 and multiplying by –0.001, that is multiplying by $0.001/2.5 = –0.0004$, and then subtracting from Row 3 gives

$$\begin{bmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & 0 & 6.002 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 2.5 \\ 6.002 \end{bmatrix}$$

<u>Back substitution</u>

$$6.002x_3 = 6.002 \quad \Rightarrow \quad x_3 = \frac{6.002}{6.002} = 1$$

Substituting the value of $x_3$ in Row 2

$$2.5x_2 + 5x_3 = 2.5 \quad \Rightarrow \quad x_2 = \frac{2.5 - 5x_2}{2.5} = \frac{2.5 - 5}{2.5} = -1$$

Substituting the value of $x_3$ and $x_2$ in Row 1

$$10x_1 - 7x_2 + 0x_2 = 7 \quad \Rightarrow$$
$$x_1 = \frac{7 + 7x_2 - 0x_3}{10} = \frac{7 + 7(-1) - 0(1)}{10} = \frac{7 - 7 - 0}{10} = \frac{0}{10} = 0$$

So the solution is

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

This, in fact, is the exact solution. By coincidence only, in this case, the round off error is fully removed.

### 2-2.2 LU Decomposition

We already studied two numerical methods of finding the solution to simultaneous linear equations – Naïve Gauss Elimination and Gaussian Elimination with Partial Pivoting. To appreciate why $LU$ Decomposition could be a better choice than the Gaussian Elimination techniques in some cases, let us discuss first what $LU$ Decomposition is about.

For any non-singular matrix $A$ on which one can conduct Naïve Gaussian Elimination or forward elimination steps, one can always write it as $A = LU$
where

$L$ = Lower triangular matrix
$U$ = Upper triangular matrix

Then if one is solving a set of equations $Ax = b$, it will imply that $LUx = b$ since $A = LU$.
Multiplying both side by $L^{-1}$, we have

$L^{-1}LUx = L^{-1}b$

$\Rightarrow\ IUx = L^{-1}b$ since $\left(L^{-1}L = I\right)$,

$\Rightarrow\ Ux = L^{-1}b$ since $\left(IU = U\right)$

Let $L^{-1}b = z$ then $Lz = b$ (1)
And $Ux = z$ (2)

So we can solve equation (1) first for $z$ and then use equation (2) to calculate $x$.

The computational time required to decompose the $A$ matrix to $LU$ form is proportional to $\frac{n^3}{3}$, where $n$ is the number of equations (size of $A$ matrix). Then to solve the $Lz = b$, the computational time is proportional to $\frac{n^2}{2}$. Then to solve the $Ux = z$, the computational time is proportional to $\frac{n^2}{2}$. So the total computational time to solve a set of equations by $LU$ decomposition is proportional to $\frac{n^3}{3} + n^2$.

In comparison, Gaussian elimination is computationally more efficient. It takes a computational time proportional to $\frac{n^3}{3} + \frac{n^2}{2}$, where the computational time for forward elimination is proportional to $\frac{n^3}{3}$ and for the back substitution the time is proportional to $\frac{n^2}{2}$.

Finding the inverse of the matrix $A$ reduces to solving $n$ sets of equations with the $n$ columns of the identity matrix as the RHS vector. For calculations of each column of the inverse of the $A$

matrix, the coefficient matrix $A$ matrix in the set of equation $Ax = b$ does not change. So if we use *LU* Decomposition method, the $A = LU$ decomposition needs to be done only once and the use of equations (1) and (2) still needs to be done '*n*' times.

So the total computational time required to find the inverse of a matrix using *LU* decomposition is proportional to $\frac{n^3}{3} + n(n^2) = \frac{4n^3}{3}$.

In comparison, if Gaussian elimination method were applied to find the inverse of a matrix, the time would be proportional to $n\left(\frac{n^3}{3} + \frac{n^2}{2}\right) = \frac{n^4}{3} + \frac{n^3}{2}$.

For large values of *n*, $\frac{n^4}{3} + \frac{n^3}{2} \rangle\rangle \frac{4n^3}{3}$

## 2-2.3 Decomposing a non-singular matrix A into the form A=LU.

***L U* Decomposition Algorithm**:

In these methods the coefficient matrix *A* of the given system of equation $Ax = b$ is written as a product of a Lower triangular matrix *L* and an Upper triangular matrix *U*, such that $A = LU$ where the elements of $L = (l_{ij} = 0;\text{ for } i < j)$ and the elements of $U = (u_{ij} = 0;\text{ for } i > j)$ that is,

$$L = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix} \text{ and } U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & u_{n-1,n} \\ 0 & \cdots & 0 & u_{nn} \end{pmatrix}.$$

Now using the rules of matrix multiplication $a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj}, \quad i, j = 1, \ldots, n$

This gives a system of $n^2$ equations for the $n^2 + n$ unknowns (the non-zero elements in *L* and *U*). To make the number of unknowns and the number of equations equal one can fix the diagonal element either in *L* or in *U* such as '1's then solve the $n^2$ equations for the remaining $n^2$ unknowns in *L* and *U*. This leads to the following algorithm:

*Algorithm*

The factorization $A = LU$, where $L = \left(l_{ij}\right)_{n\times n}$ is a lower triangular and $U = \left(u_{ij}\right)_{n\times n}$ an upper triangular, can be computed directly by the following algorithm (provided zero divisions are not encountered):

*Algorithm*

*For* $k = 1 \text{ to } n$ *do specify* $\left(l_{kk} \text{ or } u_{kk}\right)$ *and compute the other such that* $l_{kk}u_{kk} = a_{kk} - \sum_{m=1}^{k-1} l_{km}u_{mk}$ .

*Compute the* $k^{th}$ *column of* $L$ *using* $l_{ik} = \dfrac{1}{u_{kk}}\left(a_{ik} - \sum_{m=1}^{k-1} l_{im}u_{mk}\right) \ \left(k < i \le n\right)$, *and compute the* $k^{th}$ *row of* $U$ *using* $u_{kj} = \dfrac{1}{l_{kk}}\left(a_{kj} - \sum_{m=1}^{k-1} l_{km}u_{mj}\right) \ \left(k < j \le n\right)$

*End*

**Note**:

The procedure is called ***Doolittle*** or ***Crout*** Factorization when $l_{ii} = 1 \ \left(1 \le i \le n\right)$ or $u_{jj} = 1 \ \left(1 \le j \le n\right)$ respectively.

If forward elimination steps of Naïve Gauss elimination methods can be applied on a non-singular matrix, then $A$ can be decomposed into $LU$ as

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

1. The elements of the $U$ matrix are exactly the same as the coefficient matrix one obtains at the end of the forward elimination steps in Naïve Gauss Elimination.
2. The lower triangular matrix $L$ has 1 in its diagonal entries. The non-zero elements below the diagonal in $L$ are multipliers that made the corresponding entries zero in the upper triangular matrix $U$ during forward elimination.

## Solving Ax=b in pure matrix Notations

Solving systems of linear equations (AX=b) using LU factorization can be quite cumbersome, although it seem to be one of the simplest ways of finding the solution for the system, Ax=b. In pure matrix notation, the upper triangular matrix, U, can be calculated by constructing specific permutation matrices and elementary matrices to solve the Elimination process with both partial and complete pivoting.

The elimination process is equivalent to pre multiplying A by a sequence of lower-triangular matrices $M_k$ as follows:

$$M_{k-1}M_{k-2}\dots M_1A = U$$

Where $M_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \cdots & 0 & 1 \end{bmatrix}$, $\mathrm{M}_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & m_{n2} & \cdots & 1 \end{bmatrix}$ with $m_{ij} = -\frac{a_{ij}^{(j-1)}}{a_{jj}}$ known as the multiplier

In solving Gaussian elimination without partial pivoting to triangularize *A,* the process yields the factorization, $MA = U$ . In this case, the system *Ax* = *b* is equivalent to the triangular system $Ux = Mb = b'$ where $M = M_{k-1}M_{k-2}M_{k-3}\dots M_1$

The elementary matrices $M_1$ and $\mathrm{M}_2$ are called first and second Gaussian transformation matrix respectively with $M_k$ being the $k^{th}$ Gaussian transformation matrix.
Generally, to solve $Ax = b$ using Naïve Gaussian elimination without partial pivoting by this approach, a permutation matrix is introduced to perform the pivoting strategies:

First We find the factorization $MA = U$ by the triangularization algorithm using partial pivoting. We then solve the triangular system by back substitution as follows $Ux = Mb = b'$ .

Note that $M = M_{n-1}P_{n-1}M_{n-2}P_{n-2}\cdots M_2P_2M_1P_1$
The vector $b' = Mb = M_{n-1}P_{n-1}M_{n-2}P_{n-2}\cdots M_2P_2M_1P_1b$
Generally if we set $s_1 = b = (b_1, b_2, \dots, b_n)^T$
Then For $k = 1, 2, \dots, n-1$ do

$$s_{k+1} = M_kP_ks_k$$

***Example***

If $n = 3$, $P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, $M_1 = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{pmatrix}$, then $s_2 = M_1P_1s_1 = \begin{pmatrix} s_1^{(2)} \\ s_2^{(2)} \\ s_3^{(2)} \end{pmatrix}$

If $P_1 s_1 = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$ then the entries of $s^{(2)}$ are given by:

$$s_1^{(2)} = b_1$$
$$s_2^{(2)} = m_{21} b_1 + b_3$$
$$s_3^{(2)} = m_{31} b_1 + b_2$$

In the same way, to solve **Ax = b** Using Gaussian Elimination with Complete Pivoting, we modify the previous construction to include another permutation matrix, Q such that when post multiplied by A, we can perform column interchange. This results in $M(PAQ) = U$

**Example 1**

Find the *LU* decomposition of the matrix

$$A = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

**Solution**

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

The $U$ matrix is the same as found at the end of the forward elimination of Naïve Gauss elimination method, that is

Forward Elimination of Unknowns: Since there are three equations, there will be two steps of forward elimination of unknowns.

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

First step: Divide Row 1 by 25 and then multiply it by 64 and subtract the results from Row 2

$$Row\ 2 - \left[\frac{64}{25}\right] \times (Row\ 1) = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 144 & 12 & 1 \end{bmatrix}$$

Here the multiplier , $m_{21} = -\frac{64}{25}$

Divide Row 1 by 25 and then multiply it by 144 and subtract the results from Row 3

$$Row\ 3 - \left[\frac{144}{25}\right] \times (Row\ 1) = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}$$

Here the multiplier , $m_{31} = -\frac{144}{25}$ , hence the first Gaussian transformation matrix is given by:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2.56 & 1 & 0 \\ -5.76 & 0 & 1 \end{bmatrix}$$ And the corresponding product is given by:

$$A^{(1)} = M_1 A = \begin{bmatrix} 1 & 0 & 0 \\ -2.60 & 1 & 0 \\ -5.76 & 0 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix} \text{(by a single multiplication)}$$

<u>Second step</u>: We now divide Row 2 by -4.8 and then multiply by -16.8 and subtract the results from Row 3

$$Row\ 3 - \left[\frac{-16.8}{-4.8}\right] \times (Row\ 2) = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \text{ which produces } U = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

Here the multiplier, $m_{32} = -\frac{-16.8}{-4.8}$ , hence the 2nd Gaussian transformation matrix is given by:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3.5 & 1 \end{bmatrix}$$ And the corresponding product is given by:

$$A^{(2)} = M_2 A^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix} = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} = U \text{ (by a single multiplication)}$$

To find $\ell_{21}$ and $\ell_{31}$, what multiplier was used to make the $a_{21}$ and $a_{31}$ elements zero in the first step of forward elimination of Naïve Gauss Elimination Method  It was

$$\ell_{21} = -m_{21} = \frac{64}{25} = 2.56$$

$$\ell_{31} = -m_{31} = \frac{144}{25} = 5.76$$

To find $\ell_{32}$, what multiplier was used to make $a_{32}$ element zero. Remember $a_{32}$ element was made zero in the second step of forward elimination. The $A$ matrix at the beginning of the second step of forward elimination was

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}$$

So

$$\ell_{32} = -m_{32} = \frac{-16.8}{-4.8} = 3.5$$

Hence

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & -m_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix}$$

Confirm $LU = A$.

$$LU = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

**Example 2**

Use *LU* decomposition method to solve the following linear system of equations.

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

**Solution**

Recall that $Ax = b$ and if $A = LU$ then first solving $Lz = b$ and then $Ux = z$ gives the solution vector $x$.

Now in the previous example, we showed

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

First solve $Lz = b$, i.e.,

$$\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

to give

$$\begin{aligned} z_1 &= 106.8 \\ 2.56z_1 + z_2 &= 177.2 \\ 5.76z_1 + 3.5z_2 + z_3 &= 279.2 \end{aligned}$$

Forward substitution starting from the first equation gives

$$\begin{aligned} z_1 &= 106.8 \\ z_2 &= 177.2 - 2.56z_1 = 177.2 - 2.56(106.8) = -96.2 \\ z_3 &= 279.2 - 5.76z_1 - 3.5z_2 = 279.2 - 5.76(106.8) - 3.5(-96.21) = 0.735 \end{aligned}$$

Hence

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ 0.735 \end{bmatrix}$$

This matrix is same as the right hand side obtained at the end of the forward elimination steps of Naïve Gauss elimination method. Is this a coincidence?

Now solve $Ux = z$, i.e.,

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.21 \\ 0.735 \end{bmatrix}$$

$$\begin{aligned} 25x_1 + \ 5x_2 + \quad x_3 &= 106.8 \\ -4.8x_2 - 1.56x_3 &= -96.21 \\ 0.7x_3 &= 0.735 \end{aligned}$$

From the third equation $0.7x_3 = 0.735 \Rightarrow x_3 = \dfrac{0.735}{0.7} = 1.050$

Substituting the value of a$_3$ in the second equation,

$$-4.8x_2 - 1.56x_3 = -96.21 \Rightarrow x_2 = \frac{-96.21 + 1.56x_3}{-4.8} = \frac{-96.21 + 1.56(1.050)}{-4.8} = 19.70$$

Substituting the value of $x_2$ and $x_3$ in the first equation,

$$\begin{aligned} 25x_1 + 5x_2 + x_3 = 106.8 \Rightarrow x_1 &= \frac{106.8 - 5x_2 - x_3}{25} \\ &= \frac{106.8 - 5(19.70) - 1.050}{25} = 0.2900 \end{aligned}$$

The solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.2900 \\ 19.70 \\ 1.050 \end{bmatrix}$$

***Example 2.2***

Solve

$$Ax = b \text{ with } A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \text{ and } b = \begin{pmatrix} 2 \\ 6 \\ 3 \end{pmatrix}$$

(a) using partial pivoting and (b) using complete pivoting.

***Solution:***

*(a) Partial pivoting:*

We compute *U* as follows:

$Step1:$

$$P_1=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix};\quad P_1A=\begin{bmatrix}1&2&3\\0&1&1\\1&1&1\end{bmatrix};\quad M_1=\begin{bmatrix}1&0&0\\0&1&0\\-1&0&1\end{bmatrix};$$

$$A^{(1)}=M_1P_1A=\begin{bmatrix}1&2&3\\0&1&1\\0&-1&-2\end{bmatrix};$$

$Step2:$

$$P_2=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix};\quad P_2A^{(1)}=\begin{bmatrix}1&2&3\\0&1&1\\0&-1&-2\end{bmatrix};\quad M_2=\begin{bmatrix}1&0&0\\0&1&0\\0&1&1\end{bmatrix};$$

$$U=A^{(2)}=M_2P_2A^{(1)}=M_2P_2M_1P_1A=\begin{bmatrix}1&2&3\\0&1&1\\0&0&-1\end{bmatrix}$$

$Note$: Defining $P=P_2P_1$ and $L=P(M_2P_2M_1P_1)^{-1}$, we have $PA=LU$.

We compute $b'$ as follows:

$Step1:$

$$P_1=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix};\quad P_1b=\begin{bmatrix}6\\2\\3\end{bmatrix};\quad M_1=\begin{bmatrix}1&0&0\\0&1&0\\-1&0&1\end{bmatrix};\quad M_1P_1b=\begin{bmatrix}6\\2\\-3\end{bmatrix};$$

$Step2:$

$$P_2=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix};\quad P_2M_1P_1b=\begin{bmatrix}6\\2\\-3\end{bmatrix};\quad M_2=\begin{bmatrix}1&0&0\\0&1&0\\0&1&1\end{bmatrix};\quad b'=M_2P_2M_1P_1b=\begin{bmatrix}6\\2\\-1\end{bmatrix};$$

The solution of the system

$$Ux=b' \quad\Rightarrow\quad \begin{bmatrix}1&2&3\\0&1&1\\0&0&-1\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}6\\2\\-1\end{bmatrix}\text{ and } x_1=x_2=x_3=1$$

*(b) Complete pivoting:* We compute $U$ as follows:

$Step1$:

$$P_1=\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix};\quad Q_1=\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix};\quad P_1AQ_1=\begin{pmatrix}3&2&1\\1&1&0\\1&1&1\end{pmatrix};\qquad M_1=\begin{pmatrix}1&0&0\\-\frac{1}{3}&1&0\\-\frac{1}{3}&0&1\end{pmatrix};$$

$$A^{(1)}=M_1P_1AQ_1=\begin{pmatrix}1&2&3\\0&\frac{1}{3}&-\frac{1}{3}\\0&\frac{1}{3}&\frac{2}{3}\end{pmatrix}$$

$Step2$:

$$P_2=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix};\quad Q_2=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix};\quad P_2A^{(1)}Q_2=\begin{pmatrix}3&1&2\\0&\frac{2}{3}&\frac{1}{3}\\0&-\frac{1}{3}&\frac{1}{3}\end{pmatrix};\qquad M_2=\begin{pmatrix}1&0&0\\0&1&0\\0&\frac{1}{2}&1\end{pmatrix};$$

$$U=A^{(2)}=M_2P_2A^{(1)}Q_2=M_2P_2M_1P_1AQ_1Q_2=\begin{pmatrix}3&1&2\\0&\frac{2}{3}&\frac{1}{3}\\0&0&\frac{1}{2}\end{pmatrix}$$

$Note$: Defining $P=P_2P_1$, $Q=Q_1Q_2$ and $L=P(M_2P_2M_1P_1)^{-1}$, we have $PAQ=LU$.

We compute $b'$ as follows:

$Step1$:

$$P_1=\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix};\quad P_1b=\begin{pmatrix}6\\2\\3\end{pmatrix};\quad M_1=\begin{pmatrix}1&0&0\\-\frac{1}{3}&1&0\\-\frac{1}{3}&0&1\end{pmatrix};\quad M_1P_1b=\begin{pmatrix}6\\0\\1\end{pmatrix}$$

$Step2$:

$$P_2=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix};\quad P_2M_1P_1b=\begin{pmatrix}6\\1\\0\end{pmatrix};\quad M_2=\begin{pmatrix}1&0&0\\0&1&0\\0&\frac{1}{2}&1\end{pmatrix};\quad M_1P_1b=\begin{pmatrix}6\\0\\1\end{pmatrix};$$

$$b'=M_2P_2M_1P_1b=\begin{pmatrix}6\\1\\\frac{1}{2}\end{pmatrix}$$

The solution of the system

$$Uy=b'\quad\Rightarrow\quad\begin{pmatrix}3&1&2\\0&\frac{2}{3}&\frac{1}{3}\\0&0&\frac{1}{2}\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}6\\1\\\frac{1}{2}\end{pmatrix}$$

is $y_1=y_2=y_3=1$. Because $\{x_k\}$, $k=1,2,3$ is simply the rearrangement of $\{y_k\}$, we have $x_1=x_2=x_3=1$.

## 2-2.4 Finding the inverse of a square matrix using LU Decomposition

A matrix $B$ is the inverse of $A$ if $AB = I = BA$. First assume that the first column of $B$ (the inverse of $A$ is $\left[b_{11}\ b_{21}\ \cdots\ b_{n1}\right]^T$ then from the above definition of inverse and definition of matrix multiplication.

$$A\begin{bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{n1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Similarly the second column of $B$ is given by

$$A\begin{bmatrix} b_{12} \\ b_{22} \\ \vdots \\ b_{n2} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Similarly, all columns of $B$ can be found by solving *n* different sets of equations with the column of the right hand sides being the *n* columns of the identity matrix.

**Example 3**

Use *LU* decomposition to find the inverse of

$$A = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

**Solution**

Knowing that

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

We can solve for the first column of $B = A^{-1}$ by solving for

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}\begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

First solve $Lz = c$, that is

$$\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

to give

$$\begin{aligned} z_1 \qquad\qquad &= 1 \\ 2.56z_1 + \quad z_2 \qquad &= 0 \\ 5.76z_1 + 3.5z_2 + z_3 &= 0 \end{aligned}$$

Forward substitution starting from the first equation gives

$$\begin{aligned} &\mathrm{z}_1 = 1 \\ &\mathrm{z}_2 = \ 0\text{-}2.56\mathrm{z}_1 = 0 - 2.56(1) = -256 \\ &z_3 = 0 - 5.76z_1 - 3.5z_2 = 0 - 5.76(1) - 3.5(-2.56) = 3.2 \end{aligned}$$

Hence

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -2.56 \\ 3.2 \end{bmatrix}$$

Now solve $Ux = z$, that is

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ \text{- }2.56 \\ 3.2 \end{bmatrix} \Rightarrow \begin{aligned} 25b_{11} + \quad 5b_{21} + \qquad b_{31} &= 1 \\ -4.8b_{21} - 1.56b_{31} &= -2.56 \\ 0.7b_{31} &= 3.2 \end{aligned}$$

Backward substitution starting from the third equation gives

$$b_{31} = \frac{3.2}{0.7} = 4.571$$

$$b_{21} = \frac{-2.56 + 1.560 b_{31}}{-4.8} = \frac{-2.56 + 1.560(4.571)}{-4.8} = -0.9524$$

$$b_{11} = \frac{1 - 5b_{21} - b_{31}}{25} = \frac{1 - 5(-0.9524) - 4.571}{25} = 0.04762$$

Hence the first column of the inverse of $A$ is

$$\begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 0.04762 \\ -0.9524 \\ 4.571 \end{bmatrix}$$

Similarly by solving

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} -0.08333 \\ 1.417 \\ -5.000 \end{bmatrix}$$

and solving

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{b}_{13} \\ \mathrm{b}_{23} \\ \mathrm{b}_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0.03571 \\ -0.4643 \\ 1.429 \end{bmatrix}$$

Hence

$$A^{-1} = \begin{bmatrix} 0.4762 & 0.08333 & 0.0357 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.050 & 1.429 \end{bmatrix}$$

**Exercise**

Show that $AA^{-1} = I = A^{-1}A$ for the above example.

