Numerical Analysis I
Direct Methods for Solving Linear Systems

Instructor: Wei-Cheng Wang [1]

Department of Mathematics
National TsingHua University

Fall 2017

[1] These slides are based on Prof. Tsung-Ming Huang(NTNU)'s original slides

Outline

Linear systems of equations

Three operations to simplify the linear system:

1. $(\lambda E_i) \to (E_i)$: Equation E_i can be multiplied by $\lambda \neq 0$ with the resulting equation used in place of E_i.
2. $(E_i + \lambda E_j) \to (E_i)$: Equation E_j can be multiplied by $\lambda \neq 0$ and added to equation E_i with the resulting equation used in place of E_i.
3. $(E_i) \leftrightarrow (E_j)$: Equation E_i and E_j can be transposed in order.

Example

$$
\begin{array}{rrcrcrcrcr}
E_1: & x_1 & + & x_2 & & & + & 3x_4 & = & 4, \\
E_2: & 2x_1 & + & x_2 & - & x_3 & + & x_4 & = & 1, \\
E_3: & 3x_1 & - & x_2 & - & x_3 & + & 2x_4 & = & -3, \\
E_4: & -x_1 & + & 2x_2 & + & 3x_3 & - & x_4 & = & 4.
\end{array}
$$

Solution:

- $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$ and $(E_4 + E_1) \to (E_4)$:

$$\begin{array}{rrrrrrrrrl}
E_1: & x_1 & + & x_2 & & & + & 3x_4 & = & 4, \\
E_2: & & - & x_2 & - & x_3 & - & 5x_4 & = & -7, \\
E_3: & & - & 4x_2 & - & x_3 & - & 7x_4 & = & -15, \\
E_4: & & & 3x_2 & + & 3x_3 & + & 2x_4 & = & 8.
\end{array}$$

- $(E_3 - 4E_2) \to (E_3)$ and $(E_4 + 3E_2) \to (E_4)$:

$$\begin{array}{rrrrrrrrrl}
E_1: & x_1 & + & x_2 & & & + & 3x_4 & = & 4, \\
E_2: & & - & x_2 & - & x_3 & - & 5x_4 & = & -7, \\
E_3: & & & & & 3x_3 & + & 13x_4 & = & 13, \\
E_4: & & & & & & - & 13x_4 & = & -13.
\end{array}$$

- Backward-substitution process:
  1. $E_4 \;\Rightarrow\; x_4 = 1$
  2. Solve E_3 for x_3:
  $$x_3 = \frac{1}{3}(13 - 13x_4) = \frac{1}{3}(13 - 13) = 0.$$
  3. E_2 gives
  $$x_2 = -(-7 + 5x_4 + x_3) = -(-7 + 5 + 0) = 2.$$
  4. E_1 gives
  $$x_1 = 4 - 3x_4 - x_2 = 4 - 3 - 2 = -1.$$

$\square$

Solve linear systems of equations

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

Rewrite in the matrix form

$$Ax = b, \tag{1}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

and $[A, b]$ is called the augmented matrix.

Gaussian elimination with backward substitution

The augmented matrix in previous example is

$$\left[\begin{array}{rrrr|r} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right].$$

- $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$ and $(E_4 + E_1) \to (E_4)$:

$$\left[\begin{array}{rrrr|r} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right].$$

- $(E_3 - 4E_2) \to (E_3)$ and $(E_4 + 3E_2) \to (E_4)$:

$$\left[\begin{array}{rrrr|r} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right].$$

The general Gaussian elimination procedure

- Provided $a_{11} \neq 0$, for each $i = 2, 3, \ldots, n$,

$$\left(E_i - \frac{a_{i1}}{a_{11}} E_1\right) \to (E_i).$$

  Transform all the entries in the first column below the diagonal to zero. Denote the new entry in the ith row and jth column by a_{ij}.

- For $i = 2, 3 \ldots, n-1$, provided $a_{ii} \neq 0$,

$$\left(E_j - \frac{a_{ji}}{a_{ii}} E_i\right) \to (E_j), \ \forall\ j = i+1, i+2, \ldots, n.$$

  Transform all the entries in the ith column below the diagonal to zero.

- Result: An upper triangular matrix:

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & a_{nn} & b_n \end{array}\right].$$

The process of Gaussian elimination result in a sequence of matrices as follows:

$$A = A^{(1)} \to A^{(2)} \to \cdots \to A^{(n)} = \text{upper triangular matrix},$$

The matrix $A^{(k)}$ has the following form:

$$A^{(k)} = \left[\begin{array}{ccc|c|cccc}
a_{11}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \cdots & a_{1j}^{(1)} & \cdots & a_{1n}^{(1)} \\
\vdots & \ddots & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,j}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\
\hline
0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kj}^{(k)} & \cdots & a_{kn}^{(k)} \\
\hline
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & a_{ik}^{(k)} & \cdots & a_{ij}^{(k)} & \cdots & a_{in}^{(k)} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nj}^{(k)} & \cdots & a_{nn}^{(k)}
\end{array}\right]$$

The entries of $A^{(k)}$ are produced by the formula

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)}, & \text{for } i = 1, \ldots, k-1,\ j = 1, \ldots, n; \\ 0, & \text{for } i = k, \ldots, n,\ j = 1, \ldots, k-1; \\ a_{ij}^{(k-1)} - \dfrac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} \times a_{k-1,j}^{(k-1)}, & \text{for } i = k, \ldots, n,\ j = k, \ldots, n. \end{cases}$$

- The procedure will fail if one of the elements $a_{11}^{(1)}, a_{22}^{(2)}, \ldots, a_{nn}^{(n)}$ is zero.
- $a_{ii}^{(i)}$ is called the pivot element.

Backward substitution

The new linear system is triangular:

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ & & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & & & \vdots & & \\ & & & & & & a_{nn}x_n & = & b_n \end{array}$$

- Solving the nth equation for x_n gives
$$x_n = \frac{b_n}{a_{nn}}.$$
- Solving the $(n-1)$th equation for x_{n-1} and using the value for x_n yields
$$x_{n-1} = \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}}.$$
- In general,
$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}, \ \forall\ i = n-1, n-2, \ldots, 1.$$

Algorithm (Backward Substitution)

Suppose that $U \in \mathbb{R}^{n\times n}$ is nonsingular upper triangular and $b \in \mathbb{R}^n$. This algorithm computes the solution of $Ux = b$.

