Numerical Analysis

Daming Li

Department of Mathematics, Shanghai JiaoTong University,
Shanghai, 200240, China
Email: lidaming@sjtu.edu.cn

October 14, 2014

Solving Systems of Linear Equations

We want to solve a linear system

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

or its matrix form

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

Easy-to-solve system

Diagonal matrix

$$\begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}, \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1/a_{11} \\ b_2/a_{22} \\ b_3/a_{33} \\ \vdots \\ b_n/a_{nn} \end{pmatrix}$$

Lower triangular matrix: Forward substitution

$$\begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & 0 & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}, \quad x_i = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j)/a_{ii}$$

Easy-to-solve system

Upper triangular matrix: Backward substitution

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}, \quad x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii}$$

Let $(p_1, p_2, \cdots, p_n)$ be a permutation of $(1, 2, \cdots, n)$. One row of A, say row p_1, has zeros in positions $2, 3 \cdots, n$. Then another row, say row p_2, has zeros in positions $3, 4, \cdots, n$ and so on. The forward substitution for this permuted system becomes:

$$x_i \leftarrow \left(b_{p_i} - \sum_{j=1}^{i-1} a_{p_i j} x_j\right) \Big/ a_{p_i i}, \quad i = 1, \cdots, n$$

LU factorizations

Suppose that A can be factorized into the product of a lower triangular matrix L and upper triangular matrix U: $A = LU$ (it is called LU-decomposition), where

$$L = \begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & 0 & \cdots & l_{nn} \end{pmatrix}, U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{pmatrix}$$

$$Lz = b \quad \text{solve for } \ z$$

$$Ux = z \quad \text{solve for } \ x$$

L and U is not uniquely determined. If $l_{ii} = 1, i = 1, \cdots, n$, i.e., L is a unit lower triangular matrix (Doolittle's factorization). If $u_{ii} = 1, i = 1, \cdots, n$, i.e., U is a unit upper triangular matrix (Crout's factorization). If $u_{ii} = l_{ii}, i = 1, \cdots, n$, it is called Cholesky's factorization.

LU factorizations

According to $A = LU$,

$$a_{ij} = \sum_{s=1}^{n} l_{is} u_{sj} = \sum_{s=1}^{\min(i,j)} l_{is} u_{sj}$$

where we have used the fact that $l_{is} = 0$ for $s > i$ and $u_{sj} = 0$ for $s > j$. Assume that the columns $1, 2, \cdots, k-1$ in L and the rows $1, 2, \cdots, k-1$ in U has been computed, the k-th column of L and k-th row of U can be computed by

$$a_{kk} = \sum_{s=1}^{k-1} l_{ks} u_{sk} + l_{kk} u_{kk}$$

$$a_{kj} = \sum_{s=1}^{k-1} l_{ks} u_{sj} + l_{kk} u_{kj}, \quad j = k+1, \ldots, n$$

$$a_{ik} = \sum_{s=1}^{k-1} l_{is} u_{sk} + l_{ik} u_{kk}, \quad i = k+1, \ldots, n$$

LU factorizations

Find the Doolittle, Crout and Cholesky factorization of the matrix

$$A = \begin{pmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{pmatrix} \xrightarrow[\text{1st row of } U \text{ and 1st col of } L]{} \begin{pmatrix} 60 & 30 & 20 \\ 1/2 & 20 & 15 \\ 1/3 & 15 & 12 \end{pmatrix}$$

$$\xrightarrow[\text{2th row of } U \text{ and 2th col of } L]{} \begin{pmatrix} 60 & 30 & 20 \\ 1/2 & 5 & 5 \\ 1/3 & 1 & 12 \end{pmatrix}$$

$$\xrightarrow[\text{3rd row of } U]{} \begin{pmatrix} 60 & 30 & 20 \\ 1/2 & 5 & 5 \\ 1/3 & 1 & 1/3 \end{pmatrix}.$$

We have the Doolittle factorization $A = LU$, where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 60 & 30 & 20 \\ 0 & 5 & 5 \\ 0 & 0 & 1/3 \end{pmatrix}.$$

LU factorizations

$$\begin{aligned}
A &= \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{pmatrix}\begin{pmatrix} 60 & 30 & 20 \\ 0 & 5 & 5 \\ 0 & 0 & 1/3 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{pmatrix}\begin{pmatrix} 60 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1/3 \end{pmatrix}\begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 60 & 0 & 0 \\ 30 & 5 & 0 \\ 20 & 5 & 1/3 \end{pmatrix}\begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{the Crout's factorization} \\
&= LL^T \quad \text{the Cholesky factorization}
\end{aligned}$$

where

$$L = \begin{pmatrix} \sqrt{60} & 0 & 0 \\ \sqrt{60}/2 & \sqrt{5} & 0 \\ \sqrt{60}/3 & \sqrt{5} & \sqrt{1/3} \end{pmatrix}$$

LU factorizations

If all n leading principal minors of the $n \times n$ matrix A are nonsingular, then A has an LU-decomposition.
Proof. Denote the kth leading principle minors of A, L and U by A_k, L_k and U_k, respectively. We use an inductive proof. First

$$A_1 = a_{11} = l_{11}u_{11} = L_1U_1 = U_1$$

which is nonzero since A_1 is nonsingular. Assume that we implement k – 1 steps of decomposition: $A_{k-1} = L_{k-1}U_{k-1}$ where L_{k-1} is the unit lower triangular and U_{k-1} the nonsingular upper triangular. The next step of decomposition can be obtained from $A_k = L_kU_k$ since

$$\begin{pmatrix} a_{1k} \\ \vdots \\ a_{k-1,k} \end{pmatrix} = L_{k-1} \begin{pmatrix} u_{1k} \\ \vdots \\ u_{k-1,k} \end{pmatrix}$$

LU factorizations

$$(a_{k1}, \cdots, a_{k,k-1}) = (l_{k1}, \cdots, l_{k,k-1})U_{k-1}$$

$$a_{kk} = (l_{k1}, \cdots, l_{k,k-1}) \begin{pmatrix} u_{1k} \\ \vdots \\ u_{k-1,k} \end{pmatrix} + u_{kk}$$

Thus kth row of U and the kth column of L can be obtained from these equation since L_{k-1} and U_{k-1} are nonsingular.
Assume we have two LU decomposition

$$A = LU = \tilde{L}\tilde{U}$$

then

$$\tilde{L}^{-1}L = \tilde{U}U^{-1}$$

The matrix in the left side is unit lower triangular while the matrix on the right hand side is upper triangular, thus it must be identity matrix. Thus we prove the uniqueness of LU decomposition.

LU factorizations

If A is a real, symmetric, and positive definite matrix, then it has a unique factorization, $A = LL^T$, in which L is lower triangular with a positive diagonal.

