Chapter 7
Iterative Techniques in Matrix Algebra

Hung-Yuan Fan (范洪源)

Department of Mathematics,
National Taiwan Normal University, Taiwan

Spring 2016

Section 7.1
Norms of Vectors and Matrices

Def 7.1

A vector norm on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ with the following properties:

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

(ii) $\|x\| = 0 \iff x = 0$,

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

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

Note: The n-dimensional vector x is often denoted by

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_1, x_2, \cdots, x_n]^T = (x_1, x_2, \cdots, x_n)^T.$$

Def 7.2

- The l_2 and l_∞ norms for the vector $x = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^n$ are defined by

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \quad \text{and} \quad \|x\|_\infty = \max_{1 \le i \le n} |x_i|.$$

- The l_1 norm of $x \in \mathbb{R}^n$ is defined by

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

Def 7.4

Let $x = [x_1, \cdots, x_n]^T$ and $y = [y_1, \cdots, y_n]^T$ be two vectors in $\mathbb{R}^n$.

- The l_2 and l_∞ distances between x and y are defined by

$$\|x-y\|_2 = \left\{ \sum_{i=1}^{n} (x_i - y_i)^2 \right\}^{1/2} \quad \text{and} \quad \|x-y\|_\infty = \max_{1 \le i \le n} |x_i - y_i|.$$

- The l_1 distance between x and y is given by

$$\|x - y\|_1 = \sum_{i=1}^{n} |x_i - y_i|.$$

Example 2, p. 435

The 3×3 linear system

$$
\begin{aligned}
3.3330x_1 + 15920x_2 - 10.333x_3 &= 15913, \\
2.2220x_1 + 16.710x_2 + 9.6120x_3 &= 28.544, \\
1.5611x_1 + 5.1791x_2 1.6852x_3 &= 8.4254
\end{aligned}
$$

has the **exact sol.** $\mathbf{x} = [\mathbf{1}, \mathbf{1}, \mathbf{1}]^{\mathbf{T}}$. If the system is solved by GE with patrial pivoting using **5-digit rounding arithnetic**, we obtain the computed sol.

$$\tilde{x} = [\mathbf{1.2001}, \mathbf{0.99991}, \mathbf{0.92538}]^{\mathbf{T}}.$$

So, the l_∞ and l_2 distances between x and $\tilde{x}$ are

$$\|x - \tilde{x}\|_\infty = \mathbf{0.2001} \quad \text{and} \quad \|x - \tilde{x}\|_2 = \mathbf{0.21356}.$$

Def 7.5 (向量序列的收斂性)

A seq. $\{x^{(k)}\}_{k-1}^{\infty}$ of vectors in $\mathbb{R}^n$ is said to converge to $x \in \mathbb{R}^n$ with respect to the norm $\|\cdot\|$ if $\forall\, \epsilon > 0$, $\exists\, N(\epsilon) \in \mathbb{N}$ s.t.

$$\|x^{(k)} - x\| < \epsilon \qquad \forall\, k \geq N(\epsilon).$$

Thm 7.6

The seq. of vectors $\{x^{(k)}\}_{k=1}^{\infty}$ converges to $x \in \mathbb{R}^n$ **with respect to the l_∞ norm** $\Longleftrightarrow \lim\limits_{k\to\infty} x_i^{(k)} = x_i$ for $i = 1, 2, \ldots, n$.

pf: It is easily seen that $\forall\, \epsilon > 0$, $\exists\, N(\epsilon) \in \mathbb{N}$ s.t.

$$\|x^{(k)} - x\|_\infty < \epsilon \qquad \forall\, k \geq N(\epsilon)$$

$$\Longleftrightarrow |x_i^{(k)} - x_i| < \epsilon \qquad \forall\, k \geq N(\epsilon) \text{ and } 1 \leq i \leq n.$$

Example 3, p. 436

The sequence of vectors in $\mathbb{R}^4$

$$x^{(k)} = [1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k}\sin k]^T$$

converges to $x = [1, 2, 0, 0]^T \in \mathbb{R}^4$ with respect to the l_∞ norm, since

$$\lim_{k\to\infty} (2 + \frac{1}{k}) = 2, \quad \lim_{k\to\infty} \frac{3}{k^2} = 0, \quad \lim_{k\to\infty} e^{-k}\sin k = 0.$$

Question

Does the given sequence converge to x **with respect to the** l_2 **norm**?

Thm 7.7 (The Equivalence of Vector Norms)

For each $x \in \mathbb{R}^n$,

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

In this case, we say that the l_∞ and l_2 norms are equivalent.

pf: For any $x = [x_1, \cdots, x_n]^T \in \mathbb{R}^n$, let $|x_{i_0}| = \max_{1 \leq i \leq n} |x_i| = \|x\|_\infty$.

Then we see that

1. $\|x\|_\infty = |x_{i_0}| = \sqrt{x_{i_0}^2} \leq \sqrt{x_1^2 + \cdots + x_n^2} = \|x\|_2.$
2. $\|x\|_2 \leq \left(\sum_{i=1}^{n} x_{i_0}^2 \right)^{1/2} = \left(n\|x\|_\infty^2 \right)^{1/2} = \sqrt{n}\|x\|_\infty.$

So, these prove the desired inequalities.

Example 4, p. 437

Show that the sequence of vectors in Example 3

$$x^{(k)} = [1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k} \sin k]^T \in \mathbb{R}^4$$

converges to $x = [1, 2, 0, 0]^T \in \mathbb{R}^4$ with respect to the l_2 norm.

pf: In Example 3, we know that $\lim_{k \to \infty} \|x^{(k)} - x\|_\infty = 0$. So, for any $\epsilon > 0$, $\exists\, N_0 \in \mathbb{N}$ s.t.

$$\|x^{(k)} - x\|_\infty < \frac{\epsilon}{2} \qquad \forall\, k \geq N_0,$$

and furthermore, it follows from Thm 7.7 that

$$\|x^{(k)} - x\|_2 \leq \sqrt{4} \cdot \|x^{(k)} - x\|_\infty < 2 \cdot (\frac{\epsilon}{2}) = \epsilon$$

whenever $k \geq N_0$. Hence, this completes the proof.

Remarks

- Any two vector norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{R}^n$ are equivalent, i.e., $\exists\, c_1 > 0$ and $c_2 > 0$ s.t.

$$c_1\|x\|' \leq \|x\| \leq c_2\|x\|' \qquad \forall\, x \in \mathbb{R}^n.$$

- A seq. $\{x^{(k)}\}_{k=1}^{\infty}$ converges to the limit $x \in \mathbb{R}^n$ **with respect to the norm** $\|\cdot\|$ $\Longleftrightarrow$ a seq. $\{x^{(k)}\}_{k=1}^{\infty}$ converges to the limit $x \in \mathbb{R}^n$ **with respect to the norm** $\|\cdot\|'$. (向量序列的收斂性與範數無關!)
- For any $x \in \mathbb{R}^n$, the relations between l_1, l_2 and l_∞ norms are

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2,$$
$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty.$$

Def 7.8 (矩陣的範數)

A matrix norm on $\mathbb{R}^{n\times n}$ is a function $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ satisfying for all $A, B \in \mathbb{R}^{n\times n}$ and all $\alpha \in \mathbb{R}$:

(i) $\|A\| \geq 0$;

(ii) $\|A\| = 0 \iff A = 0$ (zero matrix);

(iii) $\|\alpha A\| = |\alpha| \, \|A\|$;

(iv) $\|A + B\| \leq \|A\| + \|B\|$;

(v) $\|AB\| \leq \|A\| \, \|B\|$.

