CHAPTER 4

Linear Systems: Direct Methods

Methods for solving linear systems can be categorized into two, namely, the *direct methods* and the *iterative methods*. Theoretically, direct methods give exact solution in finite number of steps and therefore these methods do not involve mathematical error. However, when we implement the direct methods on a computer, because of the presence of arithmetic error, the computed solution may be an approximate solution. On the other hand, an iterative method generates a sequence of approximate solutions to a given linear system which is expected to converge to the exact solution. In this chapter, we study some direct methods for solving system of linear equations.

4.1 Direct Methods for Linear Systems

General form of a system of n linear equations in n variables is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\cdots \\ &\cdots \\ &\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n. \end{aligned} \tag{4.1}$$

Throughout this chapter, we assume that the coefficients a_{ij} and the right hand side numbers b_i, $i, j = 1, 2, \cdots, n$ are real.

The above system of linear equations can be written in the matrix notation as

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{4.2}$$

The last equation is usually written in the short form

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{4.3}$$

where A stands for the $n \times n$ matrix with entries a_{ij}, $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$ and the right hand side vector $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)^T$.

Let us now recall a result from linear algebra concerning the solvability of the system (4.2).

Theorem 4.1.1.

Let A be an $n \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^n$. Then the following statements concerning the system of linear equations $A\boldsymbol{x} = \boldsymbol{b}$ are equivalent.

1. $\det(A) \neq 0$
2. For each right hand side vector $\boldsymbol{b}$, the system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x}$.
3. For $\boldsymbol{b} = \boldsymbol{0}$, the only solution of the system $A\boldsymbol{x} = \boldsymbol{b}$ is $\boldsymbol{x} = \boldsymbol{0}$.

Note

We always assume that the coefficient matrix A is invertible. Any discussion of what happens when A is not invertible is outside the scope of this course.

In this section, we discuss two direct methods namely, the *Gaussian elimination method* and the *LU factorization* method. We also introduce the Thomas algorithm, which is a particular case of the Gaussian elimination method for tridiagonal systems.

4.1.1 Naive Gaussian Elimination Method

Let us describe the Gaussian elimination method to solve a system of linear equations in three variables. The method for a general system is similar.

Consider the following system of three linear equations in three variables x_1, x_2, x_3:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned} \tag{4.4}$$

For convenience, we denote the first, second, and third equations by E_1, E_2, and E_3, respectively.

Step 1: If $a_{11} \neq 0$, then define

$$m_{21} = \frac{a_{21}}{a_{11}}, \quad m_{31} = \frac{a_{31}}{a_{11}}. \tag{4.5}$$

We will now obtain a new system that is equivalent to the system (4.4) as follows:

- Retain the first equation E_1 as it is.
- Replace the second equation E_2 by the equation $E_2 - m_{21}E_1$.
- Replace the third equation E_3 by the equation $E_3 - m_{31}E_1$.

The new system equivalent to (4.4) is given by

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ 0 + a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &= b_2^{(2)} \\ 0 + a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 &= b_3^{(2)}, \end{aligned} \tag{4.6}$$

where the coefficients $a_{ij}^{(2)}$, and $b_k^{(2)}$ are given by

$$\left.\begin{aligned} a_{ij}^{(2)} &= a_{ij} - m_{i1}a_{1j}, \qquad i,j = 2,3, \\ b_k^{(2)} &= b_k - m_{k1}b_1, \qquad\quad k = 2,3. \end{aligned}\right\} \tag{4.7}$$

Note that the variable x_1 has been eliminated from the last two equations.

Step 2: If $a_{22}^{(2)} \neq 0$, then define

$$m_{32} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}}. \tag{4.8}$$

We still use the same names E_1, E_2, E_3 for the first, second, and third equations of the modified system (4.6), respectively. We will now obtain a new system that is equivalent to the system (4.6) as follows:

- Retain the first two equations in (4.6) as they are.
- Replace the third equation by the equation $E_3 - m_{32}E_2$.

The new system is given by

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ 0 + a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &= b_2^{(2)} \\ 0 + 0 + a_{33}^{(3)}x_3 &= b_3^{(3)}, \end{aligned} \tag{4.9}$$

where the coefficient $a_{33}^{(3)}$, and $b_3^{(3)}$ are given by

$$\begin{aligned} a_{33}^{(3)} &= a_{33}^{(2)} - m_{32}a_{23}^{(2)}, \\ b_3^{(3)} &= b_3^{(2)} - m_{32}b_2^{(2)}. \end{aligned}$$

Note that the variable x_2 has been eliminated from the last equation. This phase of the (Naive) Gaussian elimination method is called ***forward elimination phase***.

Step 3: Observe that the system (4.9) is readily solvable for x_3 if the coefficient $a_{33}^{(3)} \neq 0$. Substituting the value of x_3 in the second equation of (4.9), we can solve for x_2. Substituting the values of x_1 and x_2 in the first equation, we can solve for x_1. This solution phase of the (Naive) Gaussian elimination method is called ***backward substitution phase***.

Note

The coefficient matrix of the system (4.9) is an upper triangular matrix given by

$$U = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & 0 & a_{33}^{(3)} \end{pmatrix}. \tag{4.10}$$

Define a lower triangular matrix L by

$$L = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{pmatrix}. \tag{4.11}$$

It is easy to verify that $LU = A$.

Remark 4.1.2 [Why Naive?].

We explain why the word "Naive" is used for the method described above.

1. First of all, we do not know if the method described here can be successfully applied for all systems of linear equations which are uniquely solvable (that is, the coefficient matrix is invertible).
2. Secondly, even when we apply the method successfully, it is not clear if the computed solution is the exact solution. In fact, it is not even clear that the computed solution is close to the exact solution.

We illustrate the above points in Example 4.1.3 and Example 4.1.4.

Example 4.1.3.

Consider the system of equations

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{4.12}$$

The Step 1 cannot be started as $a_{11} = 0$. Thus the naive Gaussian elimination method fails.

Example 4.1.4.

Let $0 < \epsilon \ll 1$. Consider the system of equations

$$\begin{pmatrix} \epsilon & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{4.13}$$

Since $\epsilon \neq 0$, after Step 1 of the Naive Gaussian elimination method, we get the system

$$\begin{pmatrix} \epsilon & 1 \\ 0 & 1 - \epsilon^{-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 - \epsilon^{-1} \end{pmatrix}. \tag{4.14}$$

Using backward substitution, the solution is obtained as

$$x_2 = \frac{2 - \epsilon^{-1}}{1 - \epsilon^{-1}}, \quad x_1 = (1 - x_2)\epsilon^{-1}. \tag{4.15}$$

Note that for a sufficiently small ϵ, the computer evaluates $2 - \epsilon^{-1}$ as $-\epsilon^{-1}$, and $1 - \epsilon^{-1}$ also as $-\epsilon^{-1}$. Thus, $x_2 \approx 1$ and as a consequence $x_1 \approx 0$. However the exact/correct solution is given by

$$x_1 = \frac{1}{1 - \epsilon} \approx 1, \quad x_2 = \frac{1 - 2\epsilon}{1 - \epsilon} \approx 1. \tag{4.16}$$

Thus, in this particular example, the solution obtained by the naive Gaussian elimination method is completely wrong.