Proof. Since A is symmetric and positive definite, all its leading principle minors are symmetric and positive definite. Thus A has an unique LU decomposition

$$A = LU = A^T = U^T L^T$$

I.e., $UL^{-T} = L^{-1}U^T$. The matrix on the left hand side is upper triangular, while the matrix on the right hand side is lower triangular, thus they are diagonal matrix D. Thus, $U = DL^T$, and $A = LDL^T$. Since A is positive definite, the diagonal matrix $D = (d_{ii})$ is also positive definite and A has the Cholesky decomposition $A = \tilde{L}\tilde{L}^T$, where $\tilde{L} = LD^{1/2}$ and $D^{-1/2} = (\sqrt{d_{ii}})$.

LU factorizations

Assume that $A = LL^T = \tilde{L}\tilde{L}^T$, where L and $\tilde{L}$ are the lower triangular with a positive diagonal. We have $\tilde{L}^{-1}L = \tilde{L}^T L^{-T} \equiv D$. The matrix on the left is lower triangular and matrix on the right hand side is upper triangular. Thus D is a diagonal matrix. Moreover the diagonal of L and $\tilde{L}$ are same. i.e, $l_{kk} = \tilde{l}_{kk} = \left(\frac{\det A_k}{\det A_{k-1}}\right)^{1/2}$. This is because $A_k = L_k L_k^T$ where A_k and L_k are the kth leading principle minors of A and L, respectively. We thus have $\det A_k = (\det L_k)^2 = (l_{11} \cdots l_{kk})^2$. Therefore D must be an identity matrix. This is the proof of uniqueness.

LU factorizations

Cholesky factorization

$$l_{kk} = \left(a_{kk} - \sum_{s=1}^{k-1} l_{ks}^2\right)^{1/2}$$

$$l_{ik} \to \left(a_{ik} - \sum_{s=1}^{k-1} l_{is} l_{ks}\right) \Big/ l_{kk}, \quad i = k+1, \cdots, n$$

$l_{kk} > 0$. Since $a_{kk} = \sum_{s=1}^{k} l_{ks}^2 \geq l_{kj}^2$, we have

$$|l_{kj}| \leq |a_{kk}|, \quad j = 1, \cdots, k$$

Exercises: P.134

1,2,3,6,7,8,12,13,15,16,19,24,29,31

Gaussian elimination

We always need to solve a linear system $Ax = b$ with $n \times n$ matrix A, $n \times 1$ column vector b and $n \times 1$ column vector x (unknown).

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

We subtract 2 times the first equation from the second.

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

The numbers 2, $1/2$ and -1 are called multiplier for the first step in the elimination process. The number 6, used as the divisor in forming these multiplier, is called the pivot element.

Gaussian elimination

After 3 steps of Gauss elimination, we obtain an equivalent linear system, which is upper triangular.

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

This system is easily solved by starting the fourth row and working backward up the rows. The first, second and third row are subtracted $5/3$, $-2/3$ and $-4/3$ times the fourth row. The final solution is $x = (1, -3, -2, 1)^T$. Define unit lower triangular matrix

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{pmatrix}$$

we have $A = LU$. This is LU decomposition of matrix A.

Gaussian elimination without pivoting

Consider linear system $A^{(1)}x = Ax = b = b^{(1)}$

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

After one step of Gaussian elimination, we have $A^{(2)}x = b^{(2)}$

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix},$$

where

$$a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i1}a_{1j}^{(1)}, \quad b_i^{(2)} = b_i^{(1)} - l_{i1}b_1^{(1)}, \qquad i, j = 2, 3, \cdots, n.$$

$$l_{i1} = a_{i1}^{(1)}/a_{11}^{(1)}, \quad i = 2, 3, \cdots, n.$$

Gaussian elimination without pivoting

Introducing

$$L_1 = \begin{pmatrix} 1 & & & & \\ -l_{21} & 1 & & & \\ -l_{31} & & 1 & & \\ \vdots & & & \ddots & \\ -l_{n1} & & & & 1 \end{pmatrix} = I - l_1 e_1^T,$$

with $l_1 = (0, l_{21}, \cdots, l_{n1})^T$, we have

$$L_1 A^{(1)} = A^{(2)}, \quad L_1 b^{(1)} = b^{(2)}$$

Gaussian elimination without pivoting

After $k-1$ steps of Gaussian elimination, we have $A^{(k)}x = b^{(k)}$

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} \\ & & \ddots & & & \vdots \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{pmatrix}$$

Calculating the multipliers

$$l_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}, \quad i = k+1, k+2, \cdots, n$$

Gaussian elimination without pivoting

After the k-th Gaussian elimination, we have $A^{(k+1)}x = b^{(k+1)}$

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & \cdots & \cdots & \cdots & a_{2n}^{(2)} \\ & & \ddots & & & & \vdots \\ & & & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ & & & \vdots & \vdots & & \vdots \\ & & & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_k^{(k)} \\ b_{k+1}^{(k+1)} \\ \vdots \\ b_n^{(k+1)} \end{pmatrix},$$

where

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)}, \quad b_i^{(k+1)} = b_i^{(k)} - l_{ik} b_k^{(k)}, \quad i, j = k+1, k+2, \cdots, n.$$

Gaussian elimination without pivoting

Introducing

$$L_k = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ & 1 & \cdots & \cdots & \cdots & \cdots & 0 \\ & & \ddots & & & & \vdots \\ & & & 1 & 0 & \cdots & 0 \\ & & & -l_{k+1,k} & 1 & \cdots & 0 \\ & & & \vdots & \vdots & & \vdots \\ & & & -l_{nk} & 0 & \cdots & 1 \end{pmatrix} = I - l_k e_k^T,$$

where

$$l_k = (0, \cdots, 0, l_{k+1,k}, \cdots, l_{nk})^T, \quad k = 1, 2, \cdots, n-1.$$

we have

$$L_k A^{(k)} = A^{(k+1)}, \quad L_k b^{(k)} = b^{(k+1)}$$

Gaussian elimination without pivoting

After the $n-1$-th Gaussian elimination, we have an upper triangular linear system $A^{(n)}x = b^{(n)}$

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{pmatrix},$$

where

$$L_{n-1}A^{(n-1)} = A^{(n)}, \quad L_{n-1}b^{(n-1)} = b^{(n)}$$

According to the above $n-1$ steps of Gaussian elimination,

$$\begin{cases} L_{n-1}\cdots L_2 L_1 A^{(1)} = A^{(n)} \equiv U \\ L_{n-1}\cdots L_2 L_1 b^{(1)} = b^{(n)} \end{cases}$$

Therefore,

$$A = LU$$

Gaussian elimination without pivoting

with the unit lower triangular matrix

$$L = L_1^{-1} \cdots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & & & 1 \end{pmatrix}$$

The upper triangular linear system can be solved by backward substitution

$$\begin{cases} x_n = b_n^{(n)} / a_{nn}^{(n)} \\ x_k = (b_k^{(k)} - \sum_{j=k+1}^{n} a_{kj}^{(k)} x_j) / a_{kk}^{(k)}, \quad (k = n-1, n-2, \cdots, 1) \end{cases}$$

in the cost of $O(n^2/2)$.

