Numerical Analysis

By Krishna Shinde
Department of Mathematics
Modern College of Arts, Science and Commerce(Autonomous)
Shivajinagar, Pune - 5

CHAPTER 2

System of Equations

GAUSSIAN ELIMINATION

In this chapter we study techniques for the solutions of the system of linear algebraic equations. The general system of n linear equations in n unknowns can be written as

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

The a_{ij} and b_i are known constants, and x_i are variables. This system can be expressed in matrix form as $Ax = b$, where A is the $n \times n$ matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ & \cdot & & & & \cdot \\ & \cdot & & & & \cdot \\ & \cdot & & & & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix}$$

and x and b are the $n-$dimensional column vector $[x_1\ x_2\ \cdot\ \cdot\ \cdot\ x_n]^T$ and $[b_1\ b_2\ \cdot\ \cdot\ \cdot\ b_n]^T$, respectively. A is called coefficient matrix, x the solution vector and b the right-hand side vector for the system.

The augmented matrix of linear equations given earlier is

$$\left[\begin{array}{cccccc|c} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} & b_2 \\ & \cdot & & & & \cdot & \cdot \\ & \cdot & & & & \cdot & \cdot \\ & \cdot & & & & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nn} & b_n \end{array}\right]$$

The object of Gaussian elimination is to transform the coefficient part of augmented matrix into upper triangular form.

Definition. A matrix U is called upper triangular if all elements below the main diagonal are zero; that is $u_{ij} = 0$, whenever $i > j$.
The transformation of the coefficient portion of the augmented matrix is carried out through the following three elementary row operations(EROs).

1. ERO_1 : Any two rows can be interchanged. The notation $R_i \leftrightarrow R_j$ indicates that row i was interchanged with row j.

2. ERO_2 : Any row can be multiplied by a non-zero constant. The notation $r_i \leftarrow mR_i$ indicates that row i was multiplied by m.

3. ERO_3 : Any multiple one row can be added to another row. The notation $r_i \leftarrow R_i + mR_j$ indicates that m times row j was added to row i.

Example 1. Solve the following system by Gaussian elimination process, consider the system

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 1 \\ x_1 + x_2 + 2x_3 + 3x_4 &= 2 \\ -x_1 + 2x_3 + x_4 &= 1 \\ 3x_1 + 2x_2 - x_3 &= 1 \end{aligned}$$

Solution. We begin by placing the pivot in the first row, first column of the augmented matrix. Now we have to replace each element below the pivot, within pivot column, with zero. This can be done by performing ERO_3 on the row below the pivot row, each time adding an appropriate multiple of pivot row. The required multiple, m, is determined by formula

$$m = -\frac{\text{Element to be replaced by 0}}{\text{Element in pivot}}.$$

Therefore, the multipliers for the second, third and fourth rows are $-1, 1$, and -3, respectively.

$$\left[\begin{array}{cccc|c} \langle 1\rangle & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 3 & 2 \\ -1 & 0 & 2 & 1 & 1 \\ 3 & 2 & -1 & 0 & 1 \end{array}\right] \xrightarrow[r_4 \leftarrow R_4 - 3R_1]{\substack{r_2 \leftarrow R_2 - R_1 \\ r_3 \leftarrow R_3 + R_1}} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 1 & 3 & 2 & 2 \\ 0 & -1 & -4 & -3 & -2 \end{array}\right]$$

After first elimination pass through the matrix the pivot is moved down one row and to right one column. That is, at second row, second column position. At this point we have 0, so solve this problem by interchanging the second and third row.

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & \langle 0\rangle & 1 & 2 & 1 \\ 0 & 1 & 3 & 2 & 2 \\ 0 & -1 & -4 & -3 & -2 \end{array}\right] \xrightarrow{R_2 \leftrightarrow R_3} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & \langle 1\rangle & 3 & 2 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & -1 & -4 & -3 & -2 \end{array}\right]$$

$$\xrightarrow{r_4 \leftarrow R_4 + R_2} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 2 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & -1 & -1 & 0 \end{array}\right]$$

For the third elimination pass we have pivot element is at third row, third column.

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 2 & 2 \\ 0 & 0 & \langle 1 \rangle & 2 & 1 \\ 0 & 0 & -1 & -1 & 0 \end{array}\right] \xrightarrow{r_4 \leftarrow R_4 + R_3} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 2 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right]$$

Now to obtain the solution we perform the back substitution and we get $x = [x_1 \ x_2 \ x_3 \ x_4]^T = [-2 \ \ 3 \ \ -1 \ \ 1]^T$.

PIVOTING STRATEGIES

Consider the system of three equations in three unknowns

$$\begin{aligned} \tfrac{2}{3}x_1 + \tfrac{2}{7}x_2 + \tfrac{1}{5}x_3 &= \tfrac{43}{15} \\ \tfrac{1}{3}x_1 + \tfrac{1}{7}x_2 - \tfrac{1}{2}x_3 &= \tfrac{5}{6} \\ \tfrac{1}{5}x_1 - \tfrac{3}{7}x_2 + \tfrac{2}{5}x_3 &= -\tfrac{12}{5} \end{aligned}$$

The first elimination pass of Gaussian elimination require that the elementary row operations $r_2 \leftarrow R_2 - \frac{1}{2}R_1$ and $r_3 \leftarrow R_3 - \frac{3}{10}R_1$. Applying these row operations we obtain the equivalent system

$$\begin{aligned} \tfrac{2}{3}x_1 + \tfrac{2}{7}x_2 + \tfrac{1}{5}x_3 &= \tfrac{43}{15} \\ -\tfrac{3}{5}x_3 &= -\tfrac{3}{5} \\ -\tfrac{36}{70}x_2 + \tfrac{17}{50}x_3 &= -\tfrac{163}{50} \end{aligned}$$

Interchanging the second and third equations produces upper triangular matrix, from which the exact solution $x_1 = 1, x_2 = 7$ and $x_3 = 1$ is obtained by back substitution.

What happens when we try to solve this system by using four decimal digits ?

