Chapter 6

Direct Methods for Solving Linear Systems

Linear systems of equations are associated with many problems in engineering and science, as well as with applications of mathematics to the social sciences and the quantitative study of business and economic problems.

In this chapter we consider direct methods for solving a linear system of $n$ equations in $n$ variables. Such a system has the form:

\[
\begin{align*}
    a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\
    a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\
    &\vdots \\
    a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n.
\end{align*}
\]  

(6.1)

In this system, the coefficients $a_{ij}$ for each $i, j = 1, 2, \ldots, n$, and $b_i$ for each $i = 1, 2, \ldots, n$, are given. We need to determine the unknowns $x_1, x_2, \ldots, x_n$.

The linear system (6.1) can be written in the form $Ax = b$, where $A$ is an invertible $n \times n$ matrix, $x$ and $b$ are two $n \times 1$ vectors. Direct methods theoretically give the exact solution to the system in a finite number of steps. In principle, using row operations, one transfer the $A$ matrix in (6.1) into upper triangular form:

\[
\begin{pmatrix}
    a_{11} & a_{12} & \ldots & a_{1,n-1} & a_{1n} \\
    a_{21} & a_{22} & \ldots & a_{2,n-1} & a_{2n} \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    a_{n-1,1} & a_{n-1,2} & \ldots & a_{n-1,n-1} & a_{n-1,n} \\
    a_{n1} & a_{n2} & \ldots & a_{n,n-1} & a_{nn}
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
    m_{11} & m_{12} & \ldots & m_{1,n-1} & m_{1n} \\
    0 & m_{22} & \ldots & m_{2,n-1} & m_{2n} \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    0 & 0 & \ldots & m_{n-1,n-1} & m_{n-1,n} \\
    0 & 0 & \ldots & 0 & m_{nn}
\end{pmatrix}.
\]

The process of transforming $A$ into upper triangular form is called elimination. In terms of the matrix, by doing elimination, we find a matrix $M$ such that $MA$ is of upper triangular form. After that, we solve the equation

\[MAx = Mb,\]

by doing back-substitution. In actual calculations, one does not derive explicitly the matrix $M$, but only the matrix $MA$ and the vector $Mb$.

In practice, of course, the solution obtained will be contaminated by the round-off error that is involved with the arithmetic being used. Analyzing the effect of this round-off error and determining ways to keep it under control will be a major component of this chapter.
6.1 Review of Linear Algebra

A linear system is often replaced by a matrix, which contains all the information about the system that is necessary to determine its solution, but in a compact way, and one that is easily represented in a computer. In this section, we recall some definitions and results of linear algebra that will be used in this Chapter.

Definition 6.1.1 (Matrix). An \( n \times m \) matrix \( A \) is a rectangular array of elements with \( n \) rows and \( m \) columns in which not only is the value of an element important, but also its position in the array. That is,

\[
A = \begin{pmatrix}
    a_{11} & a_{12} & \cdots & a_{1,m-1} & a_{1m} \\
    a_{21} & a_{22} & \cdots & a_{2,m-1} & a_{2m} \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    a_{n1} & a_{n2} & \cdots & a_{n,m-1} & a_{nm}
\end{pmatrix}.
\]

We sometimes denote \( A = (a_{ij}) \). If \( n = m \), then the matrix is called a square matrix of order \( n \).

Definition 6.1.2 (Transpose). Given an \( n \times m \) matrix \( A = (a_{ij}) \), we define its transpose as an \( m \times n \) matrix, denoted \( A^T \), as follows:

\[
A^T = \begin{pmatrix}
    a_{11} & a_{12} & \cdots & a_{1,m-1} & a_{1m} \\
    a_{21} & a_{22} & \cdots & a_{2,m-1} & a_{2m} \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    a_{n1} & a_{n2} & \cdots & a_{n,m-1} & a_{nm}
\end{pmatrix}.
\]

Definition 6.1.3 (Equivalence of matrices). Two matrices \( A \) and \( B \) are equal if they have the same number of rows and columns, say \( n \times m \), and if \( a_{ij} = b_{ij} \), for each \( i = 1, \ldots, n \), and \( j = 1, \ldots, m \).

Example 6.1.4. The following matrix

\[
A = \begin{pmatrix}
    2 & -1 & 7 \\
    3 & 1 & 0
\end{pmatrix}
\]

has two rows and three columns, so it is of size \( 2 \times 3 \). Its entries are described by \( a_{11} = 2 \), \( a_{12} = -1 \), \( a_{13} = 7 \), \( a_{21} = 3 \), \( a_{22} = 1 \), and \( a_{23} = 0 \). In this case, the transpose of \( A \) is

\[
A^T = \begin{pmatrix}
    2 & 3 \\
    -1 & 1 \\
    7 & 0
\end{pmatrix}.
\]

Note that \( A^T \) is of size \( 3 \times 2 \). Based on Definition 6.1.3, we have

\[
A = \begin{pmatrix}
    2 & -1 & 7 \\
    3 & 1 & 0
\end{pmatrix} \neq \begin{pmatrix}
    2 & 3 \\
    -1 & 1 \\
    7 & 0
\end{pmatrix} = A^T
\]

because they differ in dimension.

In general, a matrix and its transpose are of same size only if this matrix is a square matrix. We have the following definition if the matrix itself is equal to its transpose.

Definition 6.1.5 (Symmetric). A matrix \( M \) is called symmetric if \( M^T = M \).
Example 6.1.6. The following matrix

\[ M = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix} \]

is symmetric.

Two important operations performed on matrices are the sum of two matrices and the multiplication of a matrix by a real number (scalar).

Definition 6.1.7 (Sum of two matrices). If \( A \) and \( B \) are both \( n \times m \) matrices, then the sum of \( A \) and \( B \), denoted \( A + B \), is the \( n \times m \) matrix whose entries are \( a_{ij} + b_{ij} \) for each \( i = 1, \ldots, n \), and \( j = 1, \ldots, m \).

Definition 6.1.8 (Scalar multiplication). If \( A \) is an \( n \times m \) matrix and \( \lambda \) is a real number, then the scalar multiplication of \( \lambda \) and \( A \), denoted \( \lambda A \), is the \( n \times m \) matrix whose entries are \( \lambda a_{ij} \) for each \( i = 1, \ldots, n \), and \( j = 1, \ldots, m \).

Example 6.1.9. Determine \( A + B \) and \( \lambda A \) when

\[ A = \begin{pmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 4 & 2 & -8 \\ 0 & 1 & 6 \end{pmatrix}, \quad \text{and} \quad \lambda = -2. \]

Solution. We have

\[ A + B = \begin{pmatrix} 2+4 & -1+2 & 7-8 \\ 3+0 & 1+1 & 0+6 \end{pmatrix} = \begin{pmatrix} 6 & 1 & -1 \\ 3 & 2 & 6 \end{pmatrix} \]

and

\[ \lambda A = \begin{pmatrix} -2(2) & -2(-1) & -2(7) \\ -2(3) & -2(1) & -2(0) \end{pmatrix} = \begin{pmatrix} -4 & 2 & -14 \\ -6 & -2 & 0 \end{pmatrix}. \]

We have the following general properties for matrix addition and scalar multiplication. These properties are sufficient to classify the set of all \( n \times m \) matrices with real entries as a vector space over the field of real numbers. We let \( O \) denote a matrix all of whose entries are 0 and \( -A \) denote the matrix whose entries are \(-a_{ij}\).

Theorem 6.1.10. Let \( A, B, C \) be \( n \times m \) matrices and \( \lambda \) and \( \mu \) be real numbers. The following properties of addition and scalar multiplication hold:

(i) \( A + B = B + A \); (ii) \( (A + B) + C = A + (B + C) \);
(iii) \( A + O = O + A = A \); (iv) \( A + (-A) = -A + A = O \);
(v) \( \lambda(A + B) = \lambda A + \lambda B \); (vi) \( (\lambda + \mu)A = \lambda A + \mu A \);
(vii) \( \lambda(\mu A) = \lambda \mu A \); (viii) \( 1A = A \).

All these properties follow from similar results concerning the real numbers.

We introduce the notion of row and column vectors.
**Definition 6.1.11 (Vector).** The $1 \times n$ matrix
\[ x = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \]
is called an $n$-dimensional row vector, and the $n \times 1$ matrix
\[ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \]
is called an $n$-dimensional column vector. We sometimes denote column vector as $y = (y_1, y_2, \cdots, y_n)^T$. A $n$-dimensional (column or row) vector $0_n$ with all entries being zero is called $n$-dimensional zero vector. We denote $\mathbb{R}^n$ the collection of all $n$-dimensional column vectors.

**Example 6.1.12.** Any $n \times n$ matrix $A$ can be understood as an $n \times n$ array, or a collection of $n$-dimensional column vectors:
\[ A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}. \]
Each column of the matrix $A$ is a $n$-dimensional vector.

**Definition 6.1.13.** The set of $n$-dimensional column vectors $\{e_i\}_{i=1}^n$:
\[ e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots, \quad e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \]
is called a natural basis of $\mathbb{R}^n$. We remark that any $n$-dimensional column vector $x = (x_i)_{i=1}^n$ can be written in the linear combination of the basis vectors $\{e_i\}_{i=1}^n$ as follows:
\[ x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n. \]

**Definition 6.1.14 (Identity matrix of order $n$).** We define
\[ I_n := (e_1 \ e_2 \cdots \ e_n) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \]
to be the identity matrix of order $n$.

**Definition 6.1.15 (Matrix-vector multiplication).** Given an $n \times m$ matrix $A = (a_{ij})$ and an $m$-dimensional column vector $y = (y_1, y_2, \cdots, y_m)^T$, the $n$-dimensional column vector is defined to be their product $x = Ay$: for each $i = 1, \cdots, n$,
\[ x = Ay = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{where} \quad x_i = \sum_{j=1}^m a_{ij} y_j. \]
One can understand the matrix-vector multiplication from another point of view: the new vector $x = Ay$ is a linear combination of the columns of the matrix $A$. That is,

$$x = Ay = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = y_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} + y_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix} + \cdots + y_m \begin{pmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{pmatrix}.$$

**Remark.** One may easily check that $I_n x = x$ for any $n$-dimensional vector $x \in \mathbb{R}^n$.

**Example 6.1.16.** Let $A$ be a $2 \times 3$ matrix as in Example 6.1.4 and $y = (1, 2, 3)^T$. Then, their product is

$$x = Ay = \begin{pmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \cdot 1 - 1 \cdot 2 + 7 \cdot 3 \\ 3 \cdot 1 + 1 \cdot 2 + 0 \cdot 3 \end{pmatrix} = \begin{pmatrix} 21 \\ 5 \end{pmatrix}.$$

Moreover, we can understand this operation from another point of view:

$$x = Ay = 1 \cdot \begin{pmatrix} 2 \\ 3 \end{pmatrix} + 2 \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix} + 3 \cdot \begin{pmatrix} 7 \\ 0 \end{pmatrix} = \begin{pmatrix} 21 \\ 5 \end{pmatrix}.$$

We can use this matrix-vector multiplication to define general matrix-matrix multiplication.

**Definition 6.1.17 (Matrix-matrix multiplication).** Let $A$ be an $n \times m$ matrix and $B$ an $m \times p$ matrix. The **matrix product** of $A$ and $B$, denoted $AB$, is an $n \times p$ matrix $C$ whose entries $c_{ij}$ are

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{im} b_{mj},$$

for each $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$.