Gaussian elimination without pivoting

The $n-1$ steps of Gaussian elimination (called forward substitution)

$$\begin{cases} l_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}, \quad i = k+1, k+2, \cdots, n \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)}, \quad i, j = k+1, k+2, \cdots, n \\ b_i^{(k+1)} = b_i^{(k)} - l_{ik} b_k^{(k)}, \quad i = k+1, k+2, \cdots, n \end{cases}$$

with the multiplication times $\sum_{k=1}^{n-1} (n-k)^2 = O(n^3/3)$.

Gaussian elimination without pivoting

Theorem: If the pivoting elements $a_{ii}^{(i)} \neq 0$ ($i = 1, 2, \cdots, n$), we have $A = LU$, with unit lower triangular matrix L and upper triangular matrix U. Moreover this LU decomposition is unique.
Proof. One proof is just the process of the Gaussian elimination. Another proof is as follows. First we observe $a_{ij}^{(k+1)} = a_{ij}^{(k)}$ if $i \leq k$ or $j \leq k - 1$ and $u_{kj} = a_{kj}^{(n)} = a_{kj}^{(k)}$. Let $i \leq j$,

$$\begin{aligned}(LU)_{ij} &= \sum_{k=1}^{\min(i,j)} l_{ik} u_{kj} = \sum_{k=1}^{i} l_{ik} u_{kj} = \sum_{k=1}^{i-1} (a_{ik}^{(k)}/a_{kk}^{(k)}) a_{kj}^{(k)} + a_{ij}^{(i)} \\ &= \sum_{k=1}^{i-1} (a_{ij}^{(k)} - a_{ij}^{(k+1)}) + a_{ij}^{(i)} = a_{ij}^{(1)} = a_{ij}\end{aligned}$$

Similarly, if $i > j$, $(LU)_{ij} = a_{ij}$

Gaussian elimination without pivoting

Because that the leading principle minors does not changed if the elementary operations is applied on a matrix, k-th leading principal minor of A

$$\det A_k = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{kk}^{(k)}, \quad k = 1, 2, \cdots, n.$$

This result can also be obtained from the LU decomposition $A = LU$, thus $A_k = L_k U_k$.

Therefore if the first $n - 1$ leading principal minors are not zero, we have LU decomposition.

Pivoting

Example. Consider

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

The Gaussian elimination fails due to the pivoting element 0. Consider

$$\begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad 0 < \varepsilon << 1$$

After one step of Gaussian elimination, we have

$$\begin{pmatrix} \varepsilon & 1 \\ 0 & \varepsilon^{-1} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 - \varepsilon^{-1} \end{pmatrix}$$

The computer calculate the solution

$$\begin{cases} x_2 = (2 - \varepsilon^{-1})/(1 - \varepsilon^{-1}) \\ x_1 = (1 - x_2)\varepsilon^{-1} \end{cases}$$

as follows. Let $\varepsilon = 10^{-8}$.

Pivoting

For a 7-places decimal machine,

$$\begin{aligned}\varepsilon^{-1} &= 0.1000000 \times 10^9 \\ 2 &= 0.000000002 \times 10^9 \\ \varepsilon^{-1} - 2 &= 0.099999998 \times 10^9 (! =) 0.1000000 \times 10^9\end{aligned}$$

Thus the calculated x_2 is 1 by this computer and $x_1 = 0$. Although x_2 is almost exact but x_1 is totally wrong because the exact solution x_1 is close to 1. This problem is caused by the smallness of ε compared to the other elements 1 in this row.

Pivoting

If we scale the first row, then we have an equivalent system

$$\begin{pmatrix} 1 & \varepsilon^{-1} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \varepsilon^{-1} \\ 2 \end{pmatrix}, \quad 0 < \varepsilon << 1$$

After one step of Gaussian elimination, we have

$$\begin{pmatrix} 1 & \varepsilon^{-1} \\ 0 & 1 - \varepsilon^{-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \varepsilon^{-1} \\ 2 - \varepsilon^{-1} \end{pmatrix}$$

The solution

$$\begin{cases} x_2 = (2 - \varepsilon^{-1})/(1 - \varepsilon^{-1}) \\ x_1 = \varepsilon^{-1} - \varepsilon^{-1} x_2 \end{cases}$$

Let $\varepsilon = 10^{-8}$. For the 7-places decimal machine, $x_2 = 1$ and $x_1 = 0$. This is wrong.

Pivoting

If we exchange the two equation, we have an equivalent system

$$\begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad 0 < \varepsilon << 1$$

After one step of Gaussian elimination, we have

$$\begin{pmatrix} 1 & 1 \\ 0 & 1-\varepsilon \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1-2\varepsilon \end{pmatrix}$$

The solution

$$\begin{cases} x_2 = (1-2\varepsilon)/(1-\varepsilon) \approx 1 \\ x_1 = 2 - x_2 \approx 1 \end{cases}$$

It is almost exact! In fact we have the following pivot

$$\begin{pmatrix} \varepsilon & 1 & 1 \\ \boxed{1} & 1 & 2 \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 & 1-\varepsilon & 1-2\varepsilon \\ \boxed{1} & 1 & 2 \end{pmatrix}$$

Pivoting

Good algorithms must incorporate the interchanging of equations. For reasons of economy in computing, we prefer not to move the rows of the matrix around in the computer's memory. Instead, we simply choose the pivot rows in a logical manner. Suppose that instead of using the rows in the order $1, 2, ..., n - 1$ as pivot rows, we use rows $p_1, p_2, ..., p_{n-1}$. Then in the first step, multiples of row p_1 will be subtracted from the other rows. In the next step, multiples of row p_2 will be subtracted from the other rows $p_3, p_4, \cdots, p_n$ except p_1; and so on.

Pivoting

$$\begin{pmatrix} 2 & 3 & -6 \\ 1 & -6 & 8 \\ \boxed{3} & -2 & 1 \end{pmatrix}\begin{pmatrix} s_1 = 6 \\ s_2 = 8 \\ s_3 = 3 \end{pmatrix} \Longrightarrow \begin{pmatrix} 2/3 & \boxed{13/3} & -20/3 \\ 1/3 & -16/3 & 23/3 \\ 3 & -2 & 1 \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} 2/3 & 13/3 & -20/3 \\ 1/3 & -16/13 & -7/13 \\ 3 & -2 & 1 \end{pmatrix} \equiv A$$

$p_1 = 3, p_2 = 1, p_3 = 2.$

Pivoting

Define

$$(P)_{ij} = \delta_{p_i j} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

The ith row of PA is the p_ith row of A. Define $u_{ij} = a^{(3)}_{p_i,j}$ if $j \geq i$ and $l_{ij} = a^{(3)}_{p_i,j}$ if $j < i$. Thus the ith row of L and U are obtained from the p_ith row of $A = \begin{pmatrix} 2/3 & 13/3 & -20/3 \\ 1/3 & -16/13 & -7/13 \\ 3 & -2 & 1 \end{pmatrix}$

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2/3 & 1 & 0 \\ 1/3 & -16/13 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 3 & -2 & 1 \\ 0 & 13/3 & -20/3 \\ 0 & 0 & -7/13 \end{pmatrix},$$

We have $PA = LU$! This is called the Gaussian elimination with scaled row pivoting.