## SESSION 3-2: Iterative Method

### 3-1.1 Jacobi Method

Given a general set of *n* equations and *n* unknowns, we have

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

If the diagonal elements are non-zero, each equation is rewritten for the corresponding unknown, that is, the first equation is rewritten with $x_1$ on the left hand side and the second equation is rewritten with $x_2$ on the left hand side and so on as follows:

$$\begin{aligned}
x_1 &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n\right) \\
x_2 &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1 - a_{23}x_3 - \cdots - a_{2n}x_n\right) \\
&\vdots \\
x_{n-1} &= \frac{1}{a_{n-1,n-1}}\left(b_{n-1} - a_{n-1,1}x_1 - a_{n-1,2}x_2 - \cdots - a_{n-1,n}x_n\right) \\
x_n &= \frac{1}{a_{nn}}\left(b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n,n-1}x_{n-1}\right)
\end{aligned}$$

These equations can be rewritten in the summation form as

$$x_1 = \frac{1}{a_{11}}\left(b_1 - \sum_{j=1,\, j\neq 1}^{n} a_{1j}x_j\right)$$
$$x_2 = \frac{1}{a_{22}}\left(b_2 - \sum_{j=1,\, j\neq 2}^{n} a_{2j}x_j\right)$$
$$\vdots$$
$$x_{n-1} = \frac{1}{a_{n-1,n-1}}\left(b_{n-1} - \sum_{j=1,\, j\neq n-1}^{n} a_{n-1j}x_j\right)$$
$$x_n = \frac{1}{a_{nn}}\left(b_n - \sum_{j=1,\, j\neq n}^{n} a_{nj}x_j\right)$$

Hence for any row *i*, $x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij}x_j\right), \quad i = 1, 2, \ldots, n.$

By assuming an initial guess for the $x_i$ 's, one uses $x_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij}x_j^{(k-1)}\right), \quad i = 1, 2, \ldots, n.$ to calculate the new values for the $x_i$ 's. At the end of each iteration, one calculates the absolute relative approximate error for each $x_i$ as $\left|\varepsilon_a\right|_i = \left|\frac{x_i^{new} - x_i^{old}}{x_i^{new}}\right| \times 100$, where $x_i^{new}$ is the recently obtained value of $x_i$ , and $x_i^{old}$ is the previous value of $x_i$ .

When the absolute relative approximate error for each $x_i$ is less than the pre-specified tolerance, the iterations are stopped.

If the coefficient matrix $A$ of the system $Ax = b$, is written as $A = L + D + U$ where $D$ has the diagonal elements of $A$ and $L$ & $U$ respectively have the lower diagonal and upper diagonal elements of $A$ then the Jacobi scheme can be written in matrix form as

$$Dx = -(L+U)x + b$$
$$Dx^{(k)} = -(L+U)x^{(k-1)} + b$$

giving $x^{(k)} = -D^{-1}(L+U)x^{(k-1)} + D^{-1}b$, i.e., the Jacobi iterative matrix and the corresponding vector are given as $T_J = -D^{-1}(L+U)$ and $c_J = D^{-1}b$ respectively.

**Example 1**

Use Jacobi Method to solve $Ax = b$, where

$$A = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Assume an initial guess of the solution as $\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T = \begin{bmatrix} 1 & 2 & 5 \end{bmatrix}^T$. Perform two iterations only.

**Solution**

Rewriting the equations gives

$$x_1^{(k)} = \frac{106.8 - 5x_2^{(k-1)} - x_3^{(k-1)}}{25}$$
$$x_2^{(k)} = \frac{177.2 - 64x_1^{(k-1)} - x_3^{(k-1)}}{8}$$
$$x_3^{(k)} = \frac{279.2 - 144x_1^{(k-1)} - 12x_2^{(k-1)}}{1}$$

**Iteration #1**

Given the initial guess of the solution vector as

$$\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

we get

$$x_1^{(1)} = \frac{106.8 - 5(2) - (5)}{25} = 3.6720$$
$$x_2^{(1)} = \frac{177.2 - 64(1) - (5)}{8} = 13.525$$
$$x_3^{(1)} = \frac{279.2 - 144(2) - 12(5)}{1} = - - 68.8$$

The absolute relative approximate error for each $x_i$ then is

$$\left|\varepsilon_a\right|_1 = \left|\frac{3.6720 - 1.0000}{3.6720}\right| \times 100 = 72.76\%$$

$$\left|\varepsilon_a\right|_2 = \left|\frac{13.525 - 2.0000}{13.525}\right| \times 100 = 85.21\%$$

$$\left|\varepsilon_a\right|_3 = \left|\frac{-68.8 - 5.0000}{-68.8}\right| \times 100 = 108.53\%$$

At the end of the first iteration, the guess of the solution vector is

$$\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 3.6720 \\ 13.525 \\ -68.8 \end{bmatrix}$$

and the maximum absolute relative approximate error is 108.53%.

**Iteration #2**

The estimate of the solution vector at the end of iteration #1 is

$$\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 3.6720 \\ 13.525 \\ -68.8 \end{bmatrix}$$

Now we get

$$x_1^{(2)} = \frac{106.8 - 5(13.525) - 68.8}{25} = -1.185$$

$$x_2^{(2)} = \frac{177.2 - 64(3.6720) - 68.8}{8} = -15.862$$

$$x_3^{(2)} = \frac{279.2 - 144(3.6720) - 12(-68.8)}{1} = 576.032$$

The absolute relative approximate error for each $x_i$ then is

$$\left|\in_a\right|_1 = \left|\frac{-1.185-3.6720}{-1.185}\right|\times 100 = 428.37\%$$

$$\left|\in_a\right|_2 = \left|\frac{-15.862-(13.525)}{-15.862}\right|\times 100 = 185.27\%$$

$$\left|\in_a\right|_3 = \left|\frac{576.032-(-68.8)}{576.032}\right|\times 100 = 111.94\%$$

At the end of second iteration the estimate of the solution is

$$\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} -1.185 \\ -15.862 \\ 576.032 \end{bmatrix}$$

and the maximum absolute relative approximate error is 428.37%.

**Example 2**

Given the system of equations;

$$\begin{aligned} 12x_1 + 3x_2 - \ 5x_3 &= 1 \\ x_1 + 5x_2 + \ 3x_3 &= 28 \\ 3x_1 + 7x_2 + 13x_3 &= 76 \end{aligned}$$

find the solution, given $\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T$ as the initial guess.

**Solution**

Rewriting the equations, we get

$$x_1^{(k)} = \frac{1-3x_2^{(k-1)}+5x_3^{(k-1)}}{12}$$

$$x_2^{(k)} = \frac{28-x_1^{(k-1)}-3x_3^{(k-1)}}{5}$$

$$x_3^{(k)} = \frac{76-3x_1^{(k-1)}-7x_2^{(k-1)}}{13}$$

Assuming an initial guess of

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

Iteration #1:

$$x_1^{(1)} = \frac{1-3(0)+5(1)}{12} = 0.50000$$

$$x_2^{(1)} = \frac{28-(1)-3(1)}{5} = 4.80000$$

$$x_3^{(1)} = \frac{76-3(1)-7(0)}{13} = 5.6154$$

The absolute relative approximate error at the end of first iteration is

$$\left|\varepsilon_a\right|_1 = \left|\frac{0.50000-1.0000}{0.50000}\right| \times 100 = 67.662\%$$

$$\left|\varepsilon_a\right|_2 = \left|\frac{4.8000-0}{4.8000}\right| \times 100 = 100\%$$

$$\left|\varepsilon_a\right|_3 = \left|\frac{5.6154-1.0000}{5.6154}\right| \times 100 = 82.19\%$$

The maximum absolute relative approximate error is 100.000%

Iteration #2:

$$x_1^{(2)} = \frac{1-3(4.8000)+5(5.6154)}{12} = 14.677$$

$$x_2^{(2)} = \frac{28-(0.5000)-3(5.6154)}{5} = 2.13076$$

$$x_3^{(2)} = \frac{76-3(0.5000)-7(4.800)}{13} = 3.14615$$

At the end of second iteration, the absolute relative approximate error is

$$\left|\varepsilon_a\right|_1 = \left|\frac{14.677-0.50000}{14.677}\right|\times 100 = 96.59\%$$
$$\left|\varepsilon_a\right|_2 = \left|\frac{2.13076-4.8000}{2.13076}\right|\times 100 = 125.27\%$$
$$\left|\varepsilon_a\right|_3 = \left|\frac{3.14615-5.6154}{3.14615}\right|\times 100 = 78.484\%$$

The maximum absolute relative approximate error is 125.27%.

### 3-1.2 Gauss-Seidel Method

If the set of *n* equations and *n* unknowns

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

is rewritten as

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

From the above it is seen that the most recent updates of the $x_i$ 's are used immediately they are available then the above reduces to

$$x_i^{(k)} = \frac{1}{a_{ii}}\left\{b_i - \sum_{j=1}^{i-1} a_{ij}x_i^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_i^{(k-1)}\right\}, \quad i = 1,\ldots,n \qquad (*)$$

Again in the matrix notation, the coefficient matrix $A$ of the system $Ax = b$, is split into $A = L + D + U$ where $D$ has the diagonal elements of $A$ and $L$ & $U$ respectively have the lower diagonal and upper diagonal elements of $A$ then the Gauss-Seidel scheme can be written as

$(L+D)x = -Ux + b$

$(L+D)x^{(k)} = -Ux^{(k-1)} + b$

giving $x^{(k)} = -(L+D)^{-1}Ux^{(k-1)} + (L+D)^{-1}b$.

i.e., the Gauss-Seidel iterative matrix and the corresponding vector are given as $T_{GS} = -(L+D)^{-1}U$ and $c_{GS} = (L+D)^{-1}b$ respectively.

**Example 1**

Use Gauss-Seidel Method to solve $Ax = b$, where

$$A = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Assume an initial guess of the solution as $\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T = \begin{bmatrix} 1 & 2 & 5 \end{bmatrix}^T$.

**Solution**

Rewriting the equations gives

$$x_1^{(k)} = \frac{106.8 - 5x_2^{(k-1)} - x_3^{(k-1)}}{25}$$

$$x_2^{(k)} = \frac{177.2 - 64x_1^{(k)} - x_3^{(k-1)}}{8}$$

$$x_3^{(k)} = \frac{279.2 - 144x_1^{(k)} - 12x_2^{(k)}}{1}$$

Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

we get

$$x_1^{(1)} = \frac{106.8 - 5(2) - (5)}{25} = 3.6720$$

$$x_2^{(1)} = \frac{177.2 - 64(3.6720) - (5)}{8} = -7.8510$$

$$x_3^{(1)} = \frac{279.2 - 144(3.6720) - 12(-7.8510)}{1} = -155.36$$

The absolute relative approximate error for each $x_i$ then is

$$\left|\varepsilon_a\right|_1 = \left|\frac{3.6720-1.0000}{3.6720}\right| \times 100 = 72.76\%$$

$$\left|\varepsilon_a\right|_2 = \left|\frac{-7.8510-2.0000}{-7.8510}\right| \times 100 = 125.47\%$$

$$\left|\varepsilon_a\right|_3 = \left|\frac{-155.36-5.0000}{-155.36}\right| \times 100 = 103.22\%$$

At the end of the first iteration, the guess of the solution vector is

$$\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

and the maximum absolute relative approximate error is 125.47%.

Iteration #2

The estimate of the solution vector at the end of iteration #1 is

$$\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

Now we get

$$x_1^{(2)} = \frac{106.8-5(-7.8510)-155.36}{25} = 12.056$$

$$x_2^{(2)} = \frac{177.2-64(12.056)-155.36}{8} = -54.882$$

$$x_3^{(2)} = \frac{279.2-144(12.056)-12(-54.882)}{1} = -798.34$$

The absolute relative approximate error for each $x_i$ then is

$$\left|\in_a\right|_1 = \left|\frac{12.056 - 3.6720}{12.056}\right| \times 100 = 69.542\%$$

$$\left|\in_a\right|_2 = \left|\frac{-54.882 - \left(-7.8510\right)}{-54.882}\right| \times 100 = 85.695\%$$

$$\left|\in_a\right|_3 = \left|\frac{-798.34 - \left(-155.36\right)}{-798.34}\right| \times 100 = 80.54\%$$

At the end of second iteration the estimate of the solution is

$$\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} 12.056 \\ -54.882 \\ -798.34 \end{bmatrix}$$

and the maximum absolute relative approximate error is 85.695%.

Conducting more iterations gives the following values for the solution vector and the corresponding absolute relative approximate errors.

| Iteration $k$ | $x_1^{(k)}$ | $\left\|\varepsilon_a\right\|_1$ % | $x_2^{(k)}$ | $\left\|\varepsilon_a\right\|_2$ % | $x_3^{(k)}$ | $\left\|\varepsilon_a\right\|_3$ % |
|---|---|---|---|---|---|---|
| 1 | 3.672 | 72.767 | -7.8510 | 125.47 | -155.36 | 103.22 |
| 2 | 12.056 | 67.542 | -54.882 | 85.695 | -798.34 | 80.540 |
| 3 | 47.182 | 74.448 | -255.51 | 78.521 | -3448.9 | 76.852 |
| 4 | 193.33 | 75.595 | -1093.4 | 76.632 | -14440 | 76.116 |
| 5 | 800.53 | 75.850 | -4577.2 | 76.112 | -60072 | 75.962 |
| 6 | 3322.6 | 75.907 | -19049 | 75.971 | -249580 | 75.931 |

As seen in the above table, the solution is not converging to the true solution of
$x_1 = 0.29048,\ \ x_2 = 19.690,\ \ x_3 = 1.0858$

**The above system of equations does not seem to converge? Why?**
A pitfall of most iterative methods is that they may or may not converge. However, certain classes of systems of simultaneous equations do always converge to a solution using Gauss-Seidel method. This class of system of equations is where the coefficient matrix $A$ in $Ax = b$ is diagonally dominant, that is $\left|a_{ii}\right| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|a_{ij}\right|$ for all $i$ and $\left|a_{ii}\right| > \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|a_{ij}\right|$ for at least one $i$.

If a system of equations has a coefficient matrix that is not diagonally dominant, it may or may not converge. Fortunately, many physical systems that result in simultaneous linear equations have diagonally dominant coefficient matrix, which then assures convergence for iterative methods such as Gauss-Seidel method of solving simultaneous linear equations.

**Example 2**

Given the system of equations;

$$\begin{aligned} 12x_1 + 3x_2 - \ \ 5x_3 &= 1 \\ x_1 + 5x_2 + \ \ 3x_3 &= 28 \\ 3x_1 + 7x_2 + 13x_3 &= 76 \end{aligned}$$

find the solution. Given $\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T$ as the initial guess.

**Solution**

The coefficient matrix

$$A = \begin{bmatrix} 12 & 3 & -5 \\ 1 & 5 & 3 \\ 3 & 7 & 13 \end{bmatrix}$$

is diagonally dominant as

$$|a_{11}| = |12| = 12 \geq |a_{12}| + |a_{13}| = |3| + |-5| = 8$$

$$|a_{22}| = |5| = 5 \geq |a_{21}| + |a_{23}| = |1| + |3| = 4$$

$$|a_{33}| = |13| = 13 \geq |a_{31}| + |a_{32}| = |3| + |7| = 10$$

and the inequality is strictly greater than for at least one row. Hence the solution should converge using Gauss-Seidel method.

Rewriting the equations, we get

$$x_1^{(k)} = \frac{1 - 3x_2^{(k-1)} + 5x_3^{(k-1)}}{12}$$

$$x_2^{(k)} = \frac{28 - x_1^{(k)} - 3x_3^{(k-1)}}{5}$$

$$x_3^{(k)} = \frac{76 - 3x_1^{(k)} - 7x_2^{(k)}}{13}$$

Assuming an initial guess of

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

Iteration #1:

$$x_1^{(1)} = \frac{1-3(0)+5(1)}{12} = 0.50000$$
$$x_2^{(1)} = \frac{28-(0.5)-3(1)}{5} = 4.90000$$
$$x_3^{(1)} = \frac{76-3(0.50000)-7(4.9000)}{13} = 3.0923$$

The absolute relative approximate error at the end of first iteration is

$$\left|\varepsilon_a\right|_1 = \left|\frac{0.50000-1.0000}{0.50000}\right| \times 100 = 67.662\%$$
$$\left|\varepsilon_a\right|_2 = \left|\frac{4.9000-0}{4.9000}\right| \times 100 = 100\%$$
$$\left|\varepsilon_a\right|_3 = \left|\frac{3.0923-1.0000}{3.0923}\right| \times 100 = 67.662\%$$

The maximum absolute relative approximate error is 100.000%

Iteration #2:

$$x_1^{(2)} = \frac{1-3(4.9000)+5(3.0923)}{12} = 0.14679$$
$$x_2^{(2)} = \frac{28-(0.14679)-3(3.0923)}{5} = 3.7153$$
$$x_3^{(2)} = \frac{76-3(0.14679)-7(4.900)}{13} = 3.8118$$

At the end of second iteration, the absolute relative approximate error is

$$\left|\varepsilon_a\right|_1 = \left|\frac{0.14679 - 0.50000}{0.14679}\right| \times 100 = 240.62\%$$
$$\left|\varepsilon_a\right|_2 = \left|\frac{3.7153 - 4.9000}{3.7153}\right| \times 100 = 31.887\%$$
$$\left|\varepsilon_a\right|_3 = \left|\frac{3.8118 - 3.0923}{3.8118}\right| \times 100 = 18.876\%$$

The maximum absolute relative approximate error is 240.62%. This is greater than the value of 67.612% we obtained in the first iteration. Is the solution diverging? No, as you conduct more iterations, the solution converges as follows.

| Iteration *k* | $x_1^{(k)}$ | $\left\|\varepsilon_a\right\|_1$ | $x_2^{(k)}$ | $\left\|\varepsilon_a\right\|_2$ | $x_3^{(k)}$ | $\left\|\varepsilon_a\right\|_3$ |
|---|---|---|---|---|---|---|
| 1 | 0.50000 | 67.662 | 4.900 | 100.00 | 3.0923 | 67.662 |
| 2 | 0.14679 | 240.62 | 3.7153 | 31.887 | 3.8118 | 18.876 |
| 3 | 0.74275 | 80.23 | 3.1644 | 17.409 | 3.9708 | 4.0042 |
| 4 | 0.94675 | 21.547 | 3.0281 | 4.5012 | 3.9971 | 0.65798 |
| 5 | 0.99177 | 4.5394 | 3.0034 | 0.82240 | 4.0001 | 0.07499 |
| 6 | 0.99919 | 0.74260 | 3.0001 | 0.11000 | 4.0001 | 0.00000 |

This is close to the exact solution vector of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$

**Example 3**

Given the system of equations

$$\begin{aligned} 3x_1 + 7x_2 + 13x_3 &= 76 \\ x_1 + 5x_2 + \ \ 3x_3 &= 28 \\ 12x_1 + 3x_2 - \ \ 5x_3 &= 1 \end{aligned}$$

find the solution using Gauss-Seidal method. Use $\left[x_1, x_2, x_3\right]^T = \left[1 \ \ 0 \ \ 1\right]^T$ as the initial guess.

**Solution**

Rewriting the equations, we get

$$x_1^{(k)} = \frac{76 - 7x_2^{(k-1)} - 13x_3^{(k-1)}}{3}$$

$$x_2^{(k)} = \frac{28 - x_1^{(k)} - 3x_3^{(k-1)}}{5}$$

$$x_3^{(k)} = \frac{1 - 12x_1^{(k)} - 3x_3^{(k)}}{-5}$$

Assuming an initial guess of

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

the next six iterative values are given in the table below

| Iteration $k$ | $x_1^{(k)}$ | $\lvert\varepsilon_a\rvert_1$ | $x_2^{(k)}$ | $\lvert\varepsilon_a\rvert_2$ | $x_3^{(k)}$ | $\lvert\varepsilon_a\rvert_3$ |
|---|---|---|---|---|---|---|
| 1 | 21.000 | 110.71 | 0.80000 | 100.00 | 5.0680 | 98.027 |
| 2 | -196.15 | 109.83 | 14.421 | 94.453 | -462.30 | 110.96 |
| 3 | -1995.0 | 109.90 | -116.02 | 112.43 | 4718.1 | 109.80 |
| 4 | -20149 | 109.89 | 1204.6 | 109.63 | -47636 | 109.90 |
| 5 | 2.0364x$10^5$ | 109.90 | -12140 | 109.92 | 4.8144x$10^5$ | 109.89 |
| 6 | -2.0579x$10^5$ | 1.0990 | 1.2272x$10^5$ | 109.89 | -4.8653x$10^6$ | 109.89 |

You can see that this solution is not converging and the coefficient matrix is not diagonally dominant. The coefficient matrix

$$A = \begin{bmatrix} 3 & 7 & 13 \\ 1 & 5 & 3 \\ 12 & 3 & -5 \end{bmatrix}$$

is not diagonally dominant as

$$|a_{11}| = |3| = 3 \le |a_{12}| + |a_{13}| = |7| + |13| = 20$$

Hence Gauss-Seidal method may or may not converge.

However, it is the same set of equations as the previous example and that converged. The only difference is that we exchanged first and the third equation with each other and that made the coefficient matrix not diagonally dominant.

So it is possible that a system of equations can be made diagonally dominant if one exchanges the equations with each other. But it is not possible for all cases. For example, the following set of equations.

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 2x_1 + 3x_2 + 4x_3 &= 9 \\ x_1 + 7x_2 + x_3 &= 9 \end{aligned}$$

cannot be rewritten to make the coefficient matrix diagonally dominant.