The above system reduces to

$$\begin{aligned} 0.6667x_1 + 0.2857x_2 + 0.2000x_3 &= 2.8670 \\ 0.3333x_1 + 0.1429x_2 - 0.5000x_3 &= 0.8333 \\ 0.2000x_1 - 0.4286x_2 + 0.4000x_3 &= -2.4000 \end{aligned}$$

The first pass of Gaussian elimination produces

$$\begin{aligned} 0.6667x_1 + 0.2857x_2 + 0.2000x_3 &= 2.8670 \\ 0.0001x_2 - 0.6000x_3 &= -0.5997 \\ -0.5143x_2 + 0.3400x_3 &= -3.260 \end{aligned}$$

After second elimination pass of Gaussian elimination we obtain the triangular system

$$\begin{aligned} 0.6667x_1 + 0.2857x_2 + 0.2000x_3 &= 2.8670 \\ 0.0001x_2 - 0.6000x_3 &= -0.5997 \\ 3086x_3 &= 3087 \end{aligned}$$

From this we obtain the solution $x_1 = 2.715, x_2 = 3.000$ and $x_3 = 1.000$ for system by back substitution.
The error in a calculation of x_1 and x_2 is nearly 200% this is because of small pivot element 0.0001 at second elimination pass. To avoid the small pivot, we can apply pivoting strategy.

Partial Pivoting.

Partial pivoting is systematic scheme of interchanging the rows of the coefficient matrix to place a selected element in the pivot position. The simplest such scheme is called partial pivoting.
The partial pivoting during the i^{th} elimination pass of Gaussian elimination, let

$$M_i = \max_{i \le j \le n} |a_{ji}|,$$

and let j_0 be the smallest value of j for which this maximum occurs. If $j_0 > i$ then interchange row i and j_0.
In other ward, we find the element in the pivot column, starting from the i-th row and continuing to the bottom of the matrix, which is of largest magnitude, and then make that element the pivot element.

Example 1. Reconsider the system

$$\begin{aligned} 0.6667x_1 + 0.2857x_2 + 0.2000x_3 &= 2.8670 \\ 0.3333x_1 + 0.1429x_2 - 0.5000x_3 &= 0.8333 \\ 0.2000x_1 - 0.4286x_2 + 0.4000x_3 &= -2.4000 \end{aligned}$$

For the first elimination pass we proceed exactly same as before because the largest element in the first column ($i = 1$) is initially in the first equation $j_0 = 1$. Since $j_0 = i$, no interchange of equations is required. Hence after first elimination pass the system reduces to

$$\begin{aligned} 0.6667x_1 + 0.2857x_2 + 0.2000x_3 &= 2.8670 \\ 0.0001x_2 - 0.6000x_3 &= -0.5997 \\ -0.5143x_2 + 0.3400x_3 &= -3.260 \end{aligned}$$

For the second elimination pass we see that the largest element in the second column ($i = 2$) is located in the third equation ($j_0 = 3$). The partial pivoting strategy therefore

requires that second and third equations be interchanged. This yields

$$\begin{aligned} 0.6667x_1 + 0.2857x_2 + 0.2000x_3 &= 2.8670 \\ -0.5143x_2 + 0.3400x_3 &= -3.260 \\ 0.0001x_2 - 0.6000x_3 &= -0.5997 \end{aligned}$$

Now applying row transformation $r_3 \leftarrow R_3 + \frac{0.0001}{0.5143} R_2$ the above system reduces to

$$\begin{aligned} 0.6667x_1 + 0.2857x_2 + 0.2000x_3 &= 2.8670 \\ -0.5143x_2 + 0.3400x_3 &= -3.260 \\ -0.5999x_3 &= -0.6003 \end{aligned}$$

By back substitution produces the solution $x_3 = 1.01, x_2 = 7.000$ and $x_1 = 1.000$. To four digits, the values of x_1 and x_2 are exact, while the value of x_3 is in error by only one-tenth of one percent.

In the preceding example, the necessary row interchange was carried out explicitly so as not to draw attention away from the action of the pivoting strategy. This is accomplished by maintaining a vector of n elements, such that ith element of the vector indicates the row within the matrix that contains the coefficients for ith equation. Let us denote this row vector by r. The vector is initialized to

$$r = [1 \;\; 2 \;\; \ldots \;\; n]^T .$$

Example 2. Solve the system by Gaussian elimination with partial pivoting whose augmented matrix is

$$\left[\begin{array}{cccc|c} 3 & 1 & 4 & -1 & 7 \\ 2 & -2 & -1 & 2 & 1 \\ 5 & 7 & 14 & -8 & 20 \\ 1 & 3 & 2 & 4 & -4 \end{array}\right]$$

whose exact solution is $x = [1 \quad -1 \quad 1 \quad -1]^T$.

Solution. Initialize the row vector to

$$r = [1 \;\; 2 \;\; 3 \;\; 4]^T$$

To determine the location of pivot, examine the values

$$|a_{r_1 1}| = 3, |a_{r_2 1}| = 2, |a_{r_3 1}| = 5 \text{ and } |a_{r_4 1}| = 1.$$

The largest value in this list corresponds to row r_3, so $j_0 = 3$. Since $j_0 = 3 > 1 = i$ we need to interchange the first and third row. After first elimination pass we have

$$r = [3 \;\; 2 \;\; 1 \;\; 4]^T \text{ and } \left[\begin{array}{cccc|c} 0 & -3.2 & -4.4 & 3.8 & -5 \\ 0 & -4.8 & -6.6 & 5.2 & -7 \\ 5 & 7 & 14 & -8 & 20 \\ 0 & 1.6 & -0.8 & 5.6 & -8 \end{array}\right]$$

To determine the location of second pivot, examine the values

$$|a_{r_2 2}| = 4.8, |a_{r_3 2}| = 3.2 \text{ and } |a_{r_4 2}| = 1.6$$

The largest value in this list corresponds to row r_2, so $j_0 = 2$. Since $j_0 = 2 = i$, no row interchange is needed for the second pass. The second elimination pass produces

$$r = [3 \;\; 2 \;\; 1 \;\; 4]^T \text{ and } \left[\begin{array}{cccc|c} 0 & 0 & 0 & 0.3330 & -0.333 \\ 0 & -4.8 & -6.6 & 5.2 & -7 \\ 5 & 7 & 14 & -8 & 20 \\ 0 & 0 & -3 & 7.333 & -10.33 \end{array}\right]$$

The location of final pivot is determined by examining the values

$$|a_{r_3 3}| = 0 \text{ and } |a_{r_4 3}| = 3$$

The largest value here corresponds to row r_4, so $j_0 = 4$. Since $j_0 = 4 > 3 = i$, we need to interchange third and fourth row in the row vector, which gives

$$r = [3 \;\; 2 \;\; 4 \;\; 1]^T$$

Since the element in the $a_{r_4 3} = a_{13}$ position is already zero, so the third elimination pass makes no changes to the matrix. After back substitution we get, $x_4 = -1, x_3 = 0.9990, x_2 = 0.9985, x_4 = 1$.

