Gaussian Elimination for Linear Systems

Tsung-Ming Huang

Department of Mathematics
National Taiwan Normal University

October 3, 2011

Outline

Elementary matrices

Let $A \in \mathbb{C}^{n \times n}$ be a nonsingular matrix. We want to solve the linear system $Ax = b$ by

(a) **Direct methods** (finite steps);

(b) **Iterative methods** (convergence). (See Chapter 4)

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

$$\Rightarrow A_1 := L_1 A \equiv \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} A = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & -4 & -1 & -7 \\ 0 & 3 & 3 & 2 \end{bmatrix}$$

$$\Rightarrow A_2 := L_2 A_1 \equiv \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -4 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix} A_1 = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix}$$

$$= L_2 L_1 A$$

We have

$$A = L_1^{-1} L_2^{-1} A_2 = LR.$$

where L and R are lower and upper triangular, respectively.

Question

How to compute L_1^{-1} and L_2^{-1}?

$$L_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = I - \begin{bmatrix} 0 \\ 2 \\ 3 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

$$L_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -4 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix} = I - \begin{bmatrix} 0 \\ 0 \\ 4 \\ -3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$$

Definition 1

A matrix of the form

$$I - \alpha x y^* \quad (\alpha \in \mathbb{F}, x, y \in \mathbb{F}^n)$$

is called an elementary matrix.

The eigenvalues of $(I - \alpha x y^*)$ are $\{1, 1, \ldots, 1, 1 - \alpha y^* x\}$. Compute

$$(I - \alpha x y^*)(I - \beta x y^*) = I - (\alpha + \beta - \alpha\beta y^* x) x y^*.$$

If $\alpha y^* x - 1 \neq 0$ and let $\beta = \frac{\alpha}{\alpha y^* x - 1}$, then $\alpha + \beta - \alpha\beta y^* x = 0$. We have

$$(I - \alpha x y^*)^{-1} = (I - \beta x y^*),$$

where $\frac{1}{\alpha} + \frac{1}{\beta} = y^* x$.

Example 1

Let $x \in \mathbb{F}^n$, and $x^*x = 1$. Let $H = \{z : \ z^*x = 0\}$ and

$$Q = I - 2xx^* \quad (Q = Q^*, Q^{-1} = Q).$$

Then Q reflects each vector with respect to the hyperplane H. Let $y = \alpha x + w$, $w \in H$. Then, we have

$$Qy = \alpha Qx + Qw = -\alpha x + w - 2(x^*w)x = -\alpha x + w.$$

Let $y = e_i$ to be the i-th column of the unit matrix and $x = l_i = [0, \cdots, 0, l_{i+1,i}, \cdots, l_{n,i}]^T$. Then,

$$I + l_i e_i^T = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{i+1,i} & & \\ & & \vdots & \ddots & \\ & & l_{n,i} & & 1 \end{bmatrix} \tag{1}$$

Since $e_i^T l_i = 0$, we have

$$(I + l_i e_i^T)^{-1} = (I - l_i e_i^T).$$

From the equality

$$(I + l_1 e_1^T)(I + l_2 e_2^T) = \quad I + l_1 e_1^T + l_2 e_2^T + l_1(e_1^T l_2) e_2^T = \quad I + l_1 e_1^T + l_2 e_2^T$$

follows that

$$\begin{aligned} & (I + l_1 e_1^T) \cdots (I + l_i e_i^T) \cdots (I + l_{n-1} e_{n-1}^T) \\ = & I + l_1 e_1^T + l_2 e_2^T + \cdots + l_{n-1} e_{n-1}^T \\ = & \begin{bmatrix} 1 & & & \\ l_{21} & \ddots & 0 & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{bmatrix}. \end{aligned} \qquad (2)$$

Theorem 2

A lower triangular with "1" on the diagonal can be written as the product of $n-1$ elementary matrices of the form (1).

Remark: $(I + l_1 e_1^T + \cdots + l_{n-1} e_{n-1}^T)^{-1} = (I - l_{n-1} e_{n-1}^T) \cdots (I - l_1 e_1^T)$ which can not be simplified as in (2).

LR-factorization

Definition 3

Given $A \in \mathbb{C}^{n\times n}$, a lower triangular matrix L with "1" on the diagonal and an upper triangular matrix R. If $A = LR$, then the product LR is called a LR-factorization (or LR-decomposition) of A.