***Example 4***

Find an approximation to the solution of $Ax = b$ by performing two iterations of the Gauss-Seidel method where $A = \begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $x^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

**Solution**

Using $x_i^{(k)} = \dfrac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{2} a_{ij}x_j^{(k-1)}\right)$, $i = 1, 2$, $k = 1, 2, \ldots$, we have

$$D = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_{\text{off}} = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$$

$k = 1$

$$x_1^{(1)} = \frac{1}{4}\left(1 - [0,2] \times \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) = -\frac{1}{4}, \quad x_2^{(1)} = 1\left(2 - [1,0] \times \begin{bmatrix} -1/4 \\ 1 \end{bmatrix}\right) = \frac{9}{4}$$

$$\overrightarrow{x_1} = \begin{bmatrix} -1/4 \\ 9/4 \end{bmatrix}$$

$k = 2$

$$x_1^{(2)} = \frac{1}{4}\left(1 - [0,2] \times \begin{bmatrix} -1/4 \\ 9/4 \end{bmatrix}\right) = -7/8, \quad x_2^{(2)} = 1\left(2 - [1,0] \times \begin{bmatrix} -7/8 \\ 9/4 \end{bmatrix}\right) = 23/8$$

$$\overrightarrow{x_2} = \begin{bmatrix} -7/8 \\ 23/8 \end{bmatrix}$$

### 3-1.3 Successive Over-Relaxation Method

If the Gauss-Seidel iteration equations (*) is written as

$$x_i^{(k)} = x_i^{(k-1)} + \frac{1}{a_{ii}}\left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_i^{(k)} - \sum_{j=i}^{n} a_{ij} x_i^{(k-1)} \right\}, \quad i = 1, \dots, n$$

Then multiplying the second term of the right hand side of the above by $\omega$, we have

$$\begin{aligned} & x_i^{(k)} = x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_i^{(k)} - \sum_{j=i}^{n} a_{ij} x_i^{(k-1)} \right\}, \quad i = 1, \dots, n \\ \Rightarrow \quad & x_i^{(k)} = \frac{\omega}{a_{ii}}\left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_i^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_i^{(k-1)} \right\} + (1-\omega) x_i^{(k-1)}, \quad i = 1, \dots, n \end{aligned} \tag{**}$$

which gives the SOR iteration equations where the factor $\omega$ is called the acceleration parameter or relaxation factor, which generally lies in the range $0 < \omega < 2$. The determination of the optimum value of $\omega$ for maximum rate of convergence is again a matter of discussion and is not considered. When $\omega = 1$ gives the Gauss-Seidel iteration, $0 < \omega < 1$ we have **under relaxation** and $1 < \omega < 2$ we have **over relaxation**.

From (**), we can write the matrix notation for the SOR method as follows:

$$\begin{aligned} & x^{(k)} - x^{(k-1)} = \omega D^{-1}\left( -Lx^{(k)} - Ux^{(k-1)} + b - Dx^{(k-1)} \right) \\ \Rightarrow \quad & \left( I + \omega D^{-1} L \right) x^{(k)} = \omega D^{-1}\left( -U - D \right) x^{(k-1)} + x^{(k-1)} + \omega D^{-1} b \\ \Rightarrow \quad & \left( D + \omega L \right) x^{(k)} = \omega\left( -U - D \right) x^{(k-1)} + Dx^{(k-1)} + \omega b \end{aligned}$$

Therefore,

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

i.e., the SOR iterative matrix and the corresponding vector are given as
$T_{SOR} = \left( D + \omega L \right)^{-1}\left( (1-\omega) D - \omega U \right)$ and $c_{SOR} = \omega\left( D + \omega L \right)^{-1} b$ respectively.

**Example 1**

Find an approximation to the solution of $Ax = b$ by performing two iterations of the SOR method where $A = \begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $x^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with $\omega = 1.2$.

**Solution**

Using $x_i^{(k)} = \frac{\omega}{a_{ii}}\left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{2} a_{ij} x_j^{(k-1)} \right) + (1-\omega) x_i^{(k-1)}$, $i = 1, 2$, $k = 1, 2, \dots$, we have

$$D=\begin{bmatrix}4 & 0\\ 0 & 1\end{bmatrix},\ A_{\text{off}}=\begin{bmatrix}0 & 2\\ 1 & 0\end{bmatrix}$$

$k=1$

$$x_1^{(1)}=\frac{1.2}{4}\left(1-[0,2]\times\begin{bmatrix}1\\1\end{bmatrix}\right)-0.2=-0{,}5,\ \ x_2^{(1)}=1.2\left(2-[1,0]\times\begin{bmatrix}-0.5\\1\end{bmatrix}\right)-0.2=2.8$$

$$\vec{x}_1=\begin{bmatrix}-0.5\\2.8\end{bmatrix}$$

$k=2$

$$x_1^{(2)}=\frac{1.2}{4}\left(1-[0,2]\times\begin{bmatrix}-0.5\\2.8\end{bmatrix}\right)+0.1=-1.28,\ \ x_2^{(2)}=1.2\left(2-[1,0]\times\begin{bmatrix}-1.28\\2.8\end{bmatrix}\right)-0.56=3.376$$

$$\vec{x}_2=\begin{bmatrix}-1.28\\3.376\end{bmatrix}$$

## 3-1.4 A necessary and sufficient condition for the convergence of iterative schemes

Consider any iterative scheme $x^{(k)}=Tx^{(k-1)}+c$, where $T$ is the iterative matrix and $c$ is the constant vector of known values. If $e^{(k)}$ is the error in the $k^{th}$ approximation to the exact solution then it can be written as: $e^{(k)}=x-x^{(k)}$.

Similarly, $e^{(k+1)}=x-x^{(k+1)}$. Therefore $e^{(k+1)}=Te^{(k)}$ or $e^{(k)}=Te^{(k-1)}=T^2e^{(k-2)}=\cdots=T^ke^{(0)}$.

Since the sequence of approximations $\left\{x^{(0)},x^{(1)},x^{(2)},\ldots,x^{(k)},\ldots\right\}$ converges to $x$ as $n$ tends to infinity, we have $\lim\limits_{k\to\infty}e^{(k)}=0 \ \Rightarrow\ \lim\limits_{k\to\infty}T^ke^{(0)}=e^{(0)}\lim\limits_{k\to\infty}T^k=0$

If $T$ has $n$ linearly independent Eigenvectors $v_r,\ r=1,\ldots,n$ then these $n$ vectors can be used as a basis for any $n$ dimensional space and hence any vector in this $n$ dimensional space can be represented in terms of these $n$ vectors. In particular, $e^{(0)}=\sum\limits_{r=1}^{n}c_rv_r$, where $c_r,\ r=1,\ldots,n$ are scalars. Hence $e^{(1)}=Te^{(0)}=\sum\limits_{r=1}^{n}c_rTv_r$.

But $Tv_r=\lambda_rv_r$ by the definition of an eigenvalue, where $\lambda_r$ is the eigenvalue corresponding to the eigenvector $v_r$. Hence $e^{(1)}=\sum\limits_{r=1}^{n}c_r\lambda_rv_r$

Similarly $e^{(k)} = \sum_{r=1}^{n} c_r \lambda_r^k v_r$

Therefore $e^{(k)}$ will tend to the null vector as $k$ tends to infinity, for arbitrary $e^{(0)}$, if and only if $|\lambda_r| < 1$ for all $r$. In other words the iteration will converge for arbitrary $x^{(0)}$ if and only if, the spectral radius $\rho(T)$ of $T$ is less than unity.

As a corollary to this result a sufficient condition for convergence is that $\|T\| < 1$. To prove this we have that $Tv_r = \lambda_r v_r$. Hence

$$\|Tv_r\| = \|\lambda_r v_r\| = |\lambda_r| \|v_r\|$$
$$\|\lambda_r v_r\| \le \|T\| \|v_r\|$$

But for any matrix norm that is compatible with a vector norm $\|v_r\|$, $\|Tv_r\| \le \|T\| \|v_r\|$.

Therefore $\|\lambda_r v_r\| \le \|T\| \|v_r\|$so $|\lambda_r| \le \|T\|, \quad r = 1, \ldots, n$.

It follows from this that a sufficient condition for convergence is that $\|T\| < 1$. It is not a necessary condition because the norm of $T$ can exceed one even when $\rho(T) < 1$.

# SYSTEM OF NON-LINEAR EQUATIONS

## Introduction

Solutions $x = x_0$ to equations of the form $f(x) = 0$ are often required where it is impossible or infeasible to find an analytical expression for the vector $\boldsymbol{x}$. If the scalar function $f$ depends on $n$ independent variables $x_1, x_2, \ldots, x_n$, then the solution $\boldsymbol{x}_0$ will describe a surface in $n$ – 1 dimensional space. Alternatively we may consider the vector function $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$, the solutions of which typically converges to particular values of $\boldsymbol{x}$. For this course we restrict our attention to $n$ independent variables $x_1, x_2, \ldots, x_n$ and seek solutions to $F(\mathrm{x}) = 0$ where $F$ is vector valued.

## Learning Objectives

After reading this unit you should be able to:

- Explain functional fixed-point iteration,
- Explain the Gauss-Seidel and Newton's methods.
- Describe the method of undetermined coefficients or polynomial interpolation,
- Explain the Newton's divided difference method and the Lagrange interpolation method.

## content

## SESSION 1-3: Fixed Points for Functions of Several Variables

The general of a system of nonlinear equations is

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \end{aligned}$$

where each function $f_i$ maps n-dimensional space, $R^n$, into the real line *R*. The above system can be defined alternatively by defining the function $F(\mathrm{x}) = 0$, where $F : R^n \to R^n$, $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ and $F(x_1, x_2, \ldots, x_n) = \left( f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n), \ldots, f_n(x_1, x_2, \ldots, x_n) \right)$.

### 1-3.1 Functional or Fixed Point Iteration

Suppose a nonlinear system of the form $F(\mathrm{x}) = 0$ has been transformed into an equivalent fixed point problem $G(\mathrm{x}) = \mathrm{x}$. The functional or fixed point iteration process applied to *G* is as follows:

1. Select $\mathrm{x}^{(0)} = \left( x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)} \right)$.
2. Generate the sequence of vectors $\mathrm{x}^{(k)} = \left( x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)} \right)$ by
   $\mathrm{x}^{(k)} = \mathrm{G}\left( \mathrm{x}^{(k-1)} \right)$ for each $i = 1, 2, 3, \ldots$ or, component-wise,

$$\begin{aligned} x_1^{(k)} &= g_1\left( x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)} \right) \\ x_2^{(k)} &= g_2\left( x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)} \right) \\ \vdots \quad &\quad \vdots \\ x_n^{(k)} &= g_n\left( x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)} \right) \end{aligned}$$

The following theorem provides conditions for the iterative process to converge.

***Theorem***

Let $D = \left\{ (x_1, x_2, \ldots, x_n) : a_i \le x_i \le b_i, \text{ for each } i = 1, 2, \ldots, n \right\}$, for some collection of constants $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$. Suppose G is a continuous function with continuous first partial derivatives from $D \subset R^n$ into $R^n$ with the property that $G(\mathrm{x}) \in D$ whenever $\mathrm{x} \in D$. If a constant $K < 1$ exists with $\left| \frac{\partial g_i(\mathrm{x})}{\partial x_j} \right| \le \frac{K}{n}$ whenever $\mathrm{x} \in D$ for each $j = 1, 2, \ldots, n$ and each component function $g_i$, then the sequence $\left\{\mathrm{x}^{(k)}\right\}_{k=0}^{\infty}$ defined by $x^{(k)} = G(x^{(k-1)})$ for each $i$ = 1, 2, 3,... converges to the unique fixed point $\mathrm{p} \in D$, for any $\mathrm{x}^{(0)}$ in $D$, and $\left\|\mathrm{x}^{(j)} - \mathrm{p}\right\|_{\infty} \le \frac{K^j}{1-K}\left\|\mathrm{x}^{(1)} - \mathrm{x}^{(0)}\right\|_{\infty}$.

**Example 1**

$$\begin{aligned}
3x_1 - \cos(x_2x_3) \qquad\qquad & - \tfrac{1}{2} &= 0\\
x_1^2 - 81(x_2+0.1)^2 + \sin x_3 & + 1.06 &= 0\\
e^{-x_1x_2} \qquad\qquad + 20x_3 & + \tfrac{10\pi-3}{3} &= 0
\end{aligned}$$

$$\begin{aligned}
f_1(x_1, x_2, \ldots, x_n) &= 3x_1 - \cos(x_2x_3) - \tfrac{1}{2}\\
f_2(x_1, x_2, \ldots, x_n) &= x_1^2 - 81(x_2+0.1)^2 + \sin x_3 + 1.06\\
f_3(x_1, x_2, \ldots, x_n) &= e^{-x_1x_2} + 20x_3 \ + \tfrac{10\pi-3}{3}
\end{aligned}$$

$$\begin{aligned}
F(x_1, x_2, \ldots, x_n) &= \left(f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n), \ldots, f_n(x_1, x_2, \ldots, x_n)\right)\\
&= \left(3x_1 - \cos(x_2x_3) - \tfrac{1}{2}, x_1^2 - 81(x_2+0.1)^2 + \sin x_3 + 1.06, e^{-x_1x_2} + 20x_3 \ + \tfrac{10\pi-3}{3}\right)
\end{aligned}$$

**Example 2**

$$\begin{aligned}
3x_1 - \cos(x_2x_3) \qquad\qquad & - \tfrac{1}{2} &= 0\\
x_1^2 - 81(x_2+0.1)^2 + \sin x_3 & + 1.06 &= 0\\
e^{-x_1x_2} \qquad\qquad + 20x_3 & + \tfrac{10\pi-3}{3} &= 0
\end{aligned}$$

If the *i*th equation is solved for $x_i$, the system can be changed into the fixed point problem

$$\begin{aligned}
x_1 &= \tfrac{1}{3}\cos(x_2x_3) + \tfrac{1}{6},\\
x_2 &= \tfrac{1}{9}\sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1,\\
x_3 &= -\tfrac{1}{20}e^{-x_1x_2} - \tfrac{10\pi-3}{60}.
\end{aligned}$$

Let $G : R^3 \to R^3$ be defined by $G(\mathrm{x}) = \left(g_1(\mathrm{x}), g_2(\mathrm{x}), g_3(\mathrm{x})\right)$ where

$$\begin{aligned}
g_1(x_1, x_2, x_3) &= \tfrac{1}{3}\cos(x_2x_3) + \tfrac{1}{6},\\
g_2(x_1, x_2, x_3) &= \tfrac{1}{9}\sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1,\\
g_3(x_1, x_2, x_3) &= -\tfrac{1}{20}e^{-x_1x_2} - \tfrac{10\pi - 3}{60}.
\end{aligned}$$

$$\begin{aligned}
\left|g_1(x_1, x_2, x_3)\right| &\le \tfrac{1}{3}\left|\cos(x_2x_3)\right| + \tfrac{1}{6} \le \tfrac{1}{2},\\
\left|g_2(x_1, x_2, x_3)\right| &= \left|\tfrac{1}{9}\sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1\right|\\
&\le \tfrac{1}{9}\sqrt{1 + \sin 1 + 1.06} - 0.1 < 0.90,\\
\left|g_3(x_1, x_2, x_3)\right| &= \tfrac{1}{20}e^{-x_1x_2} + \tfrac{10\pi - 3}{60}\\
&\le \tfrac{1}{20}e + \tfrac{10\pi - 3}{60} < 0.61
\end{aligned}$$

so $-1 \le g_i(x_1, x_2, x_3) \le 1,$ for each $i = 1, 2, 3.$ Thus, $G(\mathrm{x}) \in D$ whenever $\mathrm{x} \in D$.

Finding bounds on the partial derivatives on *D* gives the following:

$$\left|\frac{\partial g_1}{\partial x_1}\right| = 0, \quad \left|\frac{\partial g_1}{\partial x_2}\right| \le \tfrac{1}{3}\left|x_3\right|\left|\sin(x_2x_3)\right| \le \tfrac{1}{3}\sin 1 = 0.281, \quad \left|\frac{\partial g_1}{\partial x_3}\right| \le \tfrac{1}{3}\left|x_2\right|\left|\sin(x_2x_3)\right| \le \tfrac{1}{3}\sin 1 = 0.281,$$

$$\left|\frac{\partial g_2}{\partial x_1}\right| = \frac{\left|x_1\right|}{9\sqrt{x_1^2 + \sin x_3 + 1.06}} < \frac{1}{9\sqrt{0.218}} < 0.238, \quad \left|\frac{\partial g_2}{\partial x_2}\right| \le 0,$$

$$\left|\frac{\partial g_2}{\partial x_3}\right| \le \frac{\left|\cos x_3\right|}{18\sqrt{x_1^2 + \sin x_3 + 1.06}} < \frac{1}{18\sqrt{0.218}} < 0.119,$$

$$\left|\frac{\partial g_3}{\partial x_1}\right| = \frac{\left|x_2\right|}{20}e^{-x_1x_2} \le \tfrac{1}{20}e = 0.14, \quad \left|\frac{\partial g_3}{\partial x_2}\right| = \frac{\left|x_1\right|}{20}e^{-x_1x_2}$$

Since the partial derivatives are bounded on *D*, the above Theorem implies that these functions are continuous on *D*. Consequently, *G* is continuous on *D*. Moreover, for every $x \in D$

$\left|\dfrac{\partial g_i(\mathrm{x})}{\partial x_j}\right| \le 0.281$ for each $i = 1, 2, 3$ and $j = 1, 2, 3,$ and the condition in the second part of Theorem 9.7 holds for $K = 0.843$. It can be shown that $\partial g_i(x)/\partial x_j$ for each *i* = 1, 2, 3 and *j* = 1, 2, 3 is continuous on *D*. Consequently, *G* has a unique fixed point on *D* and the nonlinear system has a solution in *D*.

***Example 3***

$$\begin{aligned} 3x_1 - \cos(x_2 x_3) \qquad\qquad & -\tfrac{1}{2} &= 0 \\ x_1^2 - 81(x_2 + 0.1)^2 + \sin x_3 + 1.06 & &= 0 \\ e^{-x_1 x_2} \qquad\qquad + 20x_3 & + \tfrac{10\pi - 3}{3} &= 0 \end{aligned}$$

**Solution**

If the i^th^ equation is solved for $x_i$, the system can be changed into the fixed point problem as

$$\begin{aligned} x_1 &= \tfrac{1}{3}\cos(x_2 x_3) + \tfrac{1}{6}, \\ x_2 &= \tfrac{1}{9}\sqrt{x_1^2 + \sin x_3 + 1.06} - 0.1, \\ x_3 &= -\tfrac{1}{20} e^{-x_1 x_2} - \tfrac{10\pi - 3}{60}. \end{aligned}$$

and write the iterative process as

$$\begin{aligned} x_1^{(k)} &= \tfrac{1}{3}\cos\left(x_2^{(k-1)} x_3^{(k-1)}\right) + \tfrac{1}{6}, \\ x_2^{(k)} &= \tfrac{1}{9}\sqrt{\left(x_1^{(k-1)}\right)^2 + \sin x_3^{(k-1)} + 1.06} - 0.1, \\ x_3^{(k)} &= -\tfrac{1}{20} e^{-x_1^{(k-1)} x_2^{(k-1)}} - \tfrac{10\pi - 3}{60}. \end{aligned}$$

$$\left\| \mathrm{x}^{(k)} - \mathrm{x}^{(k-1)} \right\|_\infty < 10^{-5}$$

| $k$ | $x_1^{(k)}$ | $x_1^{(k)}$ | $x_1^{(k)}$ | $\left\| \mathrm{x}^{(k)} - \mathrm{x}^{(k-1)} \right\|_\infty$ |
|---|---|---|---|---|
| 0 | 0.10000000 | 0.10000000 | –0.10000000 | -- |
| 1 | 0.49998333 | 0.00944115 | –0.52310127 | 0.423 |
| 2 | 0.49999593 | 0.00002557 | –0.52336331 | $9.4\times10^{-3}$ |
| 3 | 0.50000000 | 0.00001234 | –0.52359814 | $2.3\times10^{-4}$ |
| 4 | 0.50000000 | 0.00000003 | –0.52359847 | $1.2\times10^{-5}$ |
| 5 | 0.50000000 | 0.00000002 | –0.52359877 | $3.1\times10^{-7}$ |

$$\left\| \mathrm{x}^{(5)} - \mathrm{p} \right\|_\infty \le \frac{(0.843)^5}{1-0.843} 0.423 < 1.15$$

$$\mathrm{x}^{(3)} = \left(0.50000000, 1.234\times10^{-5}, -0.52359814\right)$$

$$\left\| \mathrm{x}^{(5)} - \mathrm{p} \right\|_\infty \le \frac{(0.843)^5}{1-0.843} (1.20\times10^{-5}) < 5.5\times10^{-5}$$

$$\mathrm{p} = \left(0.5, 0, -\tfrac{\pi}{6}\right) \approx \left(0.5, 0, -0.5235987757\right)$$

$$\left\| \mathrm{x}^{(5)} - \mathrm{p} \right\|_\infty \le 2\times10^{-8}$$

$$\begin{aligned}
x_1^{(k)} &= \tfrac{1}{3}\cos\left(x_2^{(k-1)}x_3^{(k-1)}\right) + \tfrac{1}{6},\\
x_2^{(k)} &= \tfrac{1}{9}\sqrt{\left(x_1^{(k)}\right)^2 + \sin x_3^{(k-1)} + 1.06} - 0.1,\\
x_3^{(k)} &= -\tfrac{1}{20}e^{-x_1^{(k)}x_2^{(k)}} - \tfrac{10\pi-3}{60}.
\end{aligned}$$

| $k$ | $x_1^{(k)}$ | $x_1^{(k)}$ | $x_1^{(k)}$ | $\left\| \mathrm{x}^{(k)} - \mathrm{x}^{(k-1)} \right\|_\infty$ |
|---|---|---|---|---|
| 0 | 0.10000000 | 0.10000000 | –0.10000000 | -- |
| 1 | 0.49998333 | 0.02222979 | –0.52304613 | 0.423 |
| 2 | 0.49997747 | 0.00002815 | –0.52359807 | $2.2\times10^{-2}$ |
| 3 | 0.50000000 | 0.00000004 | –0.52359877 | $2.8\times10^{-5}$ |
| 4 | 0.50000000 | 0.00000000 | –0.52359877 | $3.8\times10^{-8}$ |

## 1-3.2 Newton's Method

In order to construct an algorithm that led to an appropriate fixed-point method

$$A(\mathrm{x}) = \begin{pmatrix} a_{11}(\mathrm{x}) & a_{12}(\mathrm{x}) & \cdots & a_{1n}(\mathrm{x}) \\ a_{21}(\mathrm{x}) & a_{22}(\mathrm{x}) & \cdots & a_{2n}(\mathrm{x}) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(\mathrm{x}) & a_{n2}(\mathrm{x}) & \cdots & a_{nn}(\mathrm{x}) \end{pmatrix}$$

where each of the entries $a_{ij}(\mathrm{x})$ is a function from $R^n \to R$. The procedure requires that $A(\mathrm{x})$ be found so that $G(\mathrm{x}) = \mathrm{x} - A(\mathrm{x})^{-1}F(\mathrm{x})$ gives quadratic convergence to the solution $F(\mathrm{x}) = 0$, provided that $A(\mathrm{x})$ is non-singular at the fixed point.