Definition (Distances of Two Matrices)

If $A, B \in \mathbb{R}^{n\times n}$, the number $\|A - B\|$ is called the distance between A and B with respect to the matrix norm $\|\cdot\|$.

Thm 7.9 (自然矩陣範數)

If $\|\cdot\|$ is a **vector norm on** $\mathbb{R}^n$, then

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

is a matrix norm on $\mathbb{R}^{n\times n}$. (See Exercise 13 for the proof.)

pf: Only prove that $\|AB\| \leq \|A\|\,\|B\|$ for any $A, B \in \mathbb{R}^{n\times n}$ here. For any unit vector $x \in \mathbb{R}^n$, we have

$$\|A(Bx)\| = \|Bx\| \cdot \left\|A\left(\frac{Bx}{\|Bx\|}\right)\right\| \leq \|A\| \cdot \|Bx\|.$$

Thus, we conclude that

$$\begin{aligned} \|AB\| &= \max_{\|x\|=1} \|(AB)x\| = \max_{\|x\|=1} \|A(Bx)\| \\ &\leq \max_{\|x\|=1} (\|A\|\,\|Bx\|) = \|A\| \cdot \max_{\|x\|=1} \|Bx\| = \|A\|\,\|B\|. \end{aligned}$$

Remarks

- Matrix norms defined by **vector norms** are called the natural (or induced) matrix norm associated with the vector norm.
- Since $x = \frac{z}{\|z\|}$ is a unit vector for $z \neq 0$, Thm 7.9 can be rewritten as

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

Cor 7.10

For any $A \in \mathbb{R}^{n\times n}$, $0 \neq z \in \mathbb{R}^n$ and any natural norm $\|\cdot\|$,

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

Some Natural Matrix Norms

1. $\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{z\neq 0} \frac{\|Az\|_\infty}{\|z\|_\infty}$. (the l_∞ norm)
2. $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{z\neq 0} \frac{\|Az\|_2}{\|z\|_2}$. (the l_2 norm)
3. $\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{z\neq 0} \frac{\|Az\|_1}{\|z\|_1}$. (the l_1 norm)

Thm 7.11 (矩陣 ∞-範數的計算公式)

If $A = [a_{ij}] \in \mathbb{R}^{n\times n}$, then

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|. \quad (|A| = [\,|a_{ij}|\,] \text{ 的最大列和})$$

pf: The proof is separated into two parts.

(1) Assume $\|A\|_\infty = \|Ax\|_\infty$ for some $x \in \mathbb{R}^n$ with $\|x\|_\infty = \max_{1\le j\le n} |x_j| = 1$. Then we have

$$\begin{aligned}
\|A\|_\infty = \|Ax\|_\infty = \max_{1\le i\le n} |(Ax)_i| &= \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \\
&\le \max_{1\le i\le n} \sum_{j=1}^{n} \left(|a_{ij}| \cdot \|x\|_\infty \right) = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|,
\end{aligned}$$

since $\|x\|_\infty = 1$.

(2) Let $y = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^n$, where each component

$$y_j = \begin{cases} 1, & \text{if } a_{ij} \geq 0, \\ -1, & \text{if } a_{ij} < 0. \end{cases}$$

Then $\|y\|_\infty = 1$ and $a_{ij}y_j = |a_{ij}|$ for all i, j. So, we get

$$\|Ay\|_\infty = \max_{1 \leq i \leq n} |(Ay)_i| = \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} a_{ij}y_j \right| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Furthermore, it follows that

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

From the parts (1) and (2) $\Longrightarrow$ these complete the proof.

Exercise 6, p. 442 (矩陣 1-範數的計算公式)

If $A = [a_{ij}] \in \mathbb{R}^{n\times n}$, then

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|. \quad (|A| = [\,|a_{ij}|\,] \text{ 的最大行和})$$

Example 5, p. 441

For the 3×3 matrix

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

it follows that

$$\|A\|_\infty = \max_{i=1,2,3} \sum_{j=1}^{3} |a_{ij}| = \max\{\mathbf{4}, \mathbf{4}, \mathbf{7}\} = \mathbf{7},$$

$$\|A\|_1 = \max_{j=1,2,3} \sum_{i=1}^{3} |a_{ij}| = \max\{\mathbf{6}, \mathbf{6}, \mathbf{3}\} = \mathbf{6}.$$

Section 7.2
Eigenvalues and Eigenvectors

Def 7.12 (特徵多項式)

The characteristic polynomial of $A \in \mathbb{R}^{n\times n}$ is defined by

$$p(\lambda) = \det(A - \lambda I),$$

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

Note: The characteristic poly. p is an nth-degree poly. with real coefficients. So, it has at most n distinct zeros in $\mathbb{C}$.

Def 7.13 (特徵值與特徵向量)

Let $p(\lambda)$ be the characteristic poly. of $A \in \mathbb{R}^{n\times n}$.

- The number $\lambda \in \mathbb{C}$ is called an eigenvalue (or characteristic value) of A if $p(\lambda) = 0$.
- The spectrum (譜) of A, denoted by $\sigma(A)$, is the set of all eigenvalues of A.
- If $\exists 0 \neq x \in \mathbb{R}^n$ s.t. $Ax = \lambda x$ or $(A - \lambda I)x = 0$ for $\lambda \in \sigma(A)$, then x is called an eigenvector (pr characteristic vector) of A corresponding to λ.

Def 7.14

The **spectral radius (譜半徑)** of $A \in \mathbb{R}^{n\times n}$ is defined by

$$\rho(A) = \max\{|\lambda| \mid \lambda \in \sigma(A)\}.$$

(For complex $\lambda = \alpha + \beta i$, we define $|\lambda| = \sqrt{\alpha^2 + \beta^2}$.)

Thm 7.15 (矩陣 2-範數的計算公式)

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

(i) $\|A\|_2 = \sqrt{\rho(A^T A)}$.

(ii) $\rho(A) \le \|A\|$ for any natural matrix norm $\|\cdot\|$.

Review from Linear Algebra

Let $B = A^TA$ with $A \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$.

- $B^T = (A^TA)^T = A^T(A^T)^T = A^TA = B$, i.e., B is symmetric.
- For any $\lambda \in \sigma(B)$, $\lambda \geq 0$.
- B is orthogonally diagonalizable, i.e., $\exists$ orthog. $Q \in \mathbb{R}^{n\times n}$ s.t.

$$Q^TBQ = Q^T(A^TA)Q = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \equiv D,$$

  where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \geq 0$.
- Since $\|v\|_2^2 = v^Tv$, we have

$$\|Ax\|_2^2 = (Ax)^T(Ax) = x^T(A^TA)x = x^TBx$$

  for any $x \in \mathbb{R}^n$.