Basic problem

Given $b \neq 0$, $b \in \mathbb{F}^n$. Find a vector $l_1 = [0, l_{21}, \cdots, l_{n1}]^T$ and $c \in \mathbb{F}$ such that

$$(I - l_1 e_1^T)b = ce_1.$$

Solution:

$$\begin{cases} b_1 = c, \\ b_i - l_{i1}b_1 = 0, \quad i = 2, \ldots, n. \end{cases}$$

$$\begin{cases} b_1 = 0, & \text{it has no solution (since } b \neq 0), \\ b_1 \neq 0, & \text{then } c = b_1,\ l_{i1} = b_i/b_1,\ i = 2, \ldots, n. \end{cases}$$

Construction of LR-factorization:

Let $A = A^{(0)} = \left[\begin{array}{ccc} a_1^{(0)} & \cdots & a_n^{(0)} \end{array} \right]$. Apply basic problem to $a_1^{(0)}$: If $a_{11}^{(0)} \neq 0$, then there exists $L_1 = I - l_1 e_1^T$ such that

$$(I - l_1 e_1^T) a_1^{(0)} = a_{11}^{(0)} e_1.$$

Thus

$$\begin{array}{rcl} A^{(1)} & = & L_1 A^{(0)} \\ & = & \left[\begin{array}{ccc} L a_1^{(0)} & \cdots & L a_n^{(0)} \end{array} \right] \\ & = & \left[\begin{array}{cccc} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{array} \right]. \end{array}$$

The k-th step:

$$A^{(k)} = L_k A^{(k-1)} = L_k L_{k-1} \cdots L_1 A^{(0)} \tag{3}$$

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

- If $a_{kk}^{(k-1)} \neq 0$, for $k = 1, \ldots, n-1$, then the method is executable and we have that

$$A^{(n-1)} = L_{n-1} \cdots L_1 A^{(0)} = R$$

is an upper triangular matrix. Thus, $A = LR$.

- Explicit representation of L:

$$\begin{aligned} L_k &= I - l_k e_k^T, \quad L_k^{-1} = I + l_k e_k^T \\ L &= L_1^{-1} \cdots L_{n-1}^{-1} = (I + l_1 e_1^T) \cdots (I + l_{n-1} e_{n-1}^T) \\ &= I + l_1 e_1^T + \cdots + l_{n-1} e_{n-1}^T \quad \text{(by (2))}. \end{aligned}$$

Theorem 4

Let A be nonsingular. Then A has an LR-factorization ($A = LR$) if and only if $\kappa_i := det(A_i) \neq 0$, where A_i is the leading principal matrix of A, i.e.,

$$A_i = \begin{bmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} \end{bmatrix},$$

for $i = 1, \ldots, n-1$.

Proof: (**Necessity** "$\Rightarrow$"): Since $A = LR$, we have

$$\begin{bmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} \end{bmatrix} = \begin{bmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{i1} & \cdots & l_{ii} \end{bmatrix} \begin{bmatrix} r_{11} & \cdots & r_{1i} \\ & \ddots & \vdots \\ 0 & & r_{ii} \end{bmatrix}.$$

From $\det(A) \neq 0$ follows that $\det(L) \neq 0$ and $\det(R) \neq 0$. Thus, $l_{jj} \neq 0$ and $r_{jj} \neq 0$, for $j = 1, \ldots, n$. Hence $\kappa_i = l_{11} \cdots l_{ii} r_{11} \cdots r_{ii} \neq 0$.

(**Sufficiency** "$\Leftarrow$"): From (3) we have

$$A^{(0)} = (L_1^{-1} \cdots L_i^{-1}) A^{(i)}.$$

Consider the $(i+1)$-th leading principle determinant. From (3) we have

$$\begin{bmatrix} a_{11} & \cdots & a_{i,i+1} \\ \vdots & & \vdots \\ a_{i+1} & \cdots & a_{i+1,i+1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & & & & 0 \\ l_{21} & \ddots & & & \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \\ l_{i+1,1} & \cdots & \cdots & l_{i+1,i} & 1 \end{bmatrix} \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & \cdots & a_{1,i+1}^{(0)} \\ & a_{22}^{(1)} & \cdots & \cdots & a_{2,i+1}^{(1)} \\ & & \ddots & & \vdots \\ & & & a_{ii}^{(i-1)} & a_{i,i+1}^{(i-1)} \\ 0 & & & & a_{i+1,i+1}^{(i)} \end{bmatrix}.$$

Thus, $\kappa_i = 1 \cdot a_{11}^{(0)} a_{22}^{(1)} \cdots a_{i+1,i+1}^{(i)} \neq 0$ which implies $a_{i+1,i+1}^{(i)} \neq 0$.
Therefore, the LR-factorization of A exists. □

Theorem 5

If a nonsingular matrix A has an LR-factorization with $A = LR$ and $l_{11} = \cdots = l_{nn} = 1$, then the factorization is unique.

