Chapter 6
Direct Methods for Solving Linear Systems

Hung-Yuan Fan (范洪源)

Department of Mathematics,
National Taiwan Normal University, Taiwan

Spring 2016

Section 6.1
Linear Systems of Equations
(線性系統; 線性聯立方程組)

Objective

To solve a system of n **linear equations** with n unknowns:

$$\begin{aligned}
\mathbf{E_1}: &\quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
\mathbf{E_2}: &\quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
&\qquad\qquad \vdots \\
\mathbf{E_n}: &\quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n.
\end{aligned} \tag{1}$$

Definition

The vector $x = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^n$ is called **a solution** to the linear system (1).

Matrix-Vector Forms (矩陣-向量形式)

The linear system (1) can be rewritten as the followg matrix-vector form:

$$Ax = b,$$

where **coefficient matrix** $A \in \mathbb{R}^{n\times n}$ and **right-hand side vector** $b \in \mathbb{R}^n$ are defined by

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Assumption: The matrix $\mathbf{A}$ **is nonsingular** throughout the context.

The Linear Solvers (線性系統的數值方法)

1. Direct Methods: (直接法)
   - Used for solving **small- or medium-sized** linear systems with full and dense coefficient matrices. (適用於求解中小型線性系統)
   - Floating-point operation count (簡稱 flop) $\approx O(n^3)$.
   - **Gaussian Elimination** (高斯消去法，簡稱 **GE**) is an efficient and stable algorithm for solving this type of linear systems.

2. Iterative Methods: (迭代法)
   - Used for solving **large and sparse** linear systems with problem size $n \geq 10^4$. (適用於求解大型稀疏線性系統)
   - flop $\approx O(n)$ per iteration if A is a sparse matrix.
   - Jacobi, Gauss-Seidel, SOR and CG-based methods,....

Three Row Operations

In order to simplify linear system (1), we use

1. $(\lambda \mathbf{E_i}) \to (\mathbf{E_i})$: eq. E_i is replaced by $\lambda \cdot E_i$ for any $\lambda \neq 0$.
2. $(\mathbf{E_i} + \lambda \mathbf{E_j}) \to (\mathbf{E_i})$: eq. E_j is multiplied by any $\lambda \in \mathbb{R}$ and added to eq. E_i.
3. $(\mathbf{E_i}) \leftrightarrow (\mathbf{E_j})$: exchange eqs. E_i and E_j for $i \neq j$.

The Procedure of GE

Given a linear system $Ax = b$ with $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.

1. Form the **augmented matrix** $\tilde{A}^{(1)} = [A \,|\, b] \in \mathbb{R}^{n\times(n+1)}$.
2. Applying row operations continuously, we obtain a **finite** sequence of augmented matrices, i.e.,

$$\tilde{A}^{(1)} \to \tilde{A}^{(2)} \to \cdots \to \tilde{A}^{(n)},$$

   where $\tilde{A}^{(n)}$ is **upper triangular**. (上三角矩陣)
3. Use **backward substitution** (向後代入) to obtain $x_n, x_{n-1}, \cdots, x_2, x_1$.

- From the augmented matrix $\tilde{A}^{(1)} = [A \,|\, b] = [a_{ij}^{(1)}]$, we have

$$\tilde{A}^{(1)} = \left[\begin{array}{c|ccc} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,n+1}^{(1)} \\ \hline a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2,n+1}^{(1)} \\ \vdots & \vdots & & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{n,n+1}^{(1)} \end{array}\right].$$

From $\tilde{A}^{(1)}$ to $\tilde{A}^{(2)}$ (Conti'd)

- If $\mathbf{a}_{\mathbf{11}}^{(\mathbf{1})} \neq 0$, do $\left(E_i - \frac{a_{i1}^{(1)}}{\mathbf{a}_{\mathbf{11}}^{(\mathbf{1})}} E_1\right) \to (E_i)$ for $i = 2, 3, \ldots, n \Rightarrow$

$$\tilde{A}^{(2)} = \left[\begin{array}{c|ccc} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,n+1}^{(1)} \\ \hline 0 & a_{22}^{(2)} & \cdots & a_{2,n+1}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{n,n+1}^{(2)} \end{array}\right].$$

From $\tilde{A}^{(2)}$ to $\tilde{A}^{(3)}$

- Next, if $\mathbf{a_{22}^{(2)}} \neq 0$, do $\left(E_i - \frac{a_{i2}^{(2)}}{\mathbf{a_{22}^{(2)}}} E_2\right) \to (E_i)$ for $i = 3, 4, \ldots, n$

$\Rightarrow$

$$\tilde{A}^{(3)} = \left[\begin{array}{cc|ccc} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2,n+1}^{(2)} \\ \hline \vdots & 0 & a_{33}^{(3)} & \cdots & a_{3,n+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{n,n+1}^{(3)} \end{array}\right].$$

For each $k \geq 2$, suppose augmented matrix $\tilde{A}^{(k-1)}$ has the form

$$\tilde{A}^{(k-1)} = \left[\begin{array}{cccccc|ccc}
a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,k-2}^{(1)} & a_{1,k-1}^{(1)} & & a_{1k}^{(1)} & \cdots & a_{1,n+1}^{(1)} \\
0 & a_{22}^{(2)} & \cdots & a_{2,k-2}^{(2)} & a_{2,k-1}^{(2)} & & a_{2k}^{(2)} & \cdots & a_{2,n+1}^{(2)} \\
\vdots & 0 & & \vdots & \vdots & & \vdots & & \vdots \\
\vdots & \vdots & \ddots & a_{k-2,k-2}^{(k-2)} & a_{k-2,k-1}^{(k-2)} & & a_{k-2,k}^{(k-2)} & \cdots & a_{k-2,n+1}^{(k-2)} \\
\vdots & \vdots & & 0 & \mathbf{a_{k-1,k-1}^{(k-1)}} & & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n+1}^{(k-1)} \\
\hline
\vdots & \vdots & & \vdots & a_{k,k-1}^{(k-1)} & & a_{k,k}^{(k-1)} & \cdots & a_{k,n+1}^{(k-1)} \\
\vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & 0 & \cdots & 0 & a_{n,k-1}^{(k-1)} & & a_{n,k}^{(k-1)} & \cdots & a_{n,n+1}^{(k-1)}
\end{array}\right]$$

From $\tilde{A}^{(k-1)}$ to $\tilde{A}^{(k)}$, $k \geq 2$ (Conti'd)

If $\mathbf{a}_{\mathbf{k-1,k-1}}^{\mathbf{(k-1)}} \neq \mathbf{0}$, do $\left(E_i - \dfrac{a_{i,k-1}^{(k-1)}}{\mathbf{a}_{\mathbf{k-1,k-1}}^{\mathbf{(k-1)}}} E_{k-1}\right) \to (E_i)$ for $k \leq i \leq n \Rightarrow$

$$\tilde{A}^{(k)} = \left[\begin{array}{cccccc|ccc}
a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,k-2}^{(1)} & a_{1,k-1}^{(1)} & & a_{1k}^{(1)} & \cdots & a_{1,n+1}^{(1)} \\
0 & a_{22}^{(2)} & \cdots & a_{2,k-2}^{(2)} & a_{2,k-1}^{(2)} & & a_{2k}^{(2)} & \cdots & a_{2,n+1}^{(2)} \\
\vdots & 0 & & \vdots & \vdots & & \vdots & & \vdots \\
\vdots & \vdots & \ddots & a_{k-2,k-2}^{(k-2)} & a_{k-2,k-1}^{(k-2)} & & a_{k-2,k}^{(k-2)} & \cdots & a_{k-2,n+1}^{(k-2)} \\
\vdots & \vdots & & 0 & \mathbf{a}_{\mathbf{k-1,k-1}}^{\mathbf{(k-1)}} & & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n+1}^{(k-1)} \\
\hline
\vdots & \vdots & & \vdots & 0 & & a_{k,k}^{(k)} & \cdots & a_{k,n+1}^{(k)} \\
\vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & 0 & \cdots & 0 & 0 & & a_{n,k}^{(k)} & \cdots & a_{n,n+1}^{(k)}
\end{array}\right]$$

At Final Stage of GE

When $k = n$, the GE will produce an augmented matrix in **upper triangular** form, i.e., (省略右上標記號)

$$\tilde{A}^{(n)} \equiv \left[\begin{array}{ccccc|c} \mathbf{a_{11}} & a_{12} & \cdots & a_{1,n-1} & a_{1n} & a_{1,n+1} \\ 0 & \mathbf{a_{22}} & \cdots & a_{2,n-1} & a_{2n} & a_{2,n+1} \\ \vdots & 0 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{a_{n-1,n-1}} & a_{n-1,n} & a_{n-1,n+1} \\ 0 & 0 & \cdots & 0 & \mathbf{a_{nn}} & a_{n,n+1} \end{array}\right].$$

Note: Entries $\mathbf{a_{ii}} \neq \mathbf{0}$ are called **pivot elements** (軸元) for

$1 \leq i \leq n$.