Proof of Thm 7.15 (1/2)

(i) Since A^TA is symmetric, $\exists$ orthogonal $Q \in \mathbb{R}^{n\times n}$ s.t.

$$Q^T(A^TA)Q = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \equiv D,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \geq 0$. Hence,

$$\begin{aligned}\|A\|_2^2 &= \max_{\|x\|_2=1} \|Ax\|_2^2 = \max_{\|x\|_2=1} x^T(A^TA)x \\ &= \max_{\|x\|_2=1} x^TQDQ^Tx = \max_{\|y\|_2=1} y^TDy,\end{aligned}$$

where we let $y = Q^Tx$. So, $\|A\|_2^2 \leq \lambda_1$. Moreover, the maximum value of y^TDy is achieved at the vector $y^* = [1, 0, \cdots, 0]^T \in \mathbb{R}^n$ and thus $\|A\|_2^2 = \lambda_1$ or $\|A\|_2 = \sqrt{\lambda_1} = \sqrt{\rho(A^TA)}$.

Proof of Thm 7.15 (2/2)

(ii) Let $A \in \mathbb{R}^{n\times n}$ and $\|\cdot\|$ be any natural norm. For each $\lambda \in \sigma(A)$, $\exists\, 0 \neq x \in \mathbb{R}^n$ s.t.

$$Ax = \lambda x \qquad \text{with } \|x\| = 1.$$

Hence, we know that

$$|\lambda| = |\lambda| \cdot \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\|\,\|x\| = \|A\|.$$

So, the spectral radius of A satisfies $\rho(A) \leq \|A\|$.

Rematks

- If $A^T = A \in \mathbb{R}^{n\times n}$, then $\|A\|_2 = \rho(A)$.
- For any $A \in \mathbb{R}^{n\times n}$ and any $\epsilon > 0$, $\exists$ a natural norm $\|\cdot\|_\epsilon$ s.t.

$$\rho(A) < \|A\|_\epsilon < \rho(A) + \epsilon.$$

Def 7.16 (收斂矩陣的定義)

We day that a matrix $A \in \mathbb{R}^{n\times n}$ is **convergent** if

$$\lim_{k\to\infty} (A^k)_{ij} = 0, \quad \text{for } i,j = 1,2,\ldots,n.$$

Example 4, p. 448

The 2×2 matrix $A = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$ is a convergent matrix, since we have

$$A^k = \begin{bmatrix} (\frac{1}{2})^k & 0 \\ \frac{k}{2^{k+1}} & (\frac{1}{2})^k \end{bmatrix} \qquad \forall\, k \geq 1$$

by mathematical induction, $\lim_{k\to\infty} (\frac{1}{2})^k = 0$ and $\lim_{k\to\infty} \frac{k}{2^{k+1}} = 0.$

Thm 7.17 (收斂矩陣的等價條件)

Let $A \in \mathbb{R}^{n\times n}$. The following statements are equivalent.

(i) A is a convergent matrix.

(ii) $\lim_{n\to\infty} \|A^n\| = 0$ for **some** natural norm.

(iii) $\lim_{n\to\infty} \|A^n\| = 0$ for **all** natural norms.

(iv) $\rho(A) < 1$.

(v) $\lim_{n\to\infty} A^n x = 0$ for every $x \in \mathbb{R}^n$.

Section 7.3
The Jacobi and Gauss-Siedel Iterative Techniques

Basic Idea

From the ith eq. of a linear system $Ax = b$

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ii}x_i + \cdots + a_{in}x_n = b_i$$

for solving the ith component x_i, we get

$$x_i = \frac{1}{a_{ii}} \left[\sum_{\substack{j=1 \\ j \neq i}}^{n} (-a_{ij}x_j) + b_i \right],$$

provided that $a_{ii} \neq 0$ for $i = 1, 2, \ldots, n$.

Component Form (Jacobi 法的分量形式)

For each $k \geq 1$, we nay consider the Jacobi iterative method:

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[\sum_{\substack{j=1 \\ j \neq i}}^{n} (-a_{ij} x_j^{(k-1)}) + b_i \right] \quad \text{for } i = 1, \ldots, n, \qquad (1)$$

where an initial approx. $x^{(0)} = [x_1^{(0)}, \cdots, x_n^{(0)}]^T \in \mathbb{R}^n$ is given.

Example 1, p. 451

The following linear system

$$
\begin{aligned}
\mathbf{10}x_1 - x_2 + 2x_3 &= 6 \\
-x_1 + \mathbf{11}x_2 - x_3 + 3x_4 &= 25 \\
2x_1 - x_2 + \mathbf{10}x_3 - x_4 &= -11 \\
3x_2 - x_3 + \mathbf{8}x_4 &= 15
\end{aligned}
$$

has a unique solution $x = [1, 2, -1, 1]^T \in \mathbb{R}^4$. Use Jacobi's iterative technique to find an approx. $x^{(k)}$ to x starting with $x^{(0)} = [0, 0, 0, 0]^T \in \mathbb{R}^4$ until

$$
\frac{\|x^{(k)} - x^{(k-1)}\|_\infty}{\|x^{(k)}\|_\infty} < 10^{-3}.
$$

Solution (1/2)

The given linear system ca be rewritten as

$$\begin{aligned}
x_1 &= \frac{1}{10}x_2 - \frac{1}{5}x_3 + \frac{3}{5} \\
x_2 &= \frac{1}{11}x_1 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11} \\
x_3 &= \frac{-1}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}x_4 - \frac{11}{10} \\
x_4 &= \frac{-3}{8}x_2 + \frac{1}{8}x_3 + \frac{15}{8}.
\end{aligned}$$

Solution (2/2)

For each $k \geq 1$, we apply the Jacbi's method:

$$\begin{aligned}
x_1^{(k)} &= \frac{1}{10}x_2^{(k-1)} - \frac{1}{5}x_3^{(k-1)} + \frac{3}{5} \\
x_2^{(k)} &= \frac{1}{11}x_1^{(k-1)} + \frac{1}{11}x_3^{(k-1)} - \frac{3}{11}x_4^{(k-1)} + \frac{25}{11} \\
x_3^{(k)} &= \frac{-1}{5}x_1^{(k-1)} + \frac{1}{10}x_2^{(k-1)} + \frac{1}{10}x_4^{(k-1)} - \frac{11}{10} \\
x_4^{(k)} &= \frac{-3}{8}x_2^{(k-1)} + \frac{1}{8}x_3^{(k-1)} + \frac{15}{8}
\end{aligned}$$

with the initial guess $x^{(0)} = [0, 0, 0, 0]^T \in \mathbb{R}^4$.

Numerical Results

After **10** iterations of Jacobi method, we have

$$\frac{\|x^{(10)} - x^{(9)}\|_\infty}{\|x^{(10)}\|_\infty} = \frac{8.0 \times 10^{-4}}{1.9998} = \mathbf{4.0 \times 10^{-4}} < 10^{-3}.$$

In fact, the absolute error is $\|x^{(10)} - x\|_\infty = \mathbf{2 \times 10^{-4}}$.

Equivalent Matrix-Vector Forms

- As in Chapter 2, every root-finding problem $f(x) = 0$ is converted into its equivalent fixed-point form

$$x = g(x), \quad x \in I = [a, b]$$

, for some **differentiable** function g.

- Similarly, we also try to covert the original linear system $Ax = b$ into its equivalent matrix-vector form

$$x = Tx + c, \quad x \in \mathbb{R}^n,$$

where $T \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$ are fixed.