Scaled Partial Pivoting

Partial pivoting works well in many instances but does not reduce the effects of roundoff error for all problems. Consider the system

$$\begin{aligned} 0.7x_1 + 1725x_2 &= 1739 \\ 0.4352x_1 - 5.433x_2 &= 3.271 \end{aligned}$$

whose exact solution is $x_1 = 20$ and $x_2 = 1$.

After first elimination pass of partial pivoting method the above system reduces to

$$\begin{aligned} 0.7x_1 + 1725x_2 &= 1739 \\ -1077x_2 &= -1078. \end{aligned}$$

Using back substitution produces the solution $x_2 = 1.001$ and $x_1 = 17.14$. The value of x_2 is excellent agreement with the exact value, the value of x_1 is error by more than 14%. In this case, although 0.7 is larger than 0.4352, when measured relative to the other coefficients in each equation 0.7 is actually smaller than 0.4352; that is

$$\frac{0.7}{1725} < \frac{0.4352}{5.433}$$

where, 1725 and 5.433 are the absolute values of the coefficients of greatest magnitude in the first and second equations, respectively. Here we will choose the element in the pivot column which is largest in magnitude relative to the other coefficients in its equation,

then we would have to switched the order of equations in this system prior to eliminate variables:

$$\begin{aligned} 0.4352x_1 - 5.433x_2 &= 3.271 \\ 0.7x_1 + 1725x_2 &= 1739 \end{aligned}$$

Now eliminating x_1 by using Gaussian elimination method yields the equivalent system

$$\begin{aligned} 0.4352x_1 - 5.433x_2 &= 3.271 \\ 1734x_2 &= 1734 \end{aligned}$$

Back substitution from this set of equations yields $x_2 = 1$ and $x_1 = 20$ which are same as exact answers. This pivoting strategy is known as scaled partial pivoting.
Before starting scaled partial pivoting Gaussian elimination, we construct a scaled vector s as follows. For each $1 \leq i \leq n$, let

$$s_i = \max_{1 \leq i \leq n} |a_{ij}|$$

Also initialize the row vector to

$$r = [1 \;\; 2 \;\; 3 \;\; ... \;\; n]^T$$

During ith elimination pass

$$M_i = \max_{i \leq j \leq n} \left(\frac{|a_{r_j i}|}{s_{r_j}} \right)$$

and let j_0 be the smallest value of j for which maximum occurs. If $j_0 > i$, then interchange row i and j_0.
Note that while the row vector will generally change from pass to pass, then scale vector is set at the beginning of the process and is not changes thereafter.
Example 1. Solve the system using scaled partial pivoting method whose augmented matrix is given by

$$\left[\begin{array}{cccc|c} 3 & 1 & 4 & -1 & 7 \\ 2 & -2 & -1 & 2 & 1 \\ 5 & 7 & 14 & -8 & 20 \\ 1 & 3 & 2 & 4 & -4 \end{array}\right]$$

and whose exact solution is

$$x = [1 \;\; -1 \;\; 1 \;\; -1]^T$$

Set up the scale vector. Since

$$\max_{1 \leq j \leq 4} |a_{1j}| = 4, \quad \max_{1 \leq j \leq 4} |a_{2j}| = 2, \quad \max_{1 \leq j \leq 4} |a_{3j}| = 14, \quad \text{and} \quad \max_{1 \leq j \leq 4} |a_{4j}| = 4$$

Therefore,

$$s = [4 \;\; 2 \;\; 14 \;\; 4]^T$$

Next we initialize the row vector to

$$r = [1 \;\; 2 \;\; 3 \;\; 4]^T$$

To determine the location of the pivot, we examine the values

$$\frac{|a_{r_1 1}|}{s_{r_1}} = \frac{3}{4}, \frac{|a_{r_2 1}|}{s_{r_2}} = \frac{2}{2} = 1, \frac{|a_{r_3 1}|}{s_{r_3}} = \frac{5}{14}, \text{ and } \frac{|a_{r_4 1}|}{s_{r_4}} = \frac{1}{4}$$

The largest value corresponds to row r_2, so $j_0 = 2$. Since $j_0 = 2 > 1 = i$, we need to swap the first and second row. Thus, for the first elimination pass we have

$$r = [2 \;\; 1 \;\; 3 \;\; 4]^T$$

After the elimination pass, the matrix become

$$\left[\begin{array}{cccc|c} 0 & 4 & 5.5 & -4 & 5.5 \\ 2 & -2 & -1 & 2 & 1 \\ 0 & 12 & 16.5 & -13 & 17.5 \\ 0 & 4 & 2.5 & 3 & -4.5 \end{array}\right]$$

To determine the location of the next pivot, examine the values

$$\frac{|a_{r_2 2}|}{s_{r_2}} = \frac{4}{4} = 1, \frac{|a_{r_3 2}|}{s_{r_3}} = \frac{12}{14}, \text{ and } \frac{|a_{r_4 2}|}{s_{r_4}} = \frac{4}{4} = 1$$

The largest value in this list is 1, which occurs for both row r_2 and r_4. Choosing the first occurrence of the maximum value, we have $j_0 = 2 = i$. Hence, no row interchange required for second pass. The second elimination pass produces the matrix

$$\left[\begin{array}{cccc|c} 0 & 4 & 5.5 & -4 & 5.5 \\ 2 & -2 & -1 & 2 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & -3 & 7 & -10 \end{array}\right]$$

The location of the final pivot is determined by examining the values

$$\frac{|a_{r_3 3}|}{s_{r_3}} = \frac{0}{14} \text{ and } \frac{|a_{r_4 3}|}{s_{r_4}} = \frac{3}{4}$$

The largest value here corresponds to row r_4, so $j_0 = 4 > 3 = i$. We therefore need to swap the third and fourth rows, which gives the row vector

$$r = [2 \;\; 1 \;\; 4 \;\; 3]^T$$

Since the element in $a_{r_4 3} = a_{33}$ position is already zero, the third elimination pass makes no change to the matrix. Back substitution yields $x_4 = -1, x_3 = 1, x_2 = -1, x_1 = 1$ which is same as exact solution.

LU Decomposition.

Definition. The matrix L is called lower triangular if all the elements above the main diagonal are zero. That is, if $l_{ij} = 0$, whenever $i < j$.