The computation of $c_{ij}$ can be viewed as the multiplication of the entries of the $i$-th row of $A$ with corresponding entries in the $j$-th column of $B$, followed by a summation; that is,

$$c_{ij} = (a_{i1}, a_{i2}, \cdots, a_{im}) \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{mj} \end{pmatrix}.$$

This explains why the number of columns of $A$ must equal the number of rows of $B$ for the product $AB$ to be defined. Moreover, one can also understand matrix-matrix multiplication in the following sense:

$$AB = A \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{pmatrix} = A (b_1 \ b_2 \ \cdots \ b_p) = (A b_1 \ Ab_2 \ \cdots \ Ab_p),$$

where $b_j$ is the $j$-th column of the matrix $B$ for $j = 1, 2, \cdots, p$. The following example should serve to clarify the matrix multiplication process.
Example 6.1.18. Determine all possible produces of the matrices:

\[ A = \begin{pmatrix} 3 & 2 \\ -1 & 1 \\ 1 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 1 & -1 \\ 3 & 1 & 2 \end{pmatrix}, \]

\[ C = \begin{pmatrix} 2 & 1 & 0 & 1 \\ -1 & 3 & 2 & 1 \\ 1 & 1 & 2 & 0 \end{pmatrix}, \quad \text{and} \quad D = \begin{pmatrix} 1 & -1 \\ 2 & -1 \end{pmatrix}. \]

\[ \text{Solution.} \quad \text{The size of the matrices are} \]

\[ A : 3 \times 2, \quad B : 2 \times 3, \quad C : 3 \times 4, \quad \text{and} \quad D : 2 \times 2. \]

The product that can be defined, and their dimensions, are:

\[ AB : 3 \times 3, \quad BA : 2 \times 2, \quad AD : 3 \times 2, \quad BC : 2 \times 4, \quad DB : 2 \times 3, \quad \text{and} \quad DD = D^2 : 2 \times 2. \]

These produces are

\[ AB = \begin{pmatrix} 12 & 5 & 1 \\ 4 & 10 & 15 \\ 7 & 10 & 15 \end{pmatrix}, \quad BA = \begin{pmatrix} 4 & 1 \\ 10 & 1 \end{pmatrix}, \quad AD = \begin{pmatrix} 7 & -5 \\ 1 & 0 \\ 9 & -5 \end{pmatrix}, \]

\[ BC = \begin{pmatrix} 2 & 4 & 0 & 3 \\ 7 & 8 & 6 & 4 \end{pmatrix}, \quad DB = \begin{pmatrix} -1 & 0 & -3 \\ 1 & 1 & -4 \end{pmatrix}, \quad \text{and} \quad D^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}. \]

Notice that although the matrix products \( AB \) and \( BA \) are both defined, their results are very different: they do not even have the same dimension. We say that the matrix product operation is not commutative, that is, products in reverse order can differ. This is the case even when both products are defined and are of the same dimension. Almost any example will show this, for example,

\[ \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \quad \text{whereas} \quad \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}. \]

Certain important operations involving matrix product do hold, however, as indicated in the following result.

Theorem 6.1.19. Let \( A \) be \( n \times m \) matrix, \( B \) be an \( m \times k \) matrix, \( C \) be a \( k \times p \) matrix, \( D \) be an \( m \times k \) matrix, and \( \lambda \) be a real number. The following properties hold:

\[ \bullet \ A(BC) = (AB)C; \]
\[ \bullet \ A(B + D) = AB + AD; \]
\[ \bullet \ \lambda(AB) = A(\lambda B). \]

Recall that matrices that have the same number of rows as columns are called square matrix and they are important in applications. We summarize the terminology used in this Chapter.

Definition 6.1.20. \( \bullet \) A square matrix has the same number of rows as columns.
• A **diagonal** matrix $D = (d_{ij})$ is a square matrix with $d_{ij} = 0$ if $i \neq j$.

• The **identity matrix of order** $n$, is a diagonal matrix whose diagonal entries are all 1s. When the size of $I_n$ is clear, this matrix is generally written simply as $I$. (See Definition 6.1.14)

**Example 6.1.21.** Consider the identity matrix of order three,

$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$ 

If $A$ is any $3 \times 3$ matrix, then

$$AI_3 = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = A.$$ 

One can also check that $I_3A = A$. The identity matrix $I_n$ commutes with any $n \times n$ matrix $A$; that is, the order of multiplication does not matter,

$$I_nA = AI_n = A.$$ 

Keep in mind that this property is not true in general, even for square matrices.

**Definition 6.1.22 (Upper and lower triangular).** An **upper-triangular** $n \times n$ matrix $U = (u_{ij})$ has, for each $j = 1, 2, \cdots, n$, the entries

$$u_{ij} = 0, \quad \text{for each } i = j + 1, j + 2, \cdots, n;$$

and a **lower-triangular** $n \times n$ matrix $L = (\ell_{ij})$ has, for each $j = 1, 2, \cdots, n$, the entries

$$\ell_{ij} = 0, \quad \text{for each } i = 1, 2, \cdots, j - 1.$$ 

We remark that a diagonal matrix is both upper triangular and lower triangular because its only nonzero entries must lie on the main diagonal.

**Definition 6.1.23 (Invertible matrix).** Assume that a square matrix $A$ of order $n$ is given. If there exist a set of $n$-dimensional vectors $\{v_i\}_{i=1}^n$ such that

$$Av_i = e_i \quad \text{for all } i = 1, 2, \cdots, n,$$

where $e_i$’s are the natural basis of $\mathbb{R}^n$. Then, the matrix $A$ is called invertible (or nonsingular) and the matrix containing all vectors $\{v_i\}_{i=1}^n$ is called the inverse of $A$, denoting $A^{-1}$. That is,

$$A^{-1} = (v_1 \ v_2 \ \cdots \ v_n).$$

Moreover, we have $AA^{-1} = A^{-1}A = I_n$. A matrix without an inverse is called **singular** (or noninvertible).

The following properties regarding matrix inverses follow from the definition above.

**Theorem 6.1.24.** For any invertible $n \times n$ matrix $A$:

• $A^{-1}$ is unique;
• \( A^{-1} \) is nonsingular and \((A^{-1})^{-1} = A\);

• If \( B \) is also a nonsingular \( n \times n \) matrix, then \((AB)^{-1} = B^{-1}A^{-1}\).

**Example 6.1.25.** Let

\[
A = \begin{pmatrix}
1 & 2 & -1 \\
2 & 1 & 0 \\
-1 & 1 & 2
\end{pmatrix} \quad \text{and} \quad B = \frac{1}{9} \begin{pmatrix}
-2 & 5 & -1 \\
4 & -1 & 2 \\
-3 & 3 & 3
\end{pmatrix}.
\]

One can check that \( B = A^{-1} \).

Recall the transpose of a given matrix \( A \), denoted \( A^T \). The proof of the next result follows directly from the definition of the transpose.

**Theorem 6.1.26.** The following operations involving the transpose of a matrix hold whenever the operation is possible:

- \((A^T)^T = A\);
- \((AB)^T = B^TA^T\);
- \((A + B)^T = A^T + B^T\);
- If \( A^{-1} \) exists, then \((A^{-1})^T = (A^T)^{-1}\).

### 6.2 Linear Systems of Equations

Using the notations of matrix and vector, the linear system (6.1) can be written in the matrix-vector-multiplication form:

\[
Ax = b, \quad \text{where} \quad A = \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{nn}
\end{pmatrix}, \quad x = \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_n
\end{pmatrix}.
\]

**Remark.** In actual computation, the solution of a linear system \( Ax = b \), with \( A \) an invertible (square) matrix of order \( n \), is not obtained by calculating the matrix \( A^{-1} \) and then calculating \( x = A^{-1}b \). Calculating the matrix \( A^{-1} \) is effectively equivalent to solving the \( n \) linear systems \( Av_i = e_i \), for all \( i = 1, 2, \cdots, n \), with \( e_j \)'s are the natural basis of \( \mathbb{R}^n \). In order words, by such a method one would be replacing the solution of \( n \) linear systems, followed by the multiplication of \( A^{-1} \) by the vector \( b \).

**Back-substitution**

The methods which we study in this chapter are based on the following observation: If the invertible matrix \( A \) is upper triangular (or lower triangular), then, the numerical solution of a
linear system $Ax = b$ is immediate; in fact, it can be written as

$$
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1,n-1}x_{n-1} + a_{1n}x_n = b_1, $$
$$
a_{22}x_2 + \cdots + a_{2,n-1}x_{n-1} + a_{2n}x_n = b_2, $$
$$
\vdots 
$$
$$
a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1},$$
$$
a_{nn}x_n = b_n,
$$

and, since $a_{11}a_{22} \cdots a_{nn} = \det(A) \neq 0$, the system is solved by calculating $x_n$ from the last equation, then $x_{n-1}$ from the last but one, etc., giving

$$(6.3)$$

$$
x_n = a_{nn}^{-1}b_n, $$
$$
x_{n-1} = a_{n-1,n-1}^{-1}(b_{n-1} - a_{n-1,n}x_n),$$

$$
\vdots 
$$
$$
x_1 = a_{11}^{-1}(b_1 - a_{12}x_2 - \cdots - a_{1,n-1}x_{n-1} - a_{1n}x_n).$$

The method of calculation given above, called \textit{back-substitution} (or backward-substitution), requires a total of

$$
\begin{align*}
1 + 2 + \cdots + (n - 1) &= \frac{n(n - 1)}{2} \quad \text{additions/subtraction}, \\
1 + 2 + \cdots + (n - 1) &= \frac{n(n - 1)}{2} \quad \text{multiplications, and} \\
n &\quad \text{divisions.}
\end{align*}
$$

Note that each component $x_i$ appears as a linear combination of $b_i, b_{i+1}, \cdots, b_n$, this shows that the inverse of a triangular matrix is (still) a triangular matrix of the same type (upper or lower).

\textbf{Gaussian-elimination}

We briefly discuss the process of elimination, transforming a general matrix to the form of upper triangular. Recall the linear system (6.1) and we refer to the $i$-th row as equation $E_i$. We use three operations to do elimination (without changing the solution):

1. Equation $E_i$ can be multiplied by any nonzero constant $\lambda \neq 0$ with the resulting equation used in place of $E_i$. This operation is denoted $(\lambda E_i) \rightarrow (E_i)$.

2. Equation $E_j$ can be multiplied by any constant $\lambda$ and added to equation $E_i$ with the resulting equation used in place of $E_i$. This operation is denoted $(E_i + \lambda E_j) \rightarrow (E_i)$.

3. Equations $E_i$ and $E_j$ can be transposed in order. This operation is denoted $(E_i) \leftrightarrow (E_j)$.

By a sequence of these operations, a linear system will be systematically transformed into a new linear system that is upper triangular. The upper triangular system is more easily solved and has the same solutions. The sequence of operations is illustrated in the following example.

\textbf{Example 6.2.1.} Consider the following linear system:

$$(6.4)$$

$$
E_1 : \quad x_1 + x_2 + 3x_4 = 4, $$
$$
E_2 : \quad 2x_1 + x_2 - x_3 + x_4 = 1, $$
$$
E_3 : \quad 3x_1 - x_2 - x_3 + 2x_4 = -3, $$
$$
E_4 : \quad -x_1 + 2x_2 + 3x_3 - x_4 = 4.
$$
We will solve for \( x_1, x_2, x_3, \) and \( x_4 \). First use equation \( E_1 \) to eliminate the unknown \( x_1 \) from equations \( E_2, E_3, \) and \( E_4 \) by performing \( (E_2 - 2E_1) \to (E_2), (E_3 - 3E_1) \to (E_3), \) and \( (E_4 + E_1) \to (E_4) \). For example, in the second equation, we perform \( (E_2 - 2E_1) \to (E_2) \) produces

\[
(2x_1 + x_2 - x_3 + 4x_4) - 2(x_1 + x_2 + 3x_4) = 1 - 2(4) \implies -x_2 - x_3 - 5x_4 = -7.
\]

Then, the linear system becomes

\[
\begin{align*}
E_1 & : \quad x_1 + x_2 + 3x_4 = 4, \\
E_2 & : \quad -x_2 - x_3 - 5x_4 = -7, \\
E_3 & : \quad -4x_2 - x_3 - 7x_4 = -15, \\
E_4 & : \quad 3x_2 + 3x_3 + 2x_4 = 8.
\end{align*}
\]

For simplicity, the new equations are again labeled \( E_1, E_2, E_3, \) and \( E_4 \). In the new system \((6.5)\), \( E_2 \) is used to eliminate the unknown \( x_2 \) from \( E_3 \) and \( E_4 \) by performing \( (E_3 - 4E_2) \to (E_3) \) and \( (E_4 + 3E_2) \to (E_4) \). This results in

\[
\begin{align*}
E_1 & : \quad x_1 + x_2 + 3x_4 = 4, \\
E_2 & : \quad -x_2 - x_3 - 5x_4 = -7, \\
E_3 & : \quad 3x_3 + 13x_4 = 13, \\
E_4 & : \quad -13x_4 = -13.
\end{align*}
\]

The linear system \((6.6)\) is now in upper triangular form and can be solved for the unknowns by a back-substitution. Since \( E_4 \) in \((6.6)\) implies \( x_4 = 1 \), we can solve \( E_3 \) for \( x_3 \) to give

\[
x_3 = \frac{1}{3}(13 - 13x_4) = 0.
\]

Continuing, \( E_2 \) gives

\[
x_2 = -(-7 + 5x_4 + x_3) = -(7 + 5 + 0) = 2,
\]

and \( E_1 \) gives

\[
x_1 = 4 - 3x_4 - x_2 = 4 - 3 - 2 = -1.
\]

The solution to linear system \((6.4)\) is therefore \( x_1 = -1, x_2 = 2, x_3 = 0, \) and \( x_4 = 1 \).