Backward Substitution (向後代入)

From the special form of $\tilde{A}^{(n)}$, we obtain the solution x to linear system (1) as follows:

- Firstly, compute $x_n = a_{n,n+1}/a_{nn}$.
- Then compute $x_{n-1}, x_{n-2}, \cdots, x_1$ successively by

$$x_i = \frac{a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}, \quad i = n-1, n-2, \ldots, 1.$$

Note: This process is called the **backward substitution** of GE.

Backward Substitution with $n = 3$

Applying Backward Substitution for the 3×3 linear system

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

the unique solution to above linear system is computed via

$$\begin{aligned} x_3 &= b_3/a_{33}, \\ x_2 &= \frac{b_2 - a_{23}x_3}{a_{22}}, \\ x_1 &= \frac{b_3 - a_{12}x_2 - a_{13}x_3}{a_{11}}. \end{aligned}$$

Will GE break down? (1/2)

Example 2, p. 363 (GE 無法執行的例子)

Use GE to find the solution of the linear system

$$\begin{aligned} E_1: &\quad x_1 - x_2 + 2x_3 - x_4 = -8, \\ E_2: &\quad 2x_1 - 2x_2 + 3x_3 - 3x_4 = -20, \\ E_3: &\quad x_1 + x_2 + x_3 = -2, \\ E_4: &\quad x_1 - x_2 + 4x_3 + 3x_4 = 4. \end{aligned}$$

Sol: Form the augmented matrix $\tilde{A}^{(1)}$ as

$$\tilde{A}^{(1)} = \left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 2 & -2 & 3 & -3 & -20 \\ 1 & 1 & 1 & 0 & -2 \\ 1 & -1 & 4 & 3 & 4 \end{array}\right].$$

Will GE break down? (2/2)

- Since $\mathbf{a_{11}^{(1)} = 1}$, do $(E_i - a_{i1}^{(1)} E_1) \to E_i$ for $i = 2, 3, 4$, we have

$$\tilde{A}^{(2)} = \left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & \mathbf{0} & -1 & -1 & -4 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right].$$

- Because $\mathbf{a_{22}^{(2)} = 0}$, GE will break down here and **STOP**!
- How to fix it? Partial Pivoting Strategy!

Partial Pivoting (1/2)

- If we do $(\mathbf{E_2}) \leftrightarrow (\mathbf{E_3})$, then $\tilde{A}^{(2)}$ becomes

$$\tilde{A}^{(3)} = \left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & \mathbf{2} & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right].$$

- Applying $(E_4 + 2E_3) \to (E_4) \Longrightarrow$

$$\tilde{A}^{(4)} = \left[\begin{array}{cccc|c} 1 & -1 & 2 & -1 & -8 \\ 0 & \mathbf{2} & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right].$$

Partial Pivoting (2/2)

- Use **backward substitution** $\Rightarrow$

$$
\begin{aligned}
x_4 &= \frac{4}{2} = 2 \\
x_3 &= \frac{[-4-(-1)x_4]}{-1} = 2 \\
x_2 &= \frac{[6-(-1)x_3-x_4]}{2} = 3 \\
x_1 &= \frac{[-8-(-1)x_2-2x_3-(-1)x_4]}{1} = -7.
\end{aligned}
$$

- It seems that GE with partial pivoting works well in this case!

Pseudocode of Gaussian Elimination

To solve the $n \times n$ linear system (1).

Algorithm 6.1: GE with Backward Substitution

INPUT dimension n; augmented matrix $A = [a_{ij}] \in \mathbb{R}^{n \times (n+1)}$.

OUTPUT solution $x_1, x_2, \cdots, x_n$.

Step 1 For $i = 1, \ldots, n-1$ do **Steps 2–4**

- Step 2 Find **smallest** $i \le p \le n$ s.t. $\mathbf{a_{pi}} \neq \mathbf{0}$.
  If not, OUTPUT('No unique solution exists.'); **STOP**.
- Step 3 If $p \neq i$, perform $(E_p) \leftrightarrow (E_i)$.
- Step 4 For $j = i+1, \ldots, n$ do **Steps 5–6**
  - Step 5 Set $m_{ji} = a_{ji}/a_{ii}$.
  - Step 6 Perform $(E_j - m_{ji}E_i) \to (E_j)$.

Step 7 If $\mathbf{a_{nn}} = \mathbf{0}$, OUTPUT('No unique solution exists.'); **STOP**.

Step 8 Set $x_n = a_{n,n+1}/a_{nn}$. (Start **backward substitution**.)

Step 9 For $i = n-1, \ldots, 1$ set $x_i = [a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j]/a_{ii}$.

Step 10 OUTPUT$(x_1, x_2, \cdots, x_n)$; **STOP**.

Operation Counts in Steps 5 and 6

- **Multiplications/Divisions:**

$$\sum_{i=1}^{n-1}[(n-i)+(n-i)(n-i+1)] = \sum_{i=1}^{n-1}(n-i)(n-i+2)$$
$$= \sum_{i=1}^{n-1}(n^2-2ni+i^2+2n-2i) = \frac{\mathbf{2n^3+3n^2-5n}}{\mathbf{6}}.$$

- **Additions/Subtractions:**

$$\sum_{i=1}^{n-1}(n-i)(n-i+1) = \sum_{i=1}^{n-1}(n^2-2ni+i^2+n-i) = \frac{\mathbf{n^3-n}}{\mathbf{3}}.$$

Operation Counts in Steps 8 and 9

- **Multiplications/Divisions:**

$$1+\sum_{i=1}^{n-1}[(n-i)+1]=1+\sum_{i=1}^{n-1}(n-i)+(n-1)=\frac{\mathbf{n^2+n}}{\mathbf{2}}.$$

- **Additions/Subtractions:**

$$\sum_{i=1}^{n-1}[(n-i-1)+1]=\sum_{i=1}^{n-1}(n-i)=\frac{\mathbf{n^2-n}}{\mathbf{2}}.$$

Tota Number of Arithmetic Operations

- **Multiplications/Divisions:**

$$\frac{2n^3+3n^2-5n}{6}+\frac{n^2+n}{2}=\frac{\mathbf{n^3}}{\mathbf{3}}+\mathbf{n^2}-\frac{\mathbf{n}}{\mathbf{3}}.$$

- **Additions/Subtractions:**

$$\frac{n^3-n}{3}+\frac{n^2-n}{2}=\frac{\mathbf{n^3}}{\mathbf{3}}+\frac{\mathbf{n^2}}{\mathbf{2}}-\frac{\mathbf{5n}}{\mathbf{6}}.$$

- Total flop of GE $\approx O(\frac{2}{3}n^3)$ (as $n \to \infty$).

Section 6.2
Pivoting Strategies

Occurrence of Small Pivot Elements (1/2)

Example 1, p. 372

Apply GE to solve the linear system

$$\begin{aligned} E_1: &\quad \mathbf{0.003000}x_1 + 59.14x_2 = 59.17 \\ E_2: &\quad 5.291x_1 - 6.130x_2 = 46.78, \end{aligned}$$

using **4-digit rounding** arithmetic and the exact solution is $\mathbf{x_1 = 10.00}$ and $\mathbf{x_2 = 1.000}$.

Sol: Note that $m_{21} = fl(\frac{5.291}{0.003000}) = fl(1763.6\bar{6}) = \mathbf{1764}$. Then do $(E_2 - m_{21}E_1) \to (E_2) \Rightarrow$

$$fl(-6.130 + fl(-1764 \cdot 59.14))x_2 = fl(46.78 + fl(-1764 \cdot 59.17))$$

or $-104300x_2 = -104400$ and hence $\mathbf{x_2 \approx 1.001}$.

Occurrence of Small Pivot Elements (2/2)

Substituting $x_2 = 1.001$ into E_1, we obtain

$$x_1 \approx fl\Big(\frac{59.17 - (59.14)(1.001)}{0.003000}\Big) = \mathbf{-10.00},$$

which gives a totally wrong answer for x_1! Why?