**Theorem**

Suppose $\mathrm{p}$ is a solution of $G(\mathrm{x}) = \mathrm{x}$ for some function $G = (g_1, g_2, \ldots, g_n)$, mapping $R^n$ into $R^n$. If a number $\delta > 0$ exists with the property that

i) $\partial g_i / \partial x_j$ is continuous on $N_\delta = \{\mathrm{x} : \|\mathrm{x} - \mathrm{p}\| < \delta\}$ for each $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$,

ii) $\partial^2 g_i(\mathrm{x}) / \partial x_j \, \partial x_k$ is continuous, and $\|\partial^2 g_i(\mathrm{x}) / \partial x_j \, \partial x_k\| \le M$ for some constant $M$ whenever $\mathrm{x} \in N_\delta$ for each $i = 1, 2, \ldots, n,$ $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, n$,

iii) $\partial g_i(\mathrm{p}) / \partial x_j = 0$ for each $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$,

then the sequence generated by $\mathrm{x}^{(k)} = G\left(\mathrm{x}^{(k-1)}\right)$ converges quadratically to $\mathrm{p}$ for any choice of $\mathrm{x}^{(0)} \in N_\delta$ and $\left\|\mathrm{x}^{(k)} - \mathrm{p}\right\|_\infty \le \dfrac{n^2 M}{2} \left\|\mathrm{x}^{(k-1)} - \mathrm{p}\right\|_\infty^2$ for each $k \ge 1$

Since $G(\mathrm{x}) = \mathrm{x} - A(\mathrm{x})^{-1}F(\mathrm{x})$, $g_i(\mathrm{x}) = x_i - \sum_{j=1}^{n} b_{ij}(\mathrm{x}) f_j(\mathrm{x})$;

So
$$\frac{\partial g_i(\mathrm{x})}{\partial x_k} = \begin{cases} 1 - \sum_{j=1}^{n} \left( b_{ij}(\mathrm{x}) \dfrac{\partial f_j}{\partial x_k}(\mathrm{x}) + \dfrac{\partial b_{ij}}{\partial x_k}(\mathrm{x}) f_j(\mathrm{x}) \right), & \text{if } i = k, \\ -\sum_{j=1}^{n} \left( b_{ij}(\mathrm{x}) \dfrac{\partial f_j}{\partial x_k}(\mathrm{x}) + \dfrac{\partial b_{ij}}{\partial x_k}(\mathrm{x}) f_j(\mathrm{x}) \right), & \text{if } i \ne k. \end{cases}$$

$$0 = 1 - \sum_{j=1}^{n} b_{ij}(\mathrm{p}) \frac{\partial f_j}{\partial x_k}(\mathrm{p})$$

so
$$\sum_{j=1}^{n} b_{ij}(\mathrm{p}) \frac{\partial f_j}{\partial x_k}(\mathrm{p}) = 1$$

$$0 = -\sum_{j=1}^{n} b_{ij}(\mathrm{p}) \frac{\partial f_j}{\partial x_k}(\mathrm{p})$$

So
$$\sum_{j=1}^{n} b_{ij}(\mathrm{p}) \frac{\partial f_j}{\partial x_k}(\mathrm{p}) = 0$$

$$J(\mathrm{x}) = \begin{pmatrix} \dfrac{\partial f_1(\mathrm{x})}{\partial x_1} & \dfrac{\partial f_1(\mathrm{x})}{\partial x_2} & \cdots & \dfrac{\partial f_1(\mathrm{x})}{\partial x_n} \\ \dfrac{\partial f_2(\mathrm{x})}{\partial x_1} & \dfrac{\partial f_2(\mathrm{x})}{\partial x_2} & \cdots & \dfrac{\partial f_2(\mathrm{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n(\mathrm{x})}{\partial x_1} & \dfrac{\partial f_n(\mathrm{x})}{\partial x_2} & \cdots & \dfrac{\partial f_n(\mathrm{x})}{\partial x_n} \end{pmatrix}$$

$$A(\mathrm{p})^{-1} J(\mathrm{p}) = I$$

so $J(\mathrm{p}) = A(\mathrm{p})$

$$G(\mathrm{x}) = \mathrm{x} - J(\mathrm{x})^{-1} F(\mathrm{x})$$

$$\mathrm{x}^{(k)} = G\left(\mathrm{x}^{(k\text{-}1)}\right) = \mathrm{x}^{(k\text{-}1)} - J\left(\mathrm{x}^{(k\text{-}1)}\right)^{-1} F\left(\mathrm{x}^{(k\text{-}1)}\right)$$

**Example 1**

Solve the nonlinear system

$$\begin{aligned} 3x_1 - \cos(x_2 x_3) \qquad\qquad\quad - \tfrac{1}{2} &= 0 \\ x_1^2 - 81(x_2 + 0.1)^2 + \sin x_3 + 1.06 &= 0 \\ e^{-x_1 x_2} \qquad\qquad + 20x_3 \;+ \tfrac{10\pi - 3}{3} &= 0 \end{aligned}$$

**Solution**

The Jacobian matrix for the system is given by

$$J(x_1, x_2, x_3) = \begin{pmatrix} 3 & x_3 \sin x_2 x_3 & x_2 \sin x_2 x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2 e^{-x_1 x_2} & -x_1 e^{-x_1 x_2} & 20 \end{pmatrix}$$

and

$$\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{pmatrix} = \begin{pmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \end{pmatrix} + \begin{pmatrix} h_1^{(k-1)} \\ h_2^{(k-1)} \\ h_3^{(k-1)} \end{pmatrix}$$

where

$$\begin{pmatrix} h_1^{(k-1)} \\ h_2^{(k-1)} \\ h_3^{(k-1)} \end{pmatrix} = -\left[ J\left( x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)} \right) \right]^{-1} F\left( x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)} \right).$$

Thus, at the $k^{\text{th}}$ step, the linear system

$$\begin{pmatrix} 3 & x_3^{(k-1)} \sin x_2^{(k-1)} x_3^{(k-1)} & x_2 \sin x_2^{(k-1)} x_3^{(k-1)} \\ 2x_1^{(k-1)} & -162(x_2^{(k-1)} + 0.1) & \cos x_3^{(k-1)} \\ -x_2^{(k-1)} e^{-x_1^{(k-1)} x_2^{(k-1)}} & -x_1^{(k-1)} e^{-x_1^{(k-1)} x_2^{(k-1)}} & 20 \end{pmatrix} \begin{pmatrix} h_1^{(k-1)} \\ h_2^{(k-1)} \\ h_3^{(k-1)} \end{pmatrix}$$

$$\text{which is equal to } \begin{pmatrix} 3x_1^{(k-1)} - \cos(x_2^{(k-1)} x_3^{(k-1)}) - \frac{1}{2} \\ \left( x_1^{(k-1)} \right)^2 - 81(x_2^{(k-1)} + 0.1)^2 + \sin x_3^{(k-1)} + 1.06 \\ e^{-x_1^{(k-1)} x_2^{(k-1)}} + 20x_3^{(k-1)} + \frac{10\pi - 3}{3} \end{pmatrix}$$

must be solved. The results obtained using the above iterative procedure is as shown below

| $k$ | $x_1^{(k)}$ | $x_1^{(k)}$ | $x_1^{(k)}$ | $\left\Vert \mathrm{x}^{(k)} - \mathrm{x}^{(k-1)} \right\Vert_\infty$ |
|---|---|---|---|---|
| 0 | 0.10000000 | 0.10000000 | –0.10000000 | -- |
| 1 | 0.50003702 | 0.01946686 | –0.52152047 | 0.422 |
| 2 | 0.50004593 | 0.00158859 | –0.52355711 | $1.79 \times 10^{-2}$ |
| 3 | 0.50000034 | 0.00001244 | –0.52359845 | $1.58 \times 10^{-3}$ |
| 4 | 0.50000000 | 0.00000000 | –0.52359877 | $1.24 \times 10^{-5}$ |
| 5 | 0.50000000 | 0.00000000 | –0.52359877 | 0 |

$$\mathrm{p} = \left( 0.5, 0, -\tfrac{\pi}{6} \right) \approx \left( 0.5, 0, -0.5235987757 \right)$$

***Examples***

1. The nonlinear system $\underline{\mathrm{F}}(x, y) = \begin{bmatrix} x^2 - 10x + y^2 + 8 \\ xy^2 + x - 10y + 8 \end{bmatrix} = 0$ can be transformed into the fixed point problem $\mathrm{G}(x, y) = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{x^2 + y^2 + 8}{10} \\ \frac{xy^2 + x + 8}{10} \end{bmatrix}$

(a) Starting with the initial estimates $x_0 = y_0 = 0$, apply functional iteration to G to approximate the solution to an accuracy of $10^{-5}$.

(b) Does Gauss-Seidel Method accelerate convergence?

**Solution**

(a)

| i | x | y | g1(x,y) | g2(x,y) | Tol |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0.8 | 0.8 | |
| 1 | 0.8 | 0.8 | 0.928 | 0.9312 | 0.8 |
| 2 | 0.928 | 0.9312 | 0.972832 | 0.97327 | 0.1312 |
| 3 | 0.972832 | 0.97327 | 0.989366 | 0.989435 | 0.044832 |
| 4 | 0.989366 | 0.989435 | 0.995783 | 0.995794 | 0.016534 |
| 5 | 0.995783 | 0.995794 | 0.998319 | 0.998321 | 0.006417 |
| 6 | 0.998319 | 0.998321 | 0.999328 | 0.999329 | 0.002536 |
| 7 | 0.999328 | 0.999329 | 0.999732 | 0.999732 | 0.00101 |
| 8 | 0.999732 | 0.999732 | 0.999893 | 0.999893 | 0.000403 |
| 9 | 0.999893 | 0.999893 | 0.999957 | 0.999957 | 0.000161 |
| 10 | 0.999957 | 0.999957 | 0.999983 | 0.999983 | 6.44E-05 |
| 11 | 0.999983 | 0.999983 | 0.999993 | 0.999993 | 2.58E-05 |
| 12 | 0.999993 | 0.999993 | 0.999997 | 0.999997 | 1.03E-05 |
| 13 | 0.999997 | 0.999997 | 0.999999 | 0.999999 | 4.12E-06 |
| 14 | 0.999999 | 0.999999 | 1 | 1 | 1.65E-06 |

(b)

| i | x | y | g1(x,y) | g2(x,y) | Tol |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0.8 | 0.88 | |
| 1 | 0.8 | 0.88 | 0.94144 | 0.967049 | 0.88 |
| 2 | 0.94144 | 0.967049 | 0.982149 | 0.990064 | 0.14144 |
| 3 | 0.982149 | 0.990064 | 0.994484 | 0.99693 | 0.040709 |
| 4 | 0.994484 | 0.99693 | 0.998287 | 0.999045 | 0.012335 |
| 5 | 0.998287 | 0.999045 | 0.999467 | 0.999703 | 0.003803 |
| 6 | 0.999467 | 0.999703 | 0.999834 | 0.999907 | 0.00118 |
| 7 | 0.999834 | 0.999907 | 0.999948 | 0.999971 | 0.000367 |
| 8 | 0.999948 | 0.999971 | 0.999984 | 0.999991 | 0.000114 |
| 9 | 0.999984 | 0.999991 | 0.999995 | 0.999997 | 3.56E-05 |
| 10 | 0.999995 | 0.999997 | 0.999998 | 0.999999 | 1.11E-05 |
| 11 | 0.999998 | 0.999999 | 1 | 1 | 3.46E-06 |

(c)

From (a) and (b) it is seen that Gauss-Seidel Method accelerate convergence.

2. Convert the nonlinear system

$$\begin{aligned} 3x - \cos(yz) - \tfrac{1}{2} &= 0 \\ x^2 - 81\left(y + \tfrac{1}{10}\right)^2 + \sin z + 1.06 &= 0 \\ e^{-xy} + 20z \ + \tfrac{10\pi - 3}{3} &= 0 \end{aligned}$$

to a fixed point problem and use both functional iteration and the Gauss-Seidel variant of functional iteration to approximate the root to within $10^{-5}$ in the $l_\infty$ norm, starting the initial estimate $x_0 = y_0 = 0.1$ and $z_0 = -0.1$. [Note the exact root is $\left(\frac{1}{2}, 0, -\frac{\pi}{6}\right)^T$ ]

**Solution**

$$x = \frac{\cos(yz) + \frac{1}{2}}{3}, \quad y = \sqrt{\frac{x^2 + \sin z + 1.06}{81}} - \frac{1}{10}, \quad z = \frac{-e^{-xy} - \frac{10\pi - 3}{3}}{20}$$

For $k = 1, 2, \dots$, the functional iteration and the Gauss-Seidel variant of functional iteration are given as

$$\begin{aligned} x^{(k)} &= \frac{\cos(y^{(k-1)} z^{(k-1)}) + \frac{1}{2}}{3} \\ y^{(k)} &= \sqrt{\frac{(x^{(k-1)})^2 + \sin z^{(k-1)} + 1.06}{81}} - \frac{1}{10} \\ z^{(k)} &= \frac{-e^{-x^{(k-1)} y(k-1)} - \frac{10\pi - 3}{3}}{20} \end{aligned} \qquad \text{and} \qquad \begin{aligned} x^{(k)} &= \frac{\cos(y^{(k-1)} z^{(k-1)}) + \frac{1}{2}}{3} \\ y^{(k)} &= \sqrt{\frac{(x^{(k)})^2 + \sin z^{(k-1)} + 1.06}{81}} - \frac{1}{10} \\ z^{(k)} &= \frac{-e^{-x^{(k)} y^{(k)}} - \frac{10\pi - 3}{3}}{20} \end{aligned}$$

respectively for the above fixed point problem.

Using the functional iteration we have the following table:

| i | x | y | z | g1(x,y,z) | g2(x,y,z) | g3(x,y,z) | Tol |
|---|---|---|---|---|---|---|---|
| 0 | 0.1 | 0.1 | -0.1 | 0.499983 | 0.009441 | -0.5231013 | |
| 1 | 0.499983 | 0.009441 | -0.523101 | 0.499996 | 2.56E-05 | -0.5233633 | 0.399983 |
| 2 | 0.499996 | 2.56E-05 | -0.523363 | 0.5 | 1.23E-05 | -0.5235981 | 0.009154 |
| 3 | 0.5 | 1.23E-05 | -0.523598 | 0.5 | 3.42E-08 | -0.5235985 | 4.07E-06 |
| 4 | 0.5 | 3.42E-08 | -0.523598 | 0.5 | 1.65E-08 | -0.5235988 | 1.2E-05 |
| 5 | 0.5 | 1.65E-08 | -0.523599 | 0.5 | 4.57E-11 | -0.5235988 | 6.95E-12 |
| 6 | 0.5 | 4.57E-11 | -0.523599 | 0.5 | 2.2E-11 | -0.5235988 | 1.6E-08 |
| 7 | 0.5 | 2.2E-11 | -0.523599 | 0.5 | 6.1E-14 | -0.5235988 | 0 |
| 8 | 0.5 | 6.1E-14 | -0.523599 | 0.5 | 2.94E-14 | -0.5235988 | 2.14E-11 |
| 9 | 0.5 | 2.94E-14 | -0.523599 | 0.5 | 0 | -0.5235988 | 0 |
| 10 | 0.5 | 0 | -0.523599 | 0.5 | 0 | -0.5235988 | 2.87E-14 |

Using the Gauss-Seidel variant of functional iteration we have the following table:

| i | x | y | z | g1(x,y,z) | g2(x,y,z) | g3(x,y,z) | Tol |
|---|---|---|---|---|---|---|---|
| 0 | 0.1 | 0.1 | -0.1 | 0.499983 | 0.02223 | -0.5230461 | |
| 1 | 0.499983 | 0.02223 | -0.523046 | 0.499977 | 2.82E-05 | -0.5235981 | 0.399983 |
| 2 | 0.499977 | 2.82E-05 | -0.523598 | 0.5 | 3.76E-08 | -0.5235988 | 0.02165 |
| 3 | 0.5 | 3.76E-08 | -0.523599 | 0.5 | 5.03E-11 | -0.5235988 | 2.74E-05 |
| 4 | 0.5 | 5.03E-11 | -0.523599 | 0.5 | 6.72E-14 | -0.5235988 | 3.66E-08 |
| 5 | 0.5 | 6.72E-14 | -0.523599 | 0.5 | 0 | -0.5235988 | 4.9E-11 |
| 6 | 0.5 | 0 | -0.523599 | 0.5 | 0 | -0.5235988 | 6.55E-14 |

3. Starting with the initial guess $x_0 = y_0 = 1.0$, use fixed point (functional iteration to approximate the solution to the system $2x^2 + y^2 = 4.32, \quad x^2 - y^2 = 0$ by performing 5 iterations.
4. Consider the nonlinear system

$$2x + xy - 1 = 0$$
$$2y - xy + 1 = 0$$

   which has a unique root $x = (1,-1)^T$. Starting with the initial estimate $x_0 = y_0 = 0$, compare the methods of functional iteration, Gauss-Seidel Newton when approximating the root of this system (perform 5 iterations in each case).

5. Use Newton's Method to approximate the solution of the nonlinear system

$$x^2 - 2x - y + \tfrac{1}{2} = 0$$
$$x^2 + 4y^2 - 4 = 0$$

   starting with the initial estimate $(x_0, y_0) = (2, \tfrac{1}{4})$ and computing 3 iterations.

**Solution**

Using the Newton's iterative method $x^{(k)} = x^{(k-1)} - J^{-1}(x^{(k-1)})F(x^{(k-1)}), \quad \text{for } k = 1, \ldots$ with $x^{(0)} = (x_0, y_0) = (2, \tfrac{1}{4})$, we have

$$x^{(1)} = x^{(0)} - J^{-1}(x^{(0)})F(x^{(0)}) = \begin{pmatrix} 2 \\ 0.25 \end{pmatrix} - \begin{pmatrix} 2 & -1 \\ 4 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix}$$

$$= \begin{pmatrix} 2 \\ 0.25 \end{pmatrix} - \frac{1}{8} \begin{pmatrix} 2 & 1 \\ -4 & 2 \end{pmatrix} \begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 2 \\ 0.25 \end{pmatrix} - \frac{1}{8} \begin{pmatrix} 0.75 \\ -0.50 \end{pmatrix} = \begin{pmatrix} 1.90625 \\ 0.3125 \end{pmatrix}$$

$$x^{(2)} = x^{(1)} - J^{-1}(x^{(1)})F(x^{(1)}) = \begin{pmatrix} 1.90625 \\ 0.3125 \end{pmatrix} - \begin{pmatrix} 1.8125 & -1 \\ 3.8125 & 2.5 \end{pmatrix}^{-1} \begin{pmatrix} 0.00879 \\ -1.70312 \end{pmatrix}$$

$$= \begin{pmatrix} 1.90625 \\ 0.3125 \end{pmatrix} - \frac{1}{8.34375} \begin{pmatrix} 2.5 & 1 \\ -3.8125 & 1.8125 \end{pmatrix} \begin{pmatrix} 0.00879 \\ -1.70312 \end{pmatrix} = \begin{pmatrix} 2.10773 \\ 0.68232 \end{pmatrix}$$

6. Use Newton's Method to approximate the two solutions of the nonlinear system $ye^x = 2, \quad x^2 + y^2 = 4$ by computing 2 iterations for each of the given initial estimates

   (a) $(x_0, y_0) = (-0.6, +3.7)$

   (b) $(x_0, y_0) = (+1.9, +0.4)$

7. Use Newton's Method to approximate the solution of the nonlinear system

$$\begin{aligned} x^2 + y^2 + 0.6y - 0.16 &= 0 \\ x^2 - y^2 + x - 1.6y &= 0 \end{aligned}$$

   by computing 3 iterations with the initial estimate of $(x_0, y_0) = (0.6, 0.25)$.

   Using the more accurate initial estimate of $(x_0, y_0) = (0.3, 0.1)$, repeat the process using the modified Newton's method whereby the Jacobian is evaluated and held constant for subsequent iterations. Compare the two results.

8. Use the modified Newton's method (i.e., by evaluating the Jacobian and keeping at a constant value throughout) to find the root of the system $e^x + y = 0, \;\; \cosh(y) - x = \frac{7}{2}$ starting with the initial estimate $x = -2.4, \;\; y = -0.1$ computing two iterations.

# PART II

# 2 Interpolation

## 2.1 Problem Statement and Applications

Consider the following table:

| | |
|---|---|
| $x_0$ | $f_0$ |
| $x_1$ | $f_1$ |
| $x_2$ | $f_2$ |
| $\vdots$ | $\vdots$ |
| $x_k$ | $f_k$ |
| $\vdots$ | $\vdots$ |
| $x_n$ | $f_n$ |

In the above table, $f_k, k = 0, \cdots, n$ are assumed to be the values of a certain function $f(x)$, evaluated at $x_k, k = 0, \cdots, n$ in an interval containing these points. **Note that only the functional values are known, not the function $f(x)$ itself.** The problem is to find $f_u$ corresponding to a nontabulated intermediate value $x = u$.