Factorizations $PA = LU$

Let $p_1, p_2, \cdots, p_n$ be the indices of the rows in the order in which they become pivot rows. Let $A^{(1)} = A$, and define $A^{(2)}, \cdots, A^{(n)}$ recursively by the formula

$$a_{p_i j}^{(k+1)} = \begin{cases} a_{p_i j}^{(k)} & \text{if } i \le k \text{ or } i > k > j \\ a_{p_i j}^{(k)} - (a_{p_i k}^{(k)}/a_{p_k k}^{(k)})/a_{p_k j}^{(k)} & \text{if } i > k \text{ and } j > k \\ a_{p_i k}^{(k)}/a_{p_k k}^{(k)} & \text{if } i > k \text{ and } k = j \end{cases}$$

Define a permutation matrix P whose elements are $P_{ij} = \delta_{p_i j}$. Define an upper triangular matrix U whose elements are $u_{ij} = a_{p_i j}^{(n)}$ if $j \geq i$. Define a unit lower triangular matrix L whose elements are $l_{ij} = a_{p_i j}^{(n)}$ if $j < i$. Then $PA = LU$.

Factorizations $PA = LU$

Proof. From the recursive formula,

$$u_{kj} = a^{(n)}_{p_k j} = a^{(k)}_{p_k j}, \quad k \leq j$$

This is because the p_k-th row does not changed during the Gaussian elimination from $A^{(k)} \to \cdots \to A^{(n)}$.

$$l_{ik} = a^{(n)}_{p_i k} = a^{(k+1)}_{p_i k} = a^{(k)}_{p_i k} / a^{(k)}_{p_k k}, \quad k \leq j$$

This is because the k-th column does not changed during the Gaussian elimination from $A^{(k+1)} \to \cdots \to A^{(n)}$. Let $i \leq j$,

$$\begin{aligned} (LU)_{ij} &= \sum_{k=1}^{i} l_{ik} u_{kj} = \sum_{k=1}^{i-1} (a^{(k)}_{p_i k} / a^{(k)}_{p_k k}) a^{(k)}_{p_k j} + a^{(i)}_{p_i j} \\ &= \sum_{k=1}^{i-1} (a^{(k)}_{p_i j} - a^{(k+1)}_{p_i j}) + a^{(i)}_{p_i j} = a^{(1)}_{p_i j} = a_{p_i j} \end{aligned}$$

Similarly, if $i > j$, $(LU)_{ij} = a_{p_i j}$. Thus $(PA)_{ij} = (LU)_{ij}$.

Diagonally Dominant Matrices

Sometimes a system of equations has the property that Gaussian elimination without pivoting can be safely used. One class of matrices for which this is true is the class of diagonally dominant matrices.

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad (1 \leq i \leq n)$$

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

Diagonally Dominant Matrices

Proof. Let $A = A^{(1)} = (a_{ij})$ be diagonally dominant. We want to prove $A^{(2)}$, obtained by one step Gaussian elimination without pivot, is also diagonally dominant

$$|a_{ii}^{(2)}| = |a_{ii} - a_{i1}a_{1i}/a_{11}| > \sum_{i\neq j=2}^{n} |a_{ij}^{(2)}| = \sum_{i\neq j=2}^{n} |a_{ij} - a_{i1}a_{1j}/a_{11}|, \quad i = 2, \cdots, n$$

It suffices to prove that

$$|a_{ii}| - |a_{i1}a_{1i}/a_{11}| > \sum_{i\neq j=2}^{n} (|a_{ij}| + |a_{i1}a_{1j}/a_{11}|), \quad \text{i.}e.$$

$$|a_{ii}| - \sum_{i\neq j=2}^{n} |a_{ij}| > \sum_{j=2}^{n} |a_{i1}a_{1j}/a_{11}|$$

From the diagonal dominance of A, it suffice to prove the obvious result

$$|a_{i1}| > \sum_{j=2}^{n} |a_{i1}a_{1j}/a_{11}|$$

Diagonally Dominant Matrices

Every diagonally dominant matrix is nonsingular and has an LU factorization.

Proof. According to the above theorem, A has LU decomposition $A = LU$, where U is diagonally dominant and thus nonsingular and A is also nonsingular.

Diagonally Dominant Matrices

The scaled row pivoting version of Gaussian elimination is applied to a diagonally dominant matrix, then the pivots will be the natural ones: $1, 2, \cdots, n$. Hence, the work of choosing the pivots can be omitted in this case.
Proof. By the above theorem, it suffices to prove that the first pivot chosen in the algorithm is 1 $(p_1 = 1)$. Thus it is enough to prove that

$$|a_{11}|/|s_1| > |a_{i1}|/|s_i|, \quad 2 \leq i \leq n$$

where $s_i = \max_j |a_{ij}| = |a_{ii}|$. It suffices to prove that $1 > |a_{i1}|/|a_{ii}|, 2 \leq i \leq n$. It is obvious that this equality is valid.

Tridiagonal system

A is said to be tridiagonal if $a_{ij} = 0$ for all $|i - j| > 1$. Consider a tridiagonal system

$$\begin{pmatrix} d_1 & c_1 & & & \\ a_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & a_{n-1} & d_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix}$$

Forward substitution:

$$d_2 \leftarrow d_2 - a_1 c_1 / d_1, \quad b_2 \leftarrow b_2 - a_1 b_1 / d_1$$

$$d_i \leftarrow d_i - a_{i-1} c_{i-1} / d_{i-1}, \quad b_i \leftarrow b_i - a_{i-1} b_{i-1} / d_{i-1}, \quad i = 2, \cdots, n$$

Tridiagonal system

After the forward substitution,

$$\begin{pmatrix} d_1 & c_1 & & & \\ & d_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & d_{n-1} & c_{n-1} \\ & & & & d_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix}$$

Backward substitution:

$$x_n \leftarrow b_n/d_n, \quad x_i \leftarrow (b_i - c_i x_{i+1})/d_i, \quad i = n-1, \cdots, 1$$

Exercises: P.155

1(c), 3,8,10,12,17,30,41

Norms and Analysis of Errors

To discuss the errors in numerical problems involving vectors, it is useful to employ norms of vectors.
A norm in R^n is a function $||\cdot||$ from R^n to the set of non-negative reals that obeys these three postulates.

- $||x|| \geq 0$ if $x \in R^n$. $||x|| = 0$ if and only if $x = 0$
- $||\lambda x|| = |\lambda| ||x||$, if $\lambda \in R$, $x \in R^n$.
- $||x + y|| \leq ||x|| + ||y||$, if $x, y \in R^n$

We can think of $||x||$ as the length or magnitude of the vector x.