Proof: Let $A = L_1R_1 = L_2R_2$. Then $L_2^{-1}L_1 = R_2R_1^{-1} = I$. □

Corollary 6

If a nonsingular matrix A has an LR-factorization with $A = LDR$, where D is diagonal, L and R^T are unit lower triangular (with one on the diagonal) if and only if $\kappa_i \neq 0$.

Theorem 7

Let A be a nonsingular matrix. Then there exists a permutation P, such that PA has an LR-factorization.

Proof: By construction! Consider (3): There is a permutation P_k, which interchanges the k-th row with a row of index large than k, such that $0 \neq a_{kk}^{(k-1)} (\in P_k A^{(k-1)})$. This procedure is executable, for $k = 1, \ldots, n-1$. So we have

$$L_{n-1}P_{n-1} \cdots L_k P_k \cdots L_1 P_1 A^{(0)} = R. \tag{4}$$

Let P be a permutation which affects only elements $k+1, \ldots, n$. It holds

$$P(I - l_k e_k^T)P^{-1} = I - (Pl_k)e_k^T = I - \tilde{l}_k e_k^T = \tilde{L}_k, \quad (e_k^T P^{-1} = e_k^T)$$

where $\tilde{L}_k$ is lower triangular. Hence we have

$$PL_k = \tilde{L}_k P.$$

Now write all P_k in (4) to the right as

$$L_{n-1}\tilde{L}_{n-2}\cdots\tilde{L}_1 P_{n-1}\cdots P_1 A^{(0)} = R.$$

Then we have $PA = LR$ with $L^{-1} = L_{n-1}\tilde{L}_{n-2}\cdots\tilde{L}_1$ and $P = P_{n-1}\cdots P_1$. □

Gaussian elimination

Given a linear system

$$Ax = b$$

with A nonsingular. We first assume that A has an LR-factorization, i.e., $A = LR$. Thus

$$LRx = b.$$

We then (i) solve $Ly = b$; (ii) solve $Rx = y$. These imply that $LRx = Ly = b$. From (4), we have

$$L_{n-1} \cdots L_2 L_1 (A \mid b) = (R \mid L^{-1} b).$$

Algorithm: Gaussian elimination without permutation

for $k = 1, \ldots, n-1$ **do**
    if $a_{kk} = 0$ **then**
        Stop.
    else
        $\omega_j := a_{kj}\ (j = k+1, \ldots, n)$;
    end if
    for $i = k+1, \ldots, n$ **do**
        $\eta := a_{ik}/a_{kk},\ a_{ik} := \eta$;
        for $j = k+1, \ldots, n$ **do**
            $a_{ij} := a_{ij} - \eta\omega_j,\ b_j := b_j - \eta b_k$.
        end for
    end for
end for
$x_n = b_n/a_{nn}$;
for $i = n-1, n-2, \ldots, 1$ **do**
    $x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}$.
end for

Cost of computation (A flop is a floating point operation):

(i) LR-factorization: $2n^3/3$ flops;

(ii) Computation of y: n^2 flops;

(iii) Computation of x: n^2 flops.

For A^{-1}: $8/3n^3 \approx 2n^3/3 + 2kn^2$ ($k = n$ linear systems).

Pivoting: (a) Partial pivoting; (b) Complete pivoting.

From (3), we have

$$A^{(k-1)} = \begin{bmatrix} a_{11}^{(0)} & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(0)} \\ 0 & \ddots & & & & \vdots \\ \vdots & & a_{k-1,k-1}^{(k-2)} & \cdots & \cdots & a_{k-1,n}^{(k-2)} \\ \vdots & & 0 & a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{bmatrix}.$$

Partial pivoting

$$\left\{ \begin{array}{l} \text{Find a } \ p \in \{k, \ldots, n\} \text{ such that} \\ \qquad\qquad |a_{pk}| = \max_{k \le i \le n} |a_{ik}| \quad (r_k = p) \\ \text{swap } a_{kj} \text{ with } a_{pj} \text{ for } j = k, \ldots, n, \text{ and } b_k \text{ with } b_p. \end{array} \right. \tag{5}$$

- Replacing stopping step in Line 3 of Gaussian elimination Algorithm by (5), we have a new factorization of A with partial pivoting, i.e., $PA = LR$ (by Theorem 7 and $|l_{ij}| \le 1$ for $i, j = 1, \ldots, n$).
- For solving linear system $Ax = b$, we use

$$PAx = Pb \Rightarrow L(Rx) = P^T b \equiv \tilde{b}.$$

- It needs extra $n(n-1)/2$ comparisons.