Partial Pivoting (部分選軸元)

For each $i = 1, 2, \ldots, n-1$, find **smallest** integer $p \geq i$ s.t.

$$|a_{pi}^{(i)}| = \max_{i \leq j \leq n} |a_{ji}^{(i)}|.$$

Then perform the row operation

$$(\mathbf{E_i}) \leftrightarrow (\mathbf{E_p}) \quad \text{if } p \neq i.$$

This strategy is also called **maximal column pivoting**.

Improvement of Accuracy for Example 1

Example 2, p. 373 (改進例題 1 的計算精度)

Apply GE with **partial pivoting** to solve the linear system

$$\begin{aligned} E_1: &\quad \mathbf{0.003000}x_1 + 59.14x_2 = 59.17 \\ E_2: &\quad 5.291x_1 - 6.130x_2 = 46.78, \end{aligned}$$

using **4-digit rounding** arithmetic.

Sol: Since $|a_{11}| < |a_{21}|$, we first perform $(E_1) \leftrightarrow (E_2)$:

$$\begin{aligned} E_1: &\quad 5.291x_1 - 6.130x_2 = 46.78 \\ E_2: &\quad 0.003000x_1 + 59.14x_2 = 59.17. \end{aligned}$$

Then $m_{21} = fl(\frac{0.003000}{5.291}) = \mathbf{0.0005670}$. Do $(E_2 - m_{21}E_1) \to (E_2) \Rightarrow 59.14x_2 \approx 59.14$ and hence $\mathbf{x_2 = 1.000}$. Moreover, $\mathbf{x_1 = 10.00}$ is correct now!

To solve the $n \times n$ linear system (1).

Algorithm 6.2: GE with Partial Pivoting

INPUT dimension n; augmented matrix $A = [a_{ij}] \in \mathbb{R}^{n \times (n+1)}$.

OUTPUT solution $x_1, x_2, \cdots, x_n$.

Step 1 For $i = 1, \ldots, n-1$ do **Steps 2–5**

- Step 2 Find **smallest** $i \le p \le n$ s.t. $|a_{pi}| = \max_{i \le j \le n} |a_{ji}|$.
- Step 3 If $a_{pi} = 0$, OUTPUT('No unique solution exists.'); **STOP**.
- Step 4 If $p \neq i$, perform $(E_p) \leftrightarrow (E_i)$.
- Step 5 For $j = i+1, \ldots, n$ do **Steps 6–7**
  - Step 6 Set $m_{ji} = a_{ji}/a_{ii}$.
  - Step 7 Perform $(E_j - m_{ji}E_i) \to (E_j)$.

Step 8 If $\mathbf{a_{nn} = 0}$, OUTPUT('No unique solution exists.'); **STOP**.

Step 9 Set $x_n = a_{n,n+1}/a_{nn}$. (Start **backward substitution**.)

Step 10 For $i = n-1, \ldots, 1$ set $x_i = [a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j]/a_{ii}$.

Step 11 OUTPUT$(x_1, x_2, \cdots, x_n)$; **STOP**.

Example 1'

Apply GE with **partial pivoting** to solve the linear system

$$\begin{aligned} E_1: &\quad \mathbf{30.00}x_1 + \mathbf{591400}x_2 = \mathbf{591700} \\ E_2: &\quad 5.291x_1 - 6.130x_2 = 46.78, \end{aligned}$$

using **4-digit rounding** arithmetic. Exact solution is $\mathbf{x_1 = 10.00}$ and $\mathbf{x_2 = 1.000}$.

Sol: Note that

$$m_{21} = fl(\frac{5.291}{30.00}) = \mathbf{0.1764}.$$

Perform $(E_2 - m_{21}E_1) \to (E_2) \Rightarrow -104300x_2 \approx -104400$. Hence, $\mathbf{x_2 \approx 1.001}$ and $\mathbf{x_1 \approx -10.00}$! Why?

Scaled Partial Pivoting

- At the start of GE, compute n scale factors **once** as follows.

$$s_i = \max_{1 \le j \le n} |a_{ij}|, \quad i = 1, 2, \ldots, n.$$

- For each $i = 1, 2, \ldots, n-1$, find **smallest** integer $p \ge i$ s.t.

$$\frac{|a_{pi}^{(i)}|}{s_p} = \max_{i \le j \le n} \frac{|a_{ji}^{(i)}|}{s_j}.$$

- If $i \neq p$, perform $(\mathbf{E_i}) \leftrightarrow (\mathbf{E_p})$.
- This strategy is also called **scaled-column pivoting**.

Example 2'

Apply GE with **scaled partial pivoting** to solve the linear system

$$\begin{aligned} E_1: &\quad \mathbf{30.00}x_1 + \mathbf{591400}x_2 = \mathbf{591700} \\ E_2: &\quad 5.291x_1 - 6.130x_2 = 46.78, \end{aligned}$$

using **4-digit rounding** arithmetic.

Sol: First compute scale factors $s_1 = 591400$ and $ss_2 = 6.130$. For $i = 1$, we see that

$$\frac{|a_{11}|}{s_1} = fl(\frac{30.00}{591400}) = \mathbf{0.5073 \times 10^{-4}} < \frac{|a_{21}|}{s_2} = fl(\frac{5.291}{6.130}) = \mathbf{0.8631}.$$

So, perform $(E_2) \leftrightarrow (E_1) \Rightarrow$ we obtain the correct solution $x_1 = 10.00$ and $x_2 = 1.000$!

To solve the $n \times n$ linear system (1).

Algorithm 6.3: GE with Scaled Partial Pivoting

INPUT dimension n; augmented matrix $A = [a_{ij}] \in \mathbb{R}^{n\times(n+1)}$.

OUTPUT solution $x_1, x_2, \cdots, x_n$.

Step 1 For $i = 1, \ldots, n$ set $s_i = \max_{1\le j\le n} |a_{ij}|$;
If $s_i = 0$, OUTPUT('No unique solution exists.'); **STOP**.

Step 2 For $i = 1, \ldots, n-1$ do **Steps 3–6**

Step 3 Find **smallest** $i \le p \le n$ s.t. $\frac{|a_{pi}|}{s_p} = \max_{i\le j\le n} \frac{|a_{ji}|}{s_j}$.

Step 4 If $a_{pi} = 0$, OUTPUT('No unique solution exists.'); **STOP**.

Step 5 If $p \neq i$, perform $(E_p) \leftrightarrow (E_i)$.

Step 6 For $j = i+1, \ldots, n$ do **Steps 7–8**

Step 7 Set $m_{ji} = a_{ji}/a_{ii}$.

Step 8 Perform $(E_j - m_{ji}E_i) \to (E_j)$.

Step 9 If $\mathbf{a_{nn} = 0}$, OUTPUT('No unique solution exists.'); **STOP**.

Step 10 Set $x_n = a_{n,n+1}/a_{nn}$. (Start **backward substitution**.)

Step 11 For $i = n-1, \ldots, 1$ set $x_i = [a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j]/a_{ii}$.

Step 12 OUTPUT$(x_1, x_2, \cdots, x_n)$; **STOP**.

Complete Pivoting

- For each $k = 1, 2, \ldots, n-1$, find integers $k \leq p, q \leq n$ s.t.

$$|a_{pq}^{(k)}| = \max_{k \leq i,j \leq n} |a_{ij}^{(k)}|.$$

- If $p \neq i$ or $q \neq i$, **row and/or column interchanges** are performed to bring $a_{pq}^{(k)}$ to the pivot position $a_{kk}^{(k)}$.
- This strategy is also called the **maximal pivoting** at the kth step.

Section 6.5
Matrix Factorization
(矩陣分解)

Motivation

- For solving a linear system $Ax = b$, it requires $O(\frac{1}{3}n^3)$ arithmetic operations to determine $x \in \mathbb{R}^n$.
- If the right-hand vector $b \in \mathbb{R}^n$ is changed to another vector $\tilde{b}$ (and coeff. matrix A is **unchanged**), how can we solve this linear system efficiently using some matrix factorization of A generated from GE?
- In fact, if A has been factored into the **triangular form**

$$A = LU,$$

where L is lower triangular and U is upper triangular, then the operation counts can be reduced to $O(2n^2)$!

Comparison of Arithmetic Calculations

The relative rate of reduction of the operation counts $O(2n^2)$ compared with $O(\frac{1}{3}n^3)$ becomes larger and larger for $n = 10$, 10^2 and 10^3, respectively. The results are shown in the following table.