Notice that one would not need to write out the full equations at each step or to carry the variables through the calculations, if they always remained in the same column. The only variation from system to system occurs in the coefficients of the unknowns and in the values on the right side of the equations. For this reason, instead of the original equation \((6.4)\), one can consider the augmented matrix of \((6.4)\):

\[
\begin{pmatrix}
1 & 1 & 0 & 3 & 4 \\
2 & 1 & -1 & 1 & 1 \\
3 & -1 & -1 & 2 & -3 \\
-1 & 2 & 3 & -1 & 4
\end{pmatrix}.
\]

Performing the operations as described in the example produces the augmented matrices

\[
\begin{pmatrix}
1 & 1 & 0 & 3 & 4 \\
0 & -1 & -1 & -5 & -7 \\
0 & -4 & -1 & -7 & -15 \\
0 & 3 & 3 & 2 & 8
\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}
1 & 1 & 0 & 3 & 4 \\
0 & 1 & -1 & 1 & 1 \\
0 & 0 & 3 & 13 & 13 \\
0 & 0 & 0 & -13 & -13
\end{pmatrix}.
\]
One can perform back-substitution to obtain \(x_1, x_2, x_3, \text{ and } x_4\). This procedure is called Gaussian elimination with back substitution.

The general Gaussian elimination procedure applied to the linear system

\[
E_1 : \quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\
E_2 : \quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\
\vdots \\
E_n : \quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{n,n}x_n = b_n.
\]

(6.7)

is handled in a similar manner. First, form the augmented matrix:

\[
\begin{pmatrix}
  a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\
  a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  a_{n1} & a_{n2} & \cdots & a_{nn} & b_n
\end{pmatrix}.
\]

(6.8)

Provided \(a_{11} \neq 0\), we perform the operations corresponding to

\[
(E_j - (a_{ji}/a_{11})E_1) \rightarrow (E_j)
\]

to eliminate the coefficient of \(x_1\) in each of these rows. The augmented matrix reads:

\[
\begin{pmatrix}
  a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\
  0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  0 & a_{n2}^{(i-1)} & \cdots & a_{nn}^{(i-1)} & b_n^{(i-1)}
\end{pmatrix}
\]

The superscript \(^{(i)}\) means the elimination process has been done once, and so on. We then follow a sequential procedure for \(i = 2, 3, \cdots, n-1\) and perform the operation

\[
(E_j - (a_{ji}^{(i-1)}/a_{ii}^{(i-1)})E_i) \rightarrow (E_j)
\]

to eliminate \(x_i\) in each row below the \(i\)-th for all \(i = 1, 2, \cdots, n-1\). The resulting matrix has the form of upper triangular and perform back-substitution (6.3) to obtain the solution \(x_1, x_2, \cdots, x_n\).

**Example 6.2.2.** Represent the linear system

\[
E_1 : \quad x_1 - x_2 + 2x_3 - x_4 = -8, \\
E_2 : \quad 2x_1 - 2x_2 + 3x_3 - 3x_4 = -20, \\
E_3 : \quad x_1 + x_2 + x_3 = -2, \\
E_4 : \quad x_1 - x_2 + 4x_3 + 3x_4 = 4,
\]

(6.9)

as an augmented matrix and use Gaussian elimination to find its solution.

**Solution.** The augmented matrix is

\[
\tilde{A} = \tilde{A}^{(1)} = \begin{pmatrix}
1 & -1 & 2 & -1 & -8 \\
2 & -2 & 3 & -3 & -20 \\
1 & 1 & 1 & 0 & -2 \\
1 & -1 & 4 & 3 & 4
\end{pmatrix}.
\]
Performing the operations \((E_2 - 2E_1) \rightarrow (E_2), (E_3 - E_1) \rightarrow (E_3),\) and \((E_4 - E_1) \rightarrow (E_4)\) gives
\[
\tilde{A}^{(2)} = \begin{pmatrix}
1 & -1 & 2 & -1 & -8 \\
0 & 0 & -1 & -1 & -4 \\
0 & 2 & -1 & 1 & 6 \\
0 & 0 & 2 & 4 & 12
\end{pmatrix}.
\]
The diagonal entry \(a^{(1)}_{22}\), called the pivot element, is 0, so the procedure cannot continue in its present form. But operations \((E_i) \leftrightarrow (E_j)\) are permitted, so a search is made of the elements \(a^{(1)}_{32}\) and \(a^{(1)}_{42}\) for the first nonzero element. Since \(a^{(1)}_{32} = 2 \neq 0\), the operation \((E_2) \leftrightarrow (E_3)\) is performed to obtain a new matrix,
\[
\tilde{A}^{(3)} = \begin{pmatrix}
1 & -1 & 2 & -1 & -8 \\
0 & 2 & -1 & 1 & 6 \\
0 & 0 & -1 & -1 & -4 \\
0 & 0 & 2 & 4 & 12
\end{pmatrix}.
\]
Since \(x_2\) is already eliminated from \(E_3\) and \(E_4\), and the computations continue with the operation \((E_4 + 2E_3) \rightarrow (E_4)\), giving
\[
\tilde{A}^{(4)} = \begin{pmatrix}
1 & -1 & 2 & -1 & -8 \\
0 & 2 & -1 & 1 & 6 \\
0 & 0 & -1 & -1 & -4 \\
0 & 0 & 0 & 2 & 4
\end{pmatrix}.
\]
Finally, the back substitution is applied to obtain the solution:
\[
x_4 = \frac{4}{2} = 2, \quad x_3 = \frac{(-4 - (-1)x_4)}{-1} = 2, \quad x_2 = \frac{(6 - x_4 - (-1)x_3)}{2} = 3, \quad x_1 = \frac{(-8 - (-1)x_4 - 2x_3 - (-1)x_2)}{1} = -7.
\]

This example illustrate what is done if one of the pivot elements is zero. The \(k\)-th column of \(\tilde{A}^{(k)}\) from the \(k\)-th row to the \(n\)-th row is searched for the first nonzero entry. If \(a^{(k-1)}_{jk} \neq 0\) for some \(j\) with \(k + 1 \leq j \leq n\), then the operation \((E_k) \leftrightarrow (E_j)\) is performed to obtain the augmented matrix. The procedure can then be continued. If \(a^{(k-1)}_{jk} = 0\) for each \(j\), it can be shown that the linear system does not have a unique solution and the procedure stops. Finally, if \(a_{nn}^{(n-1)} = 0\), the linear system does not have a unique solution, and again the procedure stops.

Finally, let us count the number of elementary operations required in Gaussian elimination with back substitution.

(i) Elimination: in advancing from \(\tilde{A}^{(k)}\) to the augmented matrix \(\tilde{A}^{(k+1)}\), \(1 \leq k \leq n - 1\), \(n-k\) divisions are carried out and \((n-k+2)(n-k) = (n-k)^2 + 2(n-k)\) additions and
multiplications, or a total of
\[
\begin{align*}
\frac{n(n - 1)(2n + 5)}{6} & \quad \text{additions/subtraction,} \\
\frac{n(n - 1)(2n + 5)}{6} & \quad \text{multiplications, and} \\
\frac{n(n - 1)}{2} & \quad \text{divisions.}
\end{align*}
\]

(ii) The back-substitution step requires
\[
\begin{align*}
\frac{n(n - 1)}{2} & \quad \text{additions/subtraction,} \\
\frac{n(n - 1)}{2} & \quad \text{multiplications, and} \\
n & \quad \text{divisions.}
\end{align*}
\]

In the final count, Gaussian elimination with back substitution requires operations of the order of
\[
\begin{align*}
\frac{n^3}{3} & \quad \text{additions/subtraction,} \\
\frac{n^3}{3} & \quad \text{multiplications, and} \\
\frac{n^2}{2} & \quad \text{divisions.}
\end{align*}
\]

It is very instructive to compare the number of elementary operations needed for Gaussian elimination with the number of elementary operations required in applying Cramer’s rule:

\[
x_i = \frac{\det(B_i)}{\det(A)}, \quad \text{where} \quad B_i = \begin{pmatrix}
a_{11} & \cdots & a_{1,i-1} & b_1 & a_{1,i+1} & \cdots & a_{1n} \\
a_{21} & \cdots & a_{2,i-1} & b_2 & a_{2,i+1} & \cdots & a_{2n} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
a_{n1} & \cdots & a_{n,i-1} & b_n & a_{n,i+1} & \cdots & a_{nn}
\end{pmatrix},
\]

for which it is necessary to evaluate \(n + 1\) determinants and to carry out \(n\) divisions. Now the brute-force calculation of a determinant requires \(n! - 1\) additions and \((n - 1)n!\) multiplications, so that the use of Cramer’s rule requires something of the order of
\[
\begin{align*}
(n + 1)! & \quad \text{additions}, \\
(n + 2)! & \quad \text{multiplications}, \\
n & \quad \text{divisions.}
\end{align*}
\]

For \(n = 10\), for example, this gives approximately
\[
\begin{align*}
\{ & 700 \text{ operations for Gaussian elimination}, \\
& 400,000,000 \text{ operations for Cramer’s rule.}
\end{align*}
\]

It should be remembered that Gaussian elimination is the method most frequently used to solve linear systems whose matrices do not have any special properties. In particular, this method is employed for systems with full matrices.
6.3 Pivoting Strategies

In Example 6.2.2, we see that a row interchange was needed when one of the pivot elements $a_{kk}^{(k-1)} = 0$. This row interchange has the form $(E_k) \leftrightarrow (E_j)$, where $j$ is the smallest integer greater than $k$ with $a_{jk}^{(k-1)} \neq 0$. Moreover, to reduce round-off error, it is often necessary to perform row interchanges even when the pivot elements are not zero.

If the pivot element $a_{kk}^{(k-1)}$ is small in magnitude compared to $a_{jj}^{(k-1)}$, then the magnitude of the multiplier $m_{jk} = a_{jk}^{(k-1)}/a_{kk}^{(k-1)}$ will be much larger than 1. Round-off error introduced in the computation of one of the terms $a_{kl}^{(k-1)}$ is multiplied by $m_{jk}$ when computing $a_{jl}^{(k)}$, which compounds the original error. Also, when performing the back substitution for

$$x_k = \frac{b_k^{(k-1)} - \sum_{j=k+1}^{n} a_{kj}^{(k-1)} m_{jk}}{a_{kk}^{(k-1)}},$$

with a small value of $a_{kk}^{(k-1)}$, any error in the numerator can be dramatically increased because of the division by $a_{kk}^{(k-1)}$. In the next example, we see that even for small systems, round-off error can dominate the calculations.

Example 6.3.1. Apply Gaussian elimination to the system

$$E_1 : 0.003x_1 + 59.14x_2 = 59.17,$$
$$E_2 : 5.291x_1 - 6.130x_2 = 46.78,$$ (6.10)

using four-digit arithmetic with rounding, and compare the results to the exact solution $x_1 = 10.00$ and $x_2 = 1.000$.