Norms and Analysis of Errors

The Euclidean norm

$$||x||_2 = (\sum_{i=1}^{n} |x_i|^2)^{\frac{1}{2}}$$

The l_∞ norm

$$||x||_\infty = \max_{1 \leq i \leq n} |x_i|$$

The l_1 norm

$$||x||_1 = \sum_{1 \leq i \leq n} |x_i|$$

The l_p norm $(1 \leq p < \infty)$

$$||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$$

We can plot all vectors x in R^2 which satisfies $||x|| < 1$.

Norms and Analysis of Errors

Now we return the norms of matrices. Although we can define the general matrix norms, subjecting them only to the same requirements of vector norms, we usually prefer to a matrix norm to be intimately related to a vector norm. If a vector norm $||\cdot||$ is specified, the matrix norm subordinated to it is defined by

$$||A|| = \sup\{||Au|| : u \in R^n, ||u|| = 1\}$$

This is also called the matrix norm associated with the given vector norm. Here $A \in R^{n\times n}$ is an $n \times n$ matrix. We can prove that the above equation for any matrix is a norm of matrix.

Norms and Analysis of Errors

We shall verify the three axioms for a norm. First, if $A \neq 0$, then A has at least one nonzero column; say, $A^{(j)} \neq 0$. Consider the vector in which 1 is the jth component–that is, $x = (0, \cdots, 0, 1, 0, \cdots, 0)^T$. Define $v = x/\|x\|$ is of norm 1. Hence

$$\|A\| \geq \|Av\| = \frac{\|Ax\|}{\|x\|} = \frac{\|A^{(j)}\|}{\|x\|} > 0$$

Next

$$\|\lambda A\| = \sup\{\|\lambda Au\| : u \in R^n, \|u\| = 1\} = |\lambda| \|A\|$$

For the triangle inequality,

$$\begin{aligned} \|A+B\| &= \sup\{\|(A+B)u\| : u \in R^n, \|u\| = 1\} \\ &\leq \sup\{\|Au\| : u \in R^n, \|u\| = 1\} + \sup\{\|Bu\| : u \in R^n, \|u\| = 1\} \\ &= \|A\| + \|B\| \end{aligned}$$

Norms and Analysis of Errors

An important property of matrix norm which is defined above is

$$||Ax|| \leq ||A||||x||, \quad x \in R^n$$

Given a vector norm $||x||_\infty$, what is the subordinate matrix norm?

$$\begin{aligned} ||A||_\infty &= \sup_{||u||_\infty=1} ||Au||_\infty = \sup_{||u||_\infty=1} \max_{1\leq i\leq n} |(Au)_i| = \max_{1\leq i\leq n} \sup_{||u||_\infty=1} |(Au)_i| \\ &= \max_{1\leq i\leq n} \sup_{||u||_\infty=1} |\sum_j a_{ij}u_j| = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \end{aligned}$$

A subordinate matrix norm has also the following properties:

$$||I|| = 1$$

$$||AB|| \leq ||A||||B||$$

Norms and Analysis of Errors

Let us consider the linear system $Ax = b$. Suppose that the vector b is perturbed to obtain a vector $\tilde{b}$. If $\tilde{x}$ satisfies $A\tilde{x} = \tilde{b}$, how much do x and $\tilde{x}$ differ, in absolute and relative terms? Assume that A is invertible, we have

$$||x - \tilde{x}|| = ||A^{-1}b - A^{-1}\tilde{b}|| \leq ||A^{-1}|| \, ||b - \tilde{b}||$$

This gives a measure of the perturbation in x. To estimate the relative perturbation,

$$||x - \tilde{x}|| \leq ||A^{-1}|| \, ||b - \tilde{b}|| \leq ||A^{-1}|| \, ||A|| \, ||x|| \frac{||b - \tilde{b}||}{||b||}$$

where we used $||b|| = ||Ax|| \leq ||A|| \, ||x||$. Hence,

$$\frac{||x - \tilde{x}||}{||x||} \leq ||A^{-1}|| \, ||A|| \frac{||b - \tilde{b}||}{||b||}$$

The number $\kappa(A) = ||A|| \, ||A^{-1}||$ is called a condition number of a matrix A.

Norms and Analysis of Errors

Consider matrix

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

Norms and Analysis of Errors

The k-th eigenvalue of A is $4\sin^2(\frac{k\pi}{2(n+1)})$, $k=1,\cdots,n$. The spectral condition number for positive matrix is the ratio of the largest eigenvalue and the smallest eigenvalue

$$\kappa(A)_2 = \frac{\sin^2(\frac{n\pi}{2(n+1)})}{\sin^2(\frac{\pi}{2(n+1)})} \approx \left(\frac{2(n+1)}{\pi}\right)^2.$$

which is consistent with the results

n	$\kappa(A)_\infty$	$\kappa(A)_2$
8	33.909918	32.828064
16	119.500176	117.127296
32	447.885186	441.355113
64	1736.772886	1712.328166
128	6839.757524	6744.343948

Norms and Analysis of Errors

Vandermonde matrix

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}, \quad x_i = \frac{i}{n}, \quad i = 0, \cdots, n$$

The condition number

n	2	4	8
$\kappa(A)_\infty$	19.50	1277.373	4550931.708

The coefficients of Lagrangian polynomial based on x_i, $i = 0, \cdots, n$ satisfies $Ax = b$. We should not use Gaussian elimination to solve $Ax = b$.

Norms and Analysis of Errors

$$A = \begin{pmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix}, \quad A^{-1} = \varepsilon^{-2} \begin{pmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{pmatrix}$$

If the ∞-norm is employed, $||A||_\infty = 2+\varepsilon$, $||A^{-1}||_\infty = \varepsilon^{-2}(2+\varepsilon)$.
Hence $\kappa(A) = [(2+\varepsilon)/\varepsilon]^2$
If we solve $Ax = b$, numerically, we obtained an approximate solution $\tilde{x}$. We define the residual error

$$r = b - A\tilde{x}$$

and the error vector

$$e = x - \tilde{x}$$

We have the relationship between the error and residual error

$$Ae = r$$

Norms and Analysis of Errors

We can prove that

$$\frac{1}{\kappa(A)}\frac{||r||}{||b||} \leq \frac{||e||}{||x||} \leq \kappa(A)\frac{||r||}{||b||}$$

The inequality on the right hand side can be derived from

$$||e||\ ||b|| \leq ||A^{-1}r||\ ||Ax|| \leq ||A^{-1}||\ ||r||\ ||A||\ ||x||$$

The inequality on the left hand side can be derived from

$$||r||\ ||x|| = ||Ae||\ ||A^{-1}b|| \leq ||A||\ ||e||\ ||A^{-1}||\ ||b||$$

A matrix with a large condition number is said to be ill conditioned. If the condition number of a matrix is moderate size, the matrix is said to be well conditioned.