For $i = n, \ldots, 1$
  $tmp = 0$
  For $j = i + 1, \ldots, n$
    $tmp = tmp + U(i, j) * x(j)$
  End for
  $x(i) = (b(i) - tmp)/U(i, i)$
End for

Example

Solve system of linear equations.

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix}$$

Solution:

1^{st} step Use 6 as pivot element, the first row as pivot row, and multipliers $2, \frac{1}{2}, -1$ are produced to reduce the system to

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix}$$

2^{nd} step Use -4 as pivot element, the second row as pivot row, and multipliers $3, -\frac{1}{2}$ are computed to reduce the system to

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix}$$

3^{rd} step Use 2 as pivot element, the third row as pivot row, and multipliers 2 is found to reduce the system to

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -3 \end{bmatrix}$$

4th step The backward substitution is applied:

$$
\begin{aligned}
x_4 &= \frac{-3}{-3} = 1, \\
x_3 &= \frac{-9+5x_4}{2} = \frac{-9+5}{2} = -2, \\
x_2 &= \frac{10-2x_4-2x_3}{-4} = \frac{10-2+4}{-4} = -3, \\
x_1 &= \frac{12-4x_4-2x_3+2x_2}{6} = \frac{12-4+4-6}{6} = 1.
\end{aligned}
$$

□

- This example is done since $a_{kk}^{(k)} \neq 0$ for all $k = 1, 2, 3, 4$.
- How to do if $a_{kk}^{(k)} = 0$ for some k?

Example

Solve system of linear equations.

$$\begin{bmatrix} 1 & -1 & 2 & -1 \\ 2 & -2 & 3 & -3 \\ 1 & 1 & 1 & 0 \\ 1 & -1 & 4 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -8 \\ -20 \\ -2 \\ 4 \end{bmatrix}$$

Solution:

1^{st} step Use 1 as pivot element, the first row as pivot row, and multipliers $2, 1, 1$ are produced to reduce the system to

$$\begin{bmatrix} 1 & -1 & 2 & -1 \\ 0 & 0 & -1 & -1 \\ 0 & 2 & -1 & 1 \\ 0 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -8 \\ -4 \\ 6 \\ 12 \end{bmatrix}$$

2^{nd} step Since $a_{22}^{(2)} = 0$ and $a_{32}^{(2)} \neq 0$, the operation $(E_2) \leftrightarrow (E_3)$ is performed to obtain a new system

$$\begin{bmatrix} 1 & -1 & 2 & -1 \\ 0 & 2 & -1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -8 \\ 6 \\ -4 \\ 12 \end{bmatrix}$$

3^{rd} step Use -1 as pivot element, the third row as pivot row, and multipliers -2 is found to reduce the system to

$$\begin{bmatrix} 1 & -1 & 2 & -1 \\ 0 & 2 & -1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -8 \\ 6 \\ -4 \\ 4 \end{bmatrix}$$

4^{th} step The backward substitution is applied:

$$\begin{aligned} x_4 &= \frac{4}{2} = 2, \\ x_3 &= \frac{-4 + x_4}{-1} = 2, \\ x_2 &= \frac{6 - x_4 + x_3}{2} = 3, \\ x_1 &= \frac{-8 + x_4 - 2x_3 + x_2}{1} = -7. \end{aligned}$$

$\square$

- This example illustrates what is done if $a_{kk}^{(k)} = 0$ for some k.
- If $a_{pk}^{(k)} \neq 0$ for some p with $k + 1 \leq p \leq n$, then the operation $(E_k) \leftrightarrow (E_p)$ is performed to obtain new matrix.
- If $a_{pk}^{(k)} = 0$ for each p, then the linear system does not have a unique solution and the procedure stops.

Algorithm (Gaussian elimination)

Given $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, this algorithm implements the Gaussian elimination procedure to reduce A to upper triangular and modify the entries of b accordingly.

For $k = 1, \ldots, n-1$
  Let p be the smallest integer with $k \le p \le n$ and $a_{pk} \neq 0$.
  If $\nexists\ p$, then stop.
  If $p \neq k$, then perform $(E_p) \leftrightarrow (E_k)$.
  For $i = k+1, \ldots, n$
    $t = A(i,k)/A(k,k)$
    $A(i,k) = 0$
    $b(i) = b(i) - t \times b(k)$
    For $j = k+1, \ldots, n$
      $A(i,j) = A(i,j) - t \times A(k,j)$
    End for
  End for
End for

Number of floating-point arithmetic operations

Eliminate kth column

For $i = k+1, \ldots, n$
  $t = A(i,k)/A(k,k)$; $b(i) = b(i) - t \times b(k)$.
  For $j = k+1, \ldots, n$
    $A(i,j) = A(i,j) - t \times A(k,j)$
  End for
End for

- Multiplications/divisions

$$(n-k) + (n-k) + (n-k)(n-k) = (n-k)(n-k+2)$$

- Additions/subtractions

$$(n-k) + (n-k)(n-k) = (n-k)(n-k+1)$$

- Total number of operations for multiplications/divisions

$$
\begin{aligned}
\sum_{k=1}^{n-1}(n-k)(n-k+2) &= \sum_{k=1}^{n-1}(n^2-2nk+k^2+2n-2k) \\
&= (n^2+2n)\sum_{k=1}^{n-1}1 - 2(n+1)\sum_{k=1}^{n-1}k + \sum_{k=1}^{n-1}k^2 \\
&= (n2+2n)(n-1) - 2(n+1)\frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} \\
&= \frac{2n^3+3n^2-5n}{6}.
\end{aligned}
$$

- Total number of operations for additions/subtractions

$$
\begin{aligned}
\sum_{k=1}^{n-1}(n-k)(n-k+1) &= \sum_{k=1}^{n-1}(n^2-2nk+k^2+n-k) \\
&= (n^2+n)\sum_{k=1}^{n-1}1 - (2n+1)\sum_{k=1}^{n-1}k + \sum_{k=1}^{n-1}k^2 = \frac{n^3-n}{3}.
\end{aligned}
$$

Backward substitution

$$
\begin{array}{l}
x(n) = b(n)/U(n,n). \\
\text{For } i = n-1, \ldots, 1 \\
\quad tmp = U(i,i+1) \times x(i+1) \\
\quad \text{For } j = i+2, \ldots, n \\
\quad\quad tmp = tmp + U(i,j) \times x(j) \\
\quad \text{End for} \\
\quad x(i) = (b(i) - tmp)/U(i,i) \\
\text{End for}
\end{array}
$$

- Multiplications/divisions

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

- Additions/subtractions

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

The total number of arithmetic operations in Gaussian elimination with backward substitution is:

- Multiplications/divisions

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

- Additions/subtractions

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

Pivoting Strategies

- If $a_{kk}^{(k)}$ is small in magnitude compared to $a_{jk}^{(k)}$, then

$$|m_{jk}| = \left| \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}} \right| > 1.$$

Round-off error introduced in the computation of

$$a_{j\ell}^{(k+1)} = a_{j\ell}^{(k)} - m_{jk} a_{k\ell}^{(k)}, \quad \text{for} \quad \ell = k+1, \ldots, n.$$

- Error can be increased when performing the backward substitution for

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

with a small value of $a_{kk}^{(k)}$.