Such a problem is called an **Interpolation Problem**. The numbers $x_0,\ x_1, \cdots, x_n$ are called the **nodes**.

**Interpolation Problem**

Given $(n+1)$ points: $(x_0, f_0), (x_1, f_1), \cdots (x_n, f_n)$, find $f_u$ corresponding to $x_u$, where $x_0 < x_u < x_n$; assuming that $f_0, f_1, \cdots, f_n$ are the values of a certain function $f(x)$ at $x = x_0, x_1, \cdots, x_n$, respectively.

The Interpolation problem is also a classical problem and dates back to the time of **Newton** and **Kepler**, who needed to solve such a problem in analyzing data on the positions of stars and planets. It is also of interest in numerous other practical applications. Here is an example.

## 2.2 Existence and Uniqueness

It is well-known that a continuous function $f(x)$ on $[a, b]$ can be approximated as close as possible by means of a polynomial. Specifically, for each $\epsilon > 0$, there exists a polynomial $P(x)$ such that $|f(x) - P(x)| < \epsilon$ for all $x$ in $[a, b]$. This is a classical result, known as **Weierstrass Approximation Theorem**.

Knowing that $f_k, k = 0, \cdots, n$ are the values of a certain function at $x_k$, the most obvious thing then to do is to construct a polynomial $P_n(x)$ of degree at most $n$ that passes through the $(n + 1)$ points: $(x_0, f_0),\ (x_1, f_1),\ \cdots, (x_n, f_n)$.

**Indeed, if the nodes $x_0, x_1, ..., x_n$ are assumed to be distinct, then such a polynomial always does exist and is unique, as can be seen from the following.**

Let $P_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ be a polynomial of degree at most $n$. If $P_n(x)$ interpolates at $x_0, x_1, \cdots, x_n$, we must have, by definition

$$\begin{aligned}
&P_n(x_0) = f_0 = a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n \\
&P_n(x_1) = f_1 = a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n \\
&\vdots \\
&P_n(x_n) = f_n = a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_2^n
\end{aligned} \tag{2.1}$$

These equations can be written in matrix form:

$$\begin{pmatrix} 1 & x_0 & x_0^2 \cdots x_0^n \\ 1 & x_1 & x_1^2 \cdots x_1^n \\ \vdots & & \\ 1 & x_n & x_n^2 \cdots x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix}$$

Because $x_0,\ x_1, \cdots, x_n$ are distinct, it can be shown [**Exercise**] that the matrix of the above system is nonsingular. Thus, the linear system for the unknowns $a_0,\ a_1, \cdots, a_n$ has a unique solution, in view of the following well-known result, available in any linear algebra text book.

The $n \times n$ algebraic linear system $Ax = b$ has a unique solution for every $b$ if and only if $A$ is nonsingular.

This means that $P_n(x)$ exists and is unique.

**Theorem 2.1** ***(Existence and Uniqueness Theorem for Polynomial Interpolation)*** *Given $(n+1)$ distinct points $x_0,\ x_1,\cdots,x_n$ and the associated values $f_0,\ f_1,\cdots,f_n$ of a function $f(x)$ at these points (that is, $f(x_i)=f_i,\ i=0,1,\cdots,n$)), there is a* **unique polynomial** *$P_n(x)$ of degree at most $n$ such that $P_n(x_i)=f_i, i=0,1,\cdots,n$.*

The polynomial $P_n(x)$ in Theorem 3.1 is called the **interpolating polynomial**.

## 2.3 The Lagrange Interpolation

Once we know that the interpolating polynomial exists and is unique, the problem then becomes how to construct an interpolating polynomial; that is, how to construct a polynomial $P_n(x)$ of degree at most $n$, such that

$$P_n(x_i)=f_i,\ i=0,1,\cdots,n.$$

It is natural to obtain the polynomial by solving the linear system (3.1) in the previous section. Unfortunately, the matrix of this linear system, known as the **Vandermonde Matrix**, is usually **highly ill-conditioned**, and the **solution of such an ill-conditioned system, even by the use of a stable method, may not be accurate.** There are, however, several other ways to construct such a polynomial, that do not require solution of a Vandermonde system. We describe one such in the following:

Suppose $n=1$, that is, suppose that we have only two points $(x_0,f_0),\ (x_1,f_1)$, then it is easy to see that the linear polynomial

$$P_1(x)=\frac{x-x_1}{(x_0-x_1)}f_0+\frac{(x-x_0)}{(x_1-x_0)}f_1$$

is an interpolating polynomial, because

$$P_1(x_0)=f_0,\ P_1(x_1)=f_1.$$

For convenience, we shall write the polynomial $P_1(x)$ in the form

$$P_1(x) = L_0(x)f_0 + L_1(x)f_1,$$

where, $L_1(x) = \dfrac{x - x_0}{x_1 - x_0}$, and $L_1(x) = \dfrac{x - x_0}{x_1 - x_0}$.

Note that both the polynomials $L_0(x)$ and $L_1(x)$ are polynomials of degree 1.

The concept can be generalized easily for polynomials of higher degrees.

To generate polynomials of higher degrees, let's define the set of polynomials $\{L_k(x)\}$ recursively, as follows:

$$L_k(x) = \frac{(x - x_0)(x - x_1)\cdots(x - x_{k-1})(x - x_{k+1})\cdots(x - x_n)}{(x_k - x_0)(x_k - x_1)\cdots(x_k - x_{k-1})(x_k - x_{k+1})\cdots(x_k - x_n)},\ k = 0, 1, 2, \cdots, n. \tag{2.2}$$

We will now show that the polynomial $P_n(x)$ defined by

$$P_n(x) = L_0(x)f_0 + L_1(x)f_1 + \cdots + L_n(x)f_n \tag{2.3}$$

is an interpolating polynomial.

To see this, note that

$$\begin{aligned}
L_0(x) &= \frac{(x - x_1)\cdots(x - x_n)}{(x_0 - x_1)\cdots(x_0 - x_n)} \\
L_1(x) &= \frac{(x - x_0)(x - x_2)\cdots(x - x_n)}{(x_1 - x_0)(x_1 - x_2)\cdots(x_1 - x_n)} \\
&\vdots \\
L_n(x) &= \frac{(x - x_0)(x - x_1)(x - x_2)\cdots(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)(x_n - x_2)\cdots(x_n - x_{n-1})}
\end{aligned}$$

Also, note that

$$\begin{aligned}
&L_0(x_0) = 1,\ L_0(x_1) = L_0(x_2) = \cdots = L_0(x_n) = 0 \\
&L_1(x_1) = 1,\ L_1(x_0) = L_1(x_2) = \cdots = L_n(x_n) = 0
\end{aligned}$$

In general

$L_k(x_k) = 1$ and $L_k(x_i) = 0,\ i \neq k.$

Thus

$$
\begin{aligned}
&P_n(x_0) = L_0(x_0)f_0 + L_1(x_0)f_1 + \cdots + L_n(x_0)f_n = f_0\\
&P_n(x_1) = L_0(x_1)f_0 + L_1(x_1)f_1 + \cdots + L_n(x_1)f_n = 0 + f_1 + \cdots + 0 = f_1\\
&\vdots\\
&P_n(x_n) = L_0(x_n)f_0 + L_1(x_n)f_1 + \cdots + L_n(x_n)f_n = 0 + 0 + \cdots + 0 + f_n = f_n
\end{aligned}
$$

That is, the polynomial $P_n(x)$ has the property that $P_n(x_k) = f_k,\ k = 0, 1, \cdots, n$.

The polynomial $P_n(x)$ defined by (3.3) is known as the **Lagrange Interpolating Polynomial**.

**Example 2.1** *Interpolate $f(x)$ from the following table:*

| | |
|---|---|
| 0 | 7 |
| 1 | 13 |
| 2 | 21 |
| 4 | 43 |

*and find and approximation to $f(3)$.*

$$
\begin{aligned}
L_0(x) &= \frac{(x-1)(x-2)(x-4)}{(-1)(-2)(-4)}\\
L_1(x) &= \frac{(x-0)(x-2)(x-4)}{1\cdot(-1)(-3)}\\
L_2(x) &= \frac{(x-0)(x-1)(x-4)}{2\cdot 1\cdot(-2)}\\
L_3(x) &= \frac{(x-0)(x-1)(x-2)}{4\cdot 3\cdot 2}
\end{aligned}
$$

$P_3(x) = 7L_0(x) + 13L_1(x) + 21L_2(x) + 43L_3(x)$

*Thus,* $P_3(3) = 7L_0(3) + 13L_1(3) + 21L_2(3) + 43L_3(3) = 31.$

*Now,* $L_0(3) = \frac{1}{4},\ L_1(3) = -1,\ L_2(3) = \frac{3}{2},\ L_3(3) = \frac{1}{4}.$

*So,* $P_3(3) = 7L_0(3) + 13L_1(3) + 21L_2(3) + 43L_3(3) = 31.$

**Verify:** Note that $f(x)$ in this case is $f(x) = x^2 + 5x + 7$, and the exact value of $f(x)$ at $x = 3$ is 31.

**Example 2.2** *Given*

| $i$ | $x_i$ | $f(x_i)$ |
|---|---|---|
| 0 | 2 | $\frac{1}{2}$ |
| 1 | 2.5 | $\frac{1}{2.5}$ |
| 2 | 4 | $\frac{1}{4}$ |

*We want to find $f(3)$.*

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-2.5)(x-4)}{(-0.5)(-2)} = (x-2.5)(x-4)$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-2)(x-4)}{(0.5)(-1.5)} = -\frac{1}{0.75}(x-2)(x-4)$$

$$L_2(x) = \frac{(x-x_1)(x-x_0)}{(x_2-x_1)(x_2-x_0)} = \frac{1}{3}(x-2.5)(x-2)$$

$$P_2(x) = f(x_0)L_0(x) + f(x_1)L_1(x) + f(x_2)L_2(x) = \frac{1}{2}L_0(x) + \frac{1}{2.5}L_1(x) + \frac{1}{4}L_2(x)$$

$$P_2(3) = \frac{1}{2}L_0(3) + \frac{1}{2.5}L_1(3) + \frac{1}{4}L_2(3) = \frac{1}{2}(-0.5) + \frac{1}{2.5}\left(\frac{1}{0.75}\right) + \frac{1}{4}\left(\frac{.5}{3}\right) = 0.3250$$

**Verify:** (The value of $f(x)$ at $x = 3$ is $\frac{1}{3} = 0.3333$).

## 2.4 Error in Interpolation

If $f(x)$ is approximated by an interpolating polynomial $P_n(x)$, we would like to obtain an expression for the error of interpolation at a give intermediate point, say, $\bar{x}$.

That is, we would like to calculate $E(\bar{x}) = f(\bar{x}) - P_n(\bar{x})$.

Note that, since $P_n(x_i) = f(x_i),\ E(x_i) = 0,\ i = 0, 1, 2, \cdots, n)$, that is, **there are no errors of interpolating at a tabulated point.**

Here is a result for the expression of $E(\bar{x})$.

**Theorem 2.2** ***(Interpolation-Error Theorem)*** *Let $P_n(x)$ be the interpolating polynomial that interpolates at $(n+1)$ distinct numbers in $[a, b]$, and let $f(x)$ be $(n+1)$ times continuously differentiable on $[a, b]$. Then for every $\bar{x}$ in $[a, b]$, there exists a number $\xi = \xi(\bar{x})$ (depending on $\bar{x}$) such that*

$$E_n(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \frac{f^{(n+1)}(\xi(\bar{x}))}{(n+1)!} \prod_{i=0}^{n} (\bar{x} - x_i). \tag{2.4}$$

**Proof:** If $\bar{x}$ is one of the numbers $x_0, x_1, \cdots, x_n$: then the result follows trivially. Because, the error in this case is zero, and the result will hold for any arbitrary $\xi$.

Next, assume that $\bar{x}$ is not one of the numbers $x_0, x_1, \cdots, x_n$.

Define a function $g(t)$ in variable $t$ in $[a, b]$:

$$g(t) = f(t) - P_n(t) - [f(\bar{x}) - P_n(\bar{x})] * \left[ \frac{(t - x_0)(t - x_1) \cdots (t - x_n)}{(\bar{x} - x_0)(\bar{x} - x_1) \cdots (\bar{x} - x_n)} \right]. \tag{2.5}$$

Then, noting that $f(x_k) = P_n(x)$, for $k = 0, 1, 2, \cdots, n$, we have

$$\begin{aligned} g(x_k) \quad &= f(x_k) - P_n(x_k) - [f(\bar{x}) - P_n(\bar{x})] \left[ \frac{(x_k - x_0) \cdots (x_k - x_n)}{(\bar{x} - x_0) \cdots (\bar{x} - x_n)} \right] \\ &= P_n(x_k) - P_n(x_k) - [f(\bar{x}) - P_n(\bar{x})] \ \times 0 \\ &= 0 \end{aligned} \tag{2.6}$$

(Note that the numerator of the fraction appearing above contains the term $(x_k - x_k) = 0$).

Furthermore,

$$\begin{aligned} g(\bar{x}) \quad &= f(\bar{x}) - P_n(\bar{x}) - [f(\bar{x}) - P_n(\bar{x})] * \left[ \frac{(\bar{x} - x_0) \cdots (\bar{x} - x_n)}{(\bar{x} - x_0) \cdots (\bar{x} - x_n)} \right] \\ &= f(\bar{x}) - P_n(\bar{x}) - f(\bar{x}) + P_n(\bar{x}) \\ &= 0 \end{aligned} \tag{2.7}$$

Thus, $g(t)$ becomes identically zero at $(n + 2)$ distinct points: $x_0, x_1, \cdots, x_n$, and $\bar{x}$. Furthermore, $g(t)$ is $(n + 1)$ times continuously differentiable, since $f(x)$ is so.

Therefore, by **generalized Rolle's theorem**, [ ], there exists a number $\xi(\bar{x})$ in $(a, b)$ such that $g^{(n+1)}(\xi) = 0$.

Let's compute $g^{(n+1)}(t)$ now. ¿From (2.5) we have

$$g^{(n+1)}(t) = f^{(n+1)}(t) - P_n^{(n+1)}(t) - [f(\bar{x}) - P_n(\bar{x})] \frac{d^{n+1}}{d^{n+1}} \left[ \frac{(t - x_0)(t - x_1) \cdots (t - x_n)}{((\bar{x} - x_0)(\bar{x} - x_1) \cdots (\bar{x} x_n)} \right]$$

Then

$$\begin{aligned} \frac{d^{n+1}}{dt^{n+1}} \quad & \left[ \frac{(t - x_0)(t - x_1) \cdots (t - x_n)}{(\bar{x} - x_0)(\bar{x} - x_1) \cdots (\bar{x} x_n)} \right] \\ &= \frac{1}{(\bar{x} - x_0)(\bar{x} - x_1) \cdots (\bar{x} - x_n)} \cdot \frac{d^{n+1}}{dt^{n+1}} [(t - x_0)(t - x_1) \cdots (t - x_n)] \\ &= \frac{1}{(\bar{x} - x_0)(\bar{x} - x_1) \cdots (\bar{x} - x_n)} (n + 1)! \end{aligned}$$

(note that the expression within [ ] is a polynomial of degree $n+1$).

Also, $P_n^{(n+1)}(t) = 0$, because $P_n$ is a polynomial of degree at most $n$. Thus, $P_n^{(n+1)}(\xi) = 0$.

So,

$$g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{(f(\bar{x}) - P_n(\bar{x}))}{(\bar{x}-x_0)\cdots(\bar{x}-x_n)}(n+1)! \tag{2.8}$$

Since $g^{(n+1)}(\xi) = 0$, from (2.8), we have $E_n(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}(\bar{x}-x_0)\cdots(\bar{x}-x_n)$.

**Remark:** To obtain the error of interpolation using the above theorem, we need to know the $(n+1)$th derivative of the $f(x)$ or its absolute maximum value on the interval $[a,b]$. Since in practice this value is hardly known, this error formula is of limited use only.

**Example 2.3** *Let's compute the maximum absolute error for Example 3.2.*

*Here $n = 2$.*

$$\begin{aligned} E_2(\bar{x}) &= f(\bar{x}) - P_2(\bar{x}) \\ &= \frac{f^{(3)}(\xi)}{3!}(\bar{x}-x_0)(\bar{x}-x_1)(\bar{x}-x_2) \end{aligned}$$

*To know the maximum value of $E_2(\bar{x})$, we need to know $f^{(3)}(x)$.*

*Let's compute this now:*

$$f(x) = \frac{1}{x}, \;\; f'(x) = -\frac{1}{x^2}, \;\; f''(x)\frac{2}{x^3}, \;\; f^{(3)}(x) = -\frac{6}{x^4}.$$

*So,* $|f^{(3)}(\xi)| < \dfrac{6}{2^4} = \dfrac{6}{16}$ *for* $0 < x \le 2$.

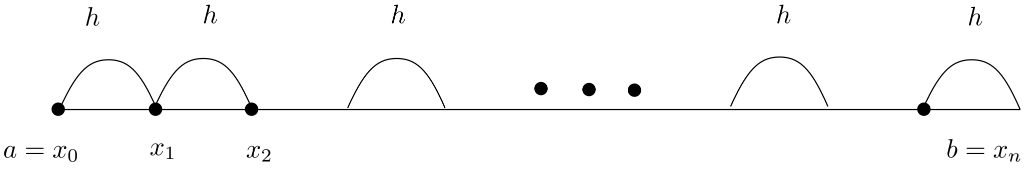
*Since* $\bar{x} = 3, \; x_0 = 2, \; x_1 = 2.5, \; x_2 = 4,$ *we have*

$$|E_2(\bar{x})| \le |\; \frac{6}{16} \times \frac{1}{6}\; (3-2)(3-2.5)(3-4)| = 0.0625.$$

*Note that in four-digit arithmetic, the difference between the value obtained by interpolation and the exact value is* $0.3333 - 0.3250 = 0.0083$.

## 2.5 Simplification of the Error Bound for Equidistant Nodes

The error formula in Theorem 3.2 can be simplified in case the tabulated points (nodes) are equally spaced; because, in this case it is possible to obtain a nice bound for the expression: $(\bar{x}-x_0)(\bar{x}-x_1)\cdots(\bar{x}-x_n)$. Suppose the nodes are equally spaced with spacing $h$; that is $x_{i+1}-x_i = h$.



Then it can be shown **[Exercise]** that

$$|(\bar{x}-x_0)(\bar{x}-x_1)\cdots(\bar{x}-x_n)| \leq \frac{h^{n+1}}{4} n!$$

If we also assume that $|f^{(n+1)}(x)| \leq M$, then we have

$$|E_n(\bar{x})| = |f(\bar{x}) - P_n(\bar{x})| \leq \frac{M}{(n+1)!}\ \frac{h^{n+1}}{4}\ n! = \frac{Mh^{n+1}}{4(n+1)}. \tag{2.9}$$

**Example 2.4** *Suppose a table of values for $f(x) = \cos x$ has to be prepared in $[0, 2\pi]$ with equal spacing nodes of spacing $h$, using* **linear interpolation***, with an error of interpolation of at most $5 \times 10^{-8}$. How small should $h$ be?*

*Here* $n = 1$.

$f(x) = \cos x,\ f'(x) = -\sin x,\ f^2(x) = f''(x) = -\cos x$

$\max |f^{(2)}(x)| = 1,$ *for* $0 \leq x \leq 2\pi$

*Thus* $M = 1$.

So, by (3.9) above we have

$$|E_1(\bar{x})| = |f(\bar{x}) - P_1(\bar{x})| \leq \frac{h^2}{8}.$$

Since the maximum error has to be $5 \times 10^{-7}$, we must have:
$\frac{h^2}{8} \leq 5 \times 10^{-7} = \frac{1}{2} \times 10^{-6}$. That is, $h\ \leq 6.3246 \times 10^{-4}$.

**Example 2.5** *Suppose a table is to be prepared for the function $f(x) = \sqrt{x}$ on $[1, 2]$. Determine the spaceing $h$ in a table such that the interpolation with a polynomial of degree 2 will give accuracy $\epsilon = 5 \times 10^{-8}$.*

*We first compute the maximum absolute error.*

*Since* $f^{(3)}(x) = \frac{3}{8}x\frac{-5}{2}$,

$$M = \left|f^{(3)}(x)\right| \leq \frac{3}{8} \quad \textit{for} \;\; 1 \leq x \leq 2.$$

Thus, taking $n = 2$ in (2.9) the maximum (absolute) error is $\frac{3}{8} \times \frac{h^3}{4x^3} = \frac{1}{32}h^3$.

Thus, to have an accuracy of $\epsilon = 5 \times 10^{-8}$, we must have $\frac{1}{32}h^3 < 5 \times 10^{-8}$ or $h^3 < 160 \times 10^{-8}$.

This means that a spacing $h$ of about $h = \sqrt[3]{160 \times 10^{-8}} = 0.0117$ will be needed in the Table to guarantee the accuracy of $5 \times 10^{-8}$.

## 2.6 Divided Differences and the Newton-Interpolation Formula

A major difficulty with the Lagrange Interpolation is that one is not sure about the degree of interpolating polynomial needed to achieve a certain accuracy. Thus, if the accuracy is not good enough with polynomial of a certain degree, one needs to increase the degree of polynomial, and **computations need to be started all over again.**

Furthermore, computing various Lagrangian polynomials is an expensive procedure. **It is, indeed, desirable to have a formula which makes use of $P_{k-1}(x)$ in computing $P_k(x)$.**

The following form of interpolation, known as **Newton's interpolation** allows us to do so.

The idea is to obtain the interpolating polynomial $P_n(x)$ in the following form:

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \tag{2.10}$$

The constants $a_0$ through $a_n$ can be determined as follows:

For $x = x_0$, $\;P_n(x_0) = a_0 = f_0$

For $x = x_1$, $\;P_n(x_1) = a_0 + a_1(x_1 - x_0) = f_1$, which gives

$$a_1 = \frac{f_1 - a_0}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0}.$$

The other numbers $a_i, i = 2, ..., n$ can similarly be obtained.