Solution. The first pivot element, $a_{11} = a_{11}^{(0)} = 0.003$, is small, and its associated multiplier,

$$m_{21} = \frac{5.291}{0.003} = 1763.66,$$

rounds to the large number 1764. Performing $(E_2 - m_{21} E_1) \rightarrow (E_2)$ and the appropriate rounding gives the system

$$E_1 : 0.003x_1 + 59.14x_2 \approx 59.17,$$
$$E_2 : -104300x_2 \approx -104400,$$

instead of the exact system, which is

$$E_1 : 0.003x_1 + 59.14x_2 = 59.17,$$
$$E_2 : -104309.376x_2 = -104309.376.$$

The disparity in the magnitudes of $m_{21} b_1$ and $b_2$ has introduced round-off error, but the round-off error has not yet been propagated. Back substitution yields $x_2 \approx 1.001$, which is a close approximation to the actual value, $x_2 = 1.000$. However, because of the small pivot $a_{11} = 0.003$, it yields

$$x_1 \approx \frac{59.17 - (59.14)(1.001)}{0.003} = \frac{59.170 - 59.200}{0.003} = -10.00.$$

This ruins the approximation to the actual value $x_1 = 10.00$. □
Partial pivoting

Example 6.3.1 shows how difficulties can raise when the pivot element is small relative to the rest of the entries in the column. To avoid this problem, pivoting is performed by selecting an element $a_{jk}^{(k-1)}$ with a larger magnitude as the pivot, and interchanging the $k$-th and $j$-th rows for $j \geq k$. The simplest strategy is to select a smallest integer $j$ such that

$$\left| a_{jk}^{(k-1)} \right| = \max_{k \leq i \leq n} \left| a_{ik}^{(k-1)} \right|$$

and perform $(E_k) \leftrightarrow (E_j)$. This technique just described is called \textbf{partial pivoting}.

\textbf{Example 6.3.2.} Consider the linear system (6.10) and apply Gaussian elimination using partial pivoting and four-digit arithmetic with rounding, and compare the results to the exact solution $x_1 = 10.00$ and $x_2 = 1.000$.

\textit{Solution.} The partial-pivoting procedure first requires finding

$$\max \left\{ \left| a_{11}^{(0)} \right|, \left| a_{21}^{(0)} \right| \right\} = \left| 5.291 \right| = \left| a_{21}^{(0)} \right|. $$

This requires that the operation $(E_2) \leftrightarrow (E_1)$ be performed to produce the equivalent system

$$E_1 : 5.291x_1 - 6.130x_2 = 46.78, $$

$$E_2 : 0.003x_1 + 59.14x_2 = 59.17. $$

The multiplier for this system is $m_{21} = a_{21}^{(0)} / a_{11}^{(0)} = 0.0005670$, and the operation $(E_2 - m_{21}E_1) \rightarrow (E_2)$ reduces the system to

$$E_1 : 5.291x_1 - 6.130x_2 \approx 46.78, $$

$$E_2 : 59.14x_2 \approx 59.12. $$

The four-digit rounding answers resulting from the back substitution are the correct values $x_1 = 10.00$ and $x_2 = 1.000$. \hfill \Box

Using the strategy of partial pivoting, each multiplier $m_{ji}$ has magnitude less than or equal to 1. Although this strategy is sufficient for many linear systems, situations do arise when it is inadequate.

\textbf{Example 6.3.3.} The linear system

$$E_1 : 30.00x_1 + 591400x_2 = 591700, $$

$$E_2 : 5.291x_1 - 6.130x_2 = 46.78. $$

is the same as that in Example 6.3.1 except that all the entries in the first equation have been multiplied by $10^4$. The partial pivoting procedure with four-digit arithmetic leads to the same results as obtained in Example 6.3.1. The maximal value in the first column is 30.00, and the multiplier $m_{21} = 5.291 / 30.00 = 0.1764$ leads to the system

$$E_1 : 30.00x_1 + 591400x_2 \approx 591700, $$

$$E_2 : -104300x_2 \approx -104400, $$

which has the same inaccurate solutions as in Example 6.3.1, $x_1 \approx -10.00$ and $x_2 \approx 1.001$. 

We introduce one more strategy for pivoting: **Scaled partial pivoting.** It places the element in the pivot position that is largest relative to the entries in its row. The first step in this procedure is to define a scale factor $s_i$ for each row:

$$s_i := \max_{1 \leq j \leq n} |a_{ij}|.$$

If we have $s_i = 0$ for some $i$, then the system has no unique solution since all entries in the $i$-th row are 0. Assuming that this is not the case, the appropriate row interchange to place zeros in the first column is determined by choosing the least integer $p$ with

$$\frac{|a_{p1}|}{s_p} = \max_{1 \leq k \leq n} \frac{|a_{k1}|}{s_k}$$

and performing $(E_1) \leftrightarrow (E_p)$. The effect of scaling is to ensure that the largest element in each row has a relative magnitude of 1 before the comparison for row interchange is performed.

In a similar manner, before eliminating the variable $x_i$ using the operations $(E_k - m_{ki}E_i)$ for $k = i + 1, \ldots, n$, we select the smallest integer $p \geq i$ with

$$\frac{|a_{pi}|}{s_p} = \max_{1 \leq k \leq n} \frac{|a_{ki}|}{s_k}$$

and perform the row interchange $(E_i) \leftrightarrow (E_p)$ if $i \neq p$. The scale factors $s_1, \ldots, s_n$ are computed only once, at the start of the procedure. They are row dependent, so they must also be interchanged when row in interchanges are performed.

**Example 6.3.4.** Consider the linear system

$$E_1 : \begin{cases} 30.00x_1 + 591400x_2 = 591700, \\ 5.291x_1 - 6.130x_2 = 46.78. \end{cases}$$

We apply scaled partial pivoting to this system. First, compute

$$s_1 = \max\{|30.00|, |591400|\} = 591400 \quad \text{and} \quad s_2 = \max\{|5.291|, |-6.130|\} = 6.130.$$ 

Consequently,

$$\frac{|a_{11}|}{s_1} = \frac{30.00}{591400} = 0.5073 \times 10^{-4} \quad \text{and} \quad \frac{|a_{21}|}{s_2} = \frac{5.291}{6.130} = 0.8631,$$

and the interchange $(E_1) \leftrightarrow (E_2)$ is made. Applying Gaussian elimination to the new system

$$E_1 : \begin{cases} 5.291x_1 - 6.130x_2 = 46.78, \\ 30.00x_1 + 591400x_2 = 591700, \end{cases}$$

produces the correct results: $x_1 = 10.00$ and $x_2 = 1.000$.

The first additional computations required for scaled partial pivoting result from the determination of the scale factors; there are $(n - 1)$ comparison for each of the $n$ rows, for a total of $n(n - 1)$ comparisons.

To determine the correct first interchange, $n$ division are performed, followed by $n - 1$ comparisons. Hence, the first interchange determination adds $n$ divisions and $(n - 1)$ comparisons.
The scaling factors are computed only once, so the second step requires: \((n - 1)\) divisions and \((n - 2)\) comparisons. We proceed in a similar manner until there are zeros below in the main diagonal in all but the \(n\)-th row. The final step requires that we perform 2 divisions and 1 comparison. As a consequence, scaled partial pivoting adds a total of

\[
n(n - 1) + \sum_{k=1}^{n-1} k = n(n - 1) + \frac{n(n - 1)}{2} = \frac{3}{2}n(n - 1) \quad \text{comparisons,}
\]

and

\[
\sum_{k=2}^{n} k = \left(\sum_{k=1}^{n} k\right) - 1 = \frac{n(n + 1)}{2} - 1 = \frac{1}{2}(n - 1)(n + 2) \quad \text{divisions}
\]

to the Gaussian elimination procedure. The time required to perform a comparison is about the same as an addition or subtraction. Since the total time to perform the basic Gaussian elimination procedure is \(O(n^3)\), scaled partial pivoting does not add significantly to the computational time required to solve a system for large values of \(n\).

### 6.4 The Determinant of a Matrix

The **determinant** of a matrix provides existence and uniqueness results for linear systems having the same number of equations and unknowns. We will denote the determinant of a square matrix \(A\) by \(\det(A)\), but it is also common to use the notation \(|A|\).

**Definition 6.4.1** (Determinant). Suppose that \(A\) is a square matrix.

- If \(A = (a)\) is a \(1 \times 1\) matrix, then \(\det(A) = a\).
- If \(A\) is an \(n \times n\) matrix, with \(n > 1\), the **minor** \(M_{ij}\) is the determinant of the \((n-1) \times (n-1)\) submatrix of \(A\) obtained by deleting the \(i\)-th row and \(j\)-th column of the matrix \(A\).
- The **cofactor** \(A_{ij}\) associated with \(M_{ij}\) is defined by \(A_{ij} = (-1)^{i+j}M_{ij}\).
- The **determinant** of the \(n \times n\) matrix \(A\), when \(n > 1\), is given either by

\[
\det(A) = \sum_{j=1}^{n} a_{ij}A_{ij}, \quad \text{for any } i = 1, 2, \cdots, n,
\]

or

\[
\det(A) = \sum_{i=1}^{n} a_{ij}A_{ij}, \quad \text{for any } j = 1, 2, \cdots, n.
\]

It can be shown that to calculate the determinant of a general \(n \times n\) matrix by this definition requires \(O(n!)\) multiplications/divisions and addition/subtractions. Even for relatively small values of \(n\), the number of calculations becomes unwieldy.

Although it appears that there are \(2n\) different definition of \(\det(A)\), depending on which row or column is chosen, all definitions give the same numerical result. The flexibility in the definition is used in the following example. It is most convenient to compute \(\det(A)\) across the row or down the column with the most zeros.
Example 6.4.2. Find the determinant of the matrix

\[
A = \begin{pmatrix}
2 & -1 & 3 & 0 \\
4 & -2 & 7 & 0 \\
-3 & -4 & 1 & 5 \\
6 & -6 & 8 & 0
\end{pmatrix}
\]

using the row or column with the most zero entries.

Solution. To compute \( \det(A) \), it is easiest to use the fourth column:

\[
\det(A) = a_{14}A_{14} + a_{24}A_{24} + a_{34}A_{34} + a_{44}A_{44} = 5A_{34} = -5M_{34}.
\]

Eliminating the third row and the fourth column gives

\[
\det(A) = -5 \det \begin{pmatrix}
2 & -1 & 3 \\
4 & -2 & 7 \\
6 & -6 & 8
\end{pmatrix} = -30.
\]

The following properties are useful in relating linear systems and Gaussian elimination to determinants.

Theorem 6.4.3. Suppose that \( A \) is an \( n \times n \) matrix.

- If any row or column of \( A \) has only zero entries, then \( \det(A) = 0 \).
- If \( A \) has two rows or two columns the same, then \( \det(A) = 0 \).
- If \( \tilde{A} \) is obtained from \( A \) by the operation \((E_i) \leftrightarrow (E_j)\), with \( i \neq j \), then \( \det(\tilde{A}) = -\det(A) \).
- If \( \tilde{A} \) is obtained from \( A \) by the operation \((\lambda E_i) \leftrightarrow (E_i)\), then \( \det(\tilde{A}) = \lambda \det(A) \).
- If \( \tilde{A} \) is obtained from \( A \) by the operation \((E_i + \lambda E_j) \leftrightarrow (E_i)\), with \( i \neq j \), then \( \det(\tilde{A}) = \det(A) \).
- If \( B \) is also an \( n \times n \) matrix, then \( \det(AB) = \det(A) \det(B) \).
- \( \det(A^T) = \det(A) \).
- When \( A^{-1} \) exists, \( \det(A^{-1}) = (\det(A))^{-1} \).
- If \( A \) is an upper triangular, lower triangular, or diagonal matrix, then \( \det(A) = \prod_{i=1}^{n} a_{ii} \).

The determinant of a triangular matrix is simply the product of its diagonal elements. By employing the row operations we can reduce a given square matrix to triangular form to find its determinant.