Complete pivoting

$$
\left\{
\begin{array}{l}
\text{Find } \; p, q \in \{k, \ldots, n\} \text{ such that} \\
\qquad\qquad |a_{pq}| \leq \max\limits_{k \leq i,j \leq n} |a_{ij}|, \; (r_k := p, c_k := q) \\
\text{swap } a_{kj} \text{ and } b_k \text{ with } a_{pj} \text{ and } b_p, \text{ resp., } (j = k, \ldots, n), \\
\text{swap } a_{ik} \text{ with } a_{iq} (i = 1, \ldots, n).
\end{array}
\right. \tag{6}
$$

- Replacing stopping step in Line 3 of Gaussian elimination Algorithm by (6), we also have a new factorization of A with complete pivoting, i.e., $PA\Pi = LR$ (by Theorem 7 and $|l_{ij}| \leq 1$, for $i, j = 1, \ldots, n$).
- For solving linear system $Ax = b$, we use

$$
PA\Pi(\Pi^T x) = Pb \Rightarrow LR\tilde{x} = \tilde{b} \Rightarrow x = \Pi\tilde{x}.
$$

- It needs $n^3/3$ comparisons.

Let

$$A = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}$$

be in three decimal-digit floating point arithmetic.

- $\kappa(A) = \|A\|_\infty \|A^{-1}\|_\infty \approx 4$. A is well-conditioned.
- Without pivoting:

$$\begin{aligned} L &= \begin{bmatrix} 1 & 0 \\ fl(1/10^{-4}) & 1 \end{bmatrix}, \quad fl(1/10^{-4}) = 10^4, \\ R &= \begin{bmatrix} 10^{-4} & 1 \\ 0 & fl(1 - 10^4 \cdot 1) \end{bmatrix}, \quad fl(1 - 10^4 \cdot 1) = -10^4. \\ LR &= \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix} \begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix} = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 0 \end{bmatrix} \neq A. \end{aligned}$$

- Here a_{22} entirely "lost" from computation. It is numerically unstable.
- Let $Ax = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Then $x \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
- But $Ly = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ solves $y_1 = 1$ and $y_2 = fl(2 - 10^4 \cdot 1) = -10^4$,
- $R\hat{x} = y$ solves $\hat{x}_2 = fl((-10^4)/(-10^4)) = 1$, $\hat{x}_1 = fl((1-1)/10^{-4}) = 0$.
- We have an erroneous solution with cond(L), cond$(R) \approx 10^8$.

Partial pivoting:

$$
\begin{aligned}
L &= \begin{bmatrix} 1 & 0 \\ fl(10^{-4}/1) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 10^{-4} & 1 \end{bmatrix}, \\
R &= \begin{bmatrix} 1 & 1 \\ 0 & fl(1-10^{-4}) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.
\end{aligned}
$$

L and R are both well-conditioned.

LDR- and LL^T-factorizations

Algorithm 2

[Crout's factorization or compact method]

For $k = 1, \ldots, n$,
    for $p = 1, 2, \ldots, k-1$,
        $r_p := d_p a_{pk}$,
        $\omega_p := a_{kp} d_p$,
        $d_k := a_{kk} - \sum_{p=1}^{k-1} a_{kp} r_p$,
    if $d_k = 0$, then stop; else
        for $i = k+1, \ldots, n$,
            $a_{ik} := (a_{ik} - \sum_{p=1}^{k-1} a_{ip} r_p)/d_k$,
            $a_{ki} := (a_{ki} - \sum_{p=1}^{k-1} \omega_p a_{pi})/d_k$.

Cost: $n^3/3$ flops.

- With partial pivoting: see Wilkinson EVP pp.225-.
- Advantage: One can use double precision for inner product.

Theorem 8

If A is nonsingular, real and symmetric, then A has a unique LDL^T factorization, where D is diagonal and L is a unit lower triangular matrix (with one on the diagonal).

Proof: $A = LDR = A^T = R^T DL^T$. It implies $L = R^T$. □

Theorem 9

If A is symmetric and positive definite, then there exists a lower triangular $G \in \mathbb{R}^{n\times n}$ with positive diagonal elements such that $A = GG^T$.