To understand this example better, we instruct the reader to solve the system (4.13) for the cases (1) $\epsilon = 10^{-3}$, and (2) $\epsilon = 10^{-5}$ using 3-digit rounding. Also, take $\epsilon = 10^{-16}$ and write a computer code to solve the given system using naive Gaussian elimination method.

In the following example, we illustrate the need of a pivoting strategy in the Gaussian elimination method.

Example 4.1.5.

Consider the linear system

$$\begin{aligned} 6x_1 + 2x_2 + 2x_3 &= -2 \\ 2x_1 + \frac{2}{3}x_2 + \frac{1}{3}x_3 &= 1 \\ x_1 + 2x_2 - x_3 &= 0. \end{aligned} \tag{4.17}$$

Let us solve this system using (naive) Gaussian elimination method using 4-digit rounding.

In 4-digit rounding approximation, the above system takes the form

$$\begin{aligned} 6.000x_1 + 2.000x_2 + 2.000x_3 &= -2.000 \\ 2.000x_1 + 0.6667x_2 + 0.3333x_3 &= 1.000 \\ 1.000x_1 + 2.000x_2 - 1.000x_3 &= 0.000 \end{aligned}$$

After eliminating x_1 from the second and third equations, we get (with $m_{21} = 0.3333$, $m_{31} = 0.1667$)

$$\begin{aligned} 6.000x_1 + 2.000x_2 + 2.000x_3 &= -2.000 \\ 0.000x_1 + 0.0001x_2 - 0.3333x_3 &= 1.667 \\ 0.000x_1 + 1.667x_2 - 1.333x_3 &= 0.3334 \end{aligned} \tag{4.18}$$

After eliminating x_2 from the third equation, we get (with $m_{32} = 16670$)

$$\begin{aligned} 6.000x_1 + 2.000x_2 + 2.000x_3 &= -2.000 \\ 0.000x_1 + 0.0001x_2 - 0.3333x_3 &= 1.667 \\ 0.000x_1 + 0.0000x_2 + 5555x_3 &= -27790 \end{aligned}$$

Using back substitution, we get $x_1 = 1.335$, $x_2 = 0$ and $x_3 = -5.003$, whereas the actual solution is $x_1 = 2.6$, $x_2 = -3.8$ and $x_3 = -5$. The difficulty with this elimination process is that the second equation in (4.18), where the coefficient of x_2 should have been zero, but rounding error prevented it and made the relative error very large.

The above examples highlight the inadequacy of the Naive Gaussian elimination method. These inadequcies can be overcome by modifying the procedure of Naive Gaussian elimination method. There are many kinds of modification. We will discuss one of the most popular modified methods which is called *modified Gaussian elimination method with partial pivoting*.

4.1.2 Modified Gaussian Elimination Method with Partial Pivoting

Consider the following system of three linear equations in three variables x_1, x_2, x_3:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned} \tag{4.19}$$

For convenience, we call the first, second, and third equations by names E_1, E_2, and E_3 respectively.

Step 1: Define $s_1 = \max\{\,|a_{11}|, |a_{21}|, |a_{31}|\,\}$. Note that $s_1 \neq 0$ (why?). Let k be the least number such that $s_1 = |a_{k1}|$. Interchange the first equation and the k^{th} equation. Let us re-write the system after this modification.

$$\begin{aligned} a^{(1)}_{11}x_1 + a^{(1)}_{12}x_2 + a^{(1)}_{13}x_3 &= b^{(1)}_1 \\ a^{(1)}_{21}x_1 + a^{(1)}_{22}x_2 + a^{(1)}_{23}x_3 &= b^{(1)}_2 \\ a^{(1)}_{31}x_1 + a^{(1)}_{32}x_2 + a^{(1)}_{33}x_3 &= b^{(1)}_3. \end{aligned} \tag{4.20}$$

where

$$a^{(1)}_{11} = a_{k1}, a^{(1)}_{12} = a_{k2}, a^{(1)}_{13} = a_{k3}, a^{(1)}_{k1} = a_{11}, a^{(1)}_{k2} = a_{12}, a^{(1)}_{k3} = a_{13}; b^{(1)}_1 = b_k, b^{(1)}_k = b_1, \tag{4.21}$$

and rest of the coefficients $a^{(1)}_{ij}$ are same as a_{ij} as all equations other than the first and k^{th} remain untouched by the interchange of first and k^{th} equation. Now eliminate the x_1 variable from the second and third equations of the system (4.20). Define

$$m_{21} = \frac{a^{(1)}_{21}}{a^{(1)}_{11}}, \quad m_{31} = \frac{a^{(1)}_{31}}{a^{(1)}_{11}}. \tag{4.22}$$

We will now obtain a new system that is equivalent to the system (4.20) as follows:

- The first equation will be retained as it is.
- Replace the second equation by the equation $E_2 - m_{21}E_1$.
- Replace the third equation by the equation $E_3 - m_{31}E_1$.

The new system is given by

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ 0 + a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &= b_2^{(2)} \\ 0 + a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 &= b_3^{(2)}, \end{aligned} \tag{4.23}$$

where the coefficients $a_{ij}^{(2)}$, and $b_k^{(2)}$ are given by

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)}, \quad i,j = 2,3 \\ b_i^{(2)} &= b_i^{(1)} - m_{i1}b_1^{(1)}, \quad i = 2,3. \end{aligned}$$

Note that the variable x_1 has been eliminated from the last two equations.

Step 2: Define $s_2 = \max\left\{ |a_{22}^{(2)}|, |a_{32}^{(2)}| \right\}$. Note that $s_2 \neq 0$ (why?). Let l be the least number such that $s_l = |a_{l2}^{(2)}|$. Interchange the second row and the l^{th} rows. Let us rewrite the system after this modification.