Example

The linear system

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

has the exact solution $x_1 = 10.00$ and $x_2 = 1.000$. Suppose Gaussian elimination is performed on this system using four-digit arithmetic with rounding.

- $a_{11} = 0.0030$ is small and

$$m_{21} = \frac{5.291}{0.0030} = 1763.6\bar{6} \approx 1764.$$

- Perform $(E_2 - m_{21}E_1) \to (E_2)$:

$$\begin{array}{rcrcr} 0.0030x_1 & + & 59.14x_2 & = & 59.17 \\ & - & 104309.37\bar{6}x_2 & = & -104309.37\bar{6}. \end{array}$$

- Rounding with four-digit arithmetic:
  Coefficient of x_2:

$$
\begin{aligned}
& -6.130 - 1764 \times 59.14 = -6.130 - 104322.96 \\
\approx\ & -6.130 - 104300 = -104306.13 \\
\approx\ & -104300.
\end{aligned}
$$

Right hand side:

$$
\begin{aligned}
& 46.78 - 1764 \times 59.17 = 46.78 - 104375.88 \\
\approx\ & 46.78 - 104400 = -104353.22 \\
\approx\ & -104400.
\end{aligned}
$$

New linear system:

$$
\begin{array}{rcrcr}
0.0030x_1 & + & 59.14x_2 & = & 59.17 \\
 & - & 104300x_2 & \approx & -104400.
\end{array}
$$

- Approximated solution:

$$\begin{aligned} x_2 &= \frac{104400}{104300} \approx 1.001, \\ x_1 &= \frac{59.17 - 59.14 \times 1.001}{0.0030} = \frac{59.17 - 59.19914}{0.0030} \\ &\approx \frac{59.17 - 59.20}{0.0030} = -10.00. \end{aligned}$$

This ruins the approximation to the actual value $x_1 = 10.00$. $\square$

Partial pivoting

- To avoid the pivot element small relative to other entries, pivoting is performed by selecting an element $a_{pq}^{(k)}$ with a larger magnitude as the pivot.
- Specifically, select pivoting $a_{pk}^{(k)}$ with

$$|a_{pk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|$$

  and perform $(E_k) \leftrightarrow (E_p)$.
- This row interchange strategy is called partial pivoting.

Example

Reconsider the linear system

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

- Find pivoting with

$$\max\{|a_{11}|, |a_{21}|\} = 5.291 = |a_{21}|.$$

- Perform $(E_2) \leftrightarrow (E_1)$:

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

- The multiplier for new system is

$$m_{21} = \frac{a_{21}}{a_{11}} = 0.0005670.$$

- The operation $(E_2 - m_{21}E_1) \to (E_2)$ reduces the system to

$$\begin{array}{rcrcl} 5.291x_1 & - & 6.130x_2 & = & 46.78, \\ & & 59.14x_2 & \approx & 59.14. \end{array}$$

- The four-digit answers resulting from the backward substitution are the correct values $x_1 = 10.00$ and $x_2 = 1.000$. $\square$

Example

The linear system

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

is the same as that in previous example except that all the entries in the first equation have been multiplied by 10^4.

The pivoting is $a_{11} = 30.00$ and the multiplier

$$m_{21} = \frac{5.291}{30.00} = 0.1764$$

leads to the system

$$\begin{array}{rcrcr} 30.00x_1 & + & 591400x_2 & = & 591700 \\ & - & 104300x_2 & \approx & -104400, \end{array}$$

which has inaccurate solution $x_2 \approx 1.001$ and $x_1 \approx -10.00$.

Scaled partial pivoting

- Define a scale factor s_i as

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

- If $s_i = 0$ for some i, then the system has no unique solution.
- In the ith column, choose the least integer $p \ge i$ with

$$\frac{|a_{pi}|}{s_p} = \max_{i \le k \le n} \frac{|a_{ki}|}{s_k}$$

and perform $(E_i) \leftrightarrow (E_p)$ if $p \ne i$.

- The scale factors $s_1, \ldots, s_n$ are computed only once and must also be interchanged when row interchanges are performed.

Example

Apply scaled partial pivoting to the linear system

$$\begin{array}{rrcrcr} E_1: & 30.00x_1 & + & 591400x_2 & = & 591700, \\ E_2: & 5.291x_1 & - & 6.130x_2 & = & 46.78. \end{array}$$

The scale factors s_1 and s_2 are

$$s_1 = \max\{|30.00|, |591400|\} = 591400$$

and

$$s_2 = \max\{|5.291|, |-6.130|\} = 6.130.$$

Consequently,

$$\begin{aligned} \frac{|a_{11}|}{s_1} &= \frac{30.00}{591400} = 0.5073 \times 10^{-4}, \\ \frac{|a_{21}|}{s_2} &= \frac{5.291}{6.130} = 0.8631, \end{aligned}$$

and the interchange $(E_1) \leftrightarrow (E_2)$ is made.

Applying Gaussian elimination to the new system

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

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

Matrix factorization

- This equation has a unique solution $x = A^{-1}b$ when the coefficient matrix A is nonsingular.
- Use Gaussian elimination to factor the coefficient matrix into a product of matrices. The factorization is called LU-factorization and has the form $A = LU$, where L is unit lower triangular and U is upper triangular.
- The solution to the original problem $Ax = LUx = b$ is then found by a two-step triangular solve process:

$$Ly = b, \qquad Ux = y.$$

- LU factorization requires $O(n^3)$ arithmetic operations. Forward substitution for solving a lower-triangular system $Ly = b$ requires $O(n^2)$. Backward substitution for solving an upper-triangular system $Ux = y$ requires $O(n^2)$ arithmetic operations.