Proof:

$$\begin{aligned} & A \text{ is symmetric positive definite} \\ \Leftrightarrow\quad & x^T A x \geq 0 \text{ for all nonzero vector } x \in \mathbb{R}^n \\ \Leftrightarrow\quad & \kappa_i \geq 0 \text{ for } i = 1, \ldots, n \\ \Leftrightarrow\quad & \text{all eigenvalues of } A \text{ are positive} \end{aligned}$$

From Corollary 6 and Theorem 8 we have $A = LDL^T$. From $L^{-1}AL^{-T} = D$ follows that

$$d_k = (e_k^T L^{-1})A(L^{-T}e_k) > 0.$$

Thus, $G = L\text{diag}\{d_1^{1/2}, \cdots, d_n^{1/2}\}$ is real, and then $A = GG^T$. □

Derive an algorithm for computing the Cholesky factorization $A = GG^T$:
Let

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

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,

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

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

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

Cholesky Factorization

Input: $n \times n$ symmetric positive definite matrix A.
Output: Cholesky factorization $A = GG^T$.

1: Initialize $G = 0$.
2: **for** $k = 1, \ldots, n$ **do**
3: $\quad G(k,k) = \sqrt{A(k,k) - \sum_{j=1}^{k-1} G(k,j)G(k,j)}$
4: $\quad$ **for** $i = k+1, \ldots, n$ **do**
5: $\quad\quad G(i,k) = \left(A(i,k) - \sum_{j=1}^{k-1} G(i,j)G(k,j)\right) \Big/ G(k,k)$
6: $\quad$ **end for**
7: **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.

For solving symmetric, indefinite systems: See Golub/ Van Loan *Matrix Computation* pp. 159-168. $PAP^T = LDL^T$, D is 1×1 or 2×2 block-diagonal matrix, P is a permutation and L is lower triangular with one on the diagonal.

Error estimation for linear systems

Consider the linear system

$$Ax = b, \tag{7}$$

and the perturbed linear system

$$(A + \delta A)(x + \delta x) = b + \delta b, \tag{8}$$

where δA and δb are errors of measure or round-off in factorization.

Definition 10

Let $\|\cdot\|$ be an operator norm and A be nonsingular. Then $\kappa \equiv \kappa(A) = \|A\|\|A^{-1}\|$ is a condition number of A corresponding to $\| \ \|$.

Theorem 11 (Forward error bound)

Let x be the solution of (7) and $x + \delta x$ be the solution of the perturbed linear system (8). If $\|\delta A\| \|A^{-1}\| < 1$, then

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa}{1 - \kappa \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right).$$

Proof: From (8) we have

$$(A + \delta A)\delta x + Ax + \delta A x = b + \delta b.$$

Thus,

$$\delta x = -(A + \delta A)^{-1}[(\delta A)x - \delta b]. \tag{9}$$

Here, Corollary 2.7 implies that $(A+\delta A)^{-1}$ exists. Now,

$$\|(A+\delta A)^{-1}\| = \|(I+A^{-1}\delta A)^{-1}A^{-1}\| \le \|A^{-1}\| \frac{1}{1-\|A^{-1}\|\|\delta A\|}.$$

On the other hand, $b = Ax$ implies $\|b\| \le \|A\|\|x\|$. So,

$$\frac{1}{\|x\|} \le \frac{\|A\|}{\|b\|}. \tag{10}$$

From (9) follows that $\|\delta x\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta A\|\|x\| + \|\delta b\|)$. By using (10), the inequality (11) is proved. □

Remark 1

If $\kappa(A)$ is large, then A (for the linear system $Ax = b$) is called ill-conditioned, else well-conditioned.

Error analysis for Gaussian algorithm

A computer in characterized by four integers:

(a) the machine base β;

(b) the precision t;

(c) the underflow limit L;

(d) the overflow limit U.

Define the set of floating point numbers.

$$F = \{f = \pm 0.d_1 d_2 \cdots d_t \times \beta^e \mid 0 \leq d_i < \beta, d_1 \neq 0, L \leq e \leq U\} \cup \{0\}.$$

Let $G = \{x \in \mathbb{R} \mid m \leq |x| \leq M\} \cup \{0\}$, where $m = \beta^{L-1}$ and $M = \beta^U(1 - \beta^{-t})$ are the minimal and maximal numbers of $F \setminus \{0\}$ in absolute value, respectively.