- For $k = 1, 2, \ldots$, compute

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

with an initial approx. $x^{(0)} \in \mathbb{R}^n$ to the unique sol. x.

A Useful Split of A

The iterative techniques for solving $Ax = b$ will be derived by first splitting A into its diagonal and off-diagonal parts, i.e.,

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ -a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

$$\equiv D - L - U. \tag{2}$$

Matlab Commands

$D = \texttt{diag}(\texttt{diag}(A));$ $L = \texttt{tril}(-A, -1);$ $U = \texttt{triu}(-A, 1);$

- From the splitting of A as in (2), the linear system $Ax = b$ is transformed to

$$(D - L - U)x = b \Leftrightarrow Dx = (L + U)x + b \Leftrightarrow x = T_j x + c_j,$$

where $T_j \equiv D^{-1}(L + U)$ and $c_j \equiv D^{-1}b$.

- It is easily seen that the component form of Jacobi method

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[\sum_{\substack{j=1 \\ j \neq i}}^{n} \left(-a_{ij} x_j^{(k-1)}\right) + b_i \right] \quad \text{for } i = 1, \ldots, n.$$

is equivalent to the following matrix-vector form

$$x^{(k)} = T_j x^{(k-1)} + c_j \qquad \forall\, k \geq 1.$$

Example 2, p. 453

The 4×4 linear system in Example 1 can be rewritten in the form

$$\begin{aligned}
x_1 &= \frac{1}{10}x_2 - \frac{1}{5}x_3 + \frac{3}{5} \\
x_2 &= \frac{1}{11}x_1 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11} \\
x_3 &= \frac{-1}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}x_4 - \frac{11}{10} \\
x_4 &= \frac{-3}{8}x_2 + \frac{1}{8}x_3 + \frac{15}{8}.
\end{aligned}$$

So, the unique sol. $x \in \mathbb{R}^4$ satisfies $x = T_j x + c_j$ with

$$T_j = \begin{bmatrix} 0 & \frac{1}{10} & \frac{-1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & \frac{-3}{11} \\ \frac{-1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & \frac{-3}{8} & \frac{1}{8} & 0 \end{bmatrix} \quad \text{and} \quad c_j = \begin{bmatrix} \frac{3}{5} \\ \frac{25}{11} \\ \frac{-11}{10} \\ \frac{15}{8} \end{bmatrix}.$$

Algorithm 7.1: Jacobi Method

INPUT dim. n; $A = [a_{ij}] \in \mathbb{R}^{n\times n}$; $b \in \mathbb{R}^n$; $X0 = x^{(0)} \in \mathbb{R}^n$; tol. TOL; max. no. of iter. N_0.

OUTPUT an approx. sol. $x_1, x_2, \ldots, x_n$ to $Ax = b$.

Step 1 Set $k = 1$.

Step 2 While ($k \leq N_0$) do **Steps 3–6**

Step 3 For $i = 1, \ldots, n$ set

$$x_i = \frac{1}{a_{ii}} \left[\sum_{\substack{j=1 \\ j\neq i}}^{n} (-a_{ij}X0_j) + b_i \right].$$

Step 4 If $\|x - X0\| < TOL$ then OUTPUT($x_1, \cdots, x_n$); **STOP**.

Step 5 Set $k = k + 1$.

Step 6 Set $X0 = x$.

Step 7 OUTPUT('Maximum number of iterations exceeded'); **STOP**.

Comments on Algorithm 7.1

- If some $a_{ii} = 0$ and A is **nonsingular**, choose $p \neq i$ s.t.

  $$|a_{pi}| \text{ is as large as possible,}$$

  and then perform $(E_p) \leftrightarrow (E_i)$ to ensure that no $a_{ii} = 0$ before applying the Jacobi method.

- In Step 4, a better stopping criterion should be

  $$\frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k)}\|} < TOL,$$

  where the vector norm $\|\cdot\|$ is the l_1, l_2 or l_∞ norm.

Jacobi Method v.s. Gauss-Seidel Metod

- For the Jacobi's method, the ith component $x_i^{(k)}$ of $x^{(k)}$ is determined by $x_1^{(k-1)}, \cdots, x_{i-1}^{(k-1)}$ and $x_{i+1}^{(k-1)}, \cdots, x_n^{(k-1)}$.
- At the kth step of Gauss-Seidel method, the ith component $x_i^{(k)}$ is computed by $x_1^{(k)}, \cdots, x_{i-1}^{(k)}$ and $x_{i+1}^{(k-1)}, \cdots, x_n^{(k-1)}$.
- Notice that the recently computed values of $x_1^{(k)}, \cdots, x_{i-1}^{(k)}$ are better approxs. to x than the values of $x_1^{(k-1)}, \cdots, x_{i-1}^{(k-1)}$.

Component Form of Gauss-Seidel Method

For each $k \geq 1$, the ith component of $x^{(k)}$ is determined by

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[\sum_{j=1}^{i-1} (-a_{ij} x_j^{(k)}) + \sum_{j=i+1}^{n} (-a_{ij} x_j^{(k-1)}) + b_i \right]$$
$$= \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} (a_{ij} x_j^{(k)}) - \sum_{j=i+1}^{n} (a_{ij} x_j^{(k-1)}) + b_i \right], \qquad (3)$$

where an initial vector $x^{(0)}$ is given and $i = 1, 2, \ldots, n$.

Algorithm 7.2: Gauss-Seidel Method

INPUT dim. n; $A = [a_{ij}] \in \mathbb{R}^{n\times n}$; $b \in \mathbb{R}^n$; $X0 = x^{(0)} \in \mathbb{R}^n$; tol. TOL; max. no. of iter. N_0.

OUTPUT an approx. sol. $x_1, x_2, \ldots, x_n$ to $Ax = b$.

Step 1 Set $k = 1$.

Step 2 While $(k \leq N_0)$ do **Steps 3–6**

Step 3 For $i = 1, \ldots, n$ set

$$x_i = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} (a_{ij} x_j) - \sum_{j=i+1}^{n} (a_{ij} X0_j) + b_i \right].$$

Step 4 If $\|x - X0\| < TOL$ then OUTPUT$(x_1, \cdots, x_n)$; **STOP**.

Step 5 Set $k = k + 1$.

Step 6 Set $X0 = x$.

Step 7 OUTPUT('Maximum number of iterations exceeded'); **STOP**.

Example 3, p. 455

The following linear system

$$\begin{aligned}
\mathbf{10}x_1 - x_2 + 2x_3 &= 6 \\
-x_1 + \mathbf{11}x_2 - x_3 + 3x_4 &= 25 \\
2x_1 - x_2 + \mathbf{10}x_3 - x_4 &= -11 \\
3x_2 - x_3 + \mathbf{8}x_4 &= 15
\end{aligned}$$

has a unique solution $x = [1, 2, -1, 1]^T \in \mathbb{R}^4$. Use **Gauss-Seidel method** to find an approx. $x^{(k)}$ to x starting with $x^{(0)} = [0, 0, 0, 0]^T \in \mathbb{R}^4$ until

$$\frac{\|x^{(k)} - x^{(k-1)}\|_\infty}{\|x^{(k)}\|_\infty} < 10^{-3}.$$

Solution

For each $k \geq 1$, we apply the Gauss-Seidel method:

$$
\begin{aligned}
x_1^{(k)} &= \frac{1}{10} x_2^{(k-1)} - \frac{1}{5} x_3^{(k-1)} + \frac{3}{5} \\
x_2^{(k)} &= \frac{1}{11} x_1^{(k)} + \frac{1}{11} x_3^{(k-1)} - \frac{3}{11} x_4^{(k-1)} + \frac{25}{11} \\
x_3^{(k)} &= \frac{-1}{5} x_1^{(k)} + \frac{1}{10} x_2^{(k)} + \frac{1}{10} x_4^{(k-1)} - \frac{11}{10} \\
x_4^{(k)} &= \frac{-3}{8} x_2^{(k)} + \frac{1}{8} x_3^{(k)} + \frac{15}{8}
\end{aligned}
$$

with the initial guess $x^{(0)} = [0, 0, 0, 0]^T \in \mathbb{R}^4$.