For a given vector $v \in \mathbb{R}^n$ with $v_k \neq 0$ for some $1 \leq k \leq n$, let

$$\ell_{ik} = \frac{v_i}{v_k}, \quad i = k+1, \ldots, n,$$
$$\ell_k = \begin{bmatrix} 0 & \cdots & 0 & \ell_{k+1,k} & \cdots & \ell_{n,k} \end{bmatrix}^T,$$

and

$$M_k = I - \ell_k e_k^T = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\ell_{k+1,k} & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\ell_{n,k} & 0 & \cdots & 1 \end{bmatrix}.$$

Then one can verify that

$$M_k v = \begin{bmatrix} v_1 & \cdots & v_k & 0 & \cdots & 0 \end{bmatrix}^T.$$

M_k is called a Gaussian transformation, the vector ℓ_k a Gauss vector. Furthermore, one can verify that

$$M_k^{-1} = (I - \ell_k e_k^T)^{-1} = I + \ell_k e_k^T = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{k+1,k} & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{n,k} & 0 & \cdots & 1 \end{bmatrix}.$$

Given a nonsingular matrix $A \in \mathbb{R}^{n\times n}$, denote $A^{(1)} \equiv [a_{ij}^{(1)}] = A$. If $a_{11}^{(1)} \neq 0$, then

$$M_1 = I - \ell_1 e_1^T,$$

where

$$\ell_1 = \begin{bmatrix} 0 & \ell_{21} & \cdots & \ell_{n1} \end{bmatrix}^T, \quad \ell_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \; i = 2, \ldots, n,$$

can be formed such that

$$A^{(2)} = M_1 A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix},$$

where

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \ell_{i1} \times a_{1j}^{(1)}, \text{ for } i = 2, \ldots, n \text{ and } j = 2, \ldots, n.$$

In general, at the k-th step, we are confronted with a matrix

$$
\begin{aligned}
A^{(k)} &= M_{k-1}\cdots M_2 M_1 A^{(1)} \\
&= \left[\begin{array}{cccc|ccc}
a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\
0 & a_{22}^{(2)} & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} \\
\vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\
0 & 0 & \cdots & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\
\hline
0 & 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\
\vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & a_{kn}^{(k)} & \cdots & a_{nn}^{(k)}
\end{array}\right].
\end{aligned}
$$

If the pivot $a_{kk}^{(k)} \neq 0$, then the multipliers

$$
\ell_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad i = k+1, \ldots, n,
$$

can be computed and the Gaussian transformation

$$M_k = I - \ell_k e_k^T, \quad \text{where} \quad \ell_k = \left[\begin{array}{cccccc} 0 & \cdots & 0 & \ell_{k+1,k} & \cdots & \ell_{nk} \end{array}\right]^T,$$

can be applied to the left of $A^{(k)}$ to obtain

$$A^{(k+1)} = M_k A^{(k)}$$

$$= \left[\begin{array}{cccc|cccc}
a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1n}^{(1)} \\
0 & a_{22}^{(2)} & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & a_{2,k+1}^{(2)} & \cdots & a_{2n}^{(2)} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & \cdots & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & a_{k-1,k+1}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\
\hline
0 & 0 & \cdots & 0 & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{kn}^{(k)} \\
\vdots & \vdots & & \vdots & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0 & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)}
\end{array}\right],$$

in which

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \ell_{ik} a_{kj}^{(k)}, \tag{2}$$

for $i = k+1, \ldots, n$, $j = k+1, \ldots, n$. Upon the completion,

$$U \equiv A^{(n)} = M_{n-1} \cdots M_2 M_1 A$$

is upper triangular. Hence

$$A = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1} U \equiv LU,$$

where

$$
\begin{aligned}
L \equiv M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1} &= (I - \ell_1 e_1^T)^{-1} (I - \ell_2 e_2^T)^{-1} \cdots (I - \ell_{n-1} e_{n-1}^T)^{-1} \\
&= (I + \ell_1 e_1^T)(I + \ell_2 e_2^T) \cdots (I + \ell_{n-1} e_{n-1}^T) \\
&= I + \ell_1 e_1^T + \ell_2 e_2^T + \cdots + \ell_{n-1} e_{n-1}^T \\
&= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{bmatrix}
\end{aligned}
$$

is unit lower triangular. This matrix factorization is called the LU-factorization of A.

Algorithm (LU Factorization with Direct Comparison)

Given $A \in \mathbb{R}^{n\times n}$ nonsingular, this algorithm computes a unit lower triangular matrix L and an upper triangular matrix U such that $A = LU$.

$i = 1$
    For $j = 1, \ldots, n$
        $U(1,j) = A(1,j)$
    End for j
    For $i' = 2, \ldots, n$
        $L(i',1) = A(i',1)/U(1,1)$
    End for i'
For $i = 2, \ldots, n-1$
    For $j = i, \ldots, n$
        $U(i,j) = A(i,j) - \sum_{k=1}^{i-1} L(i,k)U(k,j)$
    End for j
    For $i' = i+1, \ldots, n$
        $L(i',i) = (\ A(i',i) - \sum_{k=1}^{i-1} L(i',k)U(k,i)\ )/U(i,i)$
    End for i'
End for i
$U(n,n) = A(n,n) - \sum_{k=1}^{n-1} L(n,k)U(k,n)$

Algorithm (Recursive LU Factorization)

For $i = 1, \ldots, n-1$
  For $j = i, \ldots, n$
    $U(i,j) = A(i,j)$ (Compute ith row of U)
  End for j
  For $i' = i+1, \ldots, n$
    $L(i',i) = A(i',i)/A(i,i)$ (Compute ith column of L)
    For $j = i+1, \ldots, n$
      $U(i',j) = U(i',j) - L(i',i)U(i,j)$
      (Update the right-lower sub-matrix, row by row)
    End for j
  End for i'
End for i

Algorithm (Memory Saving Recursive LU Factorization)

Memory saving version of LU factorization. The matrix A is overwritten by L and U.