It is convenient to introduce the following notation, because we will show how the numbers $a_0, ..., a_n$ can be obtained using these notations.

$$f(x_i) = f[x_i] \text{ and } f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}$$

Similarly,

$$f[x_i, x_{i+1}, \cdots, x_{i+k-1}, x_{i+k}] = \frac{f[x_{i+1}, ..., x_{i+k-1}, x_{i+k}] - f[x_i, x_{i+1}, ..., x_{i+k-1}]}{x_{i+k} - x_i}$$

With these notations, we then have

$$a_0 = f_0 = f(x_0) = f[x_0]$$

$$a_1 = \frac{f_1 - f_0}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = f[x_0, x_1].$$

Continuing this, it can be shown that **[Exercise]**

$$a_k = f[x_0, x_1, \cdots, x_k]. \tag{2.11}$$

The number $f[x_0, x_1, \cdots, x_k]$ is called the $k$-th **divided difference**.

Substituting these expressions of $a_k$ in (2.11), the interpolating polynomial $P_n(x)$ now can be written in terms of the **divided differences:**

$$\begin{aligned} P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots + \\ f[x_0, x_1, \cdots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned} \tag{2.12}$$

**Notes:**

(i) Each divided difference can be obtained from two previous ones of lower orders.

For example, $f[x_0, x_1, x_2]$ can be computed from $f[x_0, x_1]$, and $f[x_1, x_2]$, and so on. Indeed, they can be arranged in form of a table as shown below:

### Table of Divided Differences ($n = 4$)

| $x$ | $f(x)$ | $1^{st}$ Divide Difference | $2^{nd}$ Divided Difference | $3^{rd}$ Divided Difference |
|---|---|---|---|---|
| $x_0$ | $f_0$ | | | |
| $x_1$ | $f_1$ | $f[x_0, x_1]$ | $f[x_0, x_1, x_2]$ | |
| $x_2$ | $f_2$ | $f[x_1, x_2]$ | $f[x_1, x_2, x_3]$ | $f[x_0, x_1, x_2, x_3]$ |
| $x_3$ | $f_3$ | $f[x_2, x_3]$ | $f[x_2, x_3, x_4]$ | $f[x_1, x_2, x_3, x_4]$ |
| $x_4$ | $f_4$ | $f[x_3, x_4]$ | | |

(ii) Note that in computing $P_n(x)$ we need only the diagonal entries of the above table; that is, we need only $f[x_0], f[x_0, x_1], \cdots, f[x_0, x_1, \cdots, x_n]$.

(iii) Since the divided differences are generated recursively, the interpolating polynomials of successively higher degrees can also be generated recursively. **Thus the work done previously can be used gainfully.**

For example,

$$P_1(x) = f[x_0] + f[x_1, x_0](x - x_0)$$

$$\begin{aligned} P_2(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &= P_1(x) + f[x_0, x_1, x_2](x - x_0)(x - x_1). \end{aligned}$$

Similarly, $P_3(x) = P_2(x) + f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2)$

$$P_4(x) = P_3(x) + f[x_0, x_1, x_2, x_3, x_4](x - x_0)(x - x_1)(x - x_2)(x - x_3)$$

and so on.

Thus, in computing $P_2(x)$, $P_1(x)$ has been gainfully used; in computing $P_3(x)$, $P_2(x)$ has been gainfully used, etc.

**Example 2.6** *Interpolate at $x = 2.5$ using the following Table, with polynomials of degree 3 and 4.*

| $n$ | $x$ | $f$ | 1st diff. | 2nd diff. | 3rd diff. | 4th diff | 5th diff |
|---|---|---|---|---|---|---|---|
| 0 | 1.0 | 0 | 0.35218 | $-0.1023$ | .0265533 | $-.006408$ | .001412 |
| 1 | 1.5 | 0.17609 | 0.24988 | $-0.0491933$ | .0105333 | $-.002172$ | |
| 2 | 2.0 | 0.30103 | 0.17609 | $-0.0281267$ | .0051033 | | |
| 3 | 3.0 | 0.47712 | 0.1339 | $-0.01792$ | | | |
| 4 | 3.5 | 0.54407 | 0.11598 | | | | |
| 5 | 4.0 | 0.60206 | | | | | |

¿From (2.12), with $n = 3$, we have

$$\begin{aligned} P_3(2.5) \quad &= 0 + (2.5 - 1.0)(.35218) + (2.5 - 1.5)(2.5 - 1.5)(-.1023) \\ &+(2.5 - 1.0)(2.5 - 1.5)(2.5 - 2.0)(.0265533) \\ &= .52827 - .15345 + .019915 \\ &= \boxed{0.394735} \end{aligned}$$

Similarly, with $n = 4$, we have

$$\begin{aligned} P_4(2.5) \quad &= P_3(2.5) + (2.5 - 1.0)(2.5 - 1.5)(2.5 - 2.0)(2.5 - 3.0)(-.006408) \\ &= .394735 + .002403 = \boxed{0.397138}. \end{aligned}$$

Note that $P_4(2.5) - P_3(2.5) = \boxed{0.002403}$.

**Verification**. The above is a table for $\log(x)$.

The exact value of $\log(2.5)$ (correct up to 5 decimal places) is 0.39794.

**(Note that in computing $P_4(2.5)$, $P_3(2.5)$ computed previously has been gainfully used).**

**Algorithm 2.1** ***Algorithm for Generating Divided Differences***

**Inputs:**

The definite numbers $x_0, x_1, \cdots, x_n$ and the values $f_0, f_1, \cdots, f_n$.

**Outputs:**

The Divided Differenced $D_{00}, D_{11}, \cdots D_{nn}$.

**Step 1:** (Initialization). Set

$d_{i,0} = 0 = f_i,\ i = 0, 1, 2, \cdots, n.$

For $i = 1, 2, \cdots, n$ do

$i = 1, 2, \cdots i$ do

$D_{ij} - \dfrac{D_{i,j-1} - D_{i-1,j-1}}{x_i - x_{i-j}}$

**End**

### A Relationship Between $n$th Divided Difference and the $n^{th}$ Derivative

The following theorem shows how the $n$th derivative of a function $f(x)$ is related to the $n$th divided difference. The proof is omitted. It can be found in any advanced numerical analysis text book (e.g., Atkins on (1978), p. 144).

**Theorem 2.3** *Suppose $f$ is $n$ times continuously differentiable and $x_0, x_1, \cdots, x_n$ are $(n+1)$ distinct numbers in $[a, b]$. Then there exists a number $\xi$ in $(a, b)$ such that*

$f[x_0, x_1, \cdots, x_n] = \dfrac{f^{(n)}(\xi)}{n!}$

### The Newton Interpolation with Equally Spaced Nodes

Suppose again that $x_0, \cdots, x_n$ are equally spaced with spacing $h$;

that is $x_{i+1} - x_i = h,\ i = 0, 1, 2, ..., n-1$. Let $x - x_0 = sh$.

Then $x - x_0 = sh$

$$x - x_1 = x - x_0 + x_0 - x_1 = (x - x_0) - (x_1 - x_0) = sh - h = (s-1)h$$

In general, $x - x_i = (s - i)h$.

So,

$$\begin{aligned} P_n(x) &= P_n(x_0+sh) = f[x_0] + f[x_0,x_1](x-x_0) + \cdots + f[x_0,x_1,\cdots,x_n](x-x_0)\cdots(x-x_{n-1}) \\ &= f[x_0] + sh\, f[x_0,x_1] + s(s-1)h^2 f[x_0,x_1,x_2] + \cdots + s(s-1)\cdots(s-h+1)h^n f[x_0,\cdots,x_h] \\ &\sum_{k=0}^{n} s(s-1)\cdots(s-k+1)h^k f[x_0,\cdots,x_k]. \end{aligned} \tag{2.13}$$

Invoke now the notation:

$$\begin{pmatrix} s \\ k \end{pmatrix} = \frac{(s)(s-1)\cdots(s-k+1)}{k!}$$

We can then write

$$P_n(x) = P_n(x_0+sh) = \sum_{k=0}^{n} \begin{pmatrix} s \\ k \end{pmatrix} h^k k! f[x_0,\cdots,x_n] \tag{2.14}$$

**The Newton Forward-Difference Formula**

Let's introduce the notations:

$\Delta f_i = f(x_{i+1}) - f(x_i)$.

Then, $\Delta f_0 = f(x_1) - f(x_0)$

So,

$$f[x_0,x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0} = \frac{\Delta f_0}{h}. \tag{2.15}$$

Also, note that

$$\begin{aligned} \Delta^2 f_0 &= \Delta(\Delta f_0) = \Delta(f(x_1)-f(x_0)) \\ &= \Delta f(x_1) - \Delta f(x_0) \\ &= f(x_2) - f(x_1) - f(x_1) + f(x_0) \\ &= f(x_2) - 2f(x_1) + f(x_0) \end{aligned} \tag{2.16}$$

So,

$$\begin{aligned} f[x_0,x_1,x_2] &= \frac{f[x_1,x_2]-f[x_0,x_1]}{(x_2-x_0)} \\ &= \frac{\frac{f(x_2)-f(x_1)}{x_2-x_1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{(x_2-x_0)} \\ &= \frac{f(x_2)-2f(x_1)-f(x_0)}{h\times 2h} = \frac{\Delta^2 f_0}{2h^2} \end{aligned} \tag{2.17}$$

In general, we have

$$\begin{aligned} & f[x_0, x_1, \ldots, x_k] \\ & \quad = \frac{1}{k!h^k}\Delta^k f_0. \end{aligned} \tag{2.18}$$

**Proof is by induction on *k*.**

For $k = 0$, the result is trivially true. We have also proved the result for $k = 1$, and $k = 2$.

Assume now that the result is true $k = m$. Then we need to show that the result is also true for $k = m + 1$.

Now, $\quad f[x_0, \cdots, x_{m+1}] = f[x_1, \cdots, x_{m+1}] - \dfrac{f[x_0, \cdots, x_m]}{x_{m+1} - x_0}$

$$= \frac{\left(\frac{\Delta^m f_1}{m!h^m} - \frac{\Delta^m f_0}{m!h^m}\right)}{(m+1)h} = \frac{\Delta^m(f_1 - f_0)}{m!(m+1)h^{m+1}} = \frac{\Delta^{m+1} f_0}{(m+1)!h^{m+1}}$$

The numbers $\Delta^k f_i$ are called $k^{th}$ order forward differences of $f$ at $x = i$.

We now show how the interpolating polynomial $P_n(x)$ given by (2.14) can be written using forward differences.

$$\begin{aligned} P_n(x) = P_n(x_0 + sh) \quad &= f[x_0] + shf[x_0, x_1] + s(s-1)h^2 f[x_0, x_1, x_2] \\ &+ \cdots + s(s-1)\cdots(s-n+1)h^n f[x_0, x_1. \cdots, x_n] \\ &= \sum_{k=0}^{n} s(s-1)\cdots(s-k+1)h^k f[x_0, x_1, \cdots, x_k] \\ &= \sum_{k=0}^{n} \begin{pmatrix} s \\ k \end{pmatrix} k!h^k f[x_0, x_1, \cdots, x_n] = \sum_{k=0}^{n} \binom{s}{k} \Delta^k f_0 \text{ using (2.18).} \end{aligned}$$

**Newton's Forward-Difference Interpolation Formula**

Let $x_0, x_1, ..., x_n$ be $(n+1)$ equidistant points with distance $h$; that is $x_{i+1} - x_i = h$. Then Newton's interpolating polynomial $P_n(x)$ of degree at most $n$, using forward-differences $\Delta^k f_0,\ k = 0, 1, 2, ..., n$ is given by

$$P_n(x) = \sum_{k=0}^{n} \begin{pmatrix} s \\ k \end{pmatrix} \Delta^k f_0,$$

where $s = \dfrac{x - x_0}{h}$.

**Illustration: The Forward Difference Table with $n = 4$**

| $x$ | $f(x)$ | $\Delta f$ | $\Delta^2 f$ | $\Delta^3 f$ | $\Delta^4 f$ |
|---|---|---|---|---|---|
| $x_0$ | $f_0$ | | | | |
| $x_1$ | $f_1$ | $\Delta f_0$ | | | |
| $x_2$ | $f_2$ | $\Delta f_1$ | $\Delta^2 f_0$ | | |
| $x_3$ | $f_3$ | $\Delta f_2$ | $\Delta^2 f_1$ | $\Delta^3 f_0$ | |
| $x_4$ | $f_4$ | $\Delta f_3$ | $\Delta^2 f_2$ | $\Delta^3 f_1$ | $\Delta^3 f_0$ |

**Example 2.7** *Let $f(x) = e^x$.*

| $x$ | $f$ | $\Delta f$ | $\Delta^2 f$ | $\Delta^3 f$ | |
|---|---|---|---|---|---|
| 0 | 1 | | | | |
| 1 | 2.7183 | 1.7183 | | | |
| 2 | 7.3891 | 4.6709 | 2.8817 | 5.2147 | |
| 3 | 20.0855 | 12.6964<br>34.5127 | 8.0964<br>21.8163 | 13.7199 | 8.5052 |
| 4 | 54.5982 | | | | |

*Let $x = 1.5$*

*Then* $s = \dfrac{x - x_0}{h} = \dfrac{1.5 - 0}{1} = 1.5$

$$P_4(1.5) = 1.7183 \times 1 + (1.5)(1.7183) + (1.5)\frac{(1.5-1)(1.5-2)}{2}$$
$$\times 2.8817 + (1.5)\frac{(1.5-1)(1.5-2)(1.5-3)}{6} \times 5.0734 = 4.0852.$$

*The correct answer up to 4 decimal digits is* 4.4817.

**Interpolation using Newton-Backward Differences**

While interpolating at some value of $x$ near the end of the difference table, it is logical to reorder the nodes so that the end-differences can be used in computation. The backward differences allow us to do so.

The backward differences are defined by

$\nabla f_n = f_n - f_{n-1}, n = 1, 2, 3, \ldots,$

and

$\nabla^k f_n = \nabla(\nabla^{k-1} f_n), k = 2, 3, \ldots$

Thus, $\nabla^2 f_n = \nabla(\nabla f_n) = \nabla(f_n - f_{n-1})$

$$= f_n - f_{n-1} - (f_{n-1} - f_{n-2})$$

$$= f_n - 2f_{n-1} + f_{n-2},$$

and so on.

The following a relationship between the backward-differences and the divided differences can be obtained.

$$f[x_{n-k}, \cdots, x_{n-1}, x_n] = \frac{1}{k!h^k} \nabla^k f_n.$$

Using these backward-differences, we can write the Newton interpolation formula as:

$$P_n(x) = f_n + s\nabla f_n + \frac{s(s+1)}{2!}\nabla^2 f_n + ... + \frac{s(s+1)...(s+h-1)}{n!}\nabla^n f_n.$$

Again, using the notation $\binom{-s}{k} = \dfrac{(-s)(-s-1)\cdots(-s-k+1)}{k!}$

we can rewrite the above formula as:

**Newton's Backward-Difference**
**Interpolations Formula**

$$P_n(x) = \sum_{k=0}^{n} (-1)^k \binom{-s}{k} \bigtriangledown^k f_n.$$

# NUMERICAL INTEGRATION

## Introduction

There are two main reasons for the need to do numerical integration: analytical integration may be impossible or infeasible, or you may wish to integrate tabulated data rather than known functions. In this section we outline the main approaches to numerical integration to evaluate integrals of the form $\int_{x_0}^{x_1} f(x)dx$, which is preferable depending in part of the results required, and in part of the function or data to be integrated.

### Learning Objectives

After reading this unit you should be able to:

- describe the Trapezoidal rule of Integration and explain how to use it to solve problems
- describe what the Simpson's 1/3rd Rule of Integration is and explain how to use it to solve problems
- state the Gaussian quadrature rules of Integration and explain how to use them to solve problems.

## Unit content

# Session 1-1 Trapezoidal Rule

## 1-4.1 Single Segment Trapezoidal Rule

Here, we will discuss the trapezoidal rule of approximating integrals of the form

$$I = \int_a^b f(x)dx \qquad (1)$$

where $f(x)$ is called the integrand, $a =$ lower limit of integration and $b =$ upper limit of integration.

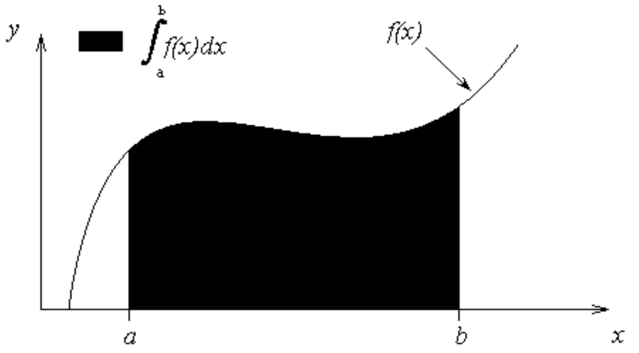


Figure 1: Integration of a function

In using the basic trapezoidal rule to approximate the value of the integral (1), one replaces the integrand $f(x)$ by a first order polynomial, i,e,.

$$I = \int_a^b f(x)dx \approx \int_a^b f_1(x)dx \qquad (2)$$

where $f_1(x) = a_0 + a_1 x$.

Hence

$$\int_a^b f(x)dx \approx \int_a^b f_1(x)dx = \int_a^b (a_0 + a_1 x)dx = a_0(b-a) + a_1\left(\frac{b^2 - a^2}{2}\right). \qquad (3)$$

Now if one chooses, $(a, f(a))$ and $(b, f(b))$ as the two points to approximate $f(x)$ by a straight line from $(a, f(a))$ to $(b, f(b))$, then

$$f(a) = f_1(a) = a_0 + a_1 a \qquad (4)$$

$$f(b) = f_1(b) = a_0 + a_1 b \qquad (5)$$

Solving the above two equations for $a_0$ and $a_1$, we have

$$a_1 = \frac{f(b) - f(a)}{b - a}, \qquad a_0 = \frac{f(a)b - f(b)a}{b - a} \qquad (6)$$

Hence from Equation (3),

$$\begin{aligned} \int_a^b f(x)dx &\approx \frac{f(a)b - f(b)a}{b - a}(b - a) + \frac{f(b) - f(a)}{b - a}\frac{b^2 - a^2}{2} \\ &= (b - a)\left[\frac{f(a) + f(b)}{2}\right] \end{aligned} \qquad (7)$$

The trapezoidal rule for evaluating the integral (1) is given as

$$\boxed{\int_a^b f(x)dx \approx (b - a)\left[\frac{f(a) + f(b)}{2}\right]}$$

The above is called the single segment form of the trapezoidal rule.

The Trapezoidal rule can also be derived from geometry. In Figure 2 the area under the curve $f_1(x)$ between $a$ and $b$ is a trapezium. The integral

$$\begin{aligned} \int_a^b f(x)dx &\approx \text{Area of trapezoid} = \frac{1}{2}(\text{Sum of parallel sides}) \times (\text{height}) \\ &= \frac{1}{2}\left(f(b) + f(a)\right)(b - a) = (b - a)\left[\frac{f(a) + f(b)}{2}\right] \end{aligned} \qquad (8)$$

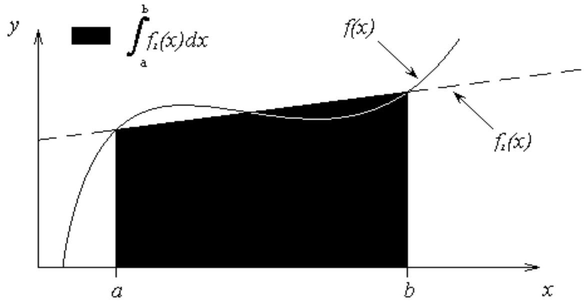


**Figure 2: Geometric Representation of Trapezoidal Rule**

**Example 1:**

Evaluate the integral $I = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$

a) Using the single segment Trapezoidal rule.
b) Find the true error, $E_t$ for part (a).
c) Find the absolute relative true error for part (a).

**Solution**

a) $I \approx (b-a)\left[\frac{f(a)+f(b)}{2}\right]$, where $a=8,\ b=30$

$$f(t) = 2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t$$

$$f(8) = 2000\ln\left[\frac{140000}{140000-2100(8)}\right]-9.8(8) = 177.27$$

$$f(30) = 2000\ln\left[\frac{140000}{140000-2100(30)}\right]-9.8(30) = 901.67$$

$$I = (30-8)\left[\frac{177.27+901.67}{2}\right] = 11868$$

b) The exact value of the above integral is $x = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt = 11061$

so the true error is $E_t = \text{True Value} - \text{Approximate Value} = 11061 - 11868 = -807$

c) The absolute relative true error, $|\in_t|$, would then be

$$|\varepsilon_t| = \left|\frac{\text{True Error}}{\text{True Value}}\right|\times 100 = \left|\frac{11061-11868}{11061}\right|\times 100 = 7.2959\%$$

### 1-4.2 Multiple-segment/Composite Trapezoidal Rule

Dividing the interval $[a,b]$ into $n$ equal segments as shown in Figure 3, we have

$$I = \int_{a}^{b} f(x)dx = \int_{a}^{a+h} f(x)dx + \int_{a+h}^{a+2h} f(x)dx + \cdots + \int_{a+(n-2)h}^{a+(n-1)h} f(x)dx + \int_{a+(n-1)h}^{b} f(x)dx \qquad (9)$$

where $h = (b-a)/n$.