Neumann Series and iterative Refinement

We say that a given sequence converges to a vector v if

$$\lim_{k\to\infty} \|v^{(k)} - v\| = 0$$

For example,

$$v^{(k)} = \begin{pmatrix} 3 - k^{-1} \\ -2 + k^{-1/2} \\ (k+1)k^{-1} \\ e^{-k} \end{pmatrix} \Longrightarrow v = \begin{pmatrix} 3 \\ -2 \\ 1 \\ 0 \end{pmatrix}$$

Any two norms $\|\cdot\|$ and $\|\cdot\|'$ in R^n are equivalent: I.e., there exists two constants c and C independent of x, such that for all $x \in R^n$

$$c\|x\| \le \|x\|' \le C\|x\|$$

Thus if a sequence in R^n converges in one norm, it must converges in the other norm of R^n. One particular convinent norm is l_∞ norm

$$\lim_{k\to\infty} \|v^{(k)} - v\|_\infty = 0$$

Neumann Series and iterative Refinement

If a sequence $\{v^{(k)}\}$ in R^n satisfies the Cauchy criterion if

$$\lim_{k \to \infty} \sum_{i,j>k} \|v^{(i)} - v^{(j)}\| = 0$$

then there necessarily exists a vector $v \in R^n$ to which this sequence converges.

Neumann Series and iterative Refinement

If A is an $n \times n$ matrix such that $\|A\| < 1$, then $I - A$ is invertible and

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

Proof. First, we shall show that $I - A$ is invertible. If it is singular, then there exists a vector x satisfying $\|x\| = 1$ and $(I - A)x = 0$. From this we have

$$1 = \|x\| = \|Ax\| \leq \|A\| \, \|x\| = \|A\|$$

which contradicts the hypothesis that $\|A\| < 1$. We shall show that the partial sums of the Neumann series converges to $(I - A)^{-1}$:

$$\sum_{k=0}^{m} A^k \to (I - A)^{-1}$$

as $m \to \infty$.

Neumann Series and iterative Refinement

It will suffice to prove that

$$(I - A)\sum_{k=0}^{m} A^k \to I$$

The left-hand side can be written as

$$(I - A)\sum_{k=0}^{m} A^k = \sum_{k=0}^{m}(A^k - A^{k+1}) = I - A^{m+1} \to I$$

since $\|A^{m+1}\| \leq \|A\|^{m+1}$.
From these result, we have

$$\|(I - A)^{-1}\| \leq \sum_{k=0}^{\infty} \|A^k\| \leq \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1 - \|A\|}$$

Neumann Series and iterative Refinement

A variant of this theorem is as follows. If A and B are $n \times n$ matrices such that $\|I - AB\| < 1$, then A and B are invertible. Furthermore,

$$A^{-1} = B \sum_{k=0}^{\infty} (I - AB)^k, \quad B^{-1} = \sum_{k=0}^{\infty} (I - AB)^k A$$

Neumann Series and iterative Refinement

If $x^{(0)}$ is an approximate solution of the equation $Ax = b$, then the precise solution x is given by

$$x = x^{(0)} + e^{(0)}$$

where $e^{(0)} = x - x^{(0)} = A^{-1}(b - Ax^{(0)})$ is the error vector of $x^{(0)}$. The residual error of $x^{(0)}$ is $r^{(0)} = b - Ax^{(0)}$ which is computable. $e^{(0)}$ can be obtained from $Ae^{(0)} = r^{(0)}$. This leads to the procedure called iterative improvement or iterative refinement.

$$\begin{cases} r^{(0)} = b - Ax^{(0)} \\ Ae^{(0)} = r^{(0)} \\ x^{(1)} = x^{(0)} + e^{(0)} \end{cases}$$

To obtain better approximate solution $x^{(0)}, x^{(1)}$, this process can be repeated. The success of this method depends on the residual $r^{(i)}$ in double precision to avoid the loss of significance expected in the substraction.

Neumann Series and iterative Refinement

To analyze this algorithm theoretically, we assume that $x^{(0)}$ is obtained by

$$x^{(0)} = Bb$$

where B is an approximate inverse of A. The iterative process can be written as

$$x^{(k+1)} = x^{(k)} + B(b - Ax^{(k)}), \quad k \geq 0$$

Here B is approximate inverse of A, which means that $\|I - BA\| < 1$. The sequence $x^{(k)}$ produces

$$x^{(m)} = B \sum_{k=0}^{m} (I - AB)^k b, \quad m \geq 0$$

which converges to the exact solution

$$x = B \sum_{k=0}^{\infty} (I - AB)^k b$$

of $Ax = b$. This is due to $\|x^{(m+1)} - x\| \leq \|I - BA\| \, \|x^{(m)} - x\|$.

Neumann Series and iterative Refinement

For solving a linear equations in extremely case, a number of refinements can be added to the factorization and solution phases. We discuss five such techniques.

- preconditioning of row equilibration
- preconditioning of column equilibration
- full pivoting
- preconditioning or scaling within each step of the elimination procedure
- iterative refinement at the end

Neumann Series and iterative Refinement

Row equilibration is the process of dividing each row of the coefficient matrix by the maximum element in absolute value in that row; That is, multiplying row i by $r_i = 1/\max_{1\le j\le n}|a_{ij}|$ for $1 \le i \le n$. After this has been done, the new elements $\tilde{a}_{ij}$ satisfying $\max_{1\le j\le n}|\tilde{a}_{ij}| \le 1$. In numerical practice on a binary computer, r_i is taken to be a number of the form 2^m as close as possible to $1/\max_{1\le j\le n}|a_{ij}|$. This is done to avoid the introduction of addition of roundoff error. Row equilibration can be written as

$$(RA)x = Rb, \quad R = diag(r_i)$$

Column equilibration can be written as

$$(AC)(C^{-1}x) = b, \quad C = diag(c_i)$$

where $c_j = 1/\max_{1\le i\le n}|a_{ij}|$.

Neumann Series and iterative Refinement

The full pivoting strategy at the initial step is to search for the largest element (in magnitude) in the matrix. This element determines the first pivot row and the first column in which zeros will be introduced by the elimination. Thus we intend to process the columns not in a natural order $1, 2, 3, \cdots, n$ but in an order determined by this more accurate pivoting strategy. Two permutations are needed, one is list the row numbers of successive pivoting elements, and the other to list the corresponding column numbers.

Neumann Series and iterative Refinement

The four technique in our list of refinements is preconditioning and scaling. It contributes to a more logical organization of the factorization phase, for each step in the algorithm will be like the first except for being applied to smaller matrices.

The fifth technique is the iterative refinements, which is discussed previously.

Exercises: P.178

2,3,4

Solutions of Equations by Iterative Methods

- The Gaussian algorithm and its invariants are termed directed methods for $Ax = b$. They proceed through a finite step of elimination to produce an exact solution if there is no roundoff error.
- An indirect method, by contrast, produces a sequence of vectors which ideally converges to the exact solution. The computation is halted when an approximate solution having a specified accuracy is obtained or after a certain number of iterations. Indirect methods are almost always iterative in nature: a simple process is applied repeatedly to generate a sequence of approximate solutions.