For $i = 1, \ldots, n-1$
    For $i' = i+1, \ldots, n$
        $A(i', i) = A(i', i)/A(i, i)$
        For $j = i+1, \ldots, n$
            $A(i', j) = A(i', j) - A(i', i)A(i, j)$
        End for
    End for
End for

Forward Substitution

When a linear system $Lx = b$ is lower triangular of the form

$$\begin{bmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

where all diagonals $\ell_{ii} \neq 0$, x_i can be obtained by the following procedure

$$\begin{aligned} x_1 &= b_1/\ell_{11}, \\ x_2 &= (b_2 - \ell_{21}x_1)/\ell_{22}, \\ x_3 &= (b_3 - \ell_{31}x_1 - \ell_{32}x_2)/\ell_{33}, \\ &\vdots \\ x_n &= (b_n - \ell_{n1}x_1 - \ell_{n2}x_2 - \cdots - \ell_{n,n-1}x_{n-1})/\ell_{nn}. \end{aligned}$$

The general formulation for computing x_i is

$$x_i = \left(b_i - \sum_{j=1}^{i-1} \ell_{ij} x_j \right) \Big/ \ell_{ii}, \qquad i = 1, 2, \ldots, n.$$

Algorithm (Forward Substitution)

Suppose that $L \in \mathbb{R}^{n \times n}$ is nonsingular lower triangular and $b \in \mathbb{R}^n$. This algorithm computes the solution of $Lx = b$.

For $i = 1, \ldots, n$
  $tmp = 0$
  For $j = 1, \ldots, i-1$
    $tmp = tmp + L(i,j) * x(j)$
  End for
  $x(i) = (b(i) - tmp)/L(i,i)$
End for

Example

$$\begin{array}{lrcrcrcrcr} E_1: & x_1 & + & x_2 & & & + & 3x_4 & = & 4, \\ E_2: & 2x_1 & + & x_2 & - & x_3 & + & x_4 & = & 1, \\ E_3: & 3x_1 & - & x_2 & - & x_3 & + & 2x_4 & = & -3, \\ E_4: & -x_1 & + & 2x_2 & + & 3x_3 & - & x_4 & = & 4. \end{array}$$

Solution:

- The sequence $\{(E_2 - 2E_1) \to (E_2), (E_3 - 3E_1) \to (E_3), (E_4 - (-1)E_1) \to (E_4), (E_3 - 4E_2) \to (E_3), (E_4 - (-3)E_2) \to (E_4)\}$ converts the system to the triangular system

$$\begin{array}{rcrcrcrcr} x_1 & + & x_2 & & & + & 3x_4 & = & 4, \\ & - & x_2 & - & x_3 & - & 5x_4 & = & -7, \\ & & & & 3x_3 & + & 13x_4 & = & 13, \\ & & & & & - & 13x_4 & = & -13. \end{array}$$

- LU factorization of A:

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

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & -4 & -1 & -7 \\ 0 & 3 & 3 & 2 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & 1 & 0 & 3 \\ 2 & -1 & -1 & -5 \\ 3 & -4 & -1 & -7 \\ -1 & 3 & 3 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & 1 & 0 & 3 \\ 2 & -1 & -1 & -5 \\ 3 & 4 & 3 & 13 \\ -1 & 3 & 0 & -13 \end{bmatrix}$$

$$= LU \text{ or } \underline{L}U.$$

- Solve $Ly = b$:

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

which implies that

$$\begin{aligned} y_1 &= 8, \\ y_2 &= 7 - 2y_1 = -9, \\ y_3 &= 14 - 3y_1 - 4y_2 = 26, \\ y_4 &= -7 + y_1 + 3y_2 = -26. \end{aligned}$$

- Solve $Ux = y$:

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

which implies that

$$\begin{aligned} x_4 &= 2, \\ x_3 &= (26 - 13x_4)/3 = 0, \\ x_2 &= (-9 + 5x_4 + x_3)/(-1) = -1, \\ x_1 &= 8 - 3x_4 - x_2 = 3. \end{aligned}$$

Partial pivoting

At the k-th step, select pivoting $a_{pk}^{(k)}$ with

$$|a_{pk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|$$

and perform $(E_k) \leftrightarrow (E_p)$. That is, choose a permutation matrix

$$P_k = \begin{bmatrix} I_{k-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & I_{p-k-1} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{n-p} \end{bmatrix}$$

so that

$$\left|(P_k A^{(k)})_{kk}\right| = \max_{k \le i \le n} \left|(A^{(k)})_{ik}\right|$$

and

$$A^{(k+1)} = M^{(k)} P_k A^{(k)}.$$

Let $P_1, \ldots, P_{k-1}$ be the permutations chosen and $M_1, \ldots M_{k-1}$ denote the Gaussian transformations performed in the first $k-1$ steps. At the k-th step, a permutation matrix P_k is chosen so that

$$|(P_k M_{k-1} \cdots M_1 P_1 A)_{kk}| = \max_{k \le i \le n} |(M_{k-1} \cdots M_1 P_1 A)_{ik}| .$$

As a consequence, $|\ell_{ij}| \le 1$ for $i = 1, \ldots, n$, $j = 1, \ldots, i$. Upon completion, we obtain an upper triangular matrix

$$U \equiv M_{n-1} P_{n-1} \cdots M_1 P_1 A. \tag{3}$$

Since any P_k is symmetric and $P_k^T P_k = P_k^2 = I$, we have

$$M_{n-1} P_{n-1} \cdots M_2 P_2 M_1 P_2 \cdots P_{n-1} P_{n-1} \cdots P_2 P_1 A = U,$$

therefore,

$$P_{n-1} \cdots P_1 A = (M_{n-1} P_{n-1} \cdots M_2 P_2 M_1 P_2 \cdots P_{n-1})^{-1} U.$$

In summary, Gaussian elimination with partial pivoting leads to the LU factorization

$$PA = LU, \tag{4}$$

where

$$P = P_{n-1} \cdots P_1$$

is a permutation matrix, and

$$\begin{aligned} L &\equiv (M_{n-1}P_{n-1} \cdots M_2 P_2 M_1 P_2 \cdots P_{n-1})^{-1} \\ &= P_{n-1} \cdots P_2 M_1^{-1} P_2 M_2^{-1} \cdots P_{n-1} M_{n-1}^{-1}. \end{aligned}$$

Since,

$$P_j = \begin{bmatrix} I_{j-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & I_{p-j-1} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{n-p} \end{bmatrix}, \quad \ell_j = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \ell_{j+1,j} \\ \vdots \\ \ell_{nj} \end{bmatrix},$$

it implies that for $i < j$,

$$e_i^T P_j = e_i^T, \quad e_i^T \ell_j = 0,$$
$$P_j \ell_i = \begin{bmatrix} 0 & \cdots & 0 & \tilde{\ell}_{i+1,i} & \cdots & \tilde{\ell}_{n,i} \end{bmatrix}^T \equiv \tilde{\ell}_i,$$