The 1st Step of GE

Let $A^{(1)} = A \in \mathbb{R}^{n\times n}$ and $b^{(1)} = b \in \mathbb{R}^n$ for a linear system.

- If $\mathbf{a}_{\mathbf{1,1}}^{(\mathbf{1})} \neq 0$, do $\left(E_i - \frac{a_{i,1}^{(1)}}{\mathbf{a}_{\mathbf{1,1}}^{(\mathbf{1})}} E_1\right) \to (E_i)$ for $i = 2, 3, \ldots, n \Rightarrow$

$$A^{(2)} = \left[\begin{array}{c|ccc} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \cdots & a_{1,n}^{(1)} \\ \hline 0 & a_{2,2}^{(2)} & \cdots & a_{2,n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n,2}^{(2)} & \cdots & a_{n,n}^{(2)} \end{array}\right], \quad b^{(2)} = \begin{bmatrix} b_1^{(2)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix}.$$

- The corresponding multipliers are given by

$$m_{i,1} = \frac{a_{i,1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2, 3, \ldots, n.$$

The 1st Step of GE (Conti'd)

- This is equivalent to

$$A^{(2)} = M^{(1)} A^{(1)} \quad \text{and} \quad b^{(2)} = M^{(1)} b^{(1)},$$

where the **first Gaussian transformation matrix** $M^{(1)} \in \mathbb{R}^{n \times n}$ is defined by

$$M^{(1)} = \left[\begin{array}{c|ccccc} 1 & 0 & \cdots & \cdots & 0 \\ \hline -m_{2,1} & 1 & 0 & \cdots & 0 \\ -m_{3,1} & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -m_{n,1} & 0 & \cdots & 0 & 1 \end{array}\right].$$

At the kth Step of GE, $2 \le k \le n-1$

- If $\mathbf{a}_{\mathbf{k},\mathbf{k}}^{(\mathbf{k})} \neq \mathbf{0}$, do $\left(E_i - \dfrac{a_{i,k}^{(k)}}{\mathbf{a}_{\mathbf{k},\mathbf{k}}^{(\mathbf{k})}} E_k\right) \to (E_i)$ for $k+1 \le i \le n \Rightarrow$

$$A^{(k+1)} = \left[\begin{array}{ccc|ccc} a_{1,1}^{(1)} & \cdots & a_{1,k}^{(1)} & a_{1,k}^{(1)} & \cdots & a_{1,n}^{(1)} \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & & \mathbf{a}_{\mathbf{k},\mathbf{k}}^{(\mathbf{k})} & a_{k,k+1}^{(k)} & \cdots & a_{k,n}^{(k)} \\ \hline & & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} \end{array}\right], \quad b^{(k+1)} = \begin{bmatrix} b_1^{(k+1)} \\ b_2^{(k+1)} \\ \vdots \\ b_n^{(k+1)} \end{bmatrix}.$$

- The corresponding multipliers are given by

$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, \quad i = k+1, k+2, \ldots, n.$$

At the kth Step of GE, $2 \leq k \leq n-1$ (Conti'd)

- This is equivalent to

$$A^{(k+1)} = M^{(k)} A^{(k)} \quad \text{and} \quad b^{(k+1)} = M^{(k)} b^{(k)},$$

where the **kth Gaussian transformation matrix** $M^{(k)} \in \mathbb{R}^{n \times n}$ is defined by

$$M^{(k)} = \left[\begin{array}{ccc|ccc} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & & 1 & 0 & & \vdots \\ \hline \vdots & & -m_{k+1,k} & 1 & \ddots & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ 0 & & -m_{n,k} & & & 1 \end{array}\right].$$

The Inverse Matrix of $M^{(k)}$

It is easily seen that the inverse matrix of $M^{(k)}$ is given by

$$L^{(k)} = [M^{(k)}]^{-1} = \left[\begin{array}{ccc|ccc} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & & 1 & 0 & & \vdots \\ \hline \vdots & & m_{k+1,k} & 1 & \ddots & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ 0 & & m_{n,k} & & & 1 \end{array}\right] \tag{2}$$

for each $k = 1, 2, \ldots, n-1$.

Check ...

$$M^{(k)} L^{(k)} = I \quad \text{or} \quad L^{(k)} M^{(k)} = I,$$

where I denotes the $n \times n$ identity matrix.

Assume the GE is performed **without row interchanges** $\Longrightarrow$

$$U \equiv A^{(n)} = M^{(n-1)} \cdots M^{(2)} M^{(1)} A = \begin{bmatrix} a_{11}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{bmatrix}$$

is an $n \times n$ **upper triangular matrix**. So, we obtain the LU factorization of A as

$$\begin{aligned} A &= [M^{(1)}]^{-1} [M^{(2)}]^{-1} \cdots [M^{(n-1)}]^{-1} U \\ &= L^{(1)} L^{(2)} \cdots L^{(n-1)} U \\ &= \begin{bmatrix} 1 & & & \\ m_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ m_{n1} & \cdots & m_{n,n-1} & 1 \end{bmatrix} \begin{bmatrix} a_{11}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{bmatrix} \equiv LU. \end{aligned}$$

Matrix Product of $L^{(1)}L^{(2)}L^{(3)}$

For example, when $n = 4$, it follows from (2) that

$$L^{(1)}L^{(2)}L^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & 0 & 1 & 0 \\ m_{41} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & m_{32} & 1 & 0 \\ 0 & m_{42} & 0 & 1 \end{bmatrix} L^{(3)}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & m_{32} & 1 & 0 \\ m_{41} & m_{42} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & m_{43} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 \\ m_{21} & \mathbf{1} & 0 & 0 \\ m_{31} & m_{32} & \mathbf{1} & 0 \\ m_{41} & m_{42} & m_{43} & \mathbf{1} \end{bmatrix} \equiv L.$$

$\Rightarrow$ L is a **lower triangular matrix** with 1s on the main diagonal.

Thm 6.19 (矩陣 A 的 LU 分解)

If the GE can be performed on $Ax = b$ **without row interchanges**, then $A \in \mathbb{R}^{n\times n}$ can be factored as $A = LU$, where

$$L = \begin{bmatrix} 1 & & & \\ m_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ m_{n1} & \cdots & m_{n,n-1} & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} a_{11}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{bmatrix}$$

are lower triangular and upper triangular matrices, respectively.

General Form of LU Factorization

Observe each entry of matrix factorization $A = LU \in \mathbb{R}^{n\times n}$, i.e.,

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & \cdots & u_{1n} \\ & \ddots & \vdots \\ 0 & & u_{nn} \end{bmatrix} = LU.$$

1. $a_{11} = l_{11}u_{11} \Rightarrow$ determine l_{11} and u_{11}.
2. determine the 1st row of U: $u_{1j} = a_{1j}/l_{11}$, and the 1st column of L: $l_{j1} = a_{j1}/u_{11}$ for $j = 2, 3, \ldots, n$.
3. For $i = 2, 3, \ldots, n-1$, we have
   - $a_{ii} = \sum_{k=1}^{i-1} l_{ik}u_{ki} + l_{ii}u_{ii} \Rightarrow$ determine l_{ii} and u_{ii} .
   - determine the ith row of U: $u_{i,j} = (a_{i,j} - \cdots)/l_{ii}$ for $j = i+1, \ldots, n$.
   - determine the ith column of L: $l_{j,i} = (a_{j,i} - \cdots)/u_{ii}$ for $j = i+1, \ldots, n$.
4. $a_{nn} = \sum_{k=1}^{n-1} l_{nk}u_{kn} + l_{nn}u_{nn} \Rightarrow$ determine l_{nn} and u_{nn} .

Algorithm 6.4: LU Factorization

INPUT dim. n; $A = [a_{ij}] \in \mathbb{R}^{n\times n}$; $l_{11} = \cdots = l_{nn} = 1$; $u_{11} = \cdots = u_{nn} = 1$.

OUTPUT lower triangular $L = [l_{ij}]$ and upper triangular $U = [u_{ij}]$.

Step 1 Select l_{11} and u_{11} satisfying $l_{11}u_{11} = a_{11}$. If $l_{11}u_{11} = 0$ then OUTPUT('Factorization impossible'); **STOP**.

Step 2 For $j = 2, \ldots, n$ set $u_{1j} = a_{1j}/l_{11}$; $l_{j1} = a_{j1}/u_{11}$.