Example 6.4.4. One can compute the determinant of the matrix

\[
A = \begin{pmatrix}
2 & 1 & -1 & 1 \\
1 & 1 & 0 & 3 \\
-1 & 2 & 3 & -1 \\
3 & -1 & -1 & 2
\end{pmatrix}
\]

and find out that \( \det(A) = 39 \) using the properties of determinant and row operations.
6.5. MATRIX FACTORIZATION

The key result relating non-singularity, Gaussian elimination, linear systems, and determinants is that the following statements are equivalent.

**Theorem 6.4.5.** The following statements are equivalent for any $n \times n$ matrix $A$.

- The equation $Ax = 0$ has the unique solution $x = 0$.
- The system $Ax = b$ has a unique solution for any $n$-dimensional column vector $b$.
- The matrix $A$ is nonsingular; that is, $A^{-1}$ exists.
- $\det(A) \neq 0$.
- Gaussian elimination with row interchanges can be performed on the system $Ax = b$ for any $n$-dimensional column vector $b$.

### 6.5 Matrix Factorization

Gaussian elimination is the principal direct technique to solve linear systems of equations. In this section, we see that the steps used to solve a system of the form $Ax = b$ can be used to factor a matrix. The factorization is particularly useful when it has the form $A = LU$, where $L$ is lower triangular and $U$ is upper triangular.

Previously, we found that Gaussian elimination applied to an arbitrary linear system requires $O(n^3)$ arithmetic operations to determine its solution. However, to solve a linear system that involves an upper-triangular system requires only backward substitution, which takes $O(n^2)$ operations. The number of operations required to solve a lower-triangular system is similar.

Suppose that $A$ has been factored into the triangular form $A = LU$, where $L$ is lower triangular and $U$ is upper triangular. Then, we can solve for $x$ more easily by using a two-step process:

- First, let $y = Ux$ and solve the lower triangular system $Ly = b$ for $y$. Since $L$ is lower-triangular, determining $y$ from this equation requires only $O(n^2)$ operations.
- Once $y$ is known, the upper triangular system $Ux = y$ requires only an additional $O(n^2)$ operations to determine the solution $x$.

Therefore, solving a linear system $Ax = b$ in factored form means that the number of operations needed to solve the system is reduced from $O(n^3/3)$ to $O(2n^2)$. Not surprisingly, the reductions from the factorization come at a cost; determining the specific matrices $L$ and $U$ requires $O(n^3/3)$ operations. But once the factorization is determined, systems involving the matrix $A$ can be solved in this simplified manner for any number of vectors $b$.

To see which matrices have an $LU$ factorization and to find how it is determined, first suppose that Gaussian elimination can be performed on the system $Ax = b$ without row interchanges. This is equivalent to having nonzero pivot elements $a_{ii}^{(i)}$ for each $i = 1, 2, \cdots, n$.

The first step in the Gaussian elimination process consists of performing, for each $j = 2, 3, \cdots, n$, the operations

$$ (E_j - m_{j,1}E_1) \rightarrow (E_j), \quad \text{where} \quad m_{j,1} = \frac{a_{j1}}{a_{11}}. \quad (6.11) $$

These operations transform the system into one in which all the entries in the first column below the diagonal are zero. The system of operations in (6.11) can be viewed in another way. It is
simultaneously accomplished by multiplying the original matrix $A$ on the left by the matrix

$$
M^{(1)} := \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
-m_{21} & 1 & 0 & \cdots & 0 \\
-m_{31} & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-m_{n1} & 0 & 0 & \cdots & 1
\end{pmatrix}.
$$

This matrix is called the first Gaussian transformation matrix. We denote $A^{(0)} = A$, $A^{(1)} = M^{(1)}A^{(0)}$, and

$$
A^{(1)}x = M^{(1)}A^{(0)}x = M^{(1)}Ax = M^{(1)}b = b^{(1)}.
$$

In a similar manner, we construct $M^{(2)}$, the identity matrix with the entries below the diagonal in the second column replaced by the negatives of the multipliers

$$
m_{j2} = \frac{a_{j2}^{(1)}}{a_{22}^{(1)}}.
$$

We denote $A^{(2)} = M^{(2)}A^{(1)}$ and $b^{(2)} = M^{(2)}b^{(1)}$. Then, we have $A^{(2)}x = b^{(2)}$. In general, with $A^{(k)}x = b^{(k)}$ already formed, multiply by the $k$-th Gaussian transformation matrix

$$
M^{(k)} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 1 & \ddots & \vdots & \vdots & \vdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & \cdots & 0 & 1 & \ddots & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & 0 & \ddots \\
0 & \cdots & 0 & -m_{n,k} & 0 & \cdots & 0 & 1
\end{pmatrix},
$$

to obtain

$$
A^{(k)}x = M^{(k)}A^{(k-1)}x = M^{(k)} \cdots M^{(1)}Ax = M^{(k)}b^{(k-1)} = b^{(k)} = M^{(k)} \cdots M^{(1)}b.
$$

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

$$
A^{(n)} = \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & a_{nn}^{(n-1)}
\end{pmatrix}
$$

given by

$$
A^{(n)} = M^{(n-1)}M^{(n-2)} \cdots M^{(1)}A.
$$

This process forms the $U = A^{(n)}$ portion of the matrix factorization $A = LU$. To determine the complementary lower triangular matrix $L$, first recall the multiplication of $A^{(k)}x = b^{(k)}$ by the Gaussian transformation of $M^{(k)}$ used to obtain (6.12):

$$
A^{(k)}x = M^{(k)}A^{(k-1)}x = M^{(k)}b^{(k-1)} = b^{(k)}.
$$
where \( M^{(k)} \) generates the row operations
\[
(E_j - m_{jk}E_k) \rightarrow (E_j) \quad \text{for} \quad j = k + 1, \ldots, n.
\]

To reverse the effects of this transformation and return to \( A^{(k)} \) requires that the operations
\[
(E_j + m_{jk}E_k) \rightarrow (E_j)
\]
be performed for each \( j = k + 1, \ldots, n \). This is equivalent to multiplying by the inverse of the matrix \( M^{(k)} \), the matrix
\[
L^{(k)} = (M^{(k)})^{-1} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 1 & \ddots & & & & \vdots \\
\vdots & \ddots & \ddots & \ddots & & & \vdots \\
\vdots & & 0 & \ddots & \ddots & & \vdots \\
\vdots & & \vdots & m_{k+1,k} & \ddots & & \vdots \\
\vdots & & \vdots & \vdots & \ddots & 0 & \vdots \\
0 & \cdots & 0 & m_{n,k} & \cdots & 0 & 1
\end{pmatrix}
\]

The lower triangular matrix \( L \) in the factorization of \( A \) is the product of the matrices \( L^{(k)} \):
\[
L = L^{(1)}L^{(2)} \cdots L^{(n-1)} = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
m_{21} & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
m_{n1} & m_{n2} & \cdots & 1
\end{pmatrix},
\]
since the product of \( L \) with the upper triangular matrix \( U = M^{(n-1)}M^{(n-2)} \cdots M^{(1)}A \) gives
\[
LU = L^{(1)}L^{(2)} \cdots L^{(n-1)}M^{(n-2)} \cdots M^{(1)}A = A.
\]

We summarize the result above in the following theorem.

**Theorem 6.5.1** (LU-factorization). If Gaussian elimination can be performed on the linear system \( Ax = b \) without row interchanges, then the matrix \( A \) can be factored into the product of a lower triangular matrix \( L \) and an upper triangular matrix \( U \), that is, \( A = LU \), with
\[
m_{ji} = a_{ji}^{(i)}/a_{ii}^{(i)}.
\]

**Example 6.5.2.** (a) Determine the \( LU \) factorization for matrix \( A \), where
\[
A = \begin{pmatrix}
1 & 0 & 3 \\
2 & 1 & -1 & 1 \\
3 & -1 & -1 & 2 \\
-1 & 2 & 3 & -1
\end{pmatrix}.
\]

(b) Use the factorization to solve \( Ax = b \) with
\[
b = \begin{pmatrix}
8 \\
7 \\
14 \\
-7
\end{pmatrix}.
\]
Solution. (a) We see that the sequence of operations with the multipliers $m_{21} = 2$, $m_{31} = 3$, $m_{41} = -1$, $m_{32} = 4$, $m_{42} = -3$, and $m_{43} = 0$,

$$
(E_2 - 2E_1) \rightarrow (E_2),
(E_3 - 3E_1) \rightarrow (E_3),
(E_4 - (-1)E_1) \rightarrow (E_4),
(E_3 - 4E_2) \rightarrow (E_3),
(E_4 - (-3)E_2) \rightarrow (E_4),
$$

converts the matrix $A$ to the upper triangular one:

$$
U = \begin{pmatrix}
1 & 1 & 0 & 3 \\
0 & -1 & -1 & -5 \\
0 & 0 & 3 & 13 \\
0 & 0 & 0 & -13 \\
\end{pmatrix}.
$$

The multipliers $m_{ij}$ and the upper triangular matrix produce the factorization:

$$
A = \begin{pmatrix}
1 & 1 & 0 & 3 \\
2 & 1 & -1 & 1 \\
3 & -1 & -1 & 2 \\
-1 & 2 & 3 & -1 \\
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 \\
3 & 4 & 1 & 0 \\
-1 & -3 & 0 & 1 \\
\end{pmatrix} \begin{pmatrix}
1 & 1 & 0 & 3 \\
0 & -1 & -1 & -5 \\
0 & 0 & 3 & 13 \\
0 & 0 & 0 & -13 \\
\end{pmatrix} = LU.
$$

(b) To solve the system

$$
Ax = LUx = b,
$$

we first introduce the substitution $y = Ux$. Then $b = L(Ux) = Ly$. That is,

$$
Ly = \begin{pmatrix}
1 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 \\
3 & 4 & 1 & 0 \\
-1 & -3 & 0 & 1 \\
\end{pmatrix} \begin{pmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4 \\
\end{pmatrix} = b = \begin{pmatrix}
8 \\
7 \\
14 \\
-7 \\
\end{pmatrix}.
$$

This system is solved for $y$ by a simple forward substitution process:

$$
y_1 = 8;
y_2 = 7 - 2y_1 = -9;
y_3 = 14 - 3y_1 - 4y_2 = 26;
y_4 = -7 + y_1 + 3y_2 = -26.
$$

We then solve $Ux = y$ for $x$, the solution of the original system; that is,

$$
\begin{pmatrix}
1 & 1 & 0 & 3 \\
0 & -1 & -1 & -5 \\
0 & 0 & 3 & 13 \\
0 & 0 & 0 & -13 \\
\end{pmatrix} \begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
\end{pmatrix} = \begin{pmatrix}
8 \\
-9 \\
26 \\
-26 \\
\end{pmatrix}.
$$

Using back substitution we obtain $x_4 = 2, x_3 = 0, x_2 = -1$, and $x_1 = 3$.

The factorization used in the example above is called Doolittle’s method and requires that 1’s be on the diagonal of $L$, which results in the factorization described above. Later on, we consider
6.5. MATRIX FACTORIZATION

Crout’s method, a factorization which requires that 1’s be on the diagonal elements of $U$, and Cholesky’s method, which requires that $\ell_{ii} = u_{ii}$ for each $i$.

Once the matrix factorization is complete, the solution to a linear system of the form $Ax = LUx = b$ is found by first letting $y = Ux$ and solving $Ly = b$ for $y$. Since $L$ is lower triangular, we have

$$y_1 = \frac{b_1}{\ell_{11}} \quad \text{and} \quad y_i = \frac{1}{\ell_{ii}} \left( b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j \right) \quad \text{for} \quad i = 2, 3, \cdots, n.$$  

After $y$ is found by this forward substitution process, the upper triangular system $Ux = y$ is solved for $x$ by back substitution using the equations

$$x_n = \frac{y_n}{u_{nn}} \quad \text{and} \quad x_i = \frac{1}{u_{ii}} \left( y_i - \sum_{j=i+1}^{n} u_{ij} x_j \right) \quad \text{for} \quad i = n-1, n-1, \cdots, 1.$$  

**Permutation matrices**

Previously, we assumed that $Ax = b$ can be solved using Gaussian elimination without row interchanges. For a practical standpoint, this factorization is useful only when row interchanges are not required to control the round-off error resulting from the use of finite-digit arithmetic. We begin the discussion with the introduction of a class of matrices that are used to rearrange, or permute, rows of a given matrix.