$\Rightarrow$

$$P_2 M_1^{-1} P_2 = P_2(I + \ell_1 e_1^T) P_2 = I + \tilde{\ell}_1 e_1^T$$

$\Rightarrow$

$$P_2 M_1^{-1} P_2 M_2^{-1} = (I + \tilde{\ell}_1 e_1^T)(I + \ell_2 e_2^T) = I + \tilde{\ell}_1 e_1^T + \ell_2 e_2^T,$$

$\Rightarrow$

$$P_3 \left(P_2 M_1^{-1} P_2 M_2^{-1}\right) P_3 = I + \hat{\ell}_1 e_1^T + \tilde{\ell}_2 e_2^T$$

$\Rightarrow \cdots$

Therefore, L is unit lower triangular.

Algorithm (LU-factorization with Partial Pivoting)

Given a nonsingular $A \in \mathbb{R}^{n\times n}$, this algorithm finds a permutation P, and computes a unit lower triangular L and an upper triangular U such that $PA = LU$. A is overwritten by L and U, and P is not formed. An integer array p is instead used for storing the row/column indices.

$p(1:n) = 1:n$
For $k = 1, \ldots, n-1$
  $m = k$
  For $i = k+1, \ldots, n$
    If $|A(p(m), k)| < |A(p(i), k)|$, then $m = i$
  End For
  $\ell = p(k)$; $p(k) = p(m)$; $p(m) = \ell$
  For $i = k+1, \ldots, n$
    $A(p(i), k) = A(p(i), k)/A(p(k), k)$
    For $j = k+1, \ldots, n$
      $A(p(i), j) = A(p(i), j) - A(p(i), k)A(p(k), j)$
    End For
  End For
End For

Since the Gaussian elimination with partial pivoting produces the factorization (4), the linear system problem should comply accordingly

$$Ax = b \Longrightarrow PAx = Pb \Longrightarrow LUx = Pb.$$

Example

Find an LU factorization of

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

- $(E_1) \leftrightarrow (E_2)$, $(E_3 + E_1) \rightarrow (E_3)$ and $(E_4 - E_1) \rightarrow (E_4)$:

$$A^{(2)} = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 0 \end{bmatrix}, P_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

- $(E_2) \leftrightarrow (E_4)$ and $(E_4 + E_3) \to (E_4)$:

$$A^{(3)} = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}, P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

- Permutation matrix P:

$$P = P_2 P_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

- Unit lower triangular matrix L:

$$L = P_2 M_1^{-1} P_2 M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

- The LU factorization of PA:

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

So

$$\begin{aligned} A &= P^{-1}LU = P^T LU \\ &= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}. \end{aligned}$$

Special types of matrices

Definition

A matrix $A \in \mathbb{R}^{n\times n}$ is said to be strictly diagonally dominant if

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

Lemma

If $A \in \mathbb{R}^{n\times n}$ is strictly diagonally dominant, then A is nonsingular.

Proof: Suppose A is singular. Then there exists $x \in \mathbb{R}^n$, $x \neq 0$ such that $Ax = 0$. Let k be the integer index such that

$$|x_k| = \max_{1\leq i\leq n} |x_i| \quad \Longrightarrow \quad \frac{|x_i|}{|x_k|} \leq 1, \quad \forall\, |x_i|.$$

Since $Ax = 0$, for the fixed k, we have

$$\sum_{j=1}^{n} a_{kj}x_j = 0 \Rightarrow a_{kk}x_k = -\sum_{j=1, j\neq k}^{n} a_{kj}x_j \Rightarrow |a_{kk}||x_k| \leq \sum_{j=1, j\neq k}^{n} |a_{kj}||x_j|,$$

which implies

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

But this contradicts the assumption that A is diagonally dominant. Therefore A must be nonsingular. $\square$

Theorem

Gaussian elimination without pivoting preserve the diagonal dominance of a matrix.

Proof: Let $A \in \mathbb{R}^{n\times n}$ be a diagonally dominant matrix and $A^{(2)} = [a_{ij}^{(2)}]$ is the result of applying one step of Gaussian elimination to $A^{(1)} = A$ without any pivoting strategy.

After one step of Gaussian elimination, $a_{i1}^{(2)} = 0$ for $i = 2, \ldots, n$, and the first row is unchanged. Therefore, the property

$$a_{11}^{(2)} > \sum_{j=2}^{n} |a_{1j}^{(2)}|$$

is preserved, and all we need to show is that

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

Using the Gaussian elimination formula (2), we have

$$
\begin{aligned}
|a_{ii}^{(2)}| &= \left| a_{ii}^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1i}^{(1)} \right| = \left| a_{ii} - \frac{a_{i1}}{a_{11}} a_{1i} \right| \\
&\geq |a_{ii}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| \\
&= |a_{ii}| - |a_{i1}| + |a_{i1}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| \\
&= |a_{ii}| - |a_{i1}| + \frac{|a_{i1}|}{|a_{11}|} \left(|a_{11}| - |a_{1i}| \right) \\
&> \sum_{j=2, j\neq i}^{n} |a_{ij}| + \frac{|a_{i1}|}{|a_{11}|} \sum_{j=2, j\neq i}^{n} |a_{1j}| \\
&= \sum_{j=2, j\neq i}^{n} |a_{ij}| + \sum_{j=2, j\neq i}^{n} \frac{|a_{i1}|}{|a_{11}|} |a_{1j}| \\
&\geq \sum_{j=2, j\neq i}^{n} \left| a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \right| = \sum_{j=2, j\neq i}^{n} |a_{ij}^{(2)}|.
\end{aligned}
$$

Thus $A^{(2)}$ is still diagonally dominant. Since the subsequent steps of Gaussian elimination mimic the first, except for being applied to submatrices of smaller size, it suffices to conclude that Gaussian elimination without pivoting preserves the diagonal dominance of a matrix. $\square$

Theorem

Let A be strictly diagonally dominant. Then Gaussian elimination can be performed on $Ax = b$ to obtain its unique solution without row or column interchanges.

Definition

A matrix A is positive definite if it is symmetric and $x^T Ax > 0 \ \forall \ x \neq 0$.