We define an operator $fl : G \to F$ by

$$fl(x) = \text{ the nearest } c \in F \text{ to } x \text{ by rounding arithmetic.}$$

One can show that fl satisfies

$$fl(x) = x(1+\varepsilon), \quad |\varepsilon| \leq eps,$$

where $eps = \frac{1}{2}\beta^{1-t}$. (If $\beta = 2$, then $eps = 2^{-t}$). It follows that

$$fl(a \circ b) = (a \circ b)(1+\varepsilon)$$

or

$$fl(a \circ b) = (a \circ b)/(1+\varepsilon),$$

where $|\varepsilon| \leq eps$ and $\circ = +, -, \times, /$.

Given $x, y \in \mathbb{R}^n$. The following algorithm computes $x^T y$ and stores the result in s.

$s = 0$,
for $k = 1, \ldots, n$,
$s = s + x_k y_k$.

Theorem 12

If $n2^{-t} \leq 0.01$, *then*

$$fl(\sum_{k=1}^{n} x_k y_k) = \sum_{k=1}^{n} x_k y_k [1 + 1.01(n + 2 - k)\theta_k 2^{-t}], \ |\theta_k| \leq 1$$

▸ Proof of Theorem 12

Let the exact LR-factorization of A be L and R ($A = LR$) and let $\tilde{L}$, $\tilde{R}$ be the LR-factorization of A by using Gaussian Algorithm (without pivoting). There are two possibilities:

(i) Forward error analysis: Estimate $|L - \tilde{L}|$ and $|R - \tilde{R}|$.

(ii) Backward error analysis: Let $\tilde{L}\tilde{R}$ be the exact LR-factorization of a perturbed matrix $\tilde{A} = A + E$. Then E will be estimated, i.e., $|E| \leq ?$.

Theorem 13

The LR-factorization $\tilde{L}$ and $\tilde{R}$ of A using Gaussian Elimination with partial pivoting satisfies

$$\tilde{L}\tilde{R} = A + E, \tag{2.6}$$

where

$$\|E\|_\infty \leq n^2 \rho \|A\|_\infty 2^{-t} \tag{2.7}$$

with

$$\rho = \max_{i,j,k} \left| a_{ij}^{(k)} \right| \Big/ \|A\|_\infty . \tag{2.8}$$

Applying Theorem 12 to the linear system $\tilde{L}y = b$ and $\tilde{R}x = y$, respectively, the solution x satisfies

$$(\tilde{L} + \delta\tilde{L})(\tilde{R} + \delta\tilde{R})x = b$$

or

$$(\tilde{L}\tilde{R} + (\delta\tilde{L})\tilde{R} + \tilde{L}(\delta\tilde{R}) + (\delta\tilde{L})(\delta\tilde{R}))x = b. \qquad (2.9)$$

Since $\tilde{L}\tilde{R} = A + E$, substituting this equation into (2.9) we get

$$[A + E + (\delta\tilde{L})\tilde{R} + \tilde{L}(\delta\tilde{R}) + (\delta\tilde{L})(\delta\tilde{R})]x = b.$$

The entries of $\tilde{L}$ and $\tilde{R}$ satisfy

$$|\widetilde{l}_{ij}| \leq 1, \text{ and } |\widetilde{r}_{ij}| \leq \rho\|A\|_\infty.$$

Therefore, we get

$$\begin{cases} \|\tilde{L}\|_\infty \leq n, \\ \|\tilde{R}\|_\infty \leq n\rho\|A\|_\infty, \\ \|\delta\tilde{L}\|_\infty \leq \frac{n(n+1)}{2}1.01 \cdot 2^{-t}, \\ \|\delta\tilde{R}\|_\infty \leq \frac{n(n+1)}{2}1.01\rho 2^{-t}. \end{cases} \tag{2.10}$$

In practical implementation we usually have $n^2 2^{-t} << 1$. So it holds

$$\|\delta\tilde{L}\|_\infty\|\delta\tilde{R}\|_\infty \leq n^2\rho\|A\|_\infty 2^{-t}.$$

Let

$$\delta A = E + (\delta\tilde{L})\tilde{R} + \tilde{L}(\delta\tilde{R}) + (\delta\tilde{L})(\delta\tilde{R}). \tag{2.11}$$

Then, from (2.7) and (2.10) we get

$$\begin{aligned} \|\delta A\|_\infty &\leq \|E\|_\infty + \|\delta\tilde{L}\|_\infty\|\tilde{R}\|_\infty + \|\tilde{L}\|_\infty\|\delta\tilde{R}\|_\infty + \|\delta\tilde{L}\|_\infty\|\delta\tilde{R}\|_\infty \\ &\leq 1.01(n^3 + 3n^2)\rho\|A\|_\infty 2^{-t} \end{aligned} \tag{2.12}$$

Theorem 14

For a linear system $Ax = b$ the solution x computed by Gaussian Elimination with partial pivoting is the exact solution of the equation $(A + \delta A)x = b$ and δA satisfies (2.12).