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ 0 + a_{22}^{(3)}x_2 + a_{23}^{(3)}x_3 &= b_2^{(3)} \\ 0 + a_{32}^{(3)}x_2 + a_{33}^{(3)}x_3 &= b_3^{(3)}, \end{aligned} \tag{4.24}$$

where the coefficients $a_{ij}^{(3)}$, and $b_i^{(3)}$ are given by

$$a_{22}^{(3)} = a_{l2}^{(2)}, a_{23}^{(3)} = a_{l3}^{(2)}, a_{l2}^{(3)} = a_{22}^{(2)}, a_{l3}^{(3)} = a_{23}^{(2)},$$
$$b_2^{(3)} = b_l^{(2)}, b_l^{(3)} = b_2^{(2)}$$

We still use the same names E_1, E_2, E_3 for the first, second, and third equations of the modified system (4.24), respectively.

In case $l = 2$, both second and third equations stay as they are. Let us now eliminate x_2 from the last equation. Define

$$m_{32} = \frac{a_{32}^{(3)}}{a_{22}^{(3)}} \tag{4.25}$$

We will now obtain a new system that is equivalent to the system (4.24) as follows:

- The first two equations in (4.24) will be retained as they are.
- Replace the third equation by the equation $E_3 - m_{32}E_2$.

The new system is given by

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ 0 + a_{22}^{(3)}x_2 + a_{23}^{(3)}x_3 &= b_2^{(3)} \\ 0 + 0 + a_{33}^{(4)}x_3 &= b_3^{(4)}, \end{aligned} \tag{4.26}$$

where the coefficient $a_{33}^{(4)}$, and $b_3^{(4)}$ are given by

$$\begin{aligned} a_{33}^{(4)} &= a_{33}^{(3)} - m_{32}a_{23}^{(3)}, \\ b_3^{(4)} &= b_3^{(3)} - m_{32}b_2^{(3)}. \end{aligned}$$

Note that the variable x_2 has been eliminated from the last equation. This phase of the modified Gaussian elimination method is called *forward elimination phase with partial pivoting*.

Step 3: Now the system (4.26) is readily solvable for x_3 if the coefficient $a_{33}^{(4)} \neq 0$. In fact, it is non-zero (why?). Substituting the value of x_3 in the second equation of (4.26), we can solve for x_2. Substituting the values of x_1 and x_2 in the first equation, we can solve for x_1. This solution phase of the modified Gaussian elimination method with partial pivoting is called *backward substitution phase*.

4.1.3 Thomas Method for Tri-diagonal System

The Gaussian elimination method can be simplified in the case of a tri-diagonal system so as to increase the efficiency. The resulting simplified method is called the ***Thomas method***.

A tri-diagonal system of linear equations is of the form

$$\begin{array}{ccccccccccc}
\beta_1 x_1 & +\gamma_1 x_2 & +0x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +0x_{n-1} & +0x_n & = b_1 \\
\alpha_2 x_1 & +\beta_2 x_2 & +\gamma_2 x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +0x_{n-1} & +0x_n & = b_2 \\
0x_1 & +\alpha_3 x_2 & +\beta_3 x_3 & +\gamma_3 x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +0x_{n-1} & +0x_n & = b_3 \\
0x_1 & +0x_2 & +\alpha_4 x_3 & +\beta_4 x_4 & +\gamma_4 x_5 & +0x_6 & +\cdots & +0x_{n-2} & +0x_{n-1} & +0x_n & = b_4 \\
 & & & & & & \cdots & & & & \\
 & & & & & & \cdots & & & & \\
 & & & & & & \cdots & & & & \\
0x_1 & +0x_2 & +0x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +\alpha_{n-1}x_{n-2} & +\beta_{n-1}x_{n-1} & +\gamma_{n-1}x_n & = b_{n-1} \\
0x_1 & +0x_2 & +0x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +\alpha_n x_{n-1} & +\beta_n x_n & = b_n
\end{array} \tag{4.27}$$

Here, we assume that all α_i, $i = 2, 3, \cdots, n$, β_i, $i = 1, 2, \cdots, n$, and γ_i, $i = 1, 2, \cdots, n-1$ are non-zero.

The first equation of the system reads

$$\beta_1 x_1 + \gamma_1 x_2 = b_1.$$

This equation can be rewritten as

$$x_1 + e_1 x_2 = f_1, \quad e_1 = \frac{\gamma_1}{\beta_1}, \quad f_1 = \frac{b_1}{\beta_1}.$$

Eliminating x_1 from the second equation of the system (4.27) by multiplying the above equation by α_2 and subtracting the resulting equation with the second equation of (4.27), we get

$$x_2 + e_2 x_3 = f_2, \quad e_2 = \frac{\gamma_2}{\beta_2 - \alpha_2 e_1}, \quad f_2 = \frac{b_2 - \alpha_2 f_1}{\beta_2 - \alpha_2 e_1}.$$

We now generalize the above procedure by assuming that the j^{th} equation is reduced to the form

$$x_j + e_j x_{j+1} = f_j,$$

where e_j and f_j are known quantity, reduce the $(j+1)^{\text{th}}$ equation in the form

$$x_{j+1} + e_{j+1} x_{j+2} = f_{j+1}, \quad e_{j+1} = \frac{\gamma_{j+1}}{\beta_{j+1} - \alpha_{j+1} e_j}, \quad f_{j+1} = \frac{b_{j+1} - \alpha_{j+1} f_j}{\beta_{j+1} - \alpha_{j+1} e_j},$$

for $j = 1, 2, \cdots, n-2$. Note that for $j = n-2$, we obtain the $j+1 = (n-1)^{\text{th}}$ equation as

$$x_{n-1} + e_{n-1} x_n = f_{n-1}, \quad e_{n-1} = \frac{\gamma_{n-1}}{\beta_{n-1} - \alpha_{n-1} e_{n-2}}, \quad f_{n-1} = \frac{b_{n-1} - \alpha_{n-1} f_{n-2}}{\beta_{n-1} - \alpha_{n-1} e_{n-2}}.$$

To obtain the reduced form of the nth equation of the system (4.27), eliminate x_{n-1} from the n^{th} equation by multiplying the above equation by α_n and subtracting the resulting equation with the nth equation of (4.27), which gives

$$(\alpha_n e_{n-1} - \beta_n) x_n = \alpha_n f_{n-1} - b_n.$$

Thus, the given tri-diagonal system (4.27) is now reduced to the system

$$\begin{array}{lllllllllll}
x_1 & +e_1 x_2 & +0x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +0x_{n-1} & +0x_n & = f_1 \\
0x_1 & +x_2 & +e_2 x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +0x_{n-1} & +0x_n & = f_2 \\
 & & & & & & \cdots & & & & \\
 & & & & & & \cdots & & & & \\
 & & & & & & \cdots & & & & \\
0x_1 & +0x_2 & +0x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +x_{n-1} & +e_{n-1} x_n & = f_{n-1} \\
0x_1 & +0x_2 & +0x_3 & +0x_4 & +0x_5 & +0x_6 & +\cdots & +0x_{n-2} & +0x_{n-1} & +(\alpha_n e_{n-1} - \beta_n) x_n & = \alpha_n f_{n-1} - b_n
\end{array}$$

which is an upper triangular matrix and hence, by back substitution we can get the solution.