Theorem

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

(a) *A has an inverse;*

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

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

(d) $(a_{ij})^2 < a_{ii}a_{jj}, \ \forall \ i \ne j$.

Proof:

(a) If x satisfies $Ax = 0$, then $x^T Ax = 0$. Since A is positive definite, this implies $x = 0$. Consequently, $Ax = 0$ has only the zero solution, and A is nonsingular.

(b) Since A is positive definite,

$$a_{ii} = e_i^T A e_i > 0,$$

where e_i is the i-th column of the $n \times n$ identify matrix.

(c) For $k \neq j$, define $x = [x_i]$ by

$$x_i = \begin{cases} 0, & \text{if } \; i \neq j \; \text{ and } \; i \neq k, \\ 1, & \text{if } \; i = j, \\ -1, & \text{if } \; i = k. \end{cases}$$

Since $x \neq 0$,

$$0 < x^T A x = a_{jj} + a_{kk} - a_{jk} - a_{kj}.$$

But $A^T = A$, so

$$2a_{kj} < a_{jj} + a_{kk}. \tag{5}$$

Now define $z = [z_i]$ by

$$z_i = \begin{cases} 0, & \text{if } \; i \neq j \; \text{ and } \; j \neq k, \\ 1, & \text{if } \; i = j \; \text{ or } \; i = k. \end{cases}$$

Then $z^T A z > 0$, so

$$-2a_{kj} < a_{jj} + a_{kk}. \tag{6}$$

Equations (5) and (6) imply that for each $k \neq j$,

$$|a_{kj}| < \frac{a_{kk} + a_{jj}}{2} \leq \max_{1\leq i\leq n} |a_{ii}|,$$

so

$$\max_{1\leq k,j\leq n} |a_{kj}| \leq \max_{1\leq i\leq n} |a_{ii}|.$$

(d) For $i \neq j$, define $x = [x_k]$ by

$$x_k = \begin{cases} 0, & \text{if } \ k \neq j \ \text{ and } \ k \neq i, \\ \alpha, & \text{if } \ k = i, \\ 1, & \text{if } \ k = j, \end{cases}$$

where α represents an arbitrary real number.

Since $x \neq 0$,

$$0 < x^T A x = a_{ii}\alpha^2 + 2a_{ij}\alpha + a_{jj} \equiv P(\alpha), \ \forall \ \alpha \in \mathbb{R}.$$

That is the quadratic polynomial $P(\alpha)$ has no real roots. It implies that

$$4a_{ij}^2 - 4a_{ii}a_{jj} < 0 \quad \text{and} \quad a_{ij}^2 < a_{ii}a_{jj}. \qquad \square$$

Definition (Leading principal minor)

Let A be an $n \times n$ matrix. The upper left $k \times k$ submatrix, denoted as

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

is called the leading $k \times k$ principal submatrix, and the determinant of A_k, $\det(A_k)$, is called the leading principal minor.

Theorem

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

Theorem

The symmetric matrix A is positive definite if and only if Gaussian elimination without row interchanges can be performed on $Ax = b$ with all pivot elements positive.

Corollary

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

Theorem

If all leading principal submatrices of $A \in \mathbb{R}^{n\times n}$ are nonsingular, then A has an LU-factorization.

Proof: Proof by mathematical induction.

1. $n = 1$, $A_1 = [a_{11}]$ is nonsingular, then $a_{11} \neq 0$. Let $L_1 = [1]$ and $U_1 = [a_{11}]$. Then $A_1 = L_1 U_1$. The theorem holds.
2. Assume that the leading principal submatrices $A_1, \ldots, A_k$ are nonsingular and A_k has an LU-factorization $A_k = L_k U_k$, where L_k is unit lower triangular and U_k is upper triangular.
3. Show that there exist an unit lower triangular matrix L_{k+1} and an upper triangular matrix U_{k+1} such that $A_{k+1} = L_{k+1} U_{k+1}$.

Write

$$A_{k+1} = \begin{bmatrix} A_k & v_k \\ w_k^T & a_{k+1,k+1} \end{bmatrix},$$

where

$$v_k = \begin{bmatrix} a_{1,k+1} \\ a_{2,k+1} \\ \vdots \\ a_{k,k+1} \end{bmatrix} \quad \text{and} \quad w_k = \begin{bmatrix} a_{k+1,1} \\ a_{k+1,2} \\ \vdots \\ a_{k+1,k} \end{bmatrix}.$$

Since A_k is nonsingular, both L_k and U_k are nonsingular. Therefore, $L_k y_k = v_k$ has a unique solution $y_k \in \mathbb{R}^k$, and $z^t U_k = w_k^T$ has a unique solution $z_k \in \mathbb{R}^k$. Let

$$L_{k+1} = \begin{bmatrix} L_k & 0 \\ z_k^T & 1 \end{bmatrix} \quad \text{and} \quad U_{k+1} = \begin{bmatrix} U_k & y_k \\ 0 & a_{k+1,k+1} - z_k^T y_k \end{bmatrix}.$$

Then L_{k+1} is unit lower triangular, U_{k+1} is upper triangular, and

$$\begin{aligned} L_{k+1}U_{k+1} &= \left[\begin{array}{cc} L_kU_k & L_ky_k \\ z_k^TU_k & z_k^Ty_k + a_{k+1,k+1} - z_k^Ty_k \end{array}\right] \\ &= \left[\begin{array}{cc} A_k & v_k \\ w_k^T & a_{k+1,k+1} \end{array}\right] = A_{k+1}. \end{aligned}$$

This proves the theorem. □

Theorem

If A is nonsingular and the LU factorization exists, then the LU factorization is unique.

Proof: Suppose both

$$A = L_1 U_1 \quad \text{and} \quad A = L_2 U_2$$

are LU factorizations. Since A is nonsingular, L_1, U_1, L_2, U_2 are all nonsingular, and

$$A = L_1 U_1 = L_2 U_2 \Longrightarrow L_2^{-1} L_1 = U_2 U_1^{-1}.$$

Since L_1 and L_2 are unit lower triangular, it implies that $L_2^{-1} L_1$ is also unit lower triangular. On the other hand, since U_1 and U_2 are upper triangular, $U_2 U_1^{-1}$ is also upper triangular. Therefore,

$$L_2^{-1} L_1 = I = U_2 U_1^{-1}$$

which implies that $L_1 = L_2$ and $U_1 = U_2$.

Lemma

If $A \in \mathbb{R}^{n \times n}$ is positive definite, then all leading principal submatrices of A are nonsingular.