An $n \times n$ permutation matrix $P = (p_{ij})$ is a matrix obtained by rearranging the rows of the identity matrix $I_n$ of order $n$. This gives a matrix with precisely one nonzero entry in each row and in each column, and each nonzero entry is a 1.

**Example 6.5.3.** The matrix

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$  

is a $3 \times 3$ permutation matrix. For any $3 \times 3$ matrix $A$, multiplying on the left by $P$ has the effect of interchanging the second and third rows of $A$:

$$PA = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}.$$  

Similarly, multiplying $A$ on the right by $P$ interchanges the second and third columns of $A$.

Two useful properties of permutation matrices relate to Gaussian elimination, the first of which is illustrated in the previous example. Suppose $k_1, \cdots, k_n$ is a permutation of the integers $1, \cdots, n$ and the permutation matrix $P = (p_{ij})$ is defined by

$$p_{ij} = \begin{cases} 1, & \text{if } j = k_i, \\ 0, & \text{otherwise}. \end{cases}$$  

Then, (i) $PA$ permutes the rows of $A$; that is,

$$PA = \begin{pmatrix} a_{k_11} & a_{k_12} & \cdots & a_{k_1n} \\ a_{k_21} & a_{k_22} & \cdots & a_{k_2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k_n1} & a_{k_n2} & \cdots & a_{k_nn} \end{pmatrix}.$$
and $P^{-1}$ exists and $P^{-1} = P^T$. For any nonsingular matrix $A$, the linear system $Ax = b$ can be solved by Gaussian elimination, with the possibility of row interchanges. If we knew the row interchanges that were required to solve the system by Gaussian elimination, we could arrange the original equations in an order that would ensure that now row interchanges are needed. Hence, there is a rearrangement of the equations in the system that permits Gaussian elimination to proceed without row interchanges. This implies that for any nonsingular matrix $A$, a permutation matrix $P$ exists for which the system $PAx = Pb$ can be solved without row interchanges. As a consequence, the matrix $PA$ can be factored into $PA = LU$, where $L$ is lower triangular and $U$ is upper triangular. Because $P^{-1} = P^T$, this produces the factorization

$$A = P^{-1}LU = P^TLU.$$  

The matrix $U$ is still upper triangular, but $P^TL$ is not lower triangular unless $P = I_n$.

**Example 6.5.4.** Determine a factorization in the form $A = P^TLU$ for the matrix

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

**Solution.** The matrix $A$ cannot have an $LU$ factorization since $a_{11} = 0$. However, using the row interchange $(E_1) \leftrightarrow (E_2)$, followed by $(E_3 - (-1)E_1) \rightarrow (E_3)$ and $(E_4 - E_1) \rightarrow (E_4)$ produces

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

Then the row interchange $(E_2) \leftrightarrow (E_4)$, followed by $(E_4 - (-1)E_3) \rightarrow (E_4)$, give the matrix

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

The permutation matrix associated with the row interchanges $(E_1) \leftrightarrow (E_2)$ and $(E_2) \leftrightarrow (E_4)$ is

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

and

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

Gaussian elimination is performed on $PA$ using the same operations as on $A$, except without the row interchanges. That is, $(E_2 - E_1) \rightarrow (E_2)$, $(E_3 - (-1)E_1) \rightarrow (E_3)$, followed by $(E_4 - (-1)E_3) \rightarrow (E_4)$. The nonzero multipliers for $PA$ are consequently, $m_{21} = 1$, $m_{31} = -1$, and $m_{43} = -1$.  

**CHAPTER 6. DIRECT METHODS FOR SOLVING LINEAR SYSTEMS**
and the \( LU \) factorization of \( PA \) is
\[
PA = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix} = LU.
\]

Multiplying by \( P^{-1} = P^T \) produces the factorization
\[
A = P^{-1}LU = P^T LU = \begin{pmatrix} 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}.
\]

\[\square\]

### 6.6 Special Types of Matrices

We now turn attention to two classes of matrices for which Gaussian elimination can be performed effectively without row interchanges.

**Diagonally Dominant Matrices**

The first class is described in the following definition.

**Definition 6.6.1.** The \( n \times n \) matrix \( A \) is said to be **diagonally dominant** when
\[
|a_{ii}| \geq \sum_{j=1, j \neq i}^{n} |a_{ij}| \quad \text{hold for each } i = 1, 2, \ldots, n.
\]

A diagonally dominant matrix is said to be **strictly diagonally dominant** when the inequality above is strict for each \( n \), that is,
\[
|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}| \quad \text{hold for each } i = 1, 2, \ldots, n.
\]

**Example 6.6.2.** Consider the matrices
\[
A = \begin{pmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & 5 & -6 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{pmatrix}.
\]

The non-symmetric matrix \( A \) is strictly diagonally dominant because
\[
|7| > |2| + |0|, \quad |5| > |3| + |-1|, \quad \text{and} \quad |-6| > |0| + |5|.
\]

The symmetric matrix \( B \) is not strictly diagonally dominant because, for example, in the first row the absolute value of the diagonal element is \(|6| < |4| + |-3| = 7\). It is interesting to note that \( A^T \) is not strictly diagonally dominant, because the middle row of \( A^T \) is \((2 \ 5 \ 5)\), nor, of course, is \( B^T \) because \( B^T = B \).

We have the following result for strictly diagonally dominant matrices.

**Theorem 6.6.3.** A strictly diagonally dominant matrix \( A \) is nonsingular. Moreover, in this case, Gaussian elimination can be performed on any linear system of the form \( Ax = b \) to obtain its unique solution without row or column interchanges.
CHAPTER 6. DIRECT METHODS FOR SOLVING LINEAR SYSTEMS

Positive Definite Matrices

The next special class of matrices is called positive definite.

**Definition 6.6.4 (Positive definite).** A square matrix $A$ of order $n$ is called positive definite if $x^T A x > 0$ for every $n$-dimensional vector $x \neq 0$. To be precise, it specifies that the $1 \times 1$ matrix generated by the operation $x^T A x$ has a positive value for its only entry since the operation is performed as follows:

$$x^T A x = (x_1 \ x_2 \ \cdots \ x_n) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= (x_1 \ x_2 \ \cdots \ x_n) \begin{pmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \sum_{j=1}^{n} a_{2j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj} x_j \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \end{pmatrix}.$$

**Example 6.6.5.** The matrix

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

is positive definite. Suppose $x$ is any three-dimensional column vector. Then

$$x^T A x = (x_1 \ x_2 \ x_3) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$= (x_1 \ x_2 \ x_3) \begin{pmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 \end{pmatrix}$$

$$= 2x_1^2 - 2x_1 x_2 + 2x_2^2 - 2x_2 x_3 + 2x_3^2.$$

Rearranging the terms gives

$$x^T A x = x_1^2 + (x_1^2 - 2x_1 x_2 + x_2^2) + (x_2^2 - 2x_2 x_3 + x_3^2) + x_3^2$$

$$= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2,$$

which implies that $x^T A x > 0$ unless $x_1 = x_2 = x_3 = 0$. Moreover, it is symmetric.

It should be clear from the example that using the definition to determine if a matrix is positive definite can be difficult. The next result provides some necessary conditions that can be used to eliminate certain matrices from consideration.

**Theorem 6.6.6.** If $A$ is an $n \times n$ positive definite matrix, then

(i) $A$ has an inverse;

(ii) $a_{ii} > 0$ for each $i = 1, 2, \ldots, n$.

(iii) $\max_{1 \leq k, j \leq n} |a_{kj}| \leq \max_{1 \leq i \leq n} |a_{ii}|$;

(iv) $(a_{ij})^2 < a_{ii}a_{jj}$ for each $i \neq j$. 
Theorem above provides some important conditions that must be true of positive definite matrices, it does not ensure that a matrix satisfying these conditions is positive definite. The following notion will be used to provide a necessary and sufficient condition.

**Definition 6.6.7.** A **leading principal submatrix** of a matrix $A$ is a matrix of the form

$$A_k = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}$$

for some $1 \leq k \leq n$.

We have the following result.

**Theorem 6.6.8.** A symmetric matrix $A$ is positive definite if and only if each of its leading principal submatrices has a positive determinant.

**Example 6.6.9.** In Example 6.6.5, we used the definition to show that the symmetric matrix

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

is positive definite. Confirm this using Theorem 6.6.8.

**Solution.** Note that $\det(A_1) = 2 > 0$,

$$\det(A_2) = \det \left( \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \right) = 4 - 1 = 3 > 0,$$

and

$$\det(A_3) = \det \left( \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \right) = 2 \det \left( \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \right) - (-1) \det \left( \begin{pmatrix} -1 & -1 \\ 0 & 2 \end{pmatrix} \right)$$

$$= 2(4 - 1) + (-2 + 0) = 4 > 0.$$

It is in agreement with Theorem 6.6.8.

**Theorem 6.6.10.** A symmetric matrix $A$ is positive definite if and only if Gaussian elimination without row interchanges can be performed on the linear system $Ax = b$ with all pivot elements positive. Moreover, in this case, the computations are stable with respect to the growth of round-off errors.

Some interesting facts that are uncovered in the constructing the proof of Theorem 6.6.10 are presented in the following corollaries.

**Corollary 6.6.11.** The matrix $A$ is positive definite if and only if $A$ can be factored in the form $A = LDL^T$, where $L$ is lower triangular with 1’s on its diagonal and $D$ is a diagonal matrix with positive diagonal entries.

**Corollary 6.6.12 (Cholesky factorization).** The matrix $A$ is positive definite if and only if $A$ can be factored in the form $A = LL^T$, where $L$ is lower triangular with nonzero diagonal entries.
Note that the matrix $L$ in two corollaries are not equal. Moreover, we have the following result when $A$ is symmetric (may not necessarily be positive definite).

**Corollary 6.6.13.** Let $A$ be a symmetric $n \times n$ matrix for which Gaussian elimination can be applied without row interchanges. Then $A$ can be factored into the form of $A = LDL^T$, where $L$ is lower triangular with 1’s on its diagonal and $D$ is the diagonal matrix with $a_{11}^{(1)}, \ldots, a_{nn}^{(1)}$ on its diagonal.