Remark 4.1.6.

If the denominator of any of the e_j's or f_j's is zero, then the Thomas method fails. This is the situation when $\beta_j - \alpha_j e_{j-1} = 0$ which is the coefficient of x_j in the reduced equation. A suitable partial pivoting as done in the modified Gaussian elimination method may sometime help us to overcome this problem.

4.1.4 LU Factorization

In Theorem 4.1.1, we have stated that when a matrix is invertible, then the corresponding linear system can be solved. Let us now ask the next question:

'Can we give examples of a class(es) of invertible matrices for which the system of linear equations (4.2) given by

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

is "easily" solvable?'

There are three types of matrices whose simple structure makes the linear system solvable "readily". These matrices are as follows:

1. **Invertible Diagonal matrices**: These matrices look like

$$\begin{pmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & d_n \end{pmatrix}$$

   and $d_i \neq 0$ for each $i = 1, 2, \cdots, n$. In this case, the solution $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^T$ is given by $\boldsymbol{x} = \left(\frac{b_1}{d_1}, \frac{b_2}{d_2}, \cdots, \frac{b_n}{d_n}\right)^T$.

2. **Invertible Lower triangular matrices**: These matrices look like

$$\begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{pmatrix}$$

and $l_{ii} \neq 0$ for each $i = 1, 2, \cdots, n$. The linear system takes the form

$$\begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{4.28}$$

From the first equation, we solve for x_1 given by

$$x_1 = \frac{b_1}{l_{11}}.$$

Substituting this value of x_1 in the second equation, we get the value of x_2 as

$$x_2 = \frac{b_2 - l_{21}\dfrac{b_1}{l_{11}}}{l_{22}}.$$

Proceeding in this manner, we solve for the vector $\boldsymbol{x}$. This procedure of obtaining solution may be called the *forward substitution*.

3. **Invertible Upper triangular matrices**: These matrices look like

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{pmatrix}$$

and $u_{ii} \neq 0$ for each $i = 1, 2, \cdots, n$. The linear system takes the form

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{4.29}$$

From the last equation, we solve for x_n given by

$$x_n = \frac{b_n}{u_{nn}}.$$

Substituting this value of x_n in the penultimate equation, we get the value of x_{n-1} as

$$x_{n-1} = \frac{b_{n-1} - u_{n-1,n}\dfrac{b_n}{u_{nn}}}{u_{n-1,n-1}}.$$

Proceeding in this manner, we solve for the vector $\boldsymbol{x}$. This procedure of obtaining solution may be called the *backward substitution*.

In general, an invertible matrix A need not be one among the simple structures listed above. However, in certain situations we can always find an invertible lower triangular matrix L and an invertible upper triangular matrix U in such a way that

$$A = LU.$$

In this case, the system $A\boldsymbol{x} = \boldsymbol{b}$ becomes

$$L\big(U\boldsymbol{x}\big) = \boldsymbol{b}.$$

To solve for $\boldsymbol{x}$, we first solve the lower triangular system

$$L\boldsymbol{z} = \boldsymbol{b}$$

for the vector $\boldsymbol{z}$, which can be obtained easily using forward substitution. After obtaining $\boldsymbol{z}$, we solve the upper triangular system

$$U\boldsymbol{x} = \boldsymbol{z}$$

for the vector $\boldsymbol{x}$, which is again obtained easily using backward substitution.

Remark 4.1.7.

In Gaussian elimination method discussed in Section 4.1.1, we have seen that a given matrix A can be reduced to an upper triangular matrix U by an elimination procedure and thereby can be solved using backward substitution. In the elimination procedure, we also obtained a lower triangular matrix L in such a way that

$$A = LU$$

as remarked in this section.

The above discussion motivates us to look for a LU factorization of a given matrix.

Definition 4.1.8 [LU factorization].

A matrix A is said to have ***LU factorization*** (or ***LU decomposition***) if there exists a lower triangular matrix L and an upper triangular matrix U such that

$$A = LU.$$

Remark 4.1.9.

Clearly if a matrix has an LU decomposition, the matrices L and U are not unique as

$$A = LU = (LD)(D^{-1}U) = \tilde{L}\tilde{U}$$

for any invertible diagonal matrix D. Note that $A = \tilde{L}\tilde{U}$ is also an LU decomposition of A as $\tilde{L}$ is a lower triangular matrix and $\tilde{U}$ is an upper triangular matrix.

We discuss three different ways of constructing LU factorization of a given matrix.

Doolittle's factorization

Definition 4.1.10 [Doolittle's factorization].

A matrix A is said to have a *Doolittle's factorization* if there exists a lower triangular matrix L with all diagonal elements as 1, and an upper triangular matrix U such that

$$A = LU.$$

We now state the sufficient condition under which the Doolittle's factorization of a given matrix exists. For this, we need the notion of leading principal minors of a given matrix, which we define first and then state the required theorem.

Definition 4.1.11 [Principal Minors of a Matrix].

Let A be an $n \times n$ matrix.

1. A ***sub-matrix*** of order k ($< n$) of the matrix A is a $k \times k$ matrix obtained by removing $n - k$ rows and $n - k$ columns from A.

   The determinant of such a sub-matrix of order k of A is called a ***minor*** of order k of the matrix A.

2. The ***principal sub-matrix*** of order k of the matrix A is obtained by removing the last $n - k$ rows and the last $n - k$ columns from A.

   The determinant of the leading principal sub-matrix of order k of A is called the ***principal minor*** of order k of the matrix A.

3. A principal sub-matrix and the corresponding principal minor are called the ***leading principal sub-matrix*** and the ***leading principal minor*** of order k, respectively, if $k < n$.

Example 4.1.12.

Consider the 3×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

There are two leading principal minors corresponding to the matrix A.

- The leading principal minor of order 1 of A is

$$|a_{11}| = a_{11}.$$

- The leading principal minor of order 2 of A is

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

We now state the sufficient condition for the existence of a Doolittle factorization of a given matrix.

Theorem 4.1.13.

Let $n \geq 2$, and A be an $n \times n$ invertible matrix such that all of its first $n-1$ leading principal minors are non-zero. Then A has an LU-decomposition where L is a unit lower triangular matrix (*i.e.* all its diagonal elements equal 1). That is, A has a Doolittle's factorization.

We omit the proof of this theorem but illustrate the construction procedure of a Doolittle factorization in the case when $n = 3$.

Example 4.1.14 [Computing the Doolittle's factorization for a 3×3 matrix].