Definition. Given a matrix A, a lower triangular matrix L and an upper triangular matrix U for which $LU = A$ are said to form an LU decomposition of A.

Example. If $A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 7 & 9 \\ 5 & 8 & -2 \end{bmatrix}$ then $U = \begin{bmatrix} 1 & 4 & 3 \\ 0 & -1 & 3 \\ 0 & 0 & -53 \end{bmatrix}$ and $L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 5 & 12 & 1 \end{bmatrix}$ form an LU decomposition for the matrix A.

Not every matrix has LU decomposition but it is possible to rearrange the rows of any non-singular matrix so that the resulting matrix does have an LU decomposition.

When a matrix has an LU decomposition, that decomposition is not unique.

For example $U_1 = \begin{bmatrix} 1 & 4 & 3 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix}$ and $L_1 = \begin{bmatrix} 1 & 0 & 0 \\ 2 & -1 & 0 \\ 5 & -12 & -53 \end{bmatrix}$, $U_2 = \begin{bmatrix} 2 & 8 & 6 \\ 0 & 3 & -9 \\ 0 & 0 & -1 \end{bmatrix}$ and $L_2 = \begin{bmatrix} 1/2 & 0 & 0 \\ 1 & -1/3 & 0 \\ 5/2 & -4 & 53 \end{bmatrix}$ forms an LU decomposition for the matrix $A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 7 & 9 \\ 5 & 8 & -2 \end{bmatrix}$.

Suppose a matrix A has two different LU decompositions $A = L_1U_1$ and $A = L_2U_2$.

$\implies L_1U_1 = L_2U_2$.

$\implies L_2^{-1}L_1 = U_2U_1^{-1}$.

The matrix on left hand side of this equation is lower triangular, while the matrix on the right hand side is upper triangular. For these two matrices to be equal, they must be diagonal matrix, call it D. Therefore, $L_1 = L_2D$ and $U_2 = DU_1$, for some diagonal matrix D. Hence we say an LU decomposition is unique up to scaling by a diagonal matrix.

Determining an LU Decomposition

Example. Determine an LU decomposition for the matrix $A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 7 & 9 \\ 5 & 8 & -2 \end{bmatrix}$ using Gaussian elimination with scaled partial pivoting.

Solution. The scaled vector associated with the matrix A is given by

$$s = [4 \;\; 7 \;\; 8]^T$$

and we initialize the row vector

$$r = [1 \;\; 2 \;\; 3]^T$$

Now examine the ratios

$$\frac{|a_{r_1 1}|}{s_{r_1}} = \frac{1}{4}, \frac{|a_{r_2 1}|}{s_{r_2}} = \frac{2}{9}, \text{ and } \frac{|a_{r_1 3}|}{s_{r_3}} = \frac{5}{8}.$$

we find the largest value corresponds to row r_3, so we need to swap the first and third row. Thus after first elimination pass we have

$$r = [3 \;\; 2 \;\; 1]^T$$

After first elimination pass, the contents of the matrix are

$$A = \begin{bmatrix} (1/2) & 12/5 & 17/5 \\ (2/5) & 19/5 & 49/5 \\ 5 & 8 & -2 \end{bmatrix}$$

Note that the opposite of each multiplier overwrites the elements which is being set to zero. To distinguish the multipliers from the other elements in the matrix, the multipliers are displayed within parenthesis.
To determine the location of second pivot, we examine the ratios

$$\frac{|a_{r_2 2}|}{s_{r_2}} = \frac{19/5}{9} = \frac{19}{45}, \text{ and } \frac{|a_{r_3 2}|}{s_{r_3}} = \frac{12/5}{4} = \frac{3}{4}.$$

The largest of these corresponds to row r_3, so we swap second and third rows, which becomes

$$r = [3 \;\; 1 \;\; 2]^T$$

After second elimination pass the contents of matrix are

$$A = \begin{bmatrix} (1/2) & 12/5 & 17/5 \\ (2/5) & (19/12) & 265/60 \\ 5 & 8 & -2 \end{bmatrix}$$

Rewriting the matrix according to the rows from row vector

$$A = \begin{bmatrix} 5 & 8 & -2 \\ (1/2) & 12/5 & 17/5 \\ (2/5) & (19/12) & 265/60 \end{bmatrix}$$

The upper triangular matrix in the decomposition is then obtained by setting the elements below the main diagonal to zero. The lower triangular matrix is obtained by setting the elements along the main diagonal to 1 and the elements above the diagonal to zero.
Therefore $U = \begin{bmatrix} 5 & 8 & -2 \\ 0 & 12/5 & 17/5 \\ 0 & 0 & 265/60 \end{bmatrix}$ and $L = \begin{bmatrix} 1 & 0 & 0 \\ 1/5 & 1 & 0 \\ 2/5 & 19/12 & 1 \end{bmatrix}$.
If we multiply the matrices L and U we obtain

$$LU = \begin{bmatrix} 5 & 8 & -2 \\ 1 & 4 & 3 \\ 2 & 7 & 9 \end{bmatrix}$$

Which is not equal to A. The rows of LU are the rows of A, but listed in different order. Observe, in particular, that the rows of LU are the rows of A listed in the order indicated by the final row vector.
Let $P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ be a matrix obtained by taking 3×3 identity matrix and reordering the rows according to the contents of the row vector $r = [3 \;\; 1 \;\; 2]^T$. If we now multiply P into the matrix A, we obtain

$$PA = \begin{bmatrix} 5 & 8 & -2 \\ 1 & 4 & 3 \\ 2 & 7 & 9 \end{bmatrix}$$

which is equal to product LU calculated above. Hence with row interchanges, we have found an LU decomposition for the matrix PA.

Solving a Linear System using an LU Decomposition

Suppose we need to solve the linear system $Ax = b$, and we have already found lower triangular matrix L and upper triangular matrix U such that $LU = PA$ for some permutation matrix P. If we multiply the linear system by P and then substitute LU for PA, we find that solving the original system is equivalent to solving $LUx = Pb$, or $L(Ux) = Pb$. Now, let $z = Ux$. Solve $Lz = Pb$ for z and then solve $Ux = z$ for x. These two substitution problems are easy to solve. Forward substitution applied to $Lz = Pb$ produce the vector z, and then back substitution applied to $Ux = z$ gives the solution vector x.