After **5** iterations of Gauss-Seidel method, we have

$$\frac{\|x^{(5)} - x^{(4)}\|_\infty}{\|x^{(5)}\|_\infty} = \frac{8.0 \times 10^{-4}}{2.000} = \mathbf{4.0 \times 10^{-4}} < 10^{-3}.$$

In fact, the absolute error is $\|x^{(5)} - x\|_\infty = \mathbf{1.0 \times 10^{-4}}$. The numerical results are shown in the following table.

Matrix-Vector Form of Gauss-Seidel Method (1/2)

- From the component form as in (3)

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} (a_{ij} x_j^{(k)}) + \sum_{j=i+1}^{n} (-a_{ij} x_j^{(k-1)}) + b_i \right],$$

we immediately obtain

$$a_{i1}x_1^{(k)} + \cdots + a_{ii}x_i^{(k)} = -a_{i,i+1}x_{i+1}^{(k-1)} - \cdots - a_{in}x_n^{(k-1)} + b_i$$

for each $i = 1, 2, \ldots, n$.

- Thus we have following matrix form for Gauss-Seidel method:

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & \cdots & a_{n,n-1} & a_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} = \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ \vdots \\ x_n^{(k-1)} \end{bmatrix} + b.$$

Matrix-Vector Form of Gauss-Seidel Method (2/2)

- For each $k \geq 1$, the above matrix equation can be rewritten as

$$
\begin{aligned}
(D - L)x^{(k)} &= Ux^{(k-1)} + b \\
\Longleftrightarrow x^{(k)} &= (D - L)^{-1}Ux^{(k-1)} + (D - L)^{-1}b \\
\Longleftrightarrow x^{(k)} &= T_g x^{(k-1)} + c_g,
\end{aligned}
$$

where $T_g \equiv (D - L)^{-1}U$ and $c_g \equiv (D - L)^{-1}b$.

- Recall the Jacobi method given by

$$x^{(k)} = T_j x^{(k-1)} + c_j, \quad k = 1, 2, \ldots,$$

where $T_j = D^{-1}(L + U)$ and $c_j = D^{-1}b$.

Some Questions

1. When does a general iteration of the form

$$x^{(k)} = Tx^{(k-1)} + c, \quad k = 1, 2, \ldots$$

   converge to a solution $x \in \mathbb{R}^n$ of the matrix equation $x = Tx + c$?
2. What is the rate of convergence for this iterative method?
3. Does the Gauss-Seidel method **always converge faster than** the Jacobi method?

Lemma 7.18

If $T \in \mathbb{R}^{n\times n}$ satisfies $\rho(T) < 1$, the $(I - T)^{-1}$ exists and

$$(I - T)^{-1} = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j$$

with $T^0 \equiv I$ being defined conventionally.

Proof of Lemma 7.18

- If $\lambda \in \sigma(T)$, then $\exists 0 \neq x \in \mathbb{R}^n$ s.t.

$$Tx = \lambda x \quad \text{or} \quad (I - T)x = (1 - \lambda)x.$$

  So, $1 - \lambda \in \sigma(I - T)$.

- Because $\rho(T) < 1$, $|\lambda| \leq \rho(T) < 1$. This means that $I - T$ does not have any zero eigenvalues and hence $(I - T)^{-1}$ exists.

- Let $S_m = \sum_{j=0}^{m} T^j$. Then we have

$$(I - T)S_m = \sum_{j=0}^{m} T^j - \sum_{j=0}^{m} T^{j+1} = I - T^{m+1}.$$

  Since $\rho(T) < 1$, T is convergent, i.e., $\lim_{m\to\infty} T^m = \mathbf{0}$ by Thm 7.17. Hence $(I - T)^{-1} = \lim_{m\to\infty} S_m = \sum_{j=0}^{\infty} T^j$.

Thm 7.19 (廣義迭代法收斂性的充要條件)

For any $x^{(0)} \in \mathbb{R}^n$, the sequence $\{x^{(k)}\}_{k=0}^{\infty}$ defined by

$$x^{(k)} = Tx^{(k-1)} + c \qquad \forall\, k \geq 1$$

converges to the **unique solution** of $x = Tx + c \Longleftrightarrow \rho(T) < 1$.

pf: The proof is illustrated as follows.

$(\Leftarrow)$ Suppose that $\rho(T) < 1$. By induction $\Longrightarrow$

$$\begin{aligned} x^{(k)} &= Tx^{(k-1)} + c \\ &= T(Tx^{(k-2)} + c) + c = T^2 x^{(k-2)} + (T + I)c \\ &\vdots \\ &= T^k x^{(0)} + (T^{k-1} + \cdots + T + I)c. \end{aligned}$$

Since $\rho(T) < 1$, $\lim_{k\to\infty} T^k x^{(0)} = 0$ by Thm 7.17. Thus, it follows from Lemma 7.18 that

$$x \equiv \lim_{k\to\infty} x^{(k)} = 0 + \left(\sum_{j=0}^{\infty} T^j \right) c = (I - T)^{-1} c.$$

Hence, the limit $x \in \mathbb{R}^n$ is the unique solution of the equation

$$x = (I - T)^{-1} c \Longleftrightarrow (I - T)x = c \Longleftrightarrow x = Tx + c.$$

$(\Rightarrow)$ Assume that $\lim_{k\to\infty} x^{(k)} = x$ for any initial vector $x^{(0)}$, where $x \in \mathbb{R}^n$ is the unique sol. of $x = Tx + c$. Now, we want to claim that

$$\rho(T) < 1 \Longleftrightarrow \lim_{k\to\infty} T^k z = 0 \quad \forall\, z \in \mathbb{R}^n$$

by applying Thm 7.17.

For any $z \in \mathbb{R}^n$, let $x^{(0)} = x - z$. Then by induction $\Longrightarrow$

$$\begin{aligned} x - x^{(k)} &= (Tx + c) - (Tx^{(k-1)} + c) \\ &= T(x - x^{(k-1)}) = \cdots = T^k(x - x^{(0)}) \\ &= T^k z, \text{ since } z = x - x^{(0)}. \end{aligned}$$

So, it follows from the assumption that

$$\lim_{k\to\infty} T^k z = x - \lim_{k\to\infty} x^{(k)} = x - x = 0.$$

Since $z \in \mathbb{R}^n$ is given arbitrarily, we have $\rho(T) < 1$.