Let us illustrate the direct computation of the factors L and U of a 3×3 matrix A, whenever its leading principal minors of order 1, 2, and 3 are non-zero. Write

$A = LU$ as

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix} \tag{4.30}$$

Let us multiply the right hand side matrices and compare the elements on both sides. First observe that

$$\begin{aligned} a_{11} = u_{11}, a_{12} = u_{12}, a_{13} = u_{13}, \\ a_{21} = l_{21}u_{11}, a_{31} = l_{31}u_{11}, \end{aligned} \tag{4.31}$$

from which we obtain the first column of L and the first row of U. Further, we can see that

$$a_{22} = l_{21}u_{12} + u_{22}, \quad a_{23} = l_{21}u_{13} + u_{23}. \tag{4.32}$$

The equations in (4.32) can be solved for u_{22} and u_{23}. Also note that we have

$$l_{31}u_{12} + l_{32}u_{22} = a_{32}, \quad l_{31}u_{13} + l_{32}u_{23} + u_{33} = a_{33}. \tag{4.33}$$

These equations yield values for l_{32} and u_{33}, completing the construction of L and U. In this process, we must have $u_{11} \neq 0$, $u_{22} \neq 0$ in order to solve for L, which is true because of the assumptions that all the leading principal minors of A are non-zero.

The decomposition we have found is the Doolittle's factorization of A.

We now illustrate the above construction for a particular matrix A.

Example 4.1.15.

Consider the matrix

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

Using (4.31), we get

$$\begin{aligned} u_{11} = 1, \ u_{12} = 1, \quad u_{13} = -1, \\ l_{21} = \frac{a_{21}}{u_{11}} = 1, l_{31} = \frac{a_{31}}{u_{11}} = -2. \end{aligned}$$

Using (4.32) and (4.33),

$$\begin{aligned} u_{22} &= a_{22} - l_{21}u_{12} = 2 - 1 \times 1 = 1 \\ u_{23} &= a_{23} - l_{21}u_{13} = -2 - 1 \times (-1) = -1 \\ l_{32} &= \frac{(a_{32} - l_{31}u_{12})}{u_{22}} = \frac{(1 - (-2) \times 1)}{1} = 3 \\ u_{33} &= a_{33} - l_{31}u_{13} - l_{32}u_{23} = 1 - (-2) \times (-1) - 3 \times (-1) = 2 \end{aligned}$$

Thus we obtain the Doolittle's factorization of A as

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -2 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \end{pmatrix}.$$

Further, taking $\boldsymbol{b} = (1, 1, 1)^T$, we now solve the system $A\boldsymbol{x} = \boldsymbol{b}$ using LU factorization, with the matrix A given above. As discussed earlier, first we have to solve the lower triangular system

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -2 & 3 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Forward substitution yields $z_1 = 1, z_2 = 0, z_3 = 3$. Keeping the vector $\boldsymbol{z} = (1, 0, 3)^T$ as the right hand side, we now solve the upper triangular system

$$\begin{pmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}.$$

Backward substitution yields $x_1 = 1, x_2 = 3/2, x_3 = 3/2$.

Crout's factorization

In Doolittle's factorization, the lower triangular matrix has special property. If we ask the upper triangular matrix to have special property in the LU decomposition, it is known as the *Crout's factorization*.

Definition 4.1.16 [Crout's factorization].

A matrix A is said to have a ***Crout's factorization*** if there exists a lower triangular matrix L, and an upper triangular matrix U with all diagonal elements as 1 such that

$$A = LU.$$

The computation of the Crout's factorization for a 3×3 matrix can be done in a similar

way as done for Doolittle's factorization in Example 4.1.14. For invertible matrices, one can easily obtain Doolittle's factorization from Crout's factorization and vice versa. See the Example 4.1.24 below.

Cholesky's factorization

Before we introduce Cholesky's factorization, we recall the concept of a positive definite matrix.

Definition 4.1.17 [Positive Definite Matrix].

A symmetric matrix A is said to be ***positive definite*** if

$$\boldsymbol{x}^T A \boldsymbol{x} > 0,$$

for every non-zero vector $\boldsymbol{x}$.

We recall (from a course on Linear Algebra) a useful theorem concerning positive definite matrices.

Lemma 4.1.18.

The following statements concerning a symmetric $n \times n$ matrix A are equivalent.

1. The matrix A is positive definite.
2. All the principal minors of the matrix A are positive.
3. All the eigenvalues of the matrix A are positive.

The statements (1) and (2) are equivalent by definition. The proof of equivalence of (3) with other two statements is out of the scope of this course.

We now define Cholesky's factorization.

Definition 4.1.19 [Cholesky's factorization].

A matrix A is said to have a ***Cholesky's factorization*** if there exists a lower triangular matrix L such that

$$A = LL^T. \tag{4.34}$$

Remark 4.1.20.

It is clear from the definition that if a matrix has a Cholesky's factorization, then the matrix has to be necessarily a symmetric matrix.

We now establish a sufficient condition for the existence of the Cholesky's factorization.

Theorem 4.1.21.

If A is an $n \times n$ symmetric and positive definite matrix, then A has a unique factorization

$$A = LL^T,$$

where L is a lower triangular matrix with positive diagonal elements.

Proof.

We prove the theorem by induction. The proof of the theorem is trivial when $n = 1$. Let us assume that the theorem holds for $n = k$ for some $k \in \mathbb{N}$. That is, we assume that for every $k \times k$ symmetric and positive definite matrix B_k, there exists a unique lower triangular $k \times k$ matrix $\tilde{L}_k$ such that

$$B_k = \tilde{L}_k \tilde{L}_k^T.$$

We prove the theorem for $n = k+1$. Let A be a $(k+1) \times (k+1)$ symmetric positive definite matrix.

We first observe that the matrix A can be written as

$$A = \begin{pmatrix} A_k & \boldsymbol{a} \\ \boldsymbol{a}^T & a_{(k+1)(k+1)} \end{pmatrix},$$

where A_k is the $k \times k$ principal sub-matrix of A and $\boldsymbol{a} = (a_{1(k+1)}, a_{2(k+1)}, \cdots, a_{k(k+1)})^T$. Observe that A_k is positive definite and therefore by our assumption, there exists a unique $k \times k$ lower triangular matrix L_k such that $A_k = L_k L_k^T$. Define

$$L = \begin{pmatrix} L_k & \boldsymbol{0} \\ \boldsymbol{l}^T & l_{(k+1)(k+1)} \end{pmatrix},$$

where the real number $l_{(k+1)(k+1)}$ and the vector $\boldsymbol{l} = (l_{1(k+1)}, l_{2(k+1)}, \cdots, l_{k(k+1)})^T$ are to be chosen such that $A = LL^T$. That is

$$\begin{pmatrix} A_k & \boldsymbol{a} \\ \boldsymbol{a}^T & a_{(k+1)(k+1)} \end{pmatrix} = \begin{pmatrix} L_k & \boldsymbol{0} \\ \boldsymbol{l}^T & l_{(k+1)(k+1)} \end{pmatrix} \begin{pmatrix} L_k^T & \boldsymbol{l} \\ \boldsymbol{0}^T & l_{(k+1)(k+1)} \end{pmatrix}. \tag{4.35}$$

Here $\mathbf{0}$ denotes the zero vector of dimension k. Clearly,

$$L_k \boldsymbol{l} = \boldsymbol{a}, \tag{4.36}$$

which by forward substitution yields the vector $\boldsymbol{l}$. Here, we need to justify that L_k is invertible, which is left as an exercise.