Solutions of Equations by Iterative Methods

- For large linear systems containing thousands of equations, iterative methods always have decisive advantages over direct methods in terms of speed and demands on computer memory. Sometimes, if the accuracy requirements are not stringent, a modest number of iterations produce an acceptable solution.
- For sparse systems (in which a large proportion of matrix are zero), the iterative methods are often very efficient. In sparse problems, the nonzero-elements are sometimes stored in a sparse-storage format. This is very common in the numerical solution of partial different equations. Another advantages of iterative methods is that they are usually stable, and they will actually dampen errors as the process continues.

Solutions of Equations by Iterative Methods

Consider the liner system

$$\begin{pmatrix} 7 & -6 \\ -8 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \end{pmatrix}$$

A straightforward procedure is to solve the *i*th equation for the unknown x_i. Jacobi iteration:

$$\begin{cases} x_1^{(k)} = \frac{6}{7} x_2^{(k-1)} + \frac{3}{7} \\ x_2^{(k)} = \frac{8}{9} x_1^{(k-1)} - \frac{4}{9} \end{cases}$$

Gauss-Seidel iteration:

$$\begin{cases} x_1^{(k)} = \frac{6}{7} x_2^{(k-1)} + \frac{3}{7} \\ x_2^{(k)} = \frac{8}{9} x_1^{(k)} - \frac{4}{9} \end{cases}$$

Both the Jacobi and Gauss-Seidel iterates seems to converges to the same limit and the latter is converging faster.

Solutions of Equations by Iterative Methods

A general type of iterative process for solving $Ax = b$ can be described as follows. It can be split to be

$$Qx = (Q - A)x + b$$

where Q is called the splitting matrix. Thus we have the following iteration

$$Qx^{(k)} = (Q - A)x^{(k-1)} + b, \quad k \geq 1$$

The initial vector $x^{(0)}$ can be arbitrary. We say that this iteration converges if it converges for any initial vector $x^{(0)}$. Here we choose Q so that (1) the sequence $x^{(k)}$ is easily computed. (2) the sequence $x^{(k)}$ converges rapidly to a solution. We always to choose that Q^{-1} approximates A^{-1}.
Once the above iteration converges, its limit vector x is the exact solution of $Ax = b$.

Solutions of Equations by Iterative Methods

If $\|I - Q^{-1}A\| < 1$ for some subordinate matrix norm, then the sequence $x^{(k)}$ produced above converges to the solution of $Ax = b$ for any initial vector $x^{(0)}$.

Proof. The proof is based on the following facts.

$$x^{(k)} = (I - Q^{-1}A)x^{(k-1)} + Q^{-1}b$$

The exact solution x satisfies

$$x = (I - Q^{-1}A)x + Q^{-1}b$$

Therefore

$$\|x^{(k)} - x\| \leq \|I - Q^{-1}A\| \, \|x^{(k-1)} - x\|$$

We can also prove that

$$\|x^{(k)} - x\| \leq \frac{\delta}{1-\delta}\|x^{(k)} - x^{(k-1)}\|$$

where $\delta = \|I - Q^{-1}A\| < 1$.

Solutions of Equations by Iterative Methods

If $Q = diag(a_{ii})$ to be the diagonal part of A, the corresponding iteration becomes Jacobi iteration. If A is diagonally dominant, then the Jacobi iteration converges to the solution of $Ax = b$ for any starting vector. This is because

$$||I - Q^{-1}A||_{\infty} = \max_{1 \le i \le n} \sum_{i \ne j = 1}^{n} |a_{ij}/a_{ii}| < 1$$

Solutions of Equations by Iterative Methods

Every square matrix is similar to a (possible complex) upper triangular matrix whose off-diagonal elements are arbitrarily small.

The spectral radius of any matrix A is

$$\rho(A) = \inf_{\|\cdot\|} \|A\|$$

in which the infimum is taken over all subordinate matrix norms.

Solutions of Equations by Iterative Methods

For the iteration formula

$$x^{(k)} = Gx^{(k-1)} + c$$

to produce a sequence converging to $(I - G)^{-1}c$, for any starting vector $x^{(0)}$, it is necessary and sufficient condition that the spectral radius of G be less than 1.
Proof. Suppose $\rho(G) < 1$, by the above theorem, there is a subordinate matrix norm such that $||G|| < 1$. For the iteration formula, we have

$$x^{(k)} = G^k x^{(0)} + \sum_{j=0}^{k-1} G^j c$$

$G^k x^{(0)} \rightarrow 0$ since $||G|| < 1$. By the theorem on the Neumann series, $\sum_{j=0}^{\infty} G^j c = (I - G)^{-1}c$. So $\lim_{k\rightarrow\infty} x^{(k)} = (I - G)^{-1}c$.

Solutions of Equations by Iterative Methods

For the converse, suppose $\rho(G) \geq 1$. Select u and λ so that

$$Gu = \lambda u, \quad |\lambda| \leq 1, \quad \|u\| \neq 0$$

If $|\lambda| = 1$, let $c = u$ and $x^{(0)} = 0$, $x^{(k)} = \sum_{j=0}^{k-1} G^j u = \frac{1-\lambda^k}{1-\lambda} u$ which diverges as $k \to \infty$. If $|\lambda| > 1$, let $c = 0$ and $x^{(0)} = u$, $x^{(k)} = G^k u = \lambda^k u$ which diverges as $k \to \infty$.

Solutions of Equations by Iterative Methods

One corollary of the above theorem: The iteration $Qx^{(k)} = (Q - A)x^{(k-1)} + b$ will produces a sequence converging to the solution of $Ax = b$, for any $x^{(0)}$, if $\rho(I - Q^{-1}A) < 1$.

Solutions of Equations by Iterative Methods

If Q is chosen to be the lower part of A (including its diagonal), the iteration is called the Gauss-Seidel iteration.
If A is diagonally dominant, then the Gauss-Seidel iteration converges for any starting vector $x^{(0)}$.
Proof. By the above corollary, it suffices to prove that

$$\rho(I - Q^{-1}A) < 1$$

Let λ be an eigenvalue of $I - Q^{-1}A$ and the corresponding eigenvectors x with $||x||_\infty = 1$. We have

$$(I - Q^{-1}A)x = \lambda x, \quad or\ Qx - Ax = \lambda Qx$$

i.e.,

$$-\sum_{j=i+1}^{n} a_{ij}x_j = \lambda \sum_{j=1}^{i} a_{ij}x_j, \quad 1 \leq i \leq n$$

Solutions of Equations by Iterative Methods

or

$$\lambda a_{ii}x_i = -\lambda \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j, \quad 1 \le i \le n$$

Select an index i such that $|x_i| = 1 \ge \|x_j\|$ for all j. Then

$$|\lambda|\ |a_{ii}| \le |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

Solving λ and using the diagonal dominance of A, we get

$$|\lambda| \le \Big(\sum_{j=i+1}^{n} |a_{ij}| \Big)\Big(|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| \Big)^{-1} < 1$$

Solutions of Equations by Iterative Methods

The space of complex n- vectors is denotes by C^n. We define the inner product

$$(x, y) = y^* x = \sum_{i=1}^{n} x_i \bar{y}_i$$

Here $y^* = (\bar{y}_1, \cdots, \bar{y}_n)$ is the conjugate transpose of y. it is easy to prove that

$$(x, x) \geq 0$$

$$(x, \lambda y) = \bar{\lambda}(x, y)$$

Then it follows that $(\alpha x + \beta y) = \alpha(x, z) + \beta(y, z)$ for any scalar α, β and vectors x, y and z. The Euclidean norm of x is

$$||x||_2 = \sqrt{(x, x)} = \Big(\sum_{i=1}^{n} |x_i|^2 \Big)^{1/2}$$

A matrix $A = (a_{ij})$ is called to be Hermitian if $A^* = A$. Here A^* is the conjugate transpose of A.