Cor 7.20 (廣義迭代法的誤差上界與收斂比率)

If $\|T\| < 1$ for any natural norm, the seq. $\{x^{(k)}\}_{k=0}^{\infty}$ defined by

$$x^{(k)} = Tx^{(k-1)} + c \qquad \forall\, k \geq 1$$

converges to the unique sol. of $x = Tx + c$, for any $x^{(0)} \in \mathbb{R}^n$. Moreover, we have

(i) $\|x^{(k)} - x\| \leq \|T\|^k \|x^{(0)} - x\| \qquad \forall\, k \geq 1$;

(ii) $\|x^{(k)} - x\| \leq \dfrac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\| \qquad \forall\, k \geq 1$.

Note: See Exercise 13 for the proof.

Thm 7.21 (Jacobi 法和 Gaussl-Seidel 法收斂的充分條件)

If $A \in \mathbb{R}^{n\times n}$ is **strictly diagonally dominant**, then both Jacobi and Gauss-Seidel methods converge to the unique sol. x of $Ax = b$, for any choice of $x^{(0)} \in \mathbb{R}^n$.

Thm 7.22 (Stein-Rosenberg)

If $a_{ij} \leq 0$ for $i \neq j$ and $a_{ii} > 0$ for $i = 1, 2, \ldots, n$, then **one and only one** of the following statements holds:

(i) $0 \leq \rho(T_g) < \rho(T_j) < 1$; (ii) $1 < \rho(T_j) < \rho(T_g)$;

(iii) $\rho(T_j) = \rho(T_g) = 0$; (iv) $\rho(T_j) = \rho(T_g) = 1$.

Note: 條件 (i) 或 (iii) 決定兩算法的收斂性，但條件 (ii) 或 (iv) 決定兩算法的發散性。

Section 7.4
Relaxation Techniques for Solving Linear Systems

Def 7.23 (殘餘向量或剩餘向量)

If $\tilde{x} \in \mathbb{R}^n$ is an approximation to the solution of a linear system $Ax = b$, then $r = b - A\tilde{x}$ is called the **residual vector** for $\tilde{x}$ with respect to the system.

Remarks

- In the Jacbi or Gauss-Seidel methods, a residual vector is associated with each calculation of an approx. component to the solution vector.
- The true objective is to generate a sequence of approximations that will cause the residual vectors to converge rapidly to zero.

- For each $i = 1, 2, \ldots, n$, let

$$\mathbf{r}_i^{(k)} = [r_{1i}^{(k)}, r_{2i}^{(k)}, \cdots, r_{ni}^{(k)}]^T \in \mathbb{R}^n$$

denote the residual vector for Gauss-Seidel method corresp. to the approx. sol. vector $\mathbf{x}_i^{(k)}$ defined by

$$\mathbf{x}_i^{(k)} = [x_1^{(k)}, x_2^{(k)}, \cdots, x_{i-1}^{(k)}, x_i^{(k-1)}, \cdots, x_n^{(k-1)}]^T \in \mathbb{R}^n.$$

- The ith component of the residual $\mathbf{r}_i^{(k)} = b - A\mathbf{x}_i^{(k)}$ is given by

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} - a_{ii}x_i^{(k-1)}.$$

- From the above equation for $r_{ii}^{(k)}$, we obtain

$$r_{ii}^{(k)} + a_{ii}x_i^{(k-1)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} \qquad (4)$$

for each $i = 1, 2, \ldots, n$.

- Note that, in the Gauss-Seidel method, we choose $x_i^{(k)}$ to be

$$x_i^{(k)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right].$$

So, we further have

$$a_{ii}x_i^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} \qquad (5)$$

for each $i = 1, 2, \ldots, n$.

- From Eqs. (4) and (5) $\Longrightarrow a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}x_i^{(k)}$.

Alternative Characterization for Gauss-Seidel Method

For each $k \geq 1$, choose the ith component of $x^{(k)}$ satisfying

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}, \quad i = 1, 2, \cdots, n.$$

- Another characterization for the Gauss-Seidel is given by

The 2nd Characterization of Gauss-Seidel Method

For each $k \geq 1$, choose $x_i^{(k)}$ satisfying

$$r_{i,i+1}^{(k)} = 0, \quad i = 1, 2, \ldots, n.$$

Relaxation Methods (鬆弛法)

For each $k \geq 1$, choose the ith component of $x^{(k)}$ satisfying

$$x_i^{(k)} = x_i^{(k-1)} + \omega \cdot \frac{r_{ii}^{(k)}}{a_{ii}}, \quad i = 1, 2, \ldots, n, \tag{6}$$

where $\omega > 0$ is a parameter. Two types of relaxation methods:

1. $0 < \omega < 1$: **under-relaxation** methods. (低鬆弛法)
2. $\omega > 1$: **over-relaxation** methods. (過度鬆弛法)

The SOR Methods

- The over-relaxation methods are also called the **Successive Over-Relaxation (SOR)** methods.
- They are often used to **accelerate** the convergence of the Gauss-Seidel method.
- These methods are particularly useful for solving the linear systems that occur in the numerical solution of certain PDEs.

Review for $r_{ii}^{(k)}$

It has been shown previously that

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k-1)}$$

for each $i = 1, 2, \ldots, n$.

Component Form of SOR Method

Combining (6) with above eq. for $r_{ii}^{(k)}$, we see that

$$\begin{aligned} x_i^{(k)} &= x_i^{(k-1)} + (\omega/a_{ii}) \cdot r_{ii}^{(k)} \\ &= x_i^{(k-1)} + \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k-1)} \right] \\ &= (1-\omega) x_i^{(k-1)} + \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right], \quad (7) \end{aligned}$$

for each $i = 1, 2, \ldots, n$.

Matrix-Vector Form of SOR Method

- From (7), the component form for SOR can be rewritten as

$$a_{ii}x_i^{(k)} - \omega \sum_{j=1}^{i-1}(-a_{ij}x_j^{(k)}) = (1-\omega)a_{ii}x_i^{(k-1)} + \omega \sum_{j=i+1}^{n}(-a_{ij}x_j^{(k-1)}) + \omega b_i$$

  for each $i = 1, 2, \ldots, n$.
- This is equivalent to the following matrix-vector form

$$(D - \omega L)x^{(k)} = [(1-\omega)D + \omega U]x^{(k-1)} + \omega b$$
$$\Longleftrightarrow x^{(k)} = T_\omega x^{(k-1)} + c_\omega,$$

  where $T_\omega \equiv (D - \omega L)^{-1}[(1-\omega)D + \omega U] \in \mathbb{R}^{n\times n}$ and the parameter-dependent vector $c_\omega \equiv \omega(D - \omega L)^{-1} b \in \mathbb{R}^n$.

Algorithm 7.3: SOR

INPUT dim. n; $A = [a_{ij}] \in \mathbb{R}^{n\times n}$; $b \in \mathbb{R}^n$; $X0 = x^{(0)} \in \mathbb{R}^n$; parameter ω; tol. TOL; max. no. of iter. N_0.

OUTPUT an approx. sol. $x_1, x_2, \ldots, x_n$ to $Ax = b$.

Step 1 Set $k = 1$.

Step 2 While $(k \le N_0)$ do **Steps 3–6**

Step 3 For $i = 1, \ldots, n$ set

$$x_i = (1-\omega)X0_i + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1}(a_{ij}x_j) - \sum_{j=i+1}^{n}(a_{ij}X0_j)\right].$$

Step 4 If $\|x - X0\| < TOL$ then OUTPUT$(x_1, \cdots, x_n)$; **STOP**.