Finally, we have $\boldsymbol{l}^T\boldsymbol{l} + l^2_{(k+1)(k+1)} = a_{(k+1)(k+1)}$ and this gives

$$l^2_{(k+1)(k+1)} = a_{(k+1)(k+1)} - \boldsymbol{l}^T\boldsymbol{l}, \tag{4.37}$$

provided the positivity of $l^2_{(k+1)(k+1)}$ is justified, which follows from taking determinant on both sides of (4.35) and using the property (3) of Lemma 4.1.18. A complete proof of this justification is left as an exercise.

We can use the inductive steps given in the proof of Theorem 4.1.21 to construct L in the Cholesky's factorization of a given symmetric positive definite matrix as illustrated in the following example.

Example 4.1.22.

Consider the matrix

$$A = \begin{pmatrix} 9 & 3 & -2 \\ 3 & 2 & 3 \\ -2 & 3 & 23 \end{pmatrix}.$$

We can check that this matrix is positive definite by any of the equivalent conditions listed in Lemma 4.1.18. Therefore, we expect a unique Cholesky's factorization for A. For the construction, we follow the proof of Theorem 4.1.21.

1. For $n = 1$, we have $A_1 = (9)$ and therefore let us take $L_1 = (3)$.
2. For $n = 2$, we have

   $$A_2 = \begin{pmatrix} 9 & 3 \\ 3 & 2 \end{pmatrix}.$$

   Therefore,

   $$L_2 = \begin{pmatrix} L_1 & 0 \\ l & l_{22} \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ l & l_{22} \end{pmatrix}.$$

   This gives $l = 1$ and $l^2 + l^2_{22} = 2$ or $l_{22} = 1$. Thus, we have

   $$L_2 = \begin{pmatrix} 3 & 0 \\ 1 & 1 \end{pmatrix}.$$

3. For $n = 3$, we take

$$L = \begin{pmatrix} L_2 & \mathbf{0} \\ \boldsymbol{l}^T & l_{33} \end{pmatrix},$$

where $\boldsymbol{l}^T = (l_{13},\ l_{23})$ and l_{33} are to be obtained in such a way that $A = LL^T$. The vector $\boldsymbol{l}$ is obtained by solving the lower triangular system (4.36) with $\boldsymbol{a} = (-2,\ \ 3)^T$ (by forward substitution), which gives $\boldsymbol{l} = (-2/3,\ 11/3)^T$. Finally, from (4.37), we have $l_{33}^2 = 82/9$. Thus, the required lower triangular matrix L is

$$L = \begin{pmatrix} 3 & 0 & 0 \\ 1 & 1 & 0 \\ -2/3 & 11/3 & \sqrt{82}/3 \end{pmatrix}.$$

It is straightforward to check that $A = LL^T$.

Remark 4.1.23 [Direct Comparison].

Observe that the constructive procedure illustrated above is similar to the direct comparison approach discussed in Example 4.1.14. Indeed, by comparing the elements on the right hand side of (4.34) with the corresponding elements on the left hand side, we see that

1. for each $i = 1, 2, \ldots, n$, the non-diagonal elements l_{ij}, for $j = 1, 2, \ldots, i-1$, are given by

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right) \Big/ \, l_{jj}; \tag{4.38}$$

2. and the diagonal elements l_{ii} are given by

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}, \qquad i = 1, 2, \ldots, n. \tag{4.39}$$

The Cholesky's factorization can also be obtained using Doolittle or Crout factorizations as illustrated by the following example.

Example 4.1.24.

Find the Doolittle, Crout, and Cholesky's factorizations of the matrix

$$A = \begin{pmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{pmatrix}.$$

1. The Doolittle factorization is given by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{pmatrix} \begin{pmatrix} 60 & 30 & 20 \\ 0 & 5 & 5 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \equiv LU.$$

2. Let us now find Crout factorization from Doolittle factorization as follows:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{pmatrix} \begin{pmatrix} 60 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \equiv LD\hat{U}.$$

   Setting $\hat{L} = LD$, we get the Crout factorization

$$A = \begin{pmatrix} 60 & 0 & 0 \\ 30 & 5 & 0 \\ 20 & 5 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \equiv \hat{L}\hat{U}.$$

3. The Cholesky factorization is obtained by splitting D as $D = D^{\frac{1}{2}} D^{\frac{1}{2}}$ in the $LD\hat{U}$ factorization above:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{pmatrix} \begin{pmatrix} 60 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

   Combining the first two matrices into one and the last two matrices into another,

we obtain the Cholesky factorization of A:

$$A = \begin{pmatrix} \sqrt{60} & 0 & 0 \\ \frac{\sqrt{60}}{2} & \sqrt{5} & 0 \\ \frac{\sqrt{60}}{3} & \sqrt{5} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \sqrt{60} & \frac{\sqrt{60}}{2} & \frac{\sqrt{60}}{3} \\ 0 & \sqrt{5} & \sqrt{5} \\ 0 & 0 & \frac{1}{\sqrt{3}} \end{pmatrix} \equiv \tilde{L}\tilde{L}^T$$

It can be observed that the idea of LU factorization method for solving a system of linear equations is parallel to the idea of Gaussian elimination method. However, LU factorization method has an advantage that once the factorization $A = LU$ is done, the solution of the linear system $A\boldsymbol{x} = \boldsymbol{b}$ is reduced to solving two triangular systems. Thus, if we want to solve a large set of linear systems where the coefficient matrix A is fixed but the right hand side vector $\boldsymbol{b}$ varies, we can do the factorization once for A and then use this factorization with different $\boldsymbol{b}$ vectors. This is obviously going to reduce the computation cost drastically as we have seen in the operation counting of Gaussian elimination method that the elimination part is the most expensive part of the computation.