Solutions of Equations by Iterative Methods

In the SOR (successive over-relaxation) method, suppose that the splitting matrix Q is chosen to be $\alpha D - C$, with α a real parameter, where D is any positive definite Hermitian matrix and C is any matrix satisfying $C + C^* = D - A$. If A is positive definite Hermitian, if Q is nonsingular, and if $\alpha > 1/2$, then the SOR iteration converges for any starting vectors.

Solutions of Equations by Iterative Methods

Suppose that A is partitioned into $A = D - L - U$ where $D = diag(A)$, L (U) is the negative of the strictly lower (upper) triangular part of A.

Jacobi iteration

$$\begin{cases} Q = D \\ G = D^{-1}(L+U) \end{cases}$$

$$Dx^{(k)} = (L+U)x^{(k-1)} + b$$

Gauss-Seidel iteration

$$\begin{cases} Q = D - L \\ G = (D-L)^{-1}U \end{cases}$$

$$(D-L)x^{(k)} = Ux^{(k-1)} + b$$

SOR iteration

$$\begin{cases} Q = \omega^{-1}(D - \omega L) \\ G = (D - \omega L)^{-1}(\omega U + (1 - \omega)D) \end{cases}$$

$$(D - \omega L)x^{(k)} = \omega(Ux^{(k-1)} + b) + (1 - \omega)Dx^{(k-1)}$$

Exercises: P.201

1,2,5,7,8,15,20,30,31,35

Analysis of Roundoff Error in the Gaussian Algorithm

We investigate the roundoff error that inevitably arise in solving $Ax = b$. An analysis, due originally to Wilkinson, is given for Gaussian elimination using unscaled row pivoting. The results are in a form of a posterior bounds on the error. Thus, at the conclusion of the computation, an assertion can be made about the magnitude of the error. We assume that the pivots are located on the diagonal.

Roundoff analysis of Gaussian elimination

The Gaussian elimination

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & \textit{if } i \le k \\ a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)} & \textit{if } i > k \textit{ and } j > k \\ 0 & \textit{if } i > k \ge j \end{cases}$$

$$l_{ik} = \begin{cases} 0 & \textit{if } i < k \\ 1 & \textit{if } i = k \\ a_{ik}^{(k)}/a_{kk}^{(k)} & \textit{if } i > k \end{cases}$$

The computed numbers in Gaussian algorithm is

$$\tilde{a}_{ij}^{(k+1)} = \begin{cases} \tilde{a}_{ij}^{(k)} & \textit{if } i \le k \\ fl(\tilde{a}_{ij}^{(k)} - fl(\tilde{l}_{ik} \tilde{a}_{kj}^{(k)})) & \textit{if } i > k \textit{ and } j > k \end{cases}$$

$$\tilde{l}_{ik} = \begin{cases} 1 & \textit{if } i = k \\ fl(\tilde{a}_{ik}^{(k)}/\tilde{a}_{kk}^{(k)}) & \textit{if } i > k \end{cases}$$

Roundoff analysis of Gaussian elimination

Denote by $\mathrm{fl}(x)$ the computer presentation (called float number) of real number x,

$$\mathrm{fl}(x \odot y) = (x \odot y)(1 - \delta) = (x \odot y)/(1 - \delta')$$

where $|\delta|, |\delta'| < \varepsilon$, ε is called machine precision

$$\varepsilon = \begin{cases} b^{1-t}/2, & \text{if the roundoff error is used} \\ b^{1-t}, & \text{if truncation error is used} \end{cases},$$

where the base b and world length t.

Roundoff analysis of Gaussian elimination

Let A be an $n \times n$ nonsingular matrix whose elements are machine numbers in a computer with unit roundoff ε. The Gaussian algorithm with row pivoting produces matrices the unit lower triangular matrix $\tilde{L}$ and upper triangular matrix $\tilde{U}$, such that

$$\tilde{L}\tilde{U} = P(A + E),$$

where $E = (e_{ij})$ satisfies

$$|e_{ij}| \leq 2n\varepsilon\rho, \quad i, j = 1, \cdots, n$$

with

$$\rho = \max_{1 \leq i,j,k \leq n} |a_{ij}^{(k)}|$$

Roundoff analysis of Gaussian elimination

Suppose that $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ are the machine numbers. The machine precision ε satisfies $n\varepsilon < 1/3$, then

$$\Big((x_1y_1 + x_2y_2) + x_3y_3\Big) + \cdots + x_ny_n$$

can be presented to be $\sum_{i=1}^{n} x_iy_i(1+\delta_i)$, where $\delta_i \le \frac{6}{5}(n+1)\varepsilon$.

Roundoff analysis of Gaussian elimination

Let L be an $n \times n$ unit lower triangular matrix whose elements are machine numbers. Let b be a vector whose components are machine numbers. The machine precision ε satisfies $(n+1)\varepsilon < 1/3$, The computed solution $\tilde{y}$ of $Ly = b$ satisfies

$$(L + \Delta)\tilde{y} = b,$$

with

$$|\Delta_{ij}| \leq \frac{6}{5}(n+1)\varepsilon |l_{ij}|, \quad i, j = 1, \cdots, n.$$

This result is also true if L is replaced by the upper triangular matrix U.

Roundoff analysis of Gaussian elimination

Let the elements of A and b are machine numbers. If the Gaussian algorithm with row pivoting is used to solve $Ax = b$, then the computed solution $\tilde{x}$ is the exact solution of a perturbed system

$$(A + F)\tilde{x} = b, \quad |f_{ij}| \leq 10n^2 \varepsilon \rho, \quad 1 \leq i, j \leq n$$

Choose $||F||_\infty$ such that

$$\kappa(A)_\infty ||F||_\infty < ||A||_\infty,$$

then

$$\frac{||\tilde{x} - x||_\infty}{||x||_\infty} \leq 10n^3 \varepsilon g_n(A) \kappa(A)_\infty,$$

with

$$g_n(A) = \frac{\max_{1 \leq i,j,k \leq n} |a_{ij}^{(k)}|}{\max_{1 \leq i,j \leq n} |a_{ij}|}$$

Here $A^{(k)} = (a_{ij}^{(k)})$ is the matrix after $k - 1$ times Gaussian eliminations.