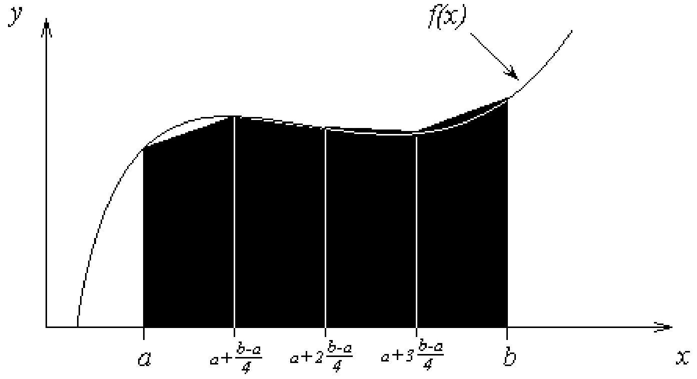


**Figure 3: Multiple (n=4) Segment Trapezoidal Rule**

Applying the single segment trapezoidal rule on Equation (9), we have

$$\boxed{\begin{aligned}\int_a^b f(x)dx &\approx \frac{h}{2}\left[f(a)+2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\}+f(b)\right]\\ &= \frac{b-a}{2n}\left[f(a)+2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\}+f(b)\right]\end{aligned}} \tag{10}$$

which is called the multiple segment or composite trapezoidal rule.

In Example 1, the true error using a single segment trapezoidal rule was large. We can divide the interval [8, 30] into [8, 19] and [19, 30] intervals and apply Trapezoidal rule over each segment.

$$f(t) = 2000\ln\left(\frac{140000}{140000-2100t}\right)-9.8t$$

$$\int_8^{30} f(t)dt = \int_8^{19} f(t)dt + \int_{19}^{30} f(t)dt = (19-8)\left[\frac{f(8)+f(19)}{2}\right]+(30-19)\left[\frac{f(19)+f(30)}{2}\right]$$

$$f(8)=177.27$$

$$f(19) = 2000\ln\left(\frac{140000}{140000-2100(19)}\right)-9.8(19) = 484.75$$

$$f(30)=901.67$$

Hence

$$\int_{8}^{30} f(t)dt = (19-8)\left[\frac{177.27+484.75}{2}\right]+(30-19)\left[\frac{484.75+901.67}{2}\right]=11266$$

The true error, $E_t = 11061 - 11266 = -205$

The true error now is reduced from 807 to 205.

**Example 2:**

Evaluate the integral $I = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$

a) Use two-segment Trapezoidal rule.
b) Find the true error, $E_t$ for part (a).
c) Find the absolute relative true error for part (a).

**Solution**

a) The solution using 2-segment Trapezoidal rule is

$$I = \frac{b-a}{2n}\left[f(a)+2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\}+f(b)\right]$$

where $n=2$, $a=8$, $b=30$, $h=\frac{b-a}{n}=\frac{30-8}{2}=11$

$$I = \frac{30-8}{2(2)}\left[f(8)+2\left\{\sum_{i=1}^{2-1} f(a+ih)\right\}+f(30)\right]=\frac{22}{4}\left[f(8)+2f(19)+f(30)\right]$$

$$=\frac{22}{4}\left[177.27+2(484.75)+901.67\right]=11266$$

b) The exact value of the above integral is

$$I = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt = 11061$$

so the true error is $E_t = \text{True Value} - \text{Approximate Value} = 11061 - 11266 = -205$

c) The absolute relative true error, $\left|\varepsilon_t\right|$, would then be

$$\left|\varepsilon_t\right| = \left|\frac{\text{True Error}}{\text{True Value}}\right| \times 100 = \left|\frac{11061 - 11266}{11061}\right| \times 100 = 1.8534\%$$

**Table 1:** Values obtained using multiple-segment Trapezoidal rule for

$$I = \int_8^{30} \left( 2000 \ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t \right) dt$$

| n | Value | E$_t$ | $\left|\in_t\right|\%$ | $\left|\in_a\right|\%$ |
|---|---|---|---|---|
| 1 | 11868 | -807 | 7.296 | --- |
| 2 | 11266 | -205 | 1.853 | 5.343 |
| 3 | 11153 | -91.4 | 0.8265 | 1.019 |
| 4 | 11113 | -51.5 | 0.4655 | 0.3594 |
| 5 | 11094 | -33.0 | 0.2981 | 0.1669 |
| 6 | 11084 | -22.9 | 0.2070 | 0.09082 |
| 7 | 11078 | -16.8 | 0.1521 | 0.05482 |
| 8 | 11074 | -12.9 | 0.1165 | 0.03560 |

**Example 3:**

Use Multiple-segment Trapezoidal Rule to find the area under the curve $f(x) = \dfrac{300x}{1+e^x}$ from $x = 0$ to $x = 10$.

**Solution**

Using two segments, we get

$$h = \frac{10-0}{2} = 5, \quad f(0) = \frac{300(0)}{1+e^0} = 0, \quad f(5) = \frac{300(5)}{1+e^5} = 10.039, \quad f(10) = \frac{300(10)}{1+e^{10}} = 0.136$$

$$I = \frac{b-a}{2n}\left[ f(a) + 2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\} + f(b)\right] = \frac{10-0}{2(2)}\left[ f(0) + 2\left\{\sum_{i=1}^{2-1} f(0+5)\right\} + f(10)\right]$$

$$= \frac{10}{4}\left[ f(0) + 2f(5) + f(10)\right] = \frac{10}{4}\left[0 + 2(10.039) + 0.136\right] = 50.535$$

So what is the true value of this integral $\int_0^{10} \frac{300x}{1+e^x}\,dx = 246.59$ ?

Making the absolute relative true error

$$\left|\varepsilon_t\right| = \left|\frac{246.59 - 50.535}{246.59}\right| \times 100\% = 79.506\%$$

Why is the true value so far away from the approximate values? Just take a look at Figure 5. As you can see, the area under the "trapezoids" (covers a small the area under the curve. As we add more segments, the approximated value quickly approaches the true value.

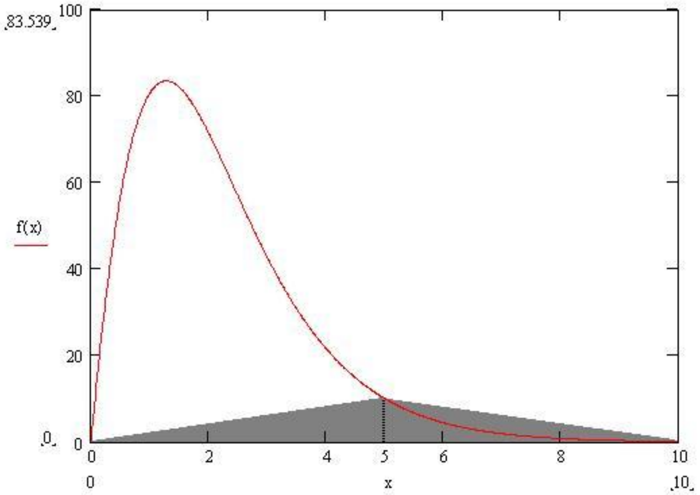


Figure 5: 2-Segment Trapezoidal Rule Approximation

**Table 2:** Values obtained using Multiple-segment Trapezoidal Rule for $\int_0^{10} \frac{300x}{1+e^x} dx$

| n | Approximate Value | $E_t$ | $\lvert \in_t \rvert$ |
|---|---|---|---|
| 1 | 0.681 | 245.91 | 99.724% |
| 2 | 50.535 | 196.05 | 79.505% |
| 4 | 170.61 | 75.978 | 30.812% |
| 8 | 227.04 | 19.546 | 7.927% |
| 16 | 241.70 | 4.887 | 1.982% |
| 32 | 245.37 | 1.222 | 0.495% |
| 64 | 246.28 | 0.305 | 0.124% |

### 1-4.3 Error in Multiple-segment Trapezoidal Rule

The true error for a single segment Trapezoidal rule is given by $E_t = \frac{(b-a)^3}{12} f''(\zeta), \quad a < \zeta < b$

where $\zeta$ is some point in $[a,b]$.

What is the error, then in the multiple-segment Trapezoidal rule? It will be simply the sum of the errors from each segment, where the error in each segment is that of the single segment Trapezoidal rule. The error in each segment is

$$E_1 = \frac{[(a+h)-a]^3}{12} f''(\zeta_1) = \frac{h^3}{12} f''(\zeta_1), \quad a < \zeta_1 < a+h$$

$$E_2 = \frac{[(a+2h)-(a+h)]^3}{12} f''(\zeta_2) = \frac{h^3}{12} f''(\zeta_2), \quad a+h < \zeta_2 < a+2h$$

$$E_i = \frac{[(a+ih)-(a+(i-1)h)]^3}{12} f''(\zeta_i) = \frac{h^3}{12} f''(\zeta_i), \quad a+(i-1)h < \zeta_i < a+ih$$

$$\ldots$$

$$E_{n-1} = \frac{\left[\{a+(n-1)h\}-\{a+(n-2)h\}\right]^3}{12} f''(\zeta_{n-1}) = \frac{h^3}{12} f''(\zeta_{n-1}), \quad a+(n-2)h < \zeta_{n-1} < a+(n-1)h$$

$$E_n = \frac{\left[b-\{a+(n-1)h\}\right]^3}{12} f''(\zeta_n) = \frac{h^3}{12} f''(\zeta_n), \quad a+(n-1)h < \zeta_n < b$$

Hence the total error in multiple-segment Trapezoidal rule is

$$E_t = \sum_{i=1}^{n} E_i = \frac{h^3}{12}\sum_{i=1}^{n} f''(\zeta_i) = \frac{(b-a)^3}{12n^3}\sum_{i=1}^{n} f''(\zeta_i) = \frac{(b-a)^3}{12n^2}\frac{\sum_{i=1}^{n} f''(\zeta_i)}{n}$$

The term $\frac{1}{n}\sum_{i=1}^{n} f''(\zeta_i)$ is an approximate average value of the second derivative $f''(x),\ a < x < b$.

Hence $E_t = \frac{(b-a)^3}{12n^2}\frac{\sum_{i=1}^{n} f''(\zeta_i)}{n}$

Below is the table for the integral $\int_8^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$ as a function of the number of segments. You can visualize that as the number of segments are doubled, the true error gets approximately quartered

**Table 4:** Values obtained using Multiple-segment Trapezoidal Rule for

$$x = \int_8^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$$

| **n** | **Value** | $E_t$ | $\lvert\varepsilon_t\rvert\%$ | $\lvert\varepsilon_a\rvert\%$ |
|---|---|---|---|---|
| 2 | 11266 | -205 | 1.854 | 5.343 |
| 4 | 11113 | -51.5 | 0.4655 | 0.3594 |
| 8 | 11074 | -12.9 | 0.1165 | 0.03560 |
| 16 | 11065 | -3.22 | 0.02913 | 0.00401 |

For example, for 2-segment Trapezoidal rule, the true error is 205, and a quarter of that error is 51.25. That is close to the true error of 51.5 for the 4-segment Trapezoidal rule.

## Session 2-1 Simpson's 1/3rd Rule:

### 2-4.1 Single Segment Simpson's 1/3rd rule

Trapezoidal rule was based on approximating the integrand by a first order polynomial, and then integrating the polynomial in the interval of integration. Simpson's 1/3rd rule is an extension of Trapezoidal rule where the integrand is non-approximated by a second order polynomial.

Hence $I = \int_a^b f(x)dx \approx \int_a^b f_2(x)dx$, where

$f_2(x)$ is a second order polynomial $f_2(x) = a_0 + a_1 x + a_2 x^2$.

Choose $(a, f(a)),\left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right),$ and $(b, f(b))$ as the three points of the function to evaluate $a_0$, $a_1$ and $a_2$.

$$f(a) = f_2(a) = a_0 + a_1 a + a_2 a^2$$

$$f\left(\frac{a+b}{2}\right) = f_2\left(\frac{a+b}{2}\right) = a_0 + a_1\left(\frac{a+b}{2}\right) + a_2\left(\frac{a+b}{2}\right)^2$$

$$f(b) = f_2(b) = a_0 + a_1 b + a_2 b^2$$

Solving the above three equations for unknowns, $a_0$, $a_1$ and $a_2$ give

$$a_0 = \frac{a^2 f(b) + abf(b) - 4abf\left(\frac{a+b}{2}\right) + abf(a) + b^2 f(a)}{a^2 - 2ab + b^2}$$

$$a_1 = -\frac{af(a) - 4af\left(\frac{a+b}{2}\right) + 3af(b) + 3bf(a) - 4bf\left(\frac{a+b}{2}\right) + bf(b)}{a^2 - 2ab + b^2}$$

$$a_2 = \frac{2\left(f(a) - 2f\left(\frac{a+b}{2}\right) + f(b)\right)}{a^2 - 2ab + b^2}$$

Then

$$I \approx \int_a^b f_2(x)dx = \int_a^b \left(a_0 + a_1 x + a_2 x^2\right)dx = \left[a_0 x + a_1 \frac{x^2}{2} + a_2 \frac{x^3}{3}\right]_a^b$$

$$= a_0(b-a) + a_1 \frac{b^2 - a^2}{2} + a_2 \frac{b^3 - a^3}{3}$$

Substituting values of $a_0$, $a_1$ and $a_2$ give

$$\int_a^b f_2(x)dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

Since for Simpson's 1/3rd Rule, the interval $[a,b]$ is broken into 2 segments, the segment width is $h = \dfrac{b-a}{2}$.

Hence the Simpson's 1/3rd rule is given by $\boxed{\int_a^b f(x)dx \cong \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]}$

Since the above form has 1/3 in its formula, it is called **Simpson's 1/3rd Rule**.

**Example 1:**

Evaluate the integral $x = \displaystyle\int_8^{30}\left(2000\ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t\right)dt$

a) Using Simpson's 1/3rd Rule to find the approximate value of $x$.

b) Find the true error, $E_t$

c) Find the absolute relative true error, $|\varepsilon_t|$.

**Solution:**

a) $$x \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

$$a = 8$$

$$b = 30$$

$$\frac{a+b}{2} = 19$$

$$f(t) = 2000\ln\left[\frac{140000}{140000-2100t}\right] - 9.8t$$

$$f(8) = 2000\ln\left[\frac{140000}{140000-2100(8)}\right] - 9.8(8) = 177.27 m/s$$

$$f(30) = 2000\ln\left[\frac{140000}{140000-2100(30)}\right] - 9.8(30) = 901.67 m/s$$

$$f(19) = 2000\ln\left(\frac{140000}{140000-2100(19)}\right) - 9.8(19) = 484.75 m/s$$

$$x = \left(\frac{b-a}{6}\right)\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] = \left(\frac{30-8}{6}\right)\left[f(8) + 4f(19) + f(30)\right]$$
$$= \left(\frac{22}{6}\right)\left[177.2667 + 4(484.7455) + 901.6740\right] = 11065.72$$

b) The exact value of the above integral is

$$x = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right] - 9.8t\right)dt = 11061.34$$

So the true error is

$$E_t = True \quad Value - Approximate \quad Value$$
$$= 11061.34 - 11065.72 = -4.38$$

c) Absolute Relative true error,

$$|\varepsilon_t| = \left|\frac{\text{True Error}}{\text{True Value}}\right| \times 100 = \left|\frac{11061.34 - 11065.72}{11061.34}\right| \times 100\% = 0.0396\%$$

## 2-4.2 Multiple Segment Simpson's 1/3rd Rule

Just like in multiple-segment Trapezoidal Rule, one can subdivide the interval $[a,b]$ into $n$ segments and apply Simpson's 1/3rd Rule repeatedly over every two segments. Note that $n$ needs to be even. Divide interval $[a,b]$ into $n$ equal segments, hence the segment width $h = \frac{b-a}{n}$.

$$\int_a^b f(x)dx = \int_{x_0}^{x_n} f(x)dx$$

where $x_0 = a$, $x_n = b$

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \cdots + \int_{x_{n-4}}^{x_{n-2}} f(x)dx + \int_{x_{n-2}}^{x_n} f(x)dx$$

Apply Simpson's 1/3rd Rule over each interval,

$$\int_a^b f(x)dx \cong (x_2 - x_0)\left[\frac{f(x_0)+4f(x_1)+f(x_2)}{6}\right] + (x_4 - x_2)\left[\frac{f(x_2)+4f(x_3)+f(x_4)}{6}\right] + \cdots$$
$$+ (x_{n-2} - x_{n-4})\left[\frac{f(x_{n-4})+4f(x_{n-3})+f(x_{n-2})}{6}\right] + (x_n - x_{n-2})\left[\frac{f(x_{n-2})+4f(x_{n-1})+f(x_n)}{6}\right]$$

Since $x_i - x_{i-2} = 2h, \ i = 2, \ 4, \ldots, \ n$

then

$$\int_a^b f(x)dx \cong 2h\left[\frac{f(x_0)+4f(x_1)+f(x_2)}{6}\right] + 2h\left[\frac{f(x_2)+4f(x_3)+f(x_4)}{6}\right] + \cdots$$
$$+ 2h\left[\frac{f(x_{n-4})+4f(x_{n-3})+f(x_{n-2})}{6}\right] + 2h\left[\frac{f(x_{n-2})+4f(x_{n-1})+f(x_n)}{6}\right]$$
$$= \frac{h}{3}\left[f(x_0) + 4\left\{f(x_1)+f(x_3)+\ldots+f(x_{n-1})\right\} + 2\left\{f(x_2)+f(x_4)+\ldots+f(x_{n-2})\right\} + f(x_n)\right]$$
$$= \frac{h}{3}\left[f(x_0) + 4\sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2\sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n)\right]$$

Hence the multiple segment Simpson's 1/3rd Rule is given by

$$\boxed{\int_a^b f(x)dx \cong \frac{b-a}{3n}\left[f(x_0) + 4\sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2\sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n)\right]}$$

**Example 2:**

Approximate the value of the integral $x = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$

a) Using the four segment Simpson's 1/3rd Rule.
b) Find the true error, $E_t$ for part (a).
c) Find the absolute relative true error for part (a).

**Solution:**

a) The $n$ segment Simpson's 1/3rd Rule is given by

$$x \approx \frac{b-a}{3n}\left[f(t_0)+4\sum_{\substack{i=1\\ i=odd}}^{n-1} f(t_i)+2\sum_{\substack{i=2\\ i=even}}^{n-2} f(t_i)+f(t_n)\right]$$

In this case $n=4$, $a=8$, $b=30$, $h=\frac{b-a}{n}=\frac{30-8}{4}=5.5$

$$f(t)=2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t$$

So

$$f(t_0)=f(8)$$

$$f(8)=2000\ln\left[\frac{140000}{140000-2100(8)}\right]-9.8(8)=177.27$$

$$f(t_1)=f(8+5.5)=f(13.5)$$

$$f(13.5)=2000\ln\left[\frac{140000}{140000-2100(13.5)}\right]-9.8(13.5)=320.25m/s$$

$$f(t_2)=f(13.5+5.5)=f(19)$$

$$f(19)=2000\ln\left(\frac{140000}{140000-2100(19)}\right)-9.8(19)=484.75m/s$$

$$f(t_3)=f(19+5.5)=f(24.5)$$

$$f(24.5) = 2000\ln\left[\frac{140000}{140000-2100(24.5)}\right] - 9.8(24.5) = 676.05 m/s$$

$$f(t_4) = f(t_n) = f(30)$$

$$f(30) = 2000\ln\left[\frac{140000}{140000-2100(30)}\right] - 9.8(30) = 901.67 m/s$$

$$x = \frac{b-a}{3n}\left[f(t_0) + 4\sum_{\substack{i=1\\ i=odd}}^{n-1} f(t_i) + 2\sum_{\substack{i=2\\ i=even}}^{n-2} f(t_i) + f(t_n)\right]$$

$$= \frac{30-8}{3(4)}\left[f(8) + 4\sum_{\substack{i=1\\ i=odd}}^{3} f(t_i) + 2\sum_{\substack{i=2\\ i=even}}^{2} f(t_i) + f(30)\right]$$

$$= \frac{11}{6}\left[177.27 + 4(320.25) + 4(676.05) + 2(484.75) + 901.67\right] = 11061.64$$

In this case, the true error is $E_t = 11061.34 - 11061.64 = -0.30$ and the absolute relative true error $|\varepsilon_t| = \left|\frac{11061.34-11061.64}{11061.34}\right| \times 100\% = 0.0027\%$

**Table 1: Values of Simpson's 1/3$^{rd}$ Rule for Example 2 with multiple segments**

| $n$ | Approximate Value | $E_t$ | $\lvert\in_t\rvert$ |
|---|---|---|---|
| 2 | 11065.72 | 4.38 | 0.0396% |
| 4 | 11061.64 | 0.30 | 0.0027% |
| 6 | 11061.40 | 0.06 | 0.0005% |
| 8 | 11061.35 | 0.01 | 0.0001% |
| 10 | 11061.34 | 0.00 | 0.0000% |

### 2-4.3 Error in Multiple Segment Simpson's 1/3rd Rule

The true error in a single application of Simpson's 1/3rd Rule is given[1]

$$E_t = -\frac{(b-a)^5}{2880} f^{(4)}(\zeta), \quad a < \zeta < b$$

In Multiple Segment Simpson's 1/3rd Rule, the error is the sum of the errors in each application of Simpson's 1/3rd Rule. The error in $n$ segment Simpson's 1/3rd Rule is given by

$$E_1 = -\frac{(x_2 - x_0)^5}{2880} f^{(4)}(\zeta_1) = -\frac{h^5}{90} f^{(4)}(\zeta_1), \quad x_0 < \zeta_1 < x_2$$

$$E_2 = -\frac{(x_4 - x_2)^5}{2880} f^{(4)}(\zeta_2) = -\frac{h^5}{90} f^{(4)}(\zeta_2), \quad x_2 < \zeta_2 < x_4$$

$$\vdots$$

$$\vdots$$

$$E_i = -\frac{(x_{2i} - x_{2(i-1)})^5}{2880} f^{(4)}(\zeta_i) = -\frac{h^5}{90} f^{(4)}(\zeta_i), \quad x_{2(i-1)} < \zeta_i < x_{2i}$$

$$\vdots$$

$$\vdots$$

$$E_{\frac{n}{2}-1} = -\frac{(x_{n-2} - x_{n-4})^5}{2880} f^{(4)}\left(\zeta_{n/2-1}\right) = -\frac{h^5}{90} f^{(4)}\left(\zeta_{n/2-1}\right), \quad x_{n-4} < \zeta_{n/2-1} < x_{n-2}$$

$$E_{\frac{n}{2}} = -\frac{(x_n - x_{n-2})^5}{2880} f^{4}\left(\zeta_{n/2}\right) = -\frac{h^5}{90} f^{(4)}\left(\zeta_{\frac{n}{2}}\right), \quad x_{n-2} < \zeta_{n/2} < x_n$$

Hence, the total error in Multiple Segment Simpson's 1/3rd Rule is

$$E_t = \sum_{i=1}^{\frac{n}{2}} E_i = -\frac{h^5}{90}\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i) = -\frac{(b-a)^5}{90n^5}\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i) = -\frac{(b-a)^5}{90n^4}\frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{n}$$

[1] The $f^{(4)}$ in the true error expression stands for the fourth derivative of *f*.