Example. Consider the linear system

$$\begin{bmatrix} 1 & 4 & 3 \\ 2 & 7 & 9 \\ 5 & 8 & -2 \end{bmatrix} x = \begin{bmatrix} -4 \\ -10 \\ 9 \end{bmatrix}.$$

Solution. Consider the system $Ax = b$, where $A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 7 & 9 \\ 5 & 8 & -2 \end{bmatrix}$ and $b = \begin{bmatrix} -4 \\ -10 \\ 9 \end{bmatrix}$.

Multiplying both side of given system by P, where $P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ we get new system $PAx = Pb$. When LU decomposition process applied to the coefficient matrix for this system produces the matrices

$$U = \begin{bmatrix} 5 & 8 & -2 \\ 0 & 12/5 & 17/5 \\ 0 & 0 & 265/60 \end{bmatrix} \text{ and } L = \begin{bmatrix} 1 & 0 & 0 \\ 1/5 & 1 & 0 \\ 2/5 & 19/12 & 1 \end{bmatrix}$$

Substituting LU for PA in new system obtained above we get $LUx = Pb$. Again substitute $Ux = z$ then the system $LUx = Pb$ reduces to $Lz = Pb$.

$$\implies \begin{bmatrix} 1 & 0 & 0 \\ 1/5 & 1 & 0 \\ 2/5 & 19/12 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -4 \\ -10 \\ 9 \end{bmatrix}$$

$$\implies \begin{bmatrix} 1 & 0 & 0 \\ 1/5 & 1 & 0 \\ 2/5 & 19/12 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 9 \\ -4 \\ -10 \end{bmatrix}$$

Using forward substitution we obtain $z_1 = 9$, $z_2 = -\frac{29}{5}$ and $z_3 = -\frac{265}{60}$. Now substituting

$z = \begin{bmatrix} 9 & -\frac{29}{5} & -\frac{265}{60} \end{bmatrix}^T$ in $Ux = z$ we get,

$$\begin{bmatrix} 5 & 8 & -2 \\ 0 & 12/5 & 17/5 \\ 0 & 0 & 265/60 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ -\frac{29}{5} \\ -\frac{265}{60} \end{bmatrix}$$

Using back substitution we obtain $x_3 = -1, x_2 = -1$ and $x_1 = 3$. Hence $x = [3 \quad -1 \quad -1]^T$ is a solution of given system.

DIRECT FACTORIZATION

Given a matrix A, the objective of an LU de composition is to determine a lower triangular matrix L and an upper triangular matrix U such that $LU = A$. LU decomposition are determined only up to a scaling by a diagonal matrix. Therefore, different factorizations may be viewed as resulting from different choices for the diagonal elements of either L or U. The two most common choices for the diagonal entries are

$$l_{ii} = 1 \text{ for each } i = 1, 2, ..., n; \text{ and}$$

$$u_{ii} = 1 \text{ for each } i = 1, 2, ..., n,$$

This gives rise to what are known as Doolittle decomposition and Crout decomposition respectively.

Crout decomposition

Let A be an $n \times n$ matrix. To obtain the Crout decomposition of A we must determine the entries $l_{ij}(i \geq j)$ and $u_{ij}(i < j)$ such that

$$\begin{bmatrix} l_{11} & 0 & \cdot & \cdot & \cdot & 0 \\ l_{21} & l_{22} & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ l_{n1} & l_{n2} & \cdot & \cdot & \cdot & l_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & \cdot & \cdot & \cdot & u_{1n} \\ 0 & 1 & \cdot & \cdot & \cdot & u_{2n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix}$$

Note that first column of U contains a single non-zero entry 1 in the firs row. Therefore, the product of ith row of L(for $i = 1, 2, ..., n$) with the first column of U is simply the element l_{i1}. The decomposition equation requires that this value be equated to a_{i1}; that is,

$$l_{i1} = a_{i1}$$

This equation determines the first column of L. Now the l_{11} entry is known multiplying the first row of L with the jth column of $U(i = 1, 2, ..., n)$ and equating the result to a_{1j} produces the equation $l_{11}u_{1j} = a_{ij}$. Dividing by l_{11} we find

$$u_{1j} = \frac{a_{1j}}{l_{11}}$$

this determines the first row of U.
Similarly, applying the same process for other rows and columns in L and U matrices respectively and equating to corresponding entries in matrix A we obtain the entries in the matrices L and U.
Example. Compute the Crout decomposition of following matrix

$$A = \begin{bmatrix} 1 & 4 & 3 \\ 2 & 7 & 9 \\ 5 & 8 & -2 \end{bmatrix}$$

Solution. The Crout decomposition of the matrix will consist of matrices

$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \text{ and } U = \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

Forming the product of first each row of L with first column of U and equating the result with the corresponding elements from A determine the elements in the first column of L:

$$l_{11} = 1 \quad l_{21} = 2 \text{ and } l_{31} = 5.$$

The first row of U is obtained by multiplying the first row of L with the second and third column of U and then equating the corresponding elements from A. This yields the equation

$$l_{11}u_{12} = 4 \text{ and } l_{11}u_{13} = 3,$$

whose solutions are

$$u_{12} = 4 \text{ and } u_{13} = 3,$$

Now multiply the second and third row of L with the second column of U. Equating each product with the corresponding elements from A generates the equations

$$l_{21}u_{12} + l_{22} = 7 \text{ and } l_{31}u_{12} + l_{32} = 8.$$

Substituting the values determined before and solving for the elements in the second column of L gives

$$l_{22} = -1 \text{ and } l_{32} = -12$$

Next we multiply the second row of L into the third column of U to derive the equation

$$l_{21}u_{13} + l_{22}u_{23} = 9 \Longrightarrow u_{23} = -3$$

Multiplying the third row of L and the third column of U generates the equation

$$l_{31}u_{13} + l_{32}u_{23} + l_{33} = -2 \Longrightarrow l_{33} = -53.$$

Therefore,

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & -1 & 0 \\ 5 & -12 & -53 \end{bmatrix} \text{ and } U = \begin{bmatrix} 1 & 4 & 3 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix}.$$

SPLITTING METHOD

Definition. Let A be a given $n \times n$ matrix. If M and N are $n \times n$ matrices with M non-singular and $A = M - N$, then pair (M, N) is called a splitting of the matrix A.
Suppose that (M, N) forms a splitting of the matrix A. Then

$$Ax = b \text{ is equivalent to } (M - N)x = b$$

$$\implies Mx = Nx + b.$$

Pre-multiplying both side by M^{-1} produces

$$x = M^{-1}Nx + M^{-1}b.$$

Hence the splitting $A = M - N$ determines the fixed point problem $x = Tx + c$ and associated iteration scheme $x^{(k+1)} = Tx^{(k)} + c$, where

$$T = M^{-1}N \text{ and } c = M^{-1}b.$$

To establish that splitting method is always consistent, first note that with $T = M^{-1}N$

$$\begin{aligned} I - T &= I - M^{-1}N \\ &= M^{-1}(M - N) \\ &= M^{-1}A \end{aligned}$$

Therefore, $(I - T)^{-1} = A^{-1}M$. Finally, with $c = M^{-1}b$

$$\begin{aligned} (I - T)^{-1}c &= (M^{-1}A)^{-1}(M^{-1})b \\ &= A^{-1}MM^{-1}b \\ &= A^{-1}b \end{aligned}$$

as required.