Remark: The quantity ρ defined by (2.9) is called a growth factor. The growth factor measures how large the numbers become during the process of elimination. In practice, ρ is usually of order 10 for partial pivot selection. But it can be as large as $\rho = 2^{n-1}$, when

$$A = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ \vdots & -1 & \ddots & \ddots & \vdots & 1 \\ \vdots & \vdots & \ddots & \ddots & 0 & 1 \\ -1 & -1 & \cdots & -1 & 1 & 1 \\ -1 & -1 & \cdots & \cdots & -1 & 1 \end{bmatrix}.$$

Better estimates hold for special types of matrices. For example in the case of upper Hessenberg matrices, that is, matrices of the form

$$A = \begin{bmatrix} \times & \cdots & \cdots & \times \\ \times & \ddots & \ddots & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & \times & \times \end{bmatrix}$$

the bound $\rho \le (n-1)$ can be shown. (Hessenberg matrices arise in eigenvalus problems.)

For tridiagonal matrices

$$A = \begin{bmatrix} \alpha_1 & \beta_2 & & & 0 \\ \gamma_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_n \\ 0 & & & \gamma_n & \alpha_n \end{bmatrix}$$

it can even be shown that $\rho \leq 2$ holds for partial pivot selection. Hence, Gaussian elimination is quite numerically stable in this case.

For complete pivot selection, Wilkinson (1965) has shown that

$$|a_{ij}^k| \leq f(k) \max_{i,j} |a_{ij}|$$

with the function

$$f(k) := k^{\frac{1}{2}} \left[2^1 \, 3^{\frac{1}{2}} \, 4^{\frac{1}{3}} \, \cdots \, k^{\frac{1}{(k-1)}} \right]^{\frac{1}{2}} .$$

This function grows relatively slowly with k:

k	10	20	50	100
$f(k)$	19	67	530	3300

Even this estimate is too pessimistic in practice. Up until now, no matrix has been found which fails to satisfy

$$|a_{ij}^{(k)}| \leq (k+1) \max_{i,j} |a_{ij}| \quad k = 1, 2, ..., n-1,$$

when complete pivot selection is used. This indicates that Gaussian elimination with complete pivot selection is usually a stable process. Despite this, partial pivot selection is preferred in practice, for the most part, because:

(i) Complete pivot selection is more costly than partial pivot selection. (To compute $A^{(i)}$, the maximum from among $(n-i+1)^2$ elements must be determined instead of $n-i+1$ elements as in partial pivot selection.)

(ii) Special structures in a matrix, i.e. the band structure of a tridiagonal matrix, are destroyed in complete pivot selection.

Iterative Improvement:

Suppose that the linear system $Ax = b$ has been solved via the LR-factorization $PA = LR$. Now we want to improve the accuracy of the computed solution x. We compute

$$\left\{ \begin{array}{rcl} r & = & b - Ax, \\ Ly & = & Pr, \quad Rz = y, \\ x_{new} & = & x + z. \end{array} \right. \tag{2.13}$$

Then in exact arithmatic we have

$$Ax_{new} = A(x + z) = (b - r) + Az = b.$$

This leads to solve

$$Az = r$$

by using $PA = LR$.

Unfortunately, $r = fl(b - Ax)$ renders an x_{new} that is no more accurate than x. It is necessary to compute the residual $b - Ax$ with extended precision floating arithmetic.

Algorithm 4

Compute $PA = LR$ (t-digit)
Repeat: $r := b - Ax$ (2t-digit)
  Solve $Ly = Pr$ for y (t-digit)
  Solve $Rz = y$ for z (t-digit)
  Update $x = x + z$ (t-digit)

From Theorem 14 we have $(A + \delta A)z = r$, i.e.,

$$A(I + F)z = r \text{ with } F = A^{-1}\delta A. \tag{2.14}$$

Theorem 15

Let $\{x_k\}$ be the sequence constructed by Algorithm 4 for solving $Ax = b$ and $x^* = A^{-1}b$ be the exact solution. Assume F_k in (2.14) satisfying $\|F_k\| \leq \sigma < 1/2$ for all k. Then $\{x_k\} \to x^*$.