4.2 Operations Count

It is important to study the computational efficiency of a numerical method. Computational efficiency of a method is measured in terms of the time taken by an 'optimally written' computer code to complete the main computational steps of the method. One obvious way to study computational efficiency is to observe the run time of the code. Another widely accepted theoretical way of estimating computational efficiency is to count the number of arithmetic operations involved in the method. This is often referred to as *operation count* or *flops count* (floating-point operations). In this section, we count the number of arithmetic operations involved in the naive Gaussian elimination method for the system $A\boldsymbol{x} = \boldsymbol{b}$.

Note that the naive Gaussian elimination method gives LU-factorization of a given matrix A. We will also perform the operation count for the Cholesky's factorization and compare it with the operation count involved in the factorization given by the naive Gaussian elimination method.

4.2.1 Naive Gaussian Elimination Method

The naive Gaussian elimination method has three steps, namely, the forward elimination phase, which involves the LHS elimination process and the RHS modification, and finally the backward substitution process. Let us count the arithmetic operations

involved in each of these steps.

Forward elimination phase: The forward elimination phase consists of (1) the modification of the coefficient matrix A and (2) the modification of the right hand side vector $\boldsymbol{b}$.

1. **Modification of the coefficient matrix:**
   We now count the additions/subtractions, multiplications and divisions in going from the given system to the triangular system.

Step	Additions/Subtractions	Multiplications	Divisions
1	$(n-1)^2$	$(n-1)^2$	$n-1$
2	$(n-2)^2$	$(n-2)^2$	$n-2$
.	.	.	.
.	.	.	.
.	.	.	.
$n-1$	1	1	1
Total	$\frac{n(n-1)(2n-1)}{6}$	$\frac{n(n-1)(2n-1)}{6}$	$\frac{n(n-1)}{2}$

   Here we use the formula

$$\sum_{j=1}^{p} j = \frac{p(p+1)}{2}, \ \sum_{j=1}^{p} j^2 = \frac{p(p+1)(2p+1)}{6}, p \geq 1.$$

   Let us explain the first row of the above table. In the first step, computation of m_{21}, m_{31}, $\cdots$, m_{n1} involve $(n-1)$ divisions. For each i, $j = 2, 3, \cdots, n$, the computation of $a_{ij}^{(2)}$ involves a multiplication and a subtraction. In total, there are $(n-1)^2$ multiplications and $(n-1)^2$ subtractions. Note that we do not count the operations involved in computing the coefficients of x_1 in the 2^{nd} to n^{th} equations (namely, $a_{i1}^{(2)}$), as we do not compute them and simply take them as zero. Similarly, other entries in the above table can be accounted for.
   Observe that the total number of the operations involved in the modification of the coefficient matrix is equal to

$$\frac{n(n-1)(4n+1)}{6} = \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n, \tag{4.40}$$

   which is of order $O(n^3)$ as $n \to \infty$.

2. **Modification of the right hand side vector:** Proceeding as before, we get

$$\text{Addition/Subtraction} = (n-1) + (n-2) + \cdots + 1 = \frac{n(n-1)}{2}$$

Multiplication/Division $= (n-1) + (n-2) + \cdots + 1 = \dfrac{n(n-1)}{2}$

Thus the total number of the operations involved in the modification of the right hand side vector is equal to $n(n-1)$, which is of order $O(n^2)$ as $n \to \infty$.

Backward substitution phase:

Addition/Subtraction $= (n-1) + (n-2) + \cdots + 1 = \dfrac{n(n-1)}{2}$

Multiplication/Division $= n + (n-1) + \cdots + 1 = \dfrac{n(n+1)}{2}$

Thus the total number of the operations involved in the modification of the right hand side vector is equal to n^2.

Total number of operations: The total number of operations in naive Gaussian elimination method for obtaining the solution is

$$\text{Total flops} = \frac{4n^3 + 9n^2 - 7n}{6},$$

which is of order $O(n^3)$ as $n \to \infty$.

4.2.2 Cholesky's factorization

To analyze the efficiency of the Cholesky's factorization, let us count the number of arithmetic operations (*flops*, abbreviation for floating-point operations) involved in the formulae (4.38) and (4.39).

We see that (4.38) involves one division, one subtraction, $j-2$ additions and $j-1$ multiplications. We therefore consider $j-1$ additions/subtrations and j multiplications/divisions involved in (4.38). There are $i-1$ such expression to be computed for the i^{th} row. Therefore, the total number of arithmetic operations involved in the computation of all the non-diagonal elements is

$$\sum_{i=1}^{n}\sum_{j=1}^{i-1}(j-1) + \sum_{i=1}^{n}\sum_{j=1}^{i-1} j \quad = \quad \frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6}.$$

Similarly, from (4.39), we can see that there are $i-1$ additions/subtractions and $i-1$ multiplications for computing the diagonal element of the i^{th} row. Therefore, the number of arithmetic operations involved in the computation of all the diagonal elements is

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

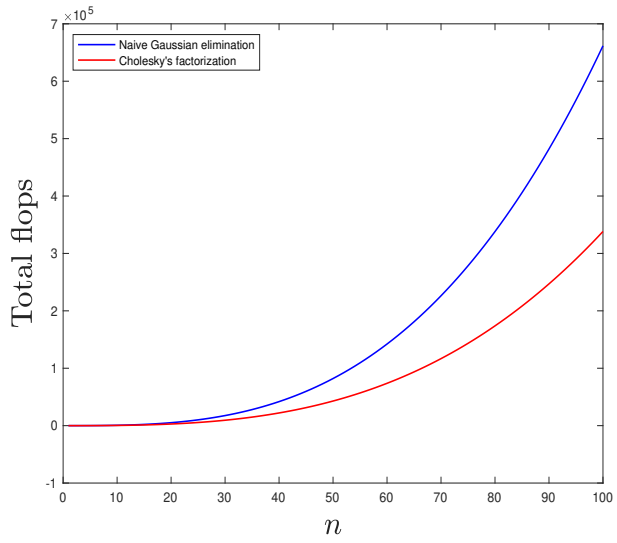

Figure 4.1: Comparison of the flops counts between naive Gaussian elimination method (blue solid line) and the Cholesky's factorization (red solid line).

Therefore, the total number of arithmetic operations involved in evaluating Cholesky's factorization is

$$\text{Total flops} = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5}{6}n. \tag{4.41}$$

In addition to the arithmetic operations, the method also involves evaluating n square roots, which does not contribute significantly to the computational cost as much as of $O(n^3/3)$. Thus, we can say that the number of arithmetic operations involved in Cholesky's factorization is $O(n^3)$ as $n \to \infty$.

Remark 4.2.1.