Step 3 For $i = 2, \ldots, n-1$ do **Steps 4–5**

Step 4 Select l_{ii} and u_{ii} satisfying $l_{ii}u_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik}u_{ki}$. If $l_{ii}u_{ii} = 0$ then OUTPUT('Factorization impossible'); **STOP**.

Step 5 For $j = i+1, \ldots, n$ set

$$u_{ij} = \frac{1}{l_{ii}}\left[a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right]; \ (i\text{th row of } U)$$

$$l_{ji} = \frac{1}{u_{ii}}\left[a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki}\right]. \ (i\text{th column of } L)$$

Step 6 Select l_{nn} and u_{nn} satisfying $l_{nn}u_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk}u_{kn}$.

Step 7 OUTPUT(L and U); **STOP**.

In Algorithm 6.4 (LU Factorization), three methods are condidered for choosing the diagonal entries of L and U.

The Choices of l_{ii} and u_{ii} $(1 \leq i \leq n)$

1. **Doolittle's method**: $l_{ii} = 1$ are required for $i = 1, 2, \ldots, n$.
2. **Crout's method**: $u_{ii} = 1$ are required for $i = 1, 2, \ldots, n$.
3. **Cholesky's method**: $l_{ii} = u_{ii}$ are required for $i = 1, 2, \ldots, n$.

Note: Doolittle's method is the same as the LU factorization given in Thm 6.19! The entries of L and U satisfy

$$l_{ij} = m_{ij}, \quad 1 \leq j < i \leq n$$

and

$$u_{ij} = a_{ij}^{(i)}, \quad 1 \leq i \leq j \leq n.$$

Example 2 (a), p. 404

Apply the **Doolittle's method** to determine the LU factorization for A in the 4×4 linear system $Ax = b$ with

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

Solution of Part (a)

- Do $(E_i - m_{i1}E_1) \to (E_i), i = 2, 3, 4$ with $m_{21} = 2$, $m_{31} = 3$, $m_{41} = -1$. Then

$$A^{(2)} = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & -4 & -1 & -7 \\ 0 & 3 & 3 & 2 \end{bmatrix}.$$

- Do $(E_i - m_{i2}E_2) \to (E_i), i = 3, 4$ with $m_{32} = 4$ and $m_{42} = -3$. Then

$$A^{(3)} = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} \equiv U.$$

Solution of Part (a)–Conti'd

- Now, if we let the lower triangular matrix L as

$$L = \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 \\ 2 & \mathbf{1} & 0 & 0 \\ 3 & 4 & \mathbf{1} & 0 \\ -1 & -3 & 0 & \mathbf{1} \end{bmatrix},$$

then it follows from Thm 6.19 that

$$A = \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 \\ 2 & \mathbf{1} & 0 & 0 \\ 3 & 4 & \mathbf{1} & 0 \\ -1 & -3 & 0 & \mathbf{1} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} = LU. \qquad (3)$$

Question

Suppose that $A \in \mathbb{R}^{n\times n}$ has been factored into its LU factorization $A = LU$, where L and U are lower and upper triangular matrices, respectively. How can we solve the linear system

$$Ax = b, \quad x \in \mathbb{R}^n$$

within $O(2n^2)$ arithmetic operations for **any** right-hand $b \in \mathbb{R}^n$?

Ans: may utilize the **triangular forms** of L and U!

Two-Step Process for Solving $Ax = b$

Since $A = LU$, we see taht

$$Ax = (LU)x = L(Ux) = b, \quad x \in \mathbb{R}^n.$$

1. Firstly, solve the **lower-triangular** system

$$Ly = b$$

for the solution $y \in \mathbb{R}^n$.

2. Next, solve the **upper-triangular** system

$$Ux = y$$

for the solution $x \in \mathbb{R}^n$ to the original linear system $Ax = b$.

Procedures of Numerical Solutions

For any right-hand vector

$$b = [b_1, b_2, \cdots, b_n]^T \in \mathbb{R}^n,$$

we shall apply

1. **Forward Substitution (向前代入)**: $y_1 = b_1/l_{11}$ and

$$y_i = \frac{1}{l_{ii}}\left(b_i - \sum_{j=1}^{i-1} l_{ij}y_j\right), \quad i = 2, 3, \ldots, n.$$

2. **Backward Substitution (向後代入)**: $x_n = y_n/u_{nn}$ and

$$x_i = \frac{1}{ul_{ii}}\left(y_i - \sum_{j=i+1}^{n} u_{ij}x_j\right), \quad i = n-1, n-2, \ldots, 1.$$

Amount of Floating-Point Arithmetic Operations

- Operation counts for the **forward substitution** $\approx O(n^2)$.
- Operation counts for the **backward substitution** $\approx O(n^2)$.
- Thus, total operation counts for solving $n \times n$ linear systems with different right-hand vectors can be reduced to $O(2n^2)$!

Example 2 (b), p. 404

Use part (a) of Example 2 to solve the 4×4 linear system $Ax = \tilde{b}$, where the coefficient matrix A is unchanged, but the right-hand vector

$$b = \begin{bmatrix} 1 \\ 1 \\ -3 \\ 4 \end{bmatrix} \text{ is changed to } \tilde{b} = \begin{bmatrix} 8 \\ 7 \\ 14 \\ -7 \end{bmatrix}.$$

Solution of Part (b)

From Eq. (3) in part (a), we have

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

(1) Apply forward substitution to solve $Ly = \tilde{b}$ for the vector y:

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

Then $y_1 = \mathbf{8}$, $y_2 = 7 - 2y_1 = \mathbf{-9}$, $y_3 = 14 - 3y_1 - 4y_2 = \mathbf{26}$ and $y_4 = -7 + y_1 + 3y_2 = \mathbf{-26}$.

Solution of Part (b)–Conti'd

(2) Solve the upper-triangular system $Ux = y$ for the sol. x:

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

Finally, we obtain $x_4 = \mathbf{2}$, $x_3 = \mathbf{0}$, $x_2 = -\mathbf{1}$ and $x_1 = \mathbf{3}$ by using the backward substitution.

Permutation Matrices

Remarks

- The LU factorization in Thm6.19 holds when the GE is performed **without row interchanges**.
- When row interchanges are required in practical computation, the LU factorization of $A \in \mathbb{R}^{n\times n}$ will be modified by applying the **permutation matrices** (置換矩陣).

Definition

- A **permutation matrix** $P = [p_{ij}] \in \mathbb{R}^{n\times n}$ is a matrix obtained rearranging rows of the identity matrix I (or I_n).
- A permutation matrix is a matrix with precisely one nonzero entry in each row and in each column, and each nonzero entry is **1**. (每列和每行只一個非零元素，且該非零元素為 **1**)

Two Useful Properties

If $k_1, k_2, \ldots, k_n$ is a permutation of integers $1, 2, \ldots, n$ and the permutation matrix $P = [p_{ij}] \in \mathbb{R}^{n \times n}$ is defined by

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

for $i, j = 1, 2, \ldots, n$, then

(1) PA permutes the rows of A , i.e.,

$$PA = \begin{bmatrix} a_{k_1,1} & a_{k_1,2} & \cdots & a_{k_1,n} \\ a_{k_2,1} & a_{k_2,2} & \cdots & a_{k_2,n} \\ \vdots & \vdots & & \vdots \\ a_{k_n,1} & a_{k_n,2} & \cdots & a_{k_n,n} \end{bmatrix}.$$

(2) P^{-1} exists and $P^{-1} = P^T$, i.e., permutation matrices are **orthogonal matrices**.

Modification of Gaussian Elimination

- Suppose that a permutation matrix $P \in \mathbb{R}^{n\times n}$ is known in advance s.t. the linear system

$$PAx = Pb$$

can be solved by the GE **without row interchanges**.

- From Thm 6.19 $\Longrightarrow \exists$ (unit) lower triangular $L \in \mathbb{R}^{n\times n}$ and upper triangular $U \in \mathbb{R}^{n\times n}$ s.t.

$$PA = LU.$$

- So, the coefficient matrix $A \in \mathbb{R}^{n\times n}$ can be factored into

$$A = P^{-1}LU = (P^TL)U \equiv \tilde{L}U,$$

where $\tilde{L} = P^TL$ is NOT a lower triangular matrix!

Thm (LU Factorization for GE with Partial Pivoting)

If GE **with partial pivoting** is used to compute the upper triangularization

$$M^{(n-1)} P_{n-1} \cdots M^{(1)} P_1 A = U$$

with P_i being the interchange permutation involving row i and row $\mu(\geq i)$ for $i = 1, 2, \ldots, n-1$, then

$$PA = LU,$$

where $P = P_{n-1} \cdots P_1$ is a permutation matrix and L is a unit lower triangular matrix with $|l_{ij}| \leq 1$.