Jacobi Method

To identify the splitting associated with the Jacobi method, first we express A in the form

$$A = D - L - U$$

Here, D is diagonal part of A, $-L$ is strictly lower triangular part of A, and $-U$ is the strictly upper triangular part.
For example, suppose

$$A = \begin{bmatrix} 5 & 1 & 2 \\ -3 & 9 & 4 \\ 1 & 2 & -7 \end{bmatrix}.$$

Then

$$D = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & -7 \end{bmatrix}, L = \begin{bmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ -1 & -2 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & -1 & -2 \\ 0 & 0 & -4 \\ 0 & 0 & 0 \end{bmatrix}.$$

Jacobi method is based on the splitting $M = D$ and $N = L + D$. In order for M to be nonsingular, it must be the case that, for each i, $d_{ii} = a_{ii} \neq 0$. If this relationship does not holds for even a single value of i, then the equations in the system must be reordered before the Jacobi method can be applied. The specific choice of splitting indicated above, the iteration scheme for the Jacobi method is defined by

$$x^{(k+1)} = T_{jac}x^{(k)} + c_{jac}$$

where

$$T_{jac} = D^{-1}(L + U) \text{ and } c_{jac} = D^{-1}b.$$

If

$$\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix}$$

Then

$$D = \begin{bmatrix} a_{11} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & a_{22} & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix}, L = \begin{bmatrix} 0 & 0 & \cdot & \cdot & \cdot & 0 \\ -a_{21} & 0 & \cdot & \cdot & \cdot & 0 \\ -a_{31} & 0 & & & & 0 \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ -a_{n1} & -a_{n2} & \cdot & \cdot & -a_{n(n-1)} & 0 \end{bmatrix} \text{ and}$$

$$U = \begin{bmatrix} 0 & -a_{12} & \cdot & \cdot & \cdot & -a_{1n} \\ 0 & 0 & -a_{23} & \cdot & \cdot & -a_{2n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & -a_{(n-1)n} \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

Substituting these values in iteration scheme we obtain the components of iteration scheme can be written as

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

Hence the Jacobi method is equivalent to solving the ith equation in the system for the unknown x_i.

Example. Solve the following system by using Jacobi method and starting with $x^{(0)} = [0 \;\; 0 \;\; 0]^T$ and which has exact solution $x = [1 \;\; -3 \;\; 4]^T$

$$\begin{aligned} 5x_1 + x_2 + 2x_3 &= 10 \\ -3x_1 + 9x_2 + 4x_3 &= -14 \\ x_1 + 2x_2 - 7x_3 &= -33. \end{aligned}$$

Solution. The Jacobi method, when applied to this system, will produce the sequence of approximations $\left\{x^{(k)}\right\}$ according to the rule

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{5}\left[10 - x_2^{(k)} - 2x_3^{(k)}\right] \\ x_2^{(k+1)} &= \frac{1}{9}\left[-14 + 3x_1^{(k)} - 4x_3^{(k)}\right] \\ x_3^{(k+1)} &= \frac{1}{-7}\left[-33 - x_1^{(k)} - 2x_2^{(k)}\right] \end{aligned}$$

The component of $x^{(1)}$ can be obtained by substituting 0 for k in above system

$$\begin{aligned} x_1^{(1)} &= \frac{1}{5}\left[10 - x_2^{(0)} - 2x_3^{(0)}\right] = 2 \\ x_2^{(1)} &= \frac{1}{9}\left[-14 + 3x_1^{(0)} - 4x_3^{(0)}\right] = -\frac{14}{3} \\ x_3^{(1)} &= \frac{1}{-7}\left[-33 - x_1^{(0)} - 2x_2^{(0)}\right] = \frac{33}{7} \end{aligned}$$

Therefore, $x^{(1)} = [2 \quad -1.555556 \quad 4.714286]^T$. Applying the similar process the following table summarizes the 14 iterations of the Jacobi method.

k	$x^{(k)}$		
0	$[0.000000$	0.000000	$0.000000]^T$
1	$[2.000000$	-1.555556	$4.714286]^T$
2	$[0.425397$	-2.984127	$4.555556]^T$
3	$[0.774603$	-3.438448	$3.922449]^T$
4	$[1.118710$	-3.040665	$3.842530]^T$
5	$[1.071121$	-2.890443	$4.005340]^T$
6	$[0.975953$	-2.978666	$4.041462]^T$
7	$[0.979148$	-3.026443	$4.002660]^T$
8	$[1.004225$	-3.008133	$3.989466]^T$
9	$[1.005840$	-2.993910	$3.998280]^T$
10	$[0.999470$	-2.997289	$4.002574]^T$
11	$[0.998428$	-3.001321	$4.000699]^T$
12	$[0.999985$	-3.000835	$3.999398]^T$
13	$[1.000408$	-2.999738	$3.999759]^T$
14	$[1.000044$	-2.999757	$4.000133]^T$

The error between 13th and 14th iteration falls below 5×10^{-4} which can be calculated by using the formula $\|x^{(k+1)} - x^{(k)}\|_\infty$.

Gauss-Seidel Method

An obvious improvement that can be made to the Jacobi method is to use the values of

$x_i^{(k+1)}$ as soon as it has been calculated in the computation of all the subsequent entries in the vector $x^{(k+1)}$, rather than writing until the next iteration. After all $x_i^{(k+1)}$ is supposed to be a better approximation to x_i than $x_i^{(k)}$. This modification amounts changing the iteration scheme to

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

Working back we find that the splitting upon which the Gauss-Seidel method is based is

$$M = D - L \text{ and } N = U$$

Thus the iteration matrix for the Gauss-Seidel method is given by

$$T_{gs} = (D-L)^{-1}U,$$

and the vector c is given by

$$c_{gs} = (D-L)^{-1}b.$$

The necessary condition for the matrix M to be invertible is the same as previous method: for each i, we must have $d_{ii} = a_{ii} \neq 0$.