Step 5 Set $k = k + 1$.

Step 6 Set $X0 = x$.

Step 7 OUTPUT('Maximum number of iterations exceeded'); **STOP**.

Example 1, p. 464

The 3×3 linear system

$$\begin{aligned} 4x_1 + 3x_2 &= 24, \\ 3x_1 + 4x_2 - x_3 &= 30, \\ -x_2 + 4x_3 &= -24 \end{aligned}$$

has the unique sol. $x = [3, 4, -5]^T \in \mathbb{R}^3$. Use the Gauss-Seidel method and SOR with $\omega = \mathbf{1.25}$ to compute an approx. sol. to x using $x^{(0)} = [1, 1, 1]^T \in \mathbb{R}^3$ for both methods.

(1) Applying the Gauss-Seidel method, we have for each $k \geq 1$,

$$\begin{aligned} x_1^{(k)} &= -0.75x_2^{(k-1)} + 6, \\ x_2^{(k)} &= -0.75x_1^{(k)} + 0.25x_3^{(k-1)} + 7.5, \\ x_3^{(k)} &= 0.25x_2^{(k)} - 6. \end{aligned}$$

- The fist 7 iterates of Gauss-Seidel method are listed below.

(2) Recall that the component form of SOR method is given by

$$x_i^{(k)} = (1-\omega)x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right]$$

for each $i = 1, 2, \ldots, n$.

- The equations for SOR method with $\omega = 1.25$ are

$$\begin{aligned} x_1^{(k)} &= -0.25x_1^{(k-1)} - 0.9375x_2^{(k-1)} + 7.5, \\ x_2^{(k)} &= -0.9375x_1^{(k)} - 0.25x_2^{(k-1)} + 0.3125x_3^{(k-1)} + 9.375, \\ x_3^{(k)} &= 0.3125x_2^{(k)} - 0.25x_3^{(k-1)} - 7.5, \end{aligned}$$

for each $k \geq 1$.

Solution (3/3)

- Again, the fist 7 iterates of SOR method are listed below.

Numerical Comparison

To obtain an approx. sol. accurate to **7 decimal places**:

- Gauss-Seidel method requires **34** iterations.
- SOR with $\omega = 1.25$ requires only **14** iterations!

Question

How to select the optimal (or suboptimal) value of the relaxation parameter $\omega > 0$?

- No complete answer to this question until now!
- Only partial results are known for certain important cases.

Thm 7.24 (Kahan)

If $a_{ii} \neq 0$ for each $i = 1, 2, \ldots, n$, then

$$\rho(T_\omega) \geq |\omega - 1|.$$

Note: From Thm 7.24 $\Longrightarrow$ the SOR converges only if $0 < \omega < 2$!

Proof of Thm 7.24

- If $\lambda_1, \cdots, \lambda_n$ are eigenvalues of $T_\omega = (D - \omega L)^{-1}[(1-\omega)D + \omega U]$, then

$$\begin{aligned}\Pi_{i=1}^{n}\lambda_i = \det(T_\omega) &= \det(D-\omega L)^{-1} \cdot \det[(1-\omega)D + \omega U] \\ &= (a_{11}a_{22}\cdots a_{nn})^{-1} \cdots (1-\omega)^n \cdot (a_{11}a_{22}\cdots a_{nn}) \\ &= (1-\omega)^n.\end{aligned}$$

- Thus, $[\rho(T_\omega)]^n \geq \Pi_{i=1}^{n}|\lambda_i| = |1-\omega|^n$ and hence $\rho(T_\omega) \geq |\omega - 1|$. Also, note that

$$|\omega - 1| \leq \rho(T_\omega) < 1 \Longrightarrow 0 < \omega < 2.$$

Thm 7.25 (Ostrowski-Reich)

If $A \in \mathbb{R}^{n\times n}$ is positive definite and $0 < \omega < 2$, then the SOR method converges **for any initial approx. vector** $x^{(0)}$.

Thm 7.26 (A 為正定且三對角線矩陣)

If $A \in \mathbb{R}^{n\times n}$ is **positive definite and tridiagonal**, then

(i) $\rho(T_g) = [\rho(T_j)]^2 < 1$, and

(ii) the **optimal choice** of ω for the SOR method is

$$\omega = \frac{2}{1+\sqrt{1-[\rho(T_j)]^2}}.$$

With this choice of ω, $\rho(T_\omega) = \omega - 1$.

Reference

The proof of above two theorems can be found in [Or2], pp. 123–133.

[Or2] J. M. Ortega, *Numerical Analysis: A Second Course*, Academic Press, New York, 1972.

Question

In Example 1, we apply the SOR method with $\omega = \mathbf{1.25}$ for solving the linear system

$$\begin{aligned} 4x_1 + 3x_2 &= 24, \\ 3x_1 + 4x_2 - x_3 &= 30, \\ -x_2 + 4x_3 &= -24. \end{aligned}$$

Is the choice of ω optimal or suboptimal for this case?

Example 2, p. 466

Find the optimal choice of ω for the SOR method for solving a linear system $Ax = b$ with the **tridiagonal** matrix

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

Sol: Note that A is symmetric and **positive definite** because

$$\det(A) = 24, \quad \det\left(\begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}\right) = 7 \quad \text{and} \quad \det([4]) = 4.$$

So, $A \in \mathbb{R}^{3\times 3}$ is a positive and tridiagonal matrix, and hence Thm 7.26 can be applied.

- Now, compute the matrix $T_j = D^{-1}(L+U)$ as

$$T_j = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & -3 & 0 \\ -3 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -0.75 & 0 \\ -0.75 & 0 & 0.25 \\ 0 & 0.25 & 0 \end{bmatrix}.$$

- Since the cha characteristic polynomial of T_j is

$$\det(T_j - \lambda I) = \det \begin{bmatrix} -\lambda & -0.75 & 0 \\ -0.75 & -\lambda & 0.25 \\ 0 & 0.25 & -\lambda \end{bmatrix} = -\lambda(\lambda^2 - 0.625),$$

we have $\rho(T_j) = \sqrt{0.625}$ or $[\rho(T_j)]^2 = \mathbf{0.625}$.

- Thus, the optimal value of ω should be

$$\omega = \frac{2}{1+\sqrt{1-[\rho(T_j)]^2}} = \frac{2}{1+\sqrt{1-\mathbf{0.625}}} \approx \mathbf{1.24!}$$

Section 7.5
Error Bounds and Iterative Refinement
(誤差上界與迭代改進)

Let $\tilde{x}$ be a computed approximation to the unique sol. $x \in \mathbb{R}^n$ of $Ax = b$ with residual vector $r = b - A\tilde{x}$.

Question

Does the small quantity of $\|r\|$ indicate that the absolute error $\|\tilde{x} - x\|$ is small as well?

No, it depends on the **conditioning** of the given problem!

Example 1 (小殘量不保證較小的絕對誤差)

The 2×2 linear system $Ax = b$ is given by

$$\begin{bmatrix} 1 & 2 \\ \mathbf{1.0001} & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ \mathbf{3.0001} \end{bmatrix}$$

has the unique solution $x = [1, 1]^T \in \mathbb{R}^2$. Find the residual vector for the **poor** approximation $\tilde{x} = [3, -0.0001]^T$.