See **Thm 3.4.1** (p. 113) in *Matrix Computations* written by G. Golub and C. F. Van Loan, 3rd ed., The Johns Hopkins University Press, 1996.

Example 3, p. 408

Determine a factorization in the form

$$PA = LU \quad \text{or} \quad A = (P^T L)U$$

for the matrix

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

Sol:

- Since $a_{11} = 0$, do $(E_1) \leftrightarrow (E_2)$, followed by $(E_3 + E_1) \to (E_3)$ and $(E_4 - E_1) \to (E_4) \Longrightarrow$

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

- Equivalently, if we let

$$P_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad M^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix},$$

  then we see that $M^{(1)} P_1 A = A^{(2)}$.
- Next, $(E_2) \leftrightarrow (E_4)$, followed by $(E_4 + E_3) \rightarrow (E_4)$, gives that

$$M^{(3)} P_2 A^{(2)} = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix} = A^{(3)} \equiv U,$$

  where permutation P_2 and unit lower triangular $M^{(3)}$ are

$$P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad M^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

- Because $M^{(1)}P_1A = A^{(2)}$ and $M^{(3)}P_2A^{(2)} = U$, we obtain

$$M^{(3)}(P_2M^{(1)}P_2)(P_2P_1)A = M^{(3)}P_2A^{(2)} = U. \qquad (4)$$

- Matrices $P \equiv P_2P_1$ and $\widetilde{M}^{(1)} \equiv P_2M^{(1)}P_2$ are computed as

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad \widetilde{M}^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then the desired lower triangular matrix L is given by

$$L = [\widetilde{M}^{(1)}]^{-1}[M^{(3)}]^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

- From (4), we immediately obtain the factorization

$$PA = LU = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

- Since P is a permutation matrix, it follows that

$$A = (P^T L)U = \begin{bmatrix} 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

Section 6.6
Special Types of Matrices

Three Types of Matrices

In practical applications, we often encounter the following types of matrices:

1. Diagonally dominant matrices. (對角佔優矩陣)
2. Positive definite matrices. (正定矩陣)
3. Band matrices. (寬帶矩陣; 帶狀矩陣)

Diagonally Dominant Matrices

Def 6.20, p. 412

Let $A = [a_{ij}]$ be an $n \times n$ matrix.

- A is called **diagonally dominant** (對角佔優) if

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

- A is called **strictly diagonally dominant** (嚴格對角佔優) if

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

Two Examples

- The nonsymmetric matrix $A = \begin{bmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & 5 & -6 \end{bmatrix}$ is **strictly diagonally dominant**.
- The symmetric matrix $B = \begin{bmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{bmatrix}$ is **NOT diagonally dominant**.

Thm 6.21 (嚴格對角佔優的性質)

Let $A \in \mathbb{R}^{n \times n}$ be **strictly diagonally dominant**. Then

(i) A is nonsingular.

(ii) The GE process can be performed on $Ax = b$ to obtain its unique solution **without row or column interchanges**.

(iii) The GE is **stable** with respect to the growth of round-off errors in this case.

Proof of Part (i)

Claim: $Ax = 0 \quad \forall\, x \in \mathbb{R}^n \Longrightarrow x = 0$.

If $x \neq 0$, $\exists\ 1 \leq k \leq n$ s.t. $0 < |x_k| = \max\limits_{1 \leq j \leq n} |x_j|$. Since $Ax = 0$, we see from the kth row of Ax that

$$0 = \sum_{j=1}^{n} a_{kj} xj \quad \text{or} \quad a_{kk} x_k = - \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj} x_j.$$

Then we obtain

$$|a_{kk}||x_k| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}||x_j| \quad \text{or} \quad |a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}|,$$

which gives rise to a contradiction!

Proof of Part (ii) (1/3)

- Since $A^{(1)} = A$ is strictly diagonally dominant, $|a_{11}^{(1)}| > 0$ and hence $A^{(2)}$ is generated by GE **without row interchanges**.
- For $i = 2, \ldots, n$, GE produces

$$a_{i1}^{(2)} = 0, \quad a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{1j}^{(1)} a_{i1}^{(1)}}{a_{11}^{(1)}} \text{ for } 2 \le j \le n.$$

Taking the absolute values on both sides $\Longrightarrow$

$$\sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}^{(2)}| \le \sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}^{(1)}| + \sum_{\substack{j=2 \\ j \neq i}}^{n} \left(|a_{1j}^{(1)}| \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|} \right). \tag{5}$$

Proof of Part (ii) (2/3)

- Since $A = A^{(1)}$ is strictly diagonally dominant, we see that

$$\sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}^{(1)}| < |a_{ii}^{(1)}| - |a_{i1}^{(1)}|, \qquad \sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{1j}^{(1)}| < |a_{11}^{(1)}| - |a_{1i}^{(1)}|. \quad (6)$$

- Combining (5) with (6) $\Longrightarrow$

$$\sum_{\substack{j=2 \\ j \neq i}}^{n} |a_{ij}^{(2)}| < |a_{ii}^{(1)}| - |a_{i1}^{(1)}| + \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|}(|a_{11}^{(1)}| - |a_{1i}^{(1)}|)$$

$$= |a_{ii}^{(1)}| - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|} \leq \left| a_{ii}^{(1)} - \frac{a_{i1}^{(1)} a_{1i}^{(1)}}{a_{11}^{(1)}} \right| = |\mathbf{a_{ii}^{(2)}}|,$$

for $i = 2, 3, \ldots, n$. Hence, $A^{(2)}$ **is strictly diagonally dominant**.

Proof of Part (ii) (3/3)

- This implies that $|a_{ii}^{(2)}| > 0$ for $i = 1, 2, \ldots, n$.
- Continue this process inductively $\Longrightarrow$ the upper triangular matrix $A^{(n)}$ is also **strictly diagonally dominant** with $|a_{ii}^{(n)}| = |a_{ii}^{(i)}| > 0$ for $i = 1, 2, \ldots, n$.
- Therefore, GE can be applied to solve the linear system $Ax = b$ **without row or column interchanges**!

Positive Definite Matrices

Def 6.22

$A \in \mathbb{R}^{n\times n}$ is called **positive definite** if it is a **symmetric** matrix satisfying $x^T Ax > 0$ for all $0 \neq x \in \mathbb{R}^n$.

Thm 6.23 (正定矩陣的必要條件)

If A is an $n \times n$ positive definite matrix, then

(i) A has an inverse;

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

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

(iv) $(a_{ij})^2 < a_{ii}a_{jj}$ for $i \neq j$.

Note: These necessary conditions are used to eliminate certain matrices from consideration.

Def 6.24

A **leading principal submatrix** of $A = [a_{ij}]$ is a matrix of the form

$$A_k = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \in \mathbb{R}^{k \times k}$$

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

Thm 6.25 (正定矩陣的充分必要條件)

$A \in \mathbb{R}^{n \times n}$ is positive definite $\iff \det(A_k) > 0$, where A_k is the kth leading principal submatrix of A for $k = 1, 2, \ldots, n$.

GE with Positive Definite Matrices

Thm 6.26

The matrix A is positive definite $\Longleftrightarrow$

- **GE without row interchanges** can be performed on $Ax = b$ with all pivot elements **positive**.
- Moreover, GE is **stable** with respect to the growth of round-off errors in this case.

Cor 6.27

$A \in \mathbb{R}^{n\times n}$ is positive definite $\iff A = LDL^T$, where $L \in \mathbb{R}^{n\times n}$ is a lower triangular matrix with 1s on its diagonal and $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix with **positive** diagonal entries.

Note: The LDL^T factorization requires

- $\frac{1}{6}\mathbf{n^3} + n^2 - \frac{7}{6}n$ multiplications/divisions,
- $\frac{1}{6}\mathbf{n^3} - \frac{1}{6}n$ additions/subtractions,
- additional $O(n^2)$ operations to obtain the solution of $Ax = b$.

So, total flop $\approx \mathbf{O}(\frac{1}{3}\mathbf{n^3})$.