Example. Solve the following system by using Gauss-Seidel method and starting with $x^{(0)} = [0 \;\; 0 \;\; 0]^T$ and which has exact solution $x^(0) = [1 \;\; -3 \;\; 4]^T$

$$\begin{aligned} 5x_1 + x_2 + 2x_3 &= 10 \\ -3x_1 + 9x_2 + 4x_3 &= -14 \\ x_1 + 2x_2 - 7x_3 &= -33. \end{aligned}$$

Solution. The Gauss-Seidel method, when applied to this system, will produce the sequence of approximations $\left\{x^{(k)}\right\}$ according to the rule

$$x_1^{(k+1)} = \frac{1}{5}\left[10 - x_2^{(k)} - 2x_3^{(k)}\right]$$

$$x_2^{(k+1)} = \frac{1}{9}\left[-14 + 3x_1^{(k+1)} - 4x_3^{(k)}\right]$$

$$x_3^{(k+1)} = \frac{1}{-7}\left[-33 - x_1^{(k+1)} - 2x_2^{(k+1)}\right]$$

The component of $x^{(1)}$ can be obtained by substituting 0 for k in above system

$$x_1^{(1)} = \frac{1}{5}\left[10 - x_2^{(0)} - 2x_3^{(0)}\right] = 2$$

$$x_2^{(1)} = \frac{1}{9}\left[-14 + 3x_1^{(1)} - 4x_3^{(0)}\right] = -\frac{8}{9}$$

$$x_3^{(1)} = \frac{1}{-7}\left[-33 - x_1^{(1)} - 2x_2^{(1)}\right] = \frac{299}{63}$$

Therefore, $x^{(1)} = [2 \;\; -0.888889 \;\; 4.746032]^T$. Applying the similar process the following table summarizes the 10 iterations of the Gauss-Seidel method.

k	$x^{(k)}$
0	$[0.000000 \quad 0.000000 \quad 0.000000]^T$
1	$[2.000000 \quad -0.888889 \quad 4.746032]^T$
2	$[0.279365 \quad -3.571781 \quad 3.733686]^T$
3	$[1.220882 \quad -2.808011 \quad 4.086409]^T$
4	$[0.927039 \quad -3.062724 \quad 3.971656]^T$
5	$[1.023883 \quad -2.979442 \quad 4.009286]^T$
6	$[0.992174 \quad -3.006736 \quad 3.996958]^T$
7	$[1.002564 \quad -2.997793 \quad 4.000997]^T$
8	$[0.999160 \quad -3.000723 \quad 3.999673]^T$
9	$[1.000275 \quad -2.999763 \quad 4.000107]^T$
10	$[0.999910 \quad -3.000078 \quad 3.999965]^T$

The error between 9th and 10th iteration falls below 5×10^{-4} which can be calculated by using the formula $\|x^{(k+1)} - x^{(k)}\|_\infty$. Note that convergence is obtained with Gauss-Seidel method in roughly 30% fewer iterations than the Jacobi method.

SOR Method.

This method attempts to improve upon the convergence of the Gauss-Seidel method by computing x_i^{k+1} as a weighted average of x_i^k and the value produced by Gauss-Seidel method. Let the weighting parameter also known as relaxation parameter, be denoted by w. Then the scheme for SOR method is given by

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

Note that when $w = 1$, the SOR method reduces to the Gauss-Seidel method. The splitting associated with the SOR method is

$$M = \frac{1}{w}D - L \text{ and } N = \left(\frac{1}{w} - 1\right)D + U.$$

Therefore,

$$T_{sor} = \left(\frac{1}{w}D - L\right)^{-1}\left[\left(\frac{1}{w} - 1\right)D + U\right]$$

and

$$c_{sor} = \left(\frac{1}{w}D - L\right)^{-1} b$$

Example. Solve the following system by using SOR method with $w = 0.9$ and initial approximation $x^{(0)} = [0 \;\; 0 \;\; 0]^T$

$$\begin{aligned} 5x_1 + x_2 + 2x_3 &= 10 \\ -3x_1 + 9x_2 + 4x_3 &= -14 \\ x_1 + 2x_2 - 7x_3 &= -33. \end{aligned}$$

Solution. The SOR method, when applied to this system, will produce the sequence of approximations $\left\{x^{(k)}\right\}$ according to the rule

$$\begin{aligned} x_1^{(k+1)} &= 0.1x_1^{(k)} + \frac{0.9}{5}\left[10 - x_2^{(k)} - 2x_3^{(k)}\right] \\ x_2^{(k+1)} &= 0.1x_2^{(k)} + \frac{0.9}{9}\left[-14 + 3x_1^{(k+1)} - 4x_3^{(k)}\right] \\ x_3^{(k+1)} &= 0.1x_3^{(k)} + \frac{0.9}{-7}\left[-33 - x_1^{(k+1)} - 2x_2^{(k+1)}\right] \end{aligned}$$

Then the components of $x^{(1)}$ are

$$\begin{aligned} x_1^{(k+1)} &= 0.1x_1^{(0)} + \frac{0.9}{5}\left[10 - x_2^{(0)} - 2x_3^{(0)}\right] = 1.8 \\ x_2^{(k+1)} &= 0.1x_2^{(0)} + \frac{0.9}{9}\left[-14 + 3x_1^{(1)} - 4x_3^{(0)}\right] = -0.86 \\ x_3^{(k+1)} &= 0.1x_3^{(0)} + \frac{0.9}{-7}\left[-33 - x_1^{(1)} - 2x_2^{(1)}\right] = 4.253143 \end{aligned}$$

Therefore, $x^{(1)} = [1.8 \quad -0.86 \quad 4.253143]^T$. Applying the similar process the following table summarizes the 6 iterations of the SOR method.

k	$x^{(k)}$
0	$[0.000000 \quad 0.000000 \quad 0.000000]^T$
1	$[1.800000 \quad -0.860000 \quad 4.253143]^T$
2	$[0.603669 \quad -3.006157 \quad 3.972774]^T$
3	$[0.971276 \quad -2.998342 \quad 3.994011]^T$
4	$[0.998985 \quad -2.997743 \quad 3.999851]^T$
5	$[0.999546 \quad -2.999851 \quad 3.999965]^T$
6	$[0.999940 \quad -2.999989 \quad 3.999992]^T$

The error between 5th and 6th iteration falls below 5×10^{-4} which can be calculated by using the formula $\|x^{(k+1)} - x^{(k)}\|_\infty = 6 \times 10^{-5}$.