▸ Proof of Theorem 15

Corollary 16

If

$$1.01(n^3 + 3n^2)\rho 2^{-t}\|A\|\,\|A^{-1}\| < 1/2,$$

then Algorithm 4 converges.

Proof: From (2.14) and (2.12) follows that

$$\|F_k\| \leq 1.01(n^3 + 3n^2)\rho 2^{-t}\kappa(A) < 1/2.$$

□

Appendix

Proof of Theorem 12: Let $s_p = fl(\sum_{k=1}^{p} x_k y_k)$ be the partial sum in Algorithm 41. Then

$$s_1 = x_1 y_1 (1 + \delta_1)$$

with $|\delta_1| \leq eps$ and for $p = 2, \dots, n$,

$$s_p = fl[s_{p-1} + fl(x_p y_p)] = [s_{p-1} + x_p y_p (1 + \delta_p)](1 + \varepsilon_p)$$

with $|\delta_p|$, $|\varepsilon_p| \leq eps$. Therefore

$$fl(x^T y) = s_n = \sum_{k=1}^{n} x_k y_k (1 + \gamma_k),$$

where $(1 + \gamma_k) = (1 + \delta_k) \prod_{j=k}^{n} (1 + \varepsilon_j)$, and $\varepsilon_1 \equiv 0$. Thus,

$$fl(\sum_{k=1}^{n} x_k y_k) = \sum_{k=1}^{n} x_k y_k [1 + 1.01(n + 2 - k)\theta_k 2^{-t}].$$

The result follows immediately from the following useful Lemma. □

Lemma 7.1

If $(1+\alpha) = \prod_{k=1}^{n}(1+\alpha_k)$, *where* $|\alpha_k| \leq 2^{-t}$ *and* $n2^{-t} \leq 0.01$, *then*

$$\prod_{k=1}^{n}(1+\alpha_k) = 1 + 1.01n\theta 2^{-t} \text{ with } |\theta| \leq 1.$$

Proof: From assumption it is easily seen that

$$(1-2^{-t})^n \leq \prod_{k=1}^{n}(1+\alpha_k) \leq (1+2^{-t})^n.$$

Expanding the Taylor expression of $(1-x)^n$ as $-1 < x < 1$, we get

$$(1-x)^n = 1 - nx + \frac{n(n-1)}{2}(1-\theta x)^{n-2}x^2 \geq 1 - nx.$$

Hence

$$(1-2^{-t})^n \geq 1 - n2^{-t}.$$

Now, we estimate the upper bound of $(1+2^{-t})^n$:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = 1 + x + \frac{x}{2}x(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots).$$

If $0 \leq x \leq 0.01$, then

$$1 + x \leq e^x \leq 1 + x + 0.01x\frac{1}{2}e^x \leq 1 + 1.01x$$

(Here, we use the fact $e^{0.01} < 2$ to the last inequality.) Let $x = 2^{-t}$. Then the left inequality of (55) implies

$$(1+2^{-t})^n \leq e^{2^{-t}n}$$

Let $x = 2^{-t}n$. Then the second inequality of (55) implies

$$e^{2^{-t}n} \leq 1 + 1.01n2^{-t}$$

From (55) and (55) we have

$$(1+2^{-t})^n \leq 1 + 1.01n2^{-t}.$$

□

Return

Proof of Theorem 15: From (2.14) and $r_k = b - Ax_k$ we have

$$A(I + F_k)z_k = b - Ax_k.$$

Since A is nonsingular, we have $(I + F_k)z_k = x^* - x_k$.

From $x_{k+1} = x_k + z_k$ we have $(I + F_k)(x_{k+1} - x_k) = x^* - x_k$, i.e.,

$$(I + F_k)x_{k+1} = F_k x_k + x^*. \tag{2.15}$$

Subtracting both sides of (2.15) from $(I + F_k)x^*$ we get

$$(I + F_k)(x_{k+1} - x^*) = F_k(x_k - x^*).$$

Then, $x_{k+1} - x^* = (I + F_k)^{-1}F_k(x_k - x^*)$. Hence,

$$\|x_{k+1} - x^*\| \leq \|F_k\| \frac{\|x_k - x^*\|}{1 - \|F_k\|} \leq \frac{\sigma}{1 - \sigma}\|x_k - x^*\|.$$

Let $\tau = \sigma/(1 - \sigma)$. Then

$$\|x_k - x^*\| \leq \tau^{k-1}\|x_1 - x^*\|.$$

But $\sigma < 1/2$ follows $\tau < 1$. This implies convergence of Algorithm 4. □

Return