The 3×3 Case

The 3×3 positive definite matrix is factored as $A = LDL^T$, i.e.,

$$\begin{bmatrix} a_{11} & \mathbf{a_{21}} & \mathbf{a_{31}} \\ a_{21} & a_{22} & \mathbf{a_{32}} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & 0 & 0 \\ l_{21} & \mathbf{1} & 0 \\ l_{31} & l_{32} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{d_1} & 0 & 0 \\ 0 & \mathbf{d_2} & 0 \\ 0 & 0 & \mathbf{d_3} \end{bmatrix} \begin{bmatrix} \mathbf{1} & l_{21} & l_{31} \\ 0 & \mathbf{1} & l_{32} \\ 0 & 0 & \mathbf{1} \end{bmatrix}$$

$$= \begin{bmatrix} d_1 & d_1 l_{21} & d_1 l_{31} \\ d_1 l_{21} & d_2 + d_1 l_{21}^2 & d_2 l_{32} + d_1 l_{21} l_{31} \\ d_1 l_{31} & d_1 l_{21} l_{31} + d_2 l_{32} & d_1 l_{31}^2 + d_2 l_{32}^2 + d_3 \end{bmatrix}.$$

Then the 6 unknowns can be obtained by

$$\begin{aligned} &\mathbf{a_{11}} : d_1 = a_{11}, \quad \mathbf{a_{21}} : l_{21} = a_{21}/d_1, \quad \mathbf{a_{31}} : l_{31} = a_{31}/d1, \\ &\mathbf{a_{22}} : d_2 = a_{22} - d_1 l_{21}^2, \quad \mathbf{a_{32}} : l_{32} = (a_{32} - d_1 l_{21} l_{31})/d_2, \\ &\mathbf{a_{33}} : d_3 = a_{33} - d_1 l_{21}^2 - d_2 l_{31}^2. \end{aligned} \tag{7}$$

Pseudocode for LDL^T Factorization

Algorithm 6.5: LDL^T Factorization

INPUT dimension n; the matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$.

OUTPUT lower triangular matrix L and diagonal matrix D.

Step 1 For $i = 1, \ldots, n$ do **Steps 2–4**

Step 2 For $j = 1, \ldots, i-1$ set $v_j = l_{ij} d_j$.

Step 3 Set $d_i = a_{ii} - \sum_{j=1}^{i-1} l_{ij} v_j$.

Step 4 For $j = i+1, \ldots, n$ set $l_{ji} = (a_{ji} - \sum_{k=1}^{i-1} l_{jk} v_k)/d_i$.

Step 5 OUTPUT(l_{ij} for $j = 1, \ldots, i-1$ and $i = 1, \ldots, n$);
OUTPUT(d_i for $i = 1, \ldots, n$); **STOP**.

Example 3, p. 418

Find the LDL^T factorization of the positive definite matrix

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

Sol: From Eqs. (7), we have

$$d_1 = a_{11} = 4, \quad l_{21} = a_{21}/d_1 = -1/4 = -0.25,$$
$$l_{31} = a_{31}/d1 = 1/4 = 0.25, \quad d_2 = a_{22} - d_1 l_{21}^2 = 4.25 - 0.25 = 4,$$
$$l_{32} = (a_{32} - d_1 l_{21} l_{31})/d_2 = 0.75, \quad d_3 = a_{33} - d_1 l_{21}^2 - d_2 l_{31}^2 = 1.$$

Hence, the matrix A can be factored as

$$A = LDL^T = \begin{bmatrix} 1 & 0 & 0 \\ -0.25 & 1 & 0 \\ 0.25 & 0.75 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -0.25 & 0.25 \\ 0 & 1 & 0.75 \\ 0 & 0 & 1 \end{bmatrix}.$$

The LL^T (or Cholesky) Factorization

Thm 6.28 (Cholesky Factorization)

A is positive definite $\Longleftrightarrow$ $A = LL^T$, where L is lower triangular with **nonzero** diagonal entries.

Note: The Cholesky factorization requires

- $\frac{1}{6}\mathbf{n^3} + \frac{1}{2}n^2 - \frac{2}{3}n$ multiplications/divisions,
- $\frac{1}{6}\mathbf{n^3} - \frac{1}{6}n$ additions/subtractions,
- The operation counts of n square roots $\approx O(n)$,
- additional $O(n^2)$ operations to obtain the solution of $Ax = b$.

So, total flop $\approx \mathbf{O}(\frac{1}{3}\mathbf{n^3})$.

The 3×3 Cholesky Factorization

The 3×3 positive definite matrix is factored as $A = LL^T$, i.e.,

$$\begin{bmatrix} a_{11} & \mathbf{a_{21}} & \mathbf{a_{31}} \\ a_{21} & a_{22} & \mathbf{a_{32}} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{l_{11}} & 0 & 0 \\ l_{21} & \mathbf{l_{22}} & 0 \\ l_{31} & l_{32} & \mathbf{l_{33}} \end{bmatrix} \begin{bmatrix} \mathbf{l_{11}} & l_{21} & l_{31} \\ 0 & \mathbf{l_{22}} & l_{32} \\ 0 & 0 & \mathbf{l_{33}} \end{bmatrix}$$

$$= \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}.$$

Then the 6 unknowns can be obtained by

$$\begin{aligned} &\mathbf{a_{11}} : l_{11} = \sqrt{a_{11}}, \quad \mathbf{a_{21}} : l_{21} = a_{21}/l_{11}, \quad \mathbf{a_{31}} : l_{31} = a_{31}/l_{11}, \\ &\mathbf{a_{22}} : l_{22} = (a_{22} - l_{21}^2)^{1/2}, \quad \mathbf{a_{32}} : l_{32} = (a_{32} - l_{21}l_{31})/l_{22}, \\ &\mathbf{a_{33}} : l_{33} = (a_{33} - l_{31}^2 - l_{32}^2)^{1/2}. \end{aligned} \tag{8}$$

Pseudocode of LL^T Factorization

Algorithm 6.6: Cholesky Factorization

INPUT dimension n; the matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$.

OUTPUT lower triangular matrix L.

Step 1 Set $l_{11} = \sqrt{a_{11}}$.

Step 2 For $j = 2, \ldots, n$ set $l_{j1} = a_{j1}/l_{11}$.

Step 3 For $i = 2, \ldots, n-1$ do **Steps 4–5**

Step 4 Set $l_{ii} = (a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2)^{1/2}$.

Step 5 For $j = i+1, \ldots, n$ set $l_{ji} = (a_{ji} - \sum_{k=1}^{i-1} l_{jk} l_{ik})/l_{ii}$.

Step 6 Set $l_{nn} = (a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2)^{1/2}$.

Step 7 OUTPUT(l_{ij} for $j = 1, \ldots, i$ and $i = 1, \ldots, n$); **STOP**.

Example 4, p. 419

Find the LL^T factorization of the positive definite matrix

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

Sol: From Eqs. (8), we have

$$l_{11} = \sqrt{a_{11}} = 2, \quad l_{21} = a_{21}/l_{11} = -1/2 = -0.5,$$
$$l_{31} = a_{31}/l_{11} = 1/2 = 0.5, \quad l_{22} = (a_{22} - l_{21}^2)^{1/2} = 2,$$
$$l_{32} = (a_{32} - l_{21}l_{31})/l_{22} = 1.5, \quad l_{33} = (a_{33} - l_{31}^2 - l_{32}^2)^{1/2} = 1.$$

Hence, the matrix A can be factored as

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

Band Matrices (帶狀矩陣)

Def 6.30

$A = [a_{ij}] \in \mathbb{R}^{n\times n}$ is called a **band matrix** if $\exists\ 1 < p, q < n$ s.t. $a_{ij} = 0$ whenever $p \le j - i$ or $q \le i - j$. The **band width** of A is defined by $\omega = p + q - 1$.

Note

- p is the number of diagonals **above, and including, the main diagonal** where **nonzero entries** may lie.
- q is the number of diagonals **below, and including, the main diagonal** where **nonzero entries** may lie.

Example

The 3×3 matrix

$$A = \begin{bmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & -5 & -6 \end{bmatrix}$$

is a band matrix with $p = q = 2$. The band width of A is $\omega = 2 + 2 - 1 = \mathbf{3}$.