**Example 6.6.14.** Determine the $LDL^T$ factorization of the positive definite matrix:

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix}.$$

**Solution.** The $LDL^T$ factorization has 1’s on the diagonal of the lower triangular matrix $L$ so we need to have

$$A = \begin{pmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} \begin{pmatrix} 1 & \ell_{21} & \ell_{31} \\ 0 & 1 & \ell_{32} \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} d_1 \ell_{21} & d_1 \ell_{31} \\ d_1 \ell_{21} & d_2 + d_1 \ell_{21}^2 & d_1 \ell_{21} \ell_{31} + d_1 \ell_{21} \ell_{32} + d_1 \ell_{21}^2 + d_3 \\ d_1 \ell_{31} & d_2 \ell_{32} + d_1 \ell_{21} \ell_{31} & d_1 \ell_{31}^2 + d_2 \ell_{32}^2 + d_3 \end{pmatrix}.$$  

Thus, we have

$a_{11} : 4 = d_1 \implies d_1 = 4, \quad a_{21} : -1 = d_1 \ell_{21} \implies \ell_{21} = -0.25,$

$a_{31} : 1 = d_1 \ell_{31} \implies \ell_{31} = 0.25, \quad a_{22} : 4.25 = d_2 + d_1 \ell_{21}^2 \implies d_2 = 4,$

$a_{32} : 2.75 = d_2 \ell_{32} + d_1 \ell_{21} \ell_{31} \implies \ell_{32} = 0.75, \quad a_{33} : 3.5 = d_1 \ell_{31}^2 + d_2 \ell_{32}^2 + d_3 \implies d_3 = 1.$

Therefore, we have

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -0.25 & 1 & 0 \\ 0.25 & 0.75 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -0.25 & 0.25 \\ 0 & 1 & 0.75 \\ 0 & 0 & 1 \end{pmatrix}.$$

\[ \square \]

**Example 6.6.15.** Determine the Cholesky $LL^T$ factorization of the positive definite matrix:

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix}.$$

**Solution.** The $LL^T$ factorization does not necessarily has 1’s on the diagonal of the lower triangular matrix $L$ so we need to have

$$A = \begin{pmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{pmatrix} \begin{pmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ 0 & \ell_{22} & \ell_{32} \\ 0 & 0 & \ell_{33} \end{pmatrix} = \begin{pmatrix} \ell_{11}^2 & \ell_{11} \ell_{21} & \ell_{11} \ell_{31} \\ \ell_{11} \ell_{21} & \ell_{21}^2 + \ell_{22}^2 & \ell_{21} \ell_{31} + \ell_{21} \ell_{32} + \ell_{21} \ell_{31} \\ \ell_{11} \ell_{31} & \ell_{21} \ell_{32} + \ell_{21} \ell_{31} & \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 \end{pmatrix}.$$  

Thus, we have
\[ a_{11} : 4 = \ell_{11}^2 \implies \ell_{11} = 2, \quad a_{21} : -1 = \ell_{11} \ell_{21} \implies \ell_{21} = -0.5, \]
\[ a_{31} : 1 = \ell_{11} \ell_{31} \implies \ell_{31} = 0.5, \quad a_{22} : 4.25 = \ell_{22}^2 + \ell_{21}^2 \implies \ell_{22} = 2, \]
\[ a_{32} : 2.75 = \ell_{22} \ell_{32} + \ell_{21} \ell_{31} \implies \ell_{32} = 1.5, \quad a_{33} : 3.5 = \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 \implies \ell_{33} = 1 \]
and
\[ A = \begin{pmatrix} 2 & 0 & 0 \\ -0.5 & 2 & 0 \\ 0.5 & 1.5 & 1 \end{pmatrix} \begin{pmatrix} 2 & -0.5 & 0.5 \\ 0 & 2 & 1.5 \\ 0 & 0 & 1 \end{pmatrix}. \]

**Tridiagonal matrices**

The last class of matrices considered are **tridiagonal matrices** having the form
\[
A = \begin{pmatrix}
    a_{11} & a_{12} & 0 & \cdots & \cdots & 0 \\
    a_{21} & a_{22} & a_{23} & \ddots & & \\
    0 & a_{32} & a_{33} & \ddots & \ddots & \\
    \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\
    \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
    0 & \cdots & \cdots & 0 & a_{n-1,n} & a_{nn}
\end{pmatrix}.
\]

Tridiagonal matrices are also considered in connection with the study of piecewise linear approximations to boundary-value problems. The factorizations can be simplified considerably in the case of tridiagonal matrices because a large number of zeros appear in these matrices in regular patterns.

To illustrate the situation, suppose a tridiagonal matrix \( A \) can be factored into the triangular matrices \( L \) and \( U \). Then \( A \) has at most \( 3n - 2 \) nonzero entries. Then there are only \( 3n - 2 \) conditions to be applied to determine the entries of \( L \) and \( U \), provided that the zero entries of \( A \) are also obtained.

Suppose that the matrices \( L \) and \( U \) also have tridiagonal form, that is,
\[
L = \begin{pmatrix}
    \ell_{11} & 0 & 0 & \cdots & \cdots & 0 \\
    \ell_{21} & \ell_{22} & 0 & \ddots & & \\
    0 & \ell_{32} & \ell_{33} & 0 & \ddots & \\
    \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\
    \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
    0 & \cdots & \cdots & 0 & \ell_{n,n-1} & \ell_{nn}
\end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix}
    1 & u_{12} & 0 & \cdots & \cdots & 0 \\
    0 & 1 & u_{23} & \ddots & & \\
    0 & 0 & 1 & u_{34} & \ddots & \\
    \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\
    \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
    0 & \cdots & \cdots & 0 & u_{n-1,n} & 1
\end{pmatrix}.
\]

There are \( 2n - 1 \) undetermined entries of \( L \) and \( n - 1 \) undetermined entries of \( U \), which totals \( 3n - 2 \), the number of possible nonzero entries of \( A \). The 0 entries of \( A \) are obtained automatically.

The multiplication involved with \( A = LU \) gives, in addition to the 0 entries,
\begin{equation}
\begin{align*}
    a_{11} &= \ell_{11}; \\
    a_{i,i-1} &= \ell_{i,i-1}, \quad \text{for } i = 2, 3, \ldots, n; \\
    a_{ii} &= \ell_{i,i-1} u_{i-1,i} + \ell_{ii}, \quad \text{for } i = 2, 3, \ldots, n; \\
    a_{i,i+1} &= \ell_{ii} u_{i,i+1}, \quad \text{for } i = 1, 2, \ldots, n - 1.
\end{align*}
\end{equation}
This is called the Crout factorization of the tridiagonal matrix.

Example 6.6.16. Determine the Crout factorization of the (symmetric) tridiagonal matrix

\[
A = \begin{pmatrix}
  2 & -1 & 0 & 0 \\
  -1 & 2 & -1 & 0 \\
  0 & -1 & 2 & -1 \\
  0 & 0 & -1 & 2
\end{pmatrix}.
\]

Use this factorization to solve the linear system \(Ax = b\) with

\[
b = \begin{pmatrix}
  1 \\
  0 \\
  0 \\
  1
\end{pmatrix}.
\]

Solution. The Crout factorization of \(A\) has the form

\[
A = \begin{pmatrix}
  a_{11} & a_{12} & 0 & 0 \\
  a_{21} & a_{22} & a_{23} & 0 \\
  0 & a_{32} & a_{33} & a_{34} \\
  0 & 0 & a_{43} & a_{44}
\end{pmatrix} = \begin{pmatrix}
  \ell_{11} & 0 & 0 & 0 \\
  \ell_{21} & \ell_{22} & 0 & 0 \\
  0 & \ell_{32} & \ell_{33} & 0 \\
  0 & 0 & \ell_{43} & \ell_{44}
\end{pmatrix} \begin{pmatrix}
  1 & u_{12} & 0 & 0 \\
  0 & 1 & u_{23} & 0 \\
  0 & 0 & 1 & u_{34} \\
  0 & 0 & 0 & 1
\end{pmatrix}
\]

Thus, we have

\[
a_{11} : 2 = \ell_{11} \implies \ell_{11} = 2, \quad a_{12} : -1 = \ell_{11}u_{12} \implies u_{12} = -\frac{1}{2},
\]

\[
a_{21} : -1 = \ell_{21} \implies \ell_{21} = -1, \quad a_{22} : 2 = \ell_{22} + \ell_{21}u_{12} \implies \ell_{22} = -\frac{3}{2},
\]

\[
a_{23} : -1 = \ell_{22}u_{23} \implies u_{23} = -\frac{2}{3}, \quad a_{32} : -1 = \ell_{32} \implies \ell_{32} = -1,
\]

\[
a_{33} : 2 = \ell_{33} + \ell_{32}u_{23} \implies \ell_{33} = \frac{4}{3}, \quad a_{34} : -1 = \ell_{33}u_{34} \implies u_{34} = -\frac{3}{4},
\]

\[
a_{43} : -1 = \ell_{43} \implies \ell_{43} = -1, \quad a_{44} : 2 = \ell_{44} + \ell_{43}u_{34} \implies \ell_{44} = -\frac{5}{4}.
\]

This gives the Crout factorization

\[
A = \begin{pmatrix}
  2 & 0 & 0 & 0 \\
  -1 & \frac{3}{2} & 0 & 0 \\
  0 & -1 & \frac{4}{3} & 0 \\
  0 & 0 & -1 & \frac{5}{4}
\end{pmatrix} \begin{pmatrix}
  1 & -\frac{1}{2} & 0 & 0 \\
  0 & 1 & -\frac{3}{2} & 0 \\
  0 & 0 & 1 & -\frac{3}{4} \\
  0 & 0 & 0 & 1
\end{pmatrix} = LU.
\]

Solving the system

\[
Lz = \begin{pmatrix}
  2 & 0 & 0 & 0 \\
  -1 & \frac{3}{2} & 0 & 0 \\
  0 & -1 & \frac{4}{3} & 0 \\
  0 & 0 & -1 & \frac{5}{4}
\end{pmatrix} \begin{pmatrix}
  z_1 \\
  z_2 \\
  z_3 \\
  z_4
\end{pmatrix} = \begin{pmatrix}
  1 \\
  0 \\
  0 \\
  1
\end{pmatrix}
\]

gives

\[
\begin{pmatrix}
  z_1 \\
  z_2 \\
  z_3 \\
  z_4
\end{pmatrix} = \begin{pmatrix}
  \frac{1}{4} \\
  \frac{3}{4} \\
  \frac{1}{4} \\
  \frac{1}{4}
\end{pmatrix}.
\]
and then solving
\[
U x = \begin{pmatrix}
1 & -\frac{1}{2} & 0 & 0 \\
0 & 1 & -\frac{2}{3} & 0 \\
0 & 0 & 1 & -\frac{3}{4} \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix}
= \begin{pmatrix}
\frac{1}{2} \\
\frac{1}{3} \\
\frac{1}{4} \\
1
\end{pmatrix}
gives \begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix} = \begin{pmatrix}
1 \\
1 \\
1 \\
1
\end{pmatrix}.
\]

\[\square\]

### 6.7 Condition of a linear system

Recall that \(\mathbb{R}^n\) denote the set of all \(n\)-dimensional column vectors with real-number components. To define a distance in \(\mathbb{R}^n\) we use the notion of a norm, which is the generalization of the absolute on \(\mathbb{R}\), the set of real numbers.

**Definition 6.7.1.** A vector norm on \(\mathbb{R}^n\) is a function \(\|\cdot\|\) from \(\mathbb{R}^n\) into \(\mathbb{R}\) with the following properties:

(i) \(\|x\| \geq 0\) for all \(x \in \mathbb{R}^n\),

(ii) \(\|x\| = 0\) if and only if \(x = 0\),

(iii) \(\|\alpha x\| = |\alpha| \|x\|\) for all \(\alpha \in \mathbb{R}\) and \(x \in \mathbb{R}^n\),

(iv) \(\|x + y\| \leq \|x\| + \|y\|\) for all \(x, y \in \mathbb{R}^n\).

**Definition 6.7.2.** The \(\ell_2\) and \(\ell_\infty\) norms for vector \(x = (x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^n\) are defined by

\[\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}\] 
and \[\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| .\]

Note that each of these norms reduces to the absolute value in the case \(n = 1\). The \(\ell_2\) norm is called the **Euclidean norm** of the vector because it represents the usual notion of distance from the origin. For example, the \(\ell_2\) norm of the vector \(x = (x_1, x_2, x_3)^T\) gives the length of the straight line joining the points \((0,0,0)\) and \((x_1,x_2,x_3)\).