The number of arithmetic operations involved in the LU-factorization of a $n \times n$ matrix using naive Gaussian elimination method is given in (4.40). Comparing this with (4.41), we can see that the Cholesky's factorization is roughly twice faster than the naive Gaussian elimination method (see Figure 4.1). Similarly, we can see that the Doolittle and the Crout factorizations also involve the number of arithmetic operations comparable with the naive Gaussian elimination method. Hence if the given matrix is a symmetric and positive definite matrix, then Cholesky's factorization with the construction procedure given in Theorem 4.1.21 or in Remark 4.1.23 is more efficient than the other factorization methods discussed in this chapter.

4.3 Exercises

Gaussian Elimination Methods

1. Solve the following systems of linear equations using Naive Gaussian elimination method, and modified Gaussian elimination method with partial pivoting.

   i)

$$\begin{aligned} 6x_1 + 2x_2 + 2x_3 &= -2, \\ 2x_1 + 0.6667x_2 + 0.3333x_3 &= 1, \\ x_1 + 2x_2 - x_3 &= 0. \end{aligned}$$

   ii)

$$\begin{aligned} 0.729x_1 + 0.81x_2 + 0.9x_3 &= 0.6867, \\ x_1 + x_2 + x_3 &= 0.8338, \\ 1.331x_1 + 1.21x_2 + 1.1x_3 &= 1 \end{aligned}$$

   iii)

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

2. Solve the system

$$0.001x_1 + x_2 = 1,\ x_1 + x_2 = 2$$

   i) using Gaussian elimination with partial pivoting with infinite precision arithmetic,
   ii) using naive Gaussian elimination with 2-digit rounding, and
   iii) using modified Gaussian elimination method with partial pivoting, using 2-digit rounding.

3. Let ϵ be such that $0 < \epsilon \ll 1$. Solve the linear system

$$\begin{aligned} x_1 + x_2 + x_3 &= 6, \\ 3x_1 + (3+\epsilon)x_2 + 4x_3 &= 20, \\ 2x_1 + x_2 + 3x_3 &= 13 \end{aligned}$$

   using naive Gaussian elimination method, and using modified Gaussian elimination method with partial pivoting. Obtain the residual error in each case on a computer for which the ϵ is less than its machine epsilon. The residual error vector corresponding to an approximate solution x^* is defined as $\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}^*$, where A and $\boldsymbol{b}$ are the coefficient matrix and the right side vector, respectively, of the given linear system.

4. A matrix $A = (a_{ij})$ is said to be *row-equilibrated* if the following condition is satisfied
$$\max_{1\le j\le n} |a_{ij}| = 1 \quad \forall i = 1, 2, \cdots, n.$$
Consider the following system of linear equations
$$\begin{aligned} 30.00x_1 + 591400x_2 &= 591700 \\ 5.291x_1 - 6.130x_2 &= 46.78 \end{aligned}$$
   i) Find the exact solution of the given system using Gaussian elimination method with partial pivoting (*i.e.,* with infinite precision arithmetic).
   ii) Solve the given system using naive Gaussian elimination method using 4-digit rounding.
   iii) Obtain a system of linear equations $\widetilde{A}\mathbf{x} = \widetilde{\mathbf{b}}$ that is equivalent to the given system, where the matrix $\widetilde{A}$ is row-equilibrated. Solve the system $\widetilde{A}\mathbf{x} = \widetilde{\mathbf{b}}$ using naive Gaussian elimination method using 4-digit rounding.
   iv) Compute the relative errors involved in the two approximations obtained above.
5. Count the number of operations involved in finding a solution using naive Gaussian elimination method to the following special class of linear systems having the form
$$\begin{aligned} a_{11}x_1 + \quad &\cdots \quad + a_{1n}x_n = b_1, \\ &\cdots \\ &\cdots \\ a_{n1}x_1 + \quad &\cdots \quad + a_{nn}x_n = b_n, \end{aligned}$$
where $a_{ij} = 0$ whenever $i - j \ge 2$. In this exercise, we assume that the naive Gaussian elimination method has been implemented successfully. You must take into account the special nature of the given system.

6. Use Thomas method to solve the tri-diagonal system of equations
$$\begin{aligned} 2x_1 + 3x_2 &= 1, \\ x_1 + 2x_2 + 3x_3 &= 4, \\ x_2 + 2x_3 + 3x_4 &= 5, \\ x_3 + 2x_4 &= 2. \end{aligned}$$

LU Decomposition

7. Prove or disprove the following statements:

i) An invertible matrix has at most one Doolittle factorization.
ii) If a singular matrix has a Doolittle factorization, then the matrix has at least two Doolittle factorizations.

8. Prove that if an invertible matrix A has an LU-factorization, then all principal minors of A are non-zero.

9. Give an example of a non-invertible 3×3 matrix A such that the leading principal minors of order 1 and 2 are non-zero, and A has Doolittle factorization.

10. Use the Doolittle's factorization to solve the system

$$\begin{aligned} 4x_1 + x_2 + x_3 &= 4,\\ x_1 + 4x_2 - 2x_3 &= 4,\\ 3x_1 + 2x_2 - 4x_3 &= 6. \end{aligned}$$

11. Show that the matrix

$$\begin{pmatrix} 2 & 2 & 1\\ 1 & 1 & 1\\ 3 & 2 & 1 \end{pmatrix}$$

is invertible but has no LU factorization. Do a suitable interchange of rows to get an invertible matrix, which has an LU factorization.

12. Consider

$$A = \begin{pmatrix} 2 & 6 & -4\\ 6 & 17 & -17\\ -4 & -17 & -20 \end{pmatrix}.$$

Determine directly the factorization $A = LDL^T$, where D is diagonal and L is a lower triangular matrix with 1's on its diagonal.

13. Let A be an $n\times n$ matrix which is positive definite. Let S be a non-empty susbset of $\{1, 2, \cdots, n\}$. Show that the submatrix $A_S = (a_{ij})_{i,j\in S}$ is positive definite.

14. Factor the matrix

$$A = \begin{pmatrix} 4 & 6 & 2\\ 6 & 10 & 3\\ 2 & 3 & 5 \end{pmatrix}$$

so that $A = LL^T$, where L is lower triangular.

15. Prove the uniqueness of the factorization $A = LL^T$, where L is a lower triangular matrix all of whose diagonal entries are positive. (**Hint:** Assume that there are lower triangular matrices L_1 and L_2 with positive diagonals. Prove that $L_1 L_2^{-1} = I$.)

16. Use Cholesky factorization to solve the system of equations

$$\begin{aligned} x_1 - 2x_2 + 2x_3 &= 4, \\ -2x_1 + 5x_2 - 3x_3 &= -7, \\ 2x_1 - 3x_2 + 6x_3 &= 10. \end{aligned}$$