NON-LINEAR SYSTEM OF EQUATIONS

Suppose we need to solve the system of three equations

$$\begin{aligned} x_1^3 - 2x_2 - 2 &= 0 \\ x_1^3 - 5x_3^2 + 7 &= 0 \\ x_2x_3^2 - 1 &= 0. \end{aligned}$$

We cannot express this system in matrix notation because the equations are nonlinear,

we can express the system in vector notation. First, define the functions

$$\begin{aligned} f_1(x_1, x_2, x_3) &= x_1^3 - 2x_2 - 2 \\ f_2(x_1, x_2, x_3) &= x_1^3 - 5x_3^2 + 7 \\ f_3(x_1, x_2, x_3) &= x_2 x_3^2 - 1 \end{aligned}$$

Note that each of these functions represent the left-hand side of one of the equations from the non-linear system. Let $x = [x_1 \;\; x_2 \;\; x_3]^T$, and construct the vector-valued function

$$F(x) = \begin{bmatrix} f_1(x_1, x_2, x_3) \\ f_2(x_1, x_2, x_3) \\ f_3(x_1, x_2, x_3) \end{bmatrix}$$

In terms of this vector-valued function, the original system of three nonlinear equations can be expressed concisely as the single vector equation $F(x) = 0$. The problem of finding a vector x for which the vector-valued function F evaluates to 0 is generalization of the rootfinding problem which was investigated in Chapter 1.

Newton's Method for System of Non-linear Equations

Given a scalar function, f, of a single scalar argument and given an initial approximation, x_0, for a root of the function, Newton's method computes a sequence of improved approximations to the root according to the rule

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Now, let F be a vector-valued function of argument x, assuming that both vectors contain m components. To apply Newton's method to the problem of approximating a solution of $F(x) = 0$, we would like to write

$$x^{(n+1)} = x^{(n)} - \frac{F(x^{(n)})}{F'(x^{(n)})}.$$

Here, $F'(x^{(n)})$ must include the derivative of each scalar component function with respect to each component of argument vector. That's m^2 individual partial derivatives. These partial derivatives should be organized so that $dF = F'(x^{(n)})\Delta x$ provides an estimate for the change in $F(x)$ when the argument changes from x to $x + \Delta x$. From multivariable calculus we know that

$$df = \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + ... + \frac{\partial f}{\partial x_m}\Delta x_m$$

for a scalar function of m arguments, which suggests that the partial derivatives in $F'(x)$ be organized into matrix form as follows:

$$F'(x) = \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & . & . & . & \partial f_1/\partial x_m \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & . & . & . & \partial f_2/\partial x_m \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \partial f_m/\partial x_1 & \partial f_m/\partial x_2 & . & . & . & \partial f_m/\partial x_m \end{bmatrix}$$

This matrix is known as the Jacobian matrix for the system and is typically denoted by $J(x)$. Having established that $F'(x)$ is a matrix, this brings up a second question: how do we divide by a matrix? We multiply by its inverse. Thus, inverse method for a system of equations takes the form

$$x^{(n+1)} = x^{(n)} - \left[J(x^{(n)})\right]^{-1} F(x^{(n)}).$$

When implementing this scheme, we will not actually compute the inverse of the Jacobian matrix. Instead, we define

$$v^{(n)} = -\left[J(x^{(n)})\right]^{-1} F(x^{(n)}),$$

and then solve the linear system of equations.

$$v^{(n)}\left[J(x^{(n)})\right] = -F(x^{(n)})$$

for $v^{(n)}$. Once $v^{(n)}$ is known, the next iterate is computed according to the rule $x^{(n+1)} = x^{(n)} + v^{(n)}$.

Example. Let us apply Newton's method to the system of three nonlinear algebraic equations

$$\begin{aligned} x_1^3 - 2x_2 - 2 &= 0 \\ x_1^3 - 5x_3^2 + 7 &= 0 \\ x_2 x_3^2 - 1 &= 0. \end{aligned}$$

This system is equivalent to the vector equation $F(x) = 0$, where

$$F(x) = \begin{bmatrix} f_1(x_1, x_2, x_3) \\ f_2(x_1, x_2, x_3) \\ f_3(x_1, x_2, x_3) \end{bmatrix} = \begin{bmatrix} x_1^3 - 2x_2 - 2 \\ x_1^3 - 5x_3^2 + 7 \\ x_2 x_3^2 - 1 \end{bmatrix}$$

The Jacobian matrix associated with $F(x)$ is easily found to be

$$J(x) = \begin{bmatrix} 3x_1^2 & -2 & 0 \\ 3x_1^2 & 0 & -10x_3 \\ 0 & x_3^2 & 2x_2 x_3 \end{bmatrix}$$

Starting from the initial vector $x^{(0)} = [1 \;\; 1 \;\; 1]^T$, we compute

$$F(x^{(0)}) = [-3 \;\; 3 \;\; 0]^T$$

and

$$J(x^{(0)}) = \begin{bmatrix} 3 & -2 & 0 \\ 3 & 0 & -10 \\ 0 & 1 & 2 \end{bmatrix}$$

Solving the linear system $\left[J(x^{(0)})\right] v^{(0)} = -F(x^{(0)})$ yields the update vector $v^{(0)} = [3/7 \;\; -6/7 \;\; 3/7]^T$, and then $x^{(1)} = x^{(0)} + v^{(0)} = [10/7 \;\; 1/7 \;\; 10/7]^T$. Continuing to iterate until the maximum norm of $v^{(n)}$ is less than 5×10^{-4}, we obtain the result listed below.

n	$x^{(n)^T}$		
0	$[1.00000000000000$	1.00000000000000	$1.00000000000000]^T$
1	$[1.42857142857143$	0.14285714285714	$1.42857142857143]^T$
2	$[1.44011117287382$	0.49305169538633	$1.41331295163980]^T$
3	$[1.44225533875822$	0.50000806218205	$1.41421499021415]^T$
4	$[1.44224957033522$	0.50000000001480	$1.41421356237591]^T$

The exact solution to this system, in the neighbourhood of the initial vector $x^{(0)} = [1 \;\; 1 \;\; 1]^T$ is $x = \left[\sqrt[3]{3} \;\; 1/2 \;\; \sqrt{2}\right]^T$. Thus four iterations of Newton's method have produced results that are correct to eight decimal places.