**Example 6.7.3.** Determine the \(\ell_2\) norm and the \(\ell_\infty\) norm of the vector \(x = (-1,1,-2)^T\).

**Solution.** The vector \(x = (-1,1,-2)^T\) in \(\mathbb{R}^3\) has norms

\[\|x\|_2 = \sqrt{(-1)^2 + 1^2 + (-2)^2} = \sqrt{6}\]

and

\[\|x\|_\infty = \max\{|-1|, |1|, |-2|\} = 2 .\]

It is easy to show that the properties in Definition 6.7.1 hold for the \(\ell_\infty\) norm because they follow from similar results for absolute values. The only property that requires much demonstration is (iv), and in this case if \(x = (x_1, x_2, \cdots, x_n)^T\) and \(y = (y_1, y_2, \cdots, y_n)^T\), then

\[\|x + y\|_\infty = \max_{1 \leq i \leq n} |x_i + y_i| \leq \max_{1 \leq i \leq n} (|x_i| + |y_i|) \max_{1 \leq i \leq n} |x_i| + \max_{1 \leq i \leq n} |y_i| = \|x\|_\infty + \|y\|_\infty .\]

The first three conditions in Definition 6.7.1 also are easy to show for the \(\ell_2\) norm. But to show (iv), we need a famous inequality.
Theorem 6.7.4 (Cauchy-Schwarz Inequality). For each $x = (x_1, x_2, \ldots, x_n)^T$ and $y = (y_1, y_2, \ldots, y_n)^T$ in $\mathbb{R}^n$,

$$x^T y = \sum_{i=1}^{n} x_i y_i \leq \left( \sum_{i=1}^{n} x_i^2 \right)^{1/2} \left( \sum_{i=1}^{n} y_i^2 \right)^{1/2} = \|x\|_2 \cdot \|y\|_2.$$  

With the Cauchy-Schwarz Inequality, we see that for $x, y \in \mathbb{R}^n$, we have

$$\|x + y\|_2^2 = \sum_{i=1}^{n} (x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \leq \|x\|_2^2 + 2 \|x\|_2 \cdot \|y\|_2 + \|y\|_2^2,$$

which gives norm property (iv):

$$\|x + y\|_2 \leq (\|x\|_2^2 + 2 \|x\|_2 \cdot \|y\|_2 + \|y\|_2^2)^{1/2} = \|x\|_2 + \|y\|_2.$$  

Theorem 6.7.5. For each $x \in \mathbb{R}^n$, we have

$$\|x\|_{\infty} \leq \|x\|_2 \leq \sqrt{n} \|x\|_{\infty}.$$  

Proof. Let $x_j$ be a coordinate of $x$ such that $\|x\|_{\infty} = \max_{1 \leq i \leq n} |x_i| = |x_j|$. Then,

$$\|x\|_\infty^2 = |x_j|^2 \leq \sum_{i=1}^{n} x_i^2 = \|x\|_2^2 \implies \|x\|_\infty \leq \|x\|_2.$$  

On the other hand,

$$\|x\|_2^2 = \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 = n x_j^2 = n \|x\|_{\infty}^2 \implies \|x\|_2 \leq \sqrt{n} \|x\|_{\infty}.$$  

Remark: It can be shown that all norms on $\mathbb{R}^n$ are equivalent with respect to convergence; that is, if $\|\cdot\|$ and $\|\cdot\|$' are any two norms on $\mathbb{R}^n$ and $(x^{(k)})_{k=1}^\infty$ has the limit $x$ with respect to $\|\cdot\|$, then the sequence also has the same limit with respect to another norm.

Matrix norms and distances

We need methods for determining the distance between $n \times n$ matrices. This again requires the use of norm. We first define the norm of matrix.

Definition 6.7.6 (Matrix norm). A matrix norm on the set of all $n \times n$ matrices is a real-valued function, $\|\cdot\|$, defined on this set, satisfying for all $n \times n$ matrices $A$ and $B$ and all real numbers $\alpha$:

1. $\|A\| \geq 0$;
2. $\|A\| = 0$ if and only if $A = O$, the matrix with all 0 entries;
3. $\|\alpha A\| = |\alpha| \|A\|$;
4. $\|A + B\| \leq \|A\| + \|B\|$;
5. $\|AB\| \leq \|A\| \cdot \|B\|$.
Although matrix norms can be obtained in various ways, the norms considered most frequently are those that are natural consequences of the vector norms $\ell_2$ and $\ell_\infty$. These norms are defined using the following results.

**Theorem 6.7.7.** If $\| \cdot \|_v$ is a vector norm on $\mathbb{R}^n$, we define $\| \cdot \|_M$ as follows:

$$\|A\|_M = \max_{\|x\|_v = 1} \|Ax\|_v.$$  

Then, $\| \cdot \|_M$ is a matrix norm.

Matrix norms defined by vector norms are called the **natural, or induced, matrix norm** associated with the vector norm. In this text, all matrix norms will be assumed to be natural matrix norms unless specified otherwise. For any $z \neq 0$, the vector $x = z / \|z\|_v$ is a unit vector with respect to the vector norm $\| \cdot \|_v$. Hence,

$$\max_{\|x\|_v = 1} \|Ax\|_v = \max_{z \neq 0} A \left( \frac{z}{\|z\|_v} \right)_v = \max_{z \neq 0} \|Az\|_v,$$

and we can alternatively write the matrix norm $\| \cdot \|_M$ as

$$\|A\|_M = \max_{z \neq 0} \|Az\|_v.$$  

The following result follows from this representation of $\|A\|_M$.

**Theorem 6.7.8.** For any vector $z \neq 0$, matrix $A$, and any natural matrix norm $\| \cdot \|_M$ induced by the vector norm $\| \cdot \|_v$, we have

$$\|Az\|_v \leq \|A\|_M \cdot \|z\|_v.$$  

The measure given to a matrix under a natural norm describes how the matrix stretches unit vectors relative to that norm. The maximum stretch is the norm of the matrix. The matrix norms we will consider have the forms

$$\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty,$$  

the $\ell_\infty$ norm,

and

$$\|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2,$$  

the $\ell_2$ norm.

The $\ell_\infty$ norm of a matrix can be easily computed from the entries of the matrix.

**Theorem 6.7.9.** If $A = (a_{ij})$ is an $n \times n$ matrix, then

$$\|A\|_\infty = \max_{1 \leq i \leq \ell} \sum_{j=1}^n |a_{ij}|.$$  

**Example 6.7.10.** Determine $\|A\|_\infty$ for the matrix

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 3 & -1 \\ 5 & -1 & 1 \end{pmatrix}.$$
CHAPTER 6. DIRECT METHODS FOR SOLVING LINEAR SYSTEMS

Solution. We have

\[ \sum_{j=1}^{3} |a_{1j}| = |1| + |2| + |-1| = 4, \quad \sum_{j=1}^{3} |a_{2j}| = |0| + |3| + |-1| = 4, \]

and

\[ \sum_{j=1}^{3} |a_{3j}| = |5| + |-1| + |1| = 7. \]

It implies that \( \|A\|_\infty = \max\{4, 4, 7\} = 7. \)

Perturbating a linear system

Consider the following linear system

\[
\begin{pmatrix}
10 & 7 & 8 & 7 \\
7 & 5 & 6 & 5 \\
8 & 6 & 10 & 9 \\
7 & 5 & 9 & 10
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{pmatrix} =
\begin{pmatrix}
32 \\
23 \\
33 \\
31
\end{pmatrix}
\text{ with solution }
\begin{pmatrix}
1 \\
1 \\
1 \\
1
\end{pmatrix}
\]

and consider the perturbed system, where the right-hand side has been very slightly modified, the matrix staying unchanged,

\[
\begin{pmatrix}
10 & 7 & 8 & 7 \\
7 & 5 & 6 & 5 \\
8 & 6 & 10 & 9 \\
7 & 5 & 9 & 10
\end{pmatrix}
\begin{pmatrix}
x_1 + \delta x_1 \\
x_2 + \delta x_2 \\
x_3 + \delta x_3 \\
x_4 + \delta x_4
\end{pmatrix} =
\begin{pmatrix}
32.1 \\
22.9 \\
33.1 \\
30.9
\end{pmatrix}
\text{ with solution }
\begin{pmatrix}
9.2 \\
-12.6 \\
4.5 \\
-1.1
\end{pmatrix}.
\]

In order words, a relative error of the order of \(1/200\) in the data (here, the components of the right-hand side) produces a relative error of the order of \(10/1\) in the result (the solution of the linear system), which represents an amplification of the relative errors of the order of 2000.

Let us now analyze this kind of phenomenon. Assume that \(A\) is invertible, and a comparison needs to be made between the two exact solutions \(x\) and \(x + \delta x\) of the systems:

\[
Ax = b, \quad A(x + \delta x) = b + \delta b.
\]

Let the same symbol \(|\cdot|\) denote any vector norm and its associated matrix norm. From the equalities \(b = Ax\) and \(\delta x = A^{-1}\delta b\), using Theorem 6.7.8 we conclude

\[ |\delta x| \leq \|A^{-1}\| |\delta b| \quad \text{and} \quad |b| \leq \|A\| |u|, \]

so that the relative error in the result, measured by the ratio \(|\delta u|/|u|\), is bounded in terms of the relative error \(|\delta b|/|b|\) in the datum \(b\), as follows:

\[ \frac{|\delta x|}{|x|} \leq \|A\| \cdot \|A^{-1}\| \frac{|\delta b|}{|b|}. \quad (6.14) \]

That is, the relative error in the result is bounded by the relative error in the data, multiplied by the number \(\|A\| \cdot \|A^{-1}\|\). In order words, for a given relative error in the data, the relative error in the corresponding result may be larger by a factor proportional to this number; in actual fact, it can be shown that this number is optimal, and the inequalities above being the best possible. These considerations lead to the following definition.
6.7. CONDITION OF A LINEAR SYSTEM

Definition 6.7.11 (Condition number). Let \( \| \cdot \| \) be a matrix norm associated with a vector norm and let \( A \) be invertible. The number

\[
\text{cond}(A) := \|A\| \|A^{-1}\|
\]

is called the condition number of the matrix \( A \), relative to the given matrix norm.

The inequality (6.14) shows that the number \( \text{cond}(A) \) measures the sensitivity of the solution of the linear system to variations in the data; a feature which is referred to as the condition of the linear system in question. The preceding, therefore, gives sense to a statement such as a linear system is well-conditioned or ill-conditioned, according as the condition number of its matrix is small or large.

Example 6.7.12. Consider the following matrix

\[
A = \begin{pmatrix}
10 & 7 & 8 & 7 \\
7 & 5 & 6 & 5 \\
8 & 6 & 10 & 9 \\
7 & 5 & 9 & 10
\end{pmatrix}.
\]

Calculate the condition number of \( A \) with respect to the \( \ell_\infty \) (vector) norm.

Solution. Recall Theorem 6.7.9 that

\[
\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.
\]

Then, we have

\[
\sum_{j=1}^{4} |a_{1j}| = 10 + 7 + 8 + 7 = 32, \quad \sum_{j=1}^{4} |a_{2j}| = 7 + 5 + 6 + 5 = 23,
\]

\[
\sum_{j=1}^{4} |a_{3j}| = 8 + 6 + 10 + 9 = 33, \quad \sum_{j=1}^{4} |a_{4j}| = 7 + 5 + 9 + 10 = 31.
\]

It implies that \( \|A\|_\infty = 33 \). On the other hand, we have

\[
A^{-1} = \begin{pmatrix}
25 & -41 & 10 & -6 \\
-41 & 68 & -17 & 10 \\
10 & -17 & 5 & -3 \\
-6 & 10 & -3 & 2
\end{pmatrix}
\]

and \( \|A^{-1}\|_\infty = 136 \).

Therefore, the condition number of \( A \) with respect to \( \ell_\infty \)-norm is \( \text{cond}(A) = 136 \cdot 33 = 4488 \).

Next, we recall the linear systems at the beginning of the section. Let

\[
x = \begin{pmatrix}
1 \\
1 \\
1 \\
1
\end{pmatrix}, \quad \delta x = \begin{pmatrix}
8.2 \\
-13.6 \\
3.5 \\
-2.1
\end{pmatrix}, \quad b = \begin{pmatrix}
32 \\
23 \\
33 \\
31
\end{pmatrix}, \quad \text{and} \quad \delta b = \begin{pmatrix}
0.1 \\
-0.1 \\
0.1 \\
-0.1
\end{pmatrix}.
\]

We compute the ratios:

\[
\frac{\|\delta x\|_\infty}{\|x\|_\infty} = 13.6, \quad \frac{\|\delta b\|_\infty}{\|b\|_\infty} = \frac{0.1}{33} \implies \text{cond}(A) \frac{\|\delta b\|_\infty}{\|b\|_\infty} = 136 \cdot 0.1 = 13.6
\]
and we have
\[
\frac{\|\delta x\|_\infty}{\|x\|_\infty} = \text{cond}(A) \frac{\|\delta b\|_\infty}{\|b\|_\infty}.
\]

In this case, the inequality (6.14) becomes an equality when the associated vector norm is \(\ell_\infty\)-norm.