Proof: For $1 \leq k \leq n$, let

$$z_k = [x_1, \ldots, x_k]^T \in \mathbb{R}^k \quad \text{and} \quad x = [x_1, \ldots, x_k, 0, \ldots, 0]^T \in \mathbb{R}^n,$$

where $x_1, \ldots, x_k \in \mathbb{R}$ are not all zero. Since A is positive definite,

$$z_k^T A_k z_k = x^T A x > 0,$$

where A_k is the $k \times k$ leading principal submatrix of A. This shows that A_k are also positive definite, hence A_k are nonsingular. $\square$

Corollary

The matrix A is positive definite if and only if

$$A = GG^T, \tag{7}$$

where G is lower triangular with positive diagonal entries.

Proof: "$\Rightarrow$"

A is positive definite
$\Rightarrow$ all leading principal submatrices of A are nonsingular
$\Rightarrow$ A has the LU factorization $A = LU$, where L is unit lower triangular and U is upper triangular.
Since A is symmetric,

$$LU = A = A^T = U^T L^T \quad \Longrightarrow \quad U(L^T)^{-1} = L^{-1}U^T.$$

$U(L^T)^{-1}$ is upper triangular and $L^{-1}U^T$ is lower triangular
$\Rightarrow$ $U(L^T)^{-1}$ to be a diagonal matrix, say, $U(L^T)^{-1} = D$.
$\Rightarrow$ $U = DL^T$. Hence

$$A = LDL^T.$$

Since A is positive definite,

$$x^T A x > 0 \quad \Longrightarrow \quad x^T L D L^T x = (L^T x)^T D (L^T x) > 0.$$

This means D is also positive definite, and hence $d_{ii} > 0$. Thus $D^{1/2}$ is well-defined and we have

$$A = LDL^T = LD^{1/2} D^{1/2} L^T \equiv GG^T,$$

where $G \equiv LD^{1/2}$. Since the LU factorization is unique, G is unique.
"$\Leftarrow$"
Since G is lower triangular with positive diagonal entries, G is nonsingular. It implies that

$$G^T x \neq 0, \ \forall \ x \neq 0.$$

Hence

$$x^T A x = x^T G G^T x = \|G^T x\|_2^2 > 0, \ \forall \ x \neq 0$$

which implies that A is positive definite.

The factorization (7) is referred to as the Cholesky factorization.

Derive an algorithm for computing the Cholesky factorization:

Let

$$A \equiv [a_{ij}] \text{ and } G = \begin{bmatrix} g_{11} & 0 & \cdots & 0 \\ g_{21} & g_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix}.$$

Assume the first $k-1$ columns of G have been determined after $k-1$ steps. By componentwise comparison with

$$[a_{ij}] = \begin{bmatrix} g_{11} & 0 & \cdots & 0 \\ g_{21} & g_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{bmatrix} \begin{bmatrix} g_{11} & g_{21} & \cdots & g_{n1} \\ 0 & g_{22} & \cdots & g_{n2} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & g_{nn} \end{bmatrix},$$

one has

$$a_{kk} = \sum_{j=1}^{k} g_{kj}^2,$$

which gives

$$g_{kk}^2 = a_{kk} - \sum_{j=1}^{k-1} g_{kj}^2.$$

Moreover,

$$a_{ik} = \sum_{j=1}^{k} g_{ij} g_{kj}, \qquad i = k+1, \ldots, n,$$

hence the k-th column of G can be computed by

$$g_{ik} = \left(a_{ik} - \sum_{j=1}^{k-1} g_{ij} g_{kj} \right) \Big/ g_{kk}, \qquad i = k+1, \ldots, n.$$

Algorithm (Cholesky Factorization)

Given an $n \times n$ symmetric positive definite matrix A, this algorithm computes the Cholesky factorization $A = GG^T$.

Initialize $G = 0$
For $k = 1, \ldots, n$
$G(k,k) = \sqrt{A(k,k) - \sum_{j=1}^{k-1} G(k,j)G(k,j)}$
For $i = k+1, \ldots, n$
$G(i,k) = \left(A(i,k) - \sum_{j=1}^{k-1} G(i,j)G(k,j)\right) \Big/ G(k,k)$
End For
End For

In addition to n square root operations, there are approximately

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

floating-point arithmetic required by the algorithm.

Band matrix

Definition

An $n \times n$ matrix A is called a band matrix if $\exists\, p$ and q with $1 < p, q < n$ such that

$$a_{ij} = 0 \ \text{ whenever } \ p \le j - i \ \text{ or } \ q \le i - j.$$

The bandwidth of a band matrix is defined as $w = p + q - 1$. That is

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1p} & 0 & \cdots & 0 \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ a_{q1} & & \ddots & & \ddots & 0 \\ 0 & \ddots & & \ddots & & a_{n-p+1,n} \\ \vdots & \ddots & \ddots & & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n,n-q+1} & \cdots & a_{nn} \end{bmatrix}.$$

Definition

A square matrix $A = [a_{ij}]$ is said to be tridiagonal if

$$A = \begin{bmatrix} a_{11} & a_{12} & & 0 \\ a_{21} & a_{22} & \ddots & \\ & \ddots & \ddots & a_{n-1,n} \\ 0 & & a_{n,n-1} & a_{n,n} \end{bmatrix}.$$

If Gaussian elimination can be applied safely without pivoting. Then L and U factors would have the form

$$L = \begin{bmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ & \ddots & \ddots & \\ 0 & & \ell_{n,n-1} & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} u_{11} & u_{12} & & 0 \\ & u_{22} & \ddots & \\ & & \ddots & u_{n-1,n} \\ & & & u_{nn} \end{bmatrix},$$

and the entries are computed by the simple algorithm which only costs $3n$ flops.

Algorithm (Tridiagonal LU Factorization)

This algorithm computes the LU factorization for a tridiagonal matrix without using pivoting strategy.

$$
\begin{array}{l}
U(1,1) = A(1,1) \\
\text{For } i = 2, \ldots, n \\
\quad U(i-1,i) = A(i-1,i) \\
\quad L(i,i-1) = A(i,i-1)/U(i-1,i-1) \\
\quad U(i,i) = A(i,i) - L(i,i-1)U(i-1,i) \\
\text{End For}
\end{array}
$$

A tridiagonal linear system arises in many applications, such as finite difference discretization to second order linear boundary-value problem and the cubic spline approximations.