Note:

- Band matrices with $p = q = 2$ are also called **tridiagonal matrices**. (三對角線矩陣)
- $\mathbf{p = q = 2}$ and $\mathbf{p = q = 4}$ are two special cases of band matrices that occur frequently in the boundary-value problems of ODEs. (微分方程的邊界值問題)

Crout Factorization of a 4×4 Tridiagonal Matrix (1/2)

The 4×4 tridiagoal matrix is factored as $A = LU$, i.e.,

$$\begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ 0 & l_{32} & l_{33} & 0 \\ 0 & 0 & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} \mathbf{1} & u_{12} & 0 & 0 \\ 0 & \mathbf{1} & u_{23} & 0 \\ 0 & 0 & \mathbf{1} & u_{34} \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{l_{11}} & l_{11}u_{12} & 0 & 0 \\ \mathbf{l_{21}} & l_{21}u_{12} + l_{22} & l_{22}u_{23} & 0 \\ 0 & \mathbf{l_{32}} & l_{32}u_{23} + l_{33} & l_{33}u_{34} \\ 0 & 0 & \mathbf{l_{43}} & l_{43}u_{34} + l_{44} \end{bmatrix}.$$

Crout Factorization of a 4×4 Tridiagonal Matrix (2/2)

Thus the **10** unknowns are determined by

$$
\begin{aligned}
&\mathbf{a_{11}}: l_{11} = a_{11}, \quad \mathbf{a_{12}}: u_{12} = a_{12}/l_{11}, \\
&\mathbf{a_{21}}: l_{21} = a_{21}, \quad \mathbf{a_{22}}: l_{22} = a_{22} - l_{21}u_{12}, \\
&\mathbf{a_{23}}: u_{23} = a_{23}/l_{22}, \quad \mathbf{a_{32}}: l_{32} = a_{32}, \\
&\mathbf{a_{33}}: l_{33} = a_{33} - l_{32}u_{23}, \quad \mathbf{a_{34}}: u_{34} = a_{34}/l_{33}, \\
&\mathbf{a_{43}}: l_{43} = a_{43}, \quad \mathbf{a_{44}}: l_{44} = a_{44} - l_{43}u_{34}.
\end{aligned} \tag{9}
$$

and $l_{43} = a_{43}$, $l_{44} = a_{44} - l_{43}u_{34}$.

The LU Factorization of Tridiagonal Matrices

Goal: the tridiagonal $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ will be factored as

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & & \vdots \\ 0 & a_{32} & a_{33} & a_{34} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & l_{n,n-1} & l_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & u_{n-1,n} \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} \equiv LU.$$

Crout (or Doolittle) Method

- $(2n-1)$ entries of L and $(n-1)$ entries of U will be determined by at most $(3n-2)$ **nonzero entries** of A.
- With the special structures of L and U, we obtain

$$
\begin{aligned}
&a_{11} = l_{11};\\
&a_{i,i-1} = l_{i,i-1}, \quad i = 2,3,\ldots,n; \ \text{(off-diagonal terms of } L)\\
&a_{ii} = l_{i,i-1}\cdot u_{i-1,i} + l_{ii}, \quad i = 2,3,\ldots,n;\\
&a_{i,i+1} = l_{ii}\cdot u_{i,i+1}, \quad i = 1,2,\ldots,n-1.
\end{aligned}
$$

Solution to Tridiagonal Linear Systems

If the tridiagonal matrix $A = LU$, where L and U are also tridiagonal matrices defined above, then $Ax = b$ can solved efficiently as follows.

1. Fist solve the **lower triangular system**

$$Lz = b$$

for the vector z with $O(n)$ operation counts.

2. Then we further solve the **upper triangular system**

$$Ux = z$$

for the solution x with $O(n)$ operation counts.

Pseudocode of Crout Method

Algorithm 6.7: Crpit Factorization for Tridiagonal Linear Systems

INPUT dimension n; augmented matrix $A = [a_{ij}] \in \mathbb{R}^{n\times(n+1)}$.

OUTPUT solution $x_1, x_2, \cdots, x_n$ to the **tridiagonal** linear system.

Step 1 Set $l_{11} = a_{11}$; $u_{12} = a_{12}/l_{11}$; $z_1 = a_{1,n+1}/l_{11}$.

Step 2 For $i = 2, \ldots, n-1$ set

$$l_{i,i-1} = a_{i,i-1}; l_{ii} = a_{ii} - l_{i,i-1}u_{i-1,i};$$
$$u_{i,i+1} = a_{i,i+1}/l_{ii}; z_i = (a_{i,n+1} - l_{i,i-1}z_{i-1})/l_{ii}.$$

Step 3 Set $l_{n,n-1} = a_{n,n-1}$; $l_{nn} = a_{nn} - l_{n,n-1}u_{n-1,n}$;
$z_n = (a_{n,n+1} - l_{n,n-1}z_{n-1})/l_{nn}$.

Step 4 Set $x_n = z_n$.

Step 5 For $i = n-1, \ldots, 1$ set $x_i = z_i - u_{i,i+1}x_{i+1}$.

Step 6 OUTPUT$(x_1, \cdots, x_n)$; **STOP**.

Operation Counts for Algorithm 6.7

Remarks

- The Crout Method requires only $(5n-4)$ multiplications/divisions and $(3n-3)$ additions/subtractions.
- Total flop $= 8n-7 = O(n)$. Very cheap!
- The Crout Factorization Alg. will **break down** if $l_{ii} = 0$ for some $1 \leq i \leq n$.

Thm 6.31

If the matrix $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ is **tridiagonal** with

$$|a_{11}| > |a_{12}|, \quad |a_{nn}| > |a_{n,n-1}|,$$
$$|a_{ii}| \geq |a_{i,i-1}| + |a_{i,i+1}| > 0 \quad \text{for } i = 2, 3, \ldots, n-1,$$

then A is nonsingular and Crout Factorization Algorithm produces

$$l_{ii} \neq 0 \quad \text{for } i = 1, 2, \ldots, n.$$

Note: A is **strictly diagonally dominant** or **positive definite** $\Longrightarrow l_{ii} \neq 0 \quad \forall\, i.$

Proof of Thm 6.31

Since $l_{11} = a_{11}$, $|l_{11}| = |a_{11}| > |a_{12}| \geq 0$ and $|u_{12}| = \frac{|a_{12}|}{|a_{11}|} < 1$.
Suppose that $|l_{jj}| > 0$ and $|u_{j,j+1}| < 1$ for $j = 1, 2, \ldots, i-1$. Then

$$\begin{aligned} |l_{ii}| &= |a_{ii} - l_{i,i-1}u_{i-1,i}| = |a_{ii} - a_{i,i-1}u_{i-1,i}| \\ &\geq |a_{ii}| - |a_{i,i-1}||u_{i-1,i}| > |a_{ii}| - |a_{i,i-1}| \geq 0, \end{aligned}$$

and

$$|u_{i,i+1}| = \frac{|a_{i,i+1}|}{|l_{ii}|} < \frac{|a_{i,i+1}|}{|a_{ii}| - |a_{i,i-1}|} \leq 1$$

for $i = 2, \ldots, n-1$. Moreover, since $|u_{n-1,n}| < 1$, we also have

$$|l_{nn}| = |a_{nn} - l_{n,n-1}u_{n-1,n}| = |a_{nn} - a_{n,n-1}u_{n-1,n}| > |a_{nn}| - |a_{n,n-1}| \geq 0.$$

So, $\det(A) = \det(L)\det(U) = l_{11} \cdot l_{22} \cdots l_{nn} \neq 0$ and hence A is nonsingular.

Example 5, p. 423

Solve the **tridiagonal** linear system $Ax = b$, where

$$A = \text{trid}(-1, 2, -1) = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

using the **Crout Factorization Algorithm**.

Sol: From Eqs. (9) and the entries of A, we obtain

$$l_{11} = 2, \quad u_{12} = -1/2, \quad l_{21} = -1, \quad l_{22} = 3/2, \quad u_{23} = -2/3,$$
$$l_{32} = -1, \quad l_{33} = 4/3, \quad u_{34} = -3/4, \quad l_{43} = -1, \quad l_{44} = 5/4.$$

So, the Crout factorization of A is given by

$$A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ -1 & 3/2 & 0 & 0 \\ 0 & -1 & 4/3 & 0 \\ 0 & 0 & -1 & 5/4 \end{bmatrix} \begin{bmatrix} 1 & -1/2 & 0 & 0 \\ 0 & 1 & -2/3 & 0 \\ 0 & 0 & 1 & -3/4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \equiv LU.$$

1. Solving $Lz = b$ by **forward substitution** $\Rightarrow z = \begin{bmatrix} 1/2 \\ 1/3 \\ 1/4 \\ 1 \end{bmatrix}$.
2. Solving $Ux = z$ by **backward substitution** $\Rightarrow x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$.

Thank you for your attention!