The term $\frac{1}{n}\sum_{i=1}^{n/2} f^{(4)}(\zeta_i)$ is an approximate average value of $f^{(4)}(x),\ a < x < b$.

Hence $E_t = -\frac{(b-a)^5}{90n^4}\overline{f}^{(4)}$, where $\overline{f}^{(4)} = \frac{1}{n}\sum_{i=1}^{n/2} f^{(4)}(\zeta_i)$.

## Session3-4 Gaussian Quadrature Rule

The general $n$-point Gaussian quadrature formula for the integral $\int_a^b f(x)dx$ is defined as $\int_a^b f(x)dx \approx \sum_{i=1}^{n} c_i f\left(x_i\right)$, where $c_i,\ x_i,\ i = 1,\ldots,n$ are the solution of the system

$$\int_a^b f_m(x)dx = \sum_{i=1}^{n} c_i f_m\left(x_i\right),\ f_m(x) = x^m,\ m = 0,1,\ldots,2n-1.$$

The 2-point Gaussian quadrature rule can be derived by assuming that the expression gives exact values for integrals of individual integrals of $\int_a^b 1dx,\ \int_a^b xdx,\ \int_a^b x^2dx,$ and $\int_a^b x^3dx$. These will give four equations as follows

$$\int_a^b 1dx = b - a = c_1 + c_2$$

$$\int_a^b xdx = \frac{b^2 - a^2}{2} = c_1x_1 + c_2x_2$$

$$\int_a^b x^2dx = \frac{b^3 - a^3}{3} = c_1{x_1}^2 + c_2{x_2}^2$$

$$\int_a^b x^3dx = \frac{b^4 - a^4}{4} = c_1{x_1}^3 + c_2{x_2}^3 \qquad (14)$$

These four simultaneous nonlinear Equations (14) can be solved with a single acceptable solution $c_1 = \frac{b-a}{2},\ c_2 = \frac{b-a}{2}$

$$x_1 = \left(\frac{b-a}{2}\right)\left(-\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}, \quad x_2 = \left(\frac{b-a}{2}\right)\left(\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2} \tag{15}$$

Hence

$$\int_a^b f(x)dx \approx \frac{b-a}{2} f\left(\frac{b-a}{2}\left(-\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right) + \frac{b-a}{2} f\left(\frac{b-a}{2}\left(\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right) \tag{16}$$

Since two points are chosen, it is called the two-point Gaussian Quadrature Rule. Higher point versions can also be developed.

### 3-4.1 Higher point Gaussian Quadrature Formulas

For example

$$\int_a^b f(x)dx \cong c_1 f(x_1) + c_2 f(x_2) + c_3 f(x_3) \tag{17}$$

is called the three-point Gauss Quadrature Rule. The coefficients $c_1$, $c_2$ and $c_3$, and the function arguments $x_1$, $x_2$ and $x_3$ are calculated by assuming the formula gives exact expressions for integrating a fifth order polynomial $\int_a^b \left(a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5\right)dx$.

General *n*-point rules would approximate the integral

$$\int_a^b f(x)dx \approx c_1 f(x_1) + c_2 f(x_2) + \ldots\ldots\ldots + c_n f(x_n) \tag{18}$$

### 3-4.2 Arguments and weighing factors for n-point Gaussian Quadrature Rules

The general *n*-point Gaussian quadrature formula for the integral $\int_{-1}^{1} g(x)dx$ is defined as $\int_{-1}^{1} g(x)dx \approx \sum_{i=1}^{n} c_i g\left(x_i\right)$, where $c_i,\ x_i,\ i = 1,\ldots,n$ are the solution of the system

$$\int_{-1}^{1} g_m(x)dx = \sum_{i=1}^{n} c_i g_m\left(x_i\right),\ g_m(x) = x^m,\ m = 0,1,\ldots,2n-1.$$

The coefficients and arguments are given for the $n$-point Gaussian Quadrature Rule for $n = 2, \ldots, 6$ in the table below

**Table 1: Weighting factors $c$ and function arguments $x$ used in Gauss Quadrature formulas**

| Points | Weighting Factors | Function Arguments |
|---|---|---|
| 2 | $c_1 = 1.000000000$ | $x_1 = -0.577350269$ |
| | $c_2 = 1.000000000$ | $x_2 = 0.577350269$ |
| 3 | $c_1 = 0.555555556$ | $x_1 = -0.774596669$ |
| | $c_2 = 0.888888889$ | $x_2 = 0.000000000$ |
| | $c_3 = 0.555555556$ | $x_3 = 0.774596669$ |
| 4 | $c_1 = 0.347854845$ | $x_1 = -0.861136312$ |
| | $c_2 = 0.652145155$ | $x_2 = -0.339981044$ |
| | $c_3 = 0.652145155$ | $x_3 = 0.339981044$ |
| | $c_4 = 0.347854845$ | $x_4 = 0.861136312$ |
| 5 | $c_1 = 0.236926885$ | $x_1 = -0.906179846$ |
| | $c_2 = 0.478628670$ | $x_2 = -0.538469310$ |
| | $c_3 = 0.568888889$ | $x_3 = 0.000000000$ |
| | $c_4 = 0.478628670$ | $x_4 = 0.538469310$ |
| | $c_5 = 0.236926885$ | $x_5 = 0.906179846$ |
| 6 | $c_1 = 0.171324492$ | $x_1 = -0.932469514$ |
| | $c_2 = 0.360761573$ | $x_2 = -0.661209386$ |
| | $c_3 = 0.467913935$ | $x_3 = -0.238619186$ |
| | $c_4 = 0.467913935$ | $x_4 = 0.238619186$ |
| | $c_5 = 0.360761573$ | $x_5 = 0.661209386$ |
| | $c_6 = 0.171324492$ | $x_6 = 0.932469514$ |

So if the table is given for $\int_{-1}^{1} g(x)dx$ integrals, how does one solve $\int_{a}^{b} f(x)dx$?

Note: $\int_{a}^{b} f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f\left((\tfrac{b-a}{2})t + (\tfrac{a+b}{2})\right)dt$

The answer lies in that any integral with limits of $[a, b]$ can be converted into an integral with limits $[-1, 1]$. Let

$$x = mt + c \qquad (20)$$

If $x = a,$ then $t = -1.$

If $x = b,$ then $t = \quad 1.$

such that

$$a = m(-1) + c\,,\ b = m(1) + c \qquad (21)$$

Solving these two simultaneous linear Equations (21) gives

$$m = \frac{b-a}{2}\,, \quad c = \frac{b+a}{2} \qquad (22)$$

Hence $x = \frac{b-a}{2}t + \frac{b+a}{2},\ \ dx = \frac{b-a}{2}dt$.

Substituting our values of $x$ and $dx$ into the integral gives us

$$\int_{a}^{b} f(x)dx = \int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)\frac{b-a}{2}dt \qquad (23)$$

**Example 1**

For an integral $\int_{-1}^{1} f(x)dx,$ show that the two-point Gauss Quadrature rule approximates to $\int_{-1}^{1} f(x)dx \cong c_1 f(x_1) + c_2 f(x_2)$, where $c_1 = 1,\ c_2 = 1,\ x_1 = -\frac{1}{\sqrt{3}}$ and $x_2 = \frac{1}{\sqrt{3}}$.

**Solution**

Assuming the formula

$$\int_{-1}^{1} f(x)dx = c_1 f(x_1) + c_2 f(x_2) \tag{E1.1}$$

gives exact values for integrals $\int_{-1}^{1} 1dx,\ \int_{-1}^{1} xdx,\ \int_{-1}^{1} x^2 dx,$ and $\int_{-1}^{1} x^3 dx$ . Then

$$\int_{-1}^{1} 1dx = 2 = c_1 + c_2 \tag{E1.2}$$

$$\int_{-1}^{1} xdx = 0 = c_1 x_1 + c_2 x_2 \tag{E1.3}$$

$$\int_{-1}^{1} x^2 dx = \frac{2}{3} = c_1 {x_1}^2 + c_2 {x_2}^2 \tag{E1.4}$$

$$\int_{-1}^{1} x^3 dx = 0 = c_1 {x_1}^3 + c_2 {x_2}^3 \tag{E1.5}$$

Multiplying Equation (E1.3) by ${x_1}^2$ and subtracting from Equation (E1.5) gives

$$c_2 x_2 \left({x_1}^2 - {x_2}^2\right) = 0 . \tag{E1.6}$$

The solution to the above equation is $c_2 = 0,$ or/and $x_2 = 0,$ or/and $x_1 = x_2,$ or/and $x_1 = -x_2$.

I. $c_2 = 0$ is not acceptable as Equations (E1.2-E1.5) reduce to $c_1 = 2;\ c_1 x_1 = 0;\ c_1 x_1^2 = \frac{2}{3};$ and $c_1 x_1^3 = 0$. But since $c_1 = 2$, then $x_1 = 0$ from $c_1 x_1 = 0$, but $x_1 = 0$ conflicts with $c_1 x_1^2 = \frac{2}{3}$.

II. $x_2 = 0$ is not acceptable as Equations (E1.2-E1.5) reduce to $c_1 + c_2 = 2$; $c_1 x_1 = 0;$ $c_1 x_1^2 = \frac{2}{3};\ c_1 x_1^3 = 0$. Since $c_1 x_1 = 0$, then $c_1$ or $x_1$ has to be zero but this violates $c_1 x_1^2 = \frac{2}{3} \neq 0$.

III. $x_1 = x_2$ is not acceptable as Equations (E1.2-E1.5) reduce to $c_1 + c_2 = 2$; $c_1x_1 + c_2x_1 = 0$; $c_1x_1^2 + c_2{x_1}^2 = \dfrac{2}{3}$; $c_1x_1^3 + c_2x_1^3 = 0$. If $x_1 \neq 0$, then $c_1x_1 + c_2x_1 = 0$ gives $c_1 + c_2 = 0$ and that violates $c_1 + c_2 = 2$. If $x_1 = 0$, then violates $c_1x_1^2 + c_2{x_1}^2 = \dfrac{2}{3} \neq 0$

That leaves the solution of $x_1 = -x_2$ as the only possible acceptable solution and in fact, it does not have violations (see it for yourself)

$$x_1 = -x_2 \tag{E1.7}$$

Substituting (E1.7) in Equation (E1.3) gives

$$c_1 = c_2 \tag{E1.8}$$

From Equations (E1.2) and (E1.8),

$$c_1 = c_2 = 1 \tag{E1.9}$$

Equations (E1.4) and (E1.9) gives

$${x_1}^2 + {x_2}^2 = \frac{2}{3} \tag{E1.10}$$

Since Equation (E1.7) requires that the two results be of opposite sign, we get

$x_1 = -\dfrac{1}{\sqrt{3}}$, $x_2 = \dfrac{1}{\sqrt{3}}$

Hence

$$\int_{-1}^{1} f(x)dx = c_1 f(x_1) + c_2 f(x_2) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \tag{E1.11}$$

**Example 2**

For an integral, $\int_a^b f(x)dx,$ derive the one-point Gaussian Quadrature Rule.

**Solution**

The one-point Gaussian quadrature rule is

$$\int_a^b f(x)dx \approx c_1 f\left(x_1\right) \tag{E2.1}$$

Assuming the formula gives exact values for integrals $\int_{-1}^{1} 1dx,$ and $\int_{-1}^{1} xdx$

$$\begin{aligned} \int_a^b 1dx &= b-a = c_1 \\ \int_a^b xdx &= \frac{b^2-a^2}{2} = c_1 x_1 \end{aligned} \tag{E2.2}$$

Since $c_1 = b-a,$ the other equation becomes

$$\begin{aligned} (b-a)x_1 &= \frac{b^2-a^2}{2} \\ x_1 &= \frac{b+a}{2} \end{aligned} \tag{E2.3}$$

Therefore, one-point Gauss Quadrature Rule can be expressed as

$$\int_a^b f(x)dx \approx (b-a)f\left(\frac{b+a}{2}\right) \tag{E2.4}$$

**Example 3**

What would be the formula for $\int_a^b f(x)dx = c_1 f(a) + c_2 f(b)$, if you want the above formula to give you exact values of $\int_a^b \left(a_0 x + b_0 x^2\right)dx,$ that is a linear combination of $x$ and $x^2$.

**Solution**

If the formula is exact for linear combination of $x$ and $x^2$, then

$$\begin{aligned}\int_a^b x\,dx &= \frac{b^2 - a^2}{2} = c_1 a + c_2 b \\ \int_a^b x^2\,dx &= \frac{b^3 - a^3}{3} = c_1 a^2 + c_2 b^2\end{aligned} \qquad \text{(E3.1)}$$

Solving the two Equations (E3.1) simultaneously gives

$$\begin{bmatrix} a & b \\ a^2 & b^2 \end{bmatrix} \begin{bmatrix} \mathrm{c}_1 \\ \mathrm{c}_2 \end{bmatrix} = \begin{bmatrix} \dfrac{b^2 - a^2}{2} \\ \dfrac{b^3 - a^3}{3} \end{bmatrix}$$

$$c_1 = -\frac{1}{6}\frac{-ab - b^2 + 2a^2}{a}, \quad c_2 = -\frac{1}{6}\frac{a^2 + ab - 2b^2}{b} \qquad \text{(E3.2)}$$

So

$$\int_a^b f(x)dx = -\frac{1}{6}\frac{-ab - b^2 + 2a^2}{a} f(a) - \frac{1}{6}\frac{a^2 + ab - 2b^2}{b} f(b) \qquad \text{(E3.3)}$$

To see if the formula works;

Evaluate $\int_2^5 \left(2x^2 - 3x\right)dx$ using the above formula.

$$\int_2^5 \left(2x^2 - 3x\right)dx \cong c_1 f(a) + c_2 f(b)$$

$$= -\frac{1}{6}\frac{-(2)(5) - 5^2 + 2(2)^2}{2}\left[2(2)^2 - 3(2)\right] - \frac{1}{6}\frac{2^2 + 2(5) - 2(5)^2}{5}[2(5)^2 - 3(5)] = 46.5$$

The exact value of $\int_2^5 \left(2x^2 - 3x\right)dx$ is given by $\int_2^5 \left(2x^2 - 3x\right)dx = \left[\frac{2x^3}{3} - \frac{3x^2}{2}\right]_2^5 = 46.5$

Now evaluate $\int_2^5 3dx$ using the above formula

$$\int_2^5 3dx \approx c_1 f(a) + c_2 f(b) = -\frac{1}{6}\frac{-2(5) - 5^2 + 2(2)^2}{2}(3) - \frac{1}{6}\frac{2^2 + 2(5) - 2(5)^2}{5}(3) = 10.35$$

The exact value of $\int_2^5 3dx$ is given by $\int_2^5 3dx = [3x]_2^5 = 9$

Because the formula will only give exact values for linear combinations of $x$ and $x^2$, it does not work exactly even for a simple integral of $\int_2^5 3dx$.

Do you see now why we choose $a_0 + a_1 x$ as the integrand for which the formula $\int_a^b f(x)dx \approx c_1 f(a) + c_2 f(b)$ gives us exact values?

## Example 4

Use the two-point Gaussian Quadrature Rule to approximate the value of the integral

$$x = \int_8^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$$

Also, find the absolute relative true error.

### Solution

First, change the limits of integration from $[8, 30]$ to $[-1, 1]$ using Equation 23 gives

$$\int_8^{30} f(t)dt = \frac{30-8}{2}\int_{-1}^{1} f\left(\frac{30-8}{2}x+\frac{30+8}{2}\right)dx = 11\int_{-1}^{1} f\left(11x+19\right)dx$$

Next, get weighting factors and function argument values from Table 1 for the two point rule,

$$c_1 = 1.000000000.$$

$$x_1 = -0.577350269$$

$$c_2 = 1.000000000$$

$$x_2 = 0.577350269$$

Now we can use the Gaussian Quadrature formula

$$
\begin{aligned}
11\int_{-1}^{1} f\left(11x+19\right)dx &\approx 11\left[c_1 f\left(11x_1+19\right)+c_2 f\left(11x_2+19\right)\right] \\
&= 11\left[ f\left(11(-0.5773503)+19\right)+ f\left(11(0.5773503)+19\right)\right] \\
&= 11\left[ f(12.64915)+ f(25.35085)\right] = 11\left[(296.8317)+(708.4811)\right] = 11058.44
\end{aligned}
$$

since

$$f(12.64915) = 2000\ln\left[\frac{140000}{140000-2100(12.64915)}\right]-9.8(12.64915) = 296.8317$$

$$f(25.35085) = 2000\ln\left[\frac{140000}{140000-2100(25.35085)}\right]-9.8(25.35085) = 708.4811$$

The absolute relative true error, $\left|\in_t\right|$, is (Exact value = 11061.34m)

$$\left|\varepsilon_t\right| = \left|\frac{11061.34-11058.44}{11061.34}\right|\times 100\% = 0.0262\%$$

**Example 5**

Use the three-point Gauss Quadrature Rule to approximate the value of the integral

$$x = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$$

Also, find the absolute relative true error.

**Solution**

First, change the limits of integration from $\left[8,\ 30\right]$ to $\left[-1,\ 1\right]$ using Equation (23) gives

$$\int_{8}^{30} f(t)dt = \frac{30-8}{2}\int_{-1}^{1} f\left(\frac{30-8}{2}x+\frac{30+8}{2}\right)dx = 11\int_{-1}^{1} f\left(11x+19\right)dx$$

The weighting factors and function argument values are

$$c_1 = \ 0.555555556$$

$$x_1 = -0.774596669$$

$$c_2 = \ 0.888888889$$

$$x_2 = \ 0.000000000$$

$$c_3 = \ 0.555555556$$

$$x_3 = \ 0.774596669$$

and the formula is

$$\begin{aligned}
11\int_{-1}^{1} f\left(11x+19\right)dx &\approx 11\left[c_1 f\left(11x_1+19\right)+c_2 f\left(11x_2+19\right)+c_3 f\left(11x_3+19\right)\right] \\
&= 11\begin{bmatrix} 0.5555556 f\left(11(-.7745967)+19\right)+0.8888889 f\left(11(0.0000000)+19\right) \\ +0.5555556 f\left(11(0.7745967)+19\right) \end{bmatrix} \\
&= 11\left[0.55556 f(10.47944)+0.88889 f(19.00000) \quad +0.55556 f(27.52056)\right] \\
&= 11\left[0.55556\times 239.3327+0.88889\times 484.7455+0.55556\times 795.1069\right] = 11061.31 \quad m
\end{aligned}$$

since

$$f(10.47944) = 2000\ln\left[\frac{140000}{140000-2100(10.47944)}\right]-9.8(10.47944) = 239.3327$$

$$f(19.00000) = 2000\ln\left[\frac{140000}{140000-2100(19.00000)}\right]-9.8(19.00000) = 484.7455$$

$$f(27.52056) = 2000\ln\left[\frac{140000}{140000-2100(27.52056)}\right]-9.8(27.52056) = 795.1069$$

The absolute relative true error, $\left|\in_t\right|$, is (Exact value = 11061.34m)

$$\left|\varepsilon_t\right| = \left|\frac{11061.34-11061.31}{11061.34}\right|\times 100\% = 0.0003\%$$

The following procedure provides the $LU$ factorization with ***partial pivoting*** of a square matrix, $A = (a_{ij})_{n\times n}$, where $L = (l_{ij})_{n\times n}$ is a unit lower triangular and $U = (u_{ij})_{n\times n}$ an upper triangular.

PROCEDURE

*Starting with $k = 1$ to $n-1$, scan the entries of the kth column of the matrix $A^{(k-1)}$ below the row $(k-1)$ to identify the pivot $a_{r_k,k}$, $r_k$ such that $|a_{r_k,k}| = \max\limits_{k\le i\le n} |a_{ik}|$. Next form the permutation matrix, $P_k$, and find an elementary matrix $M_k$ such that $A^{(k)} = M_kP_kA^{(k-1)}$ has zeros below the $(k,k)$ entry on the kth column.*

$$\text{Write} \begin{cases} L = P\left(M_{n-1}P_{n-1}M_{n-2}P_{n-2}\cdots M_2P_2M_1P_1\right)^{-1} \\ U = M_{n-1}P_{n-1}M_{n-2}P_{n-2}\cdots M_2P_2M_1P_1A \end{cases}$$

where $P = (P_{n-1}P_{n-2}\cdots P_2P_1)^{-1}$ and $PA = LU$.

Given that $A = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix}$. Using partial pivoting:

Now answer the following questions

1. Find $A^{(1)}$:
2. Find $A^{(2)}$.
3. Find $A^{(3)}$.
4. Find $\left(M_3P_3M_2P_2M_1P_1\right)^{-1}$
5. Find $P$.
6. Find $L$.

7. What is the most appropriate direct method in terms of complexity that can best be used to factorize the system into a unit lower triangular matrix and its corresponding upper triangular matrix? Explain your answer.

8. To determine the surface shape of an object from images taken of a surface from three different directions, one needs to solve a set of three equations. The coefficient matrix, A of the system formed is dependent on the light source directions with respect to the camera. The unknowns, x,y,z are therefore the incident intensities that will determine the shape of the object. The right hand side values are the light intensities from the middle of the images.

   Given the following arrangement, $\begin{bmatrix} 0.2425 & 0 & -0.9701 \\ 0 & 0.2425 & -0.9701 \\ -0.2357 & -0.2357 & -0.9428 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 247 \\ 248 \\ 239 \end{bmatrix}$, solve for the incident intencities using partial pivoting.

Due Date: Friday 4th November, 2011