Sol: The residual vector is

$$r = b - A\tilde{x} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -0.0001 \end{bmatrix} = \begin{bmatrix} 0.0002 \\ 0 \end{bmatrix}.$$

Then $\|r\|_\infty = \mathbf{0.0002}$ is small, but its absolute error $\|\tilde{x} - x\|_\infty = \mathbf{2}$ is quite large!

- The exact sol. x is just the intersection of two **nearly parallel** lines

$$\mathbf{l_1}: \quad x_1 + 2x_2 = 3,$$
$$\mathbf{l_2}: \quad 1.0001x_1 + 2x_2 = 3.0001.$$

- The poor approximation $\tilde{x}$ lies on $\mathbf{l_2}$, but **lies close to** $\mathbf{l_1}$. So, small quantity of $\|r\|_\infty$ is obtained.

Thm 7.27 (殘量與相對誤差的關係)

Let A be **nonsingular** and $\tilde{x}$ be an approx. to the sol. x of the linear system $Ax = b$ with residual vector $r = b - A\tilde{x}$. If $\|\cdot\|$ denotes any natural norm, then

- $\|\tilde{x} - x\| \leq \|A^{-1}\| \cdot \|r\|$, and
- $\frac{\|\tilde{x}-x\|}{\|x\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \cdot \frac{\|r\|}{\|b\|}$, provided that $x \neq 0$ and $b \neq 0$.

pf: Since $Ax = b$ and $A\tilde{x} = b - r$, we see that

$$A(\tilde{x} - x) = (b - r) - b = -r \quad \text{or} \quad \tilde{x} - x = -A^{-1}r.$$

Taking norm $\|\cdot\| \Longrightarrow \|\tilde{x} - x\| = \|A^{-1r}\| \leq \|A^{-1}\|\|r\|$. Because $\|b\| \leq \|A\|\|x\|$, we have $1/\|x\| \leq \|A\|/\|b\|$ and the above inequality becomes

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|r\|}{\|x\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|r\|}{\|b\|}.$$

Def 7.28, p. 470

The **condition number** (條件數) of a nonsingular matrix A is defined by

$$K(A) = \kappa(A) = \|A\|\|A^{-1}\|,$$

where $\|\cdot\|$ is any natural matrix norm.

The Conditioning of A

Since $I = AA^{-1}$, it is easily seen that

$$1 = \|I\| \leq \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = K(A).$$

Thus we say that

- A is **well-conditioned** (良態的) if $K(A)$ is close to 1.
- A is **illl-conditioned** (病態的) if $K(A)$ is significantly greater than 1.

Rewriting Thm 7.27 as ...

- $\|\tilde{x} - x\| \leq \|A^{-1}\| \cdot \|r\| = K(A) \cdot \frac{\|r\|}{\|A\|}$, and
- $\frac{\|\tilde{x}-x\|}{\|x\|} \leq K(A) \cdot \frac{\|r\|}{\|b\|}$. (相對誤差 $\leq$ 條件數 $\times$ 相對殘量)

Notes

- The condition number $K(A)$ can be viewed as a **magnifying factor (放大因子)** of the absolute or relative error.
- A is well-conditioned and $\|r\|$ is small $\Longrightarrow$ absolute or relative error is small as well.

Example 2, p. 471

Use l_∞ norm to determine the condition number of the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}$$

given in Example 1.

Sol: Note that $\|A\|_\infty = \mathbf{3.0001}$. But its inverse is

$$A^{-1} = \begin{bmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{bmatrix}, \text{ so } \|A^{-1}\|_\infty = \mathbf{20000}.$$

Hence, the condition number of A is $K(A) = K_\infty(A) = \mathbf{60002}$.

Question

For any nonsingular $A \in \mathbb{R}^{n \times n}$, how to estimate its condition number

$$K(A) = \|A\| \, \|A^{-1}\|$$

efficiently using t-**digit arithmetic**?

The Estimation of Condition Numbers (1/2)

- Assume that **t-digit arithmetic** is used in the process of GE for solving the approx. sol. $\tilde{x}$ to the linear system $Ax = b$.
- It can be shown from pp. 45–47 of [FM, 1967] that

$$\|r\| \approx 10^{-t}\|A\|\|\tilde{x}\| \quad \text{with } r = b - A\tilde{x}.$$

Reference

[FM] G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

- Use **2t-digit arithmetic** (double precision) to evaluate the residual vector $r = b - A\tilde{x}$.

- Apply LU factorization generated from GE to obtain an approx. sol. $\tilde{y}$ of the linear system

$$Ay = r$$

using the **t-digit arithmetic**.

- Thus we then have

$$\tilde{y} \approx A^{-1}r = A^{-1}(b - A\tilde{x}) = x - \tilde{x} \quad \text{or} \quad x \approx \tilde{x} + \tilde{y}.$$

- Taking norm $\|\cdot\| \Longrightarrow$

$$\|\tilde{y}\| \approx \|A^{-1}r\| \leq \|A^{-1}\| \cdot \|r\| \approx 10^{-t}\|\tilde{x}\| \cdot K(A).$$

- Finally, the condition number $K(A)$ can be estimated by

$$K(A) \approx 10^{t} \cdot \frac{\|\tilde{y}\|}{\|\tilde{x}\|}.$$

An Illustrative Example (1/2)

Example

The 3×3 linear system $Ax = b$ with

$$A = \begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$$

has the unique solution $\mathbf{x} = [\mathbf{1}, \mathbf{1}, \mathbf{1}]^{\mathbf{T}} \in \mathbb{R}^3$.

- Applying GE with **5-digit rounding** arithmetic $\Longrightarrow$ computed solution $\tilde{x} = [1.2001, 0.99991, 0.92538]$.
- Use **10-digit rounding arithmetic** to obtain

$$r = fl(b - A\tilde{x}) = \begin{bmatrix} -0.00518 \\ 0.27412914 \\ -0.186160367 \end{bmatrix}.$$

An Illustrative Example (2/2)

- Solving $Ay = r$ with **5-digit rounding arithmetic** for $\tilde{y} \Longrightarrow$

$$\tilde{y} = [-0.20008, 8.9987 \times 10^{-5}, 0.074607]^T.$$

- So, the condition number of A is estimated by

$$K_\infty(A) \approx \frac{\|\tilde{y}\|_\infty}{\|\tilde{x}\|_\infty} 10^5 = \mathbf{16672}$$

without computing the inverse matrix A^{-1} explicitly!

- Furthermore, the *exact* condition number is

$$K_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty = (1.0041)(15934) = \mathbf{15999}$$

using **5-digit rounding arithmetic**.

Question

Are the residual bounds in Thm 7.27

- $\|\tilde{x} - x\| \le \|A^{-1}\| \cdot \|r\| = K(A) \cdot \frac{\|r\|}{\|A\|}$, and
- $\frac{\|\tilde{x}-x\|}{\|x\|} \le K(A) \cdot \frac{\|r\|}{\|b\|}$

sharp (or tight) for this Example?

Is the residual bounds in Thm 7.27 sharp?

- Since the exact sol. $x = [1, 1, 1]^T$ is known, we compute

$$\|\tilde{x} - x\|_\infty = \mathbf{0.2001} \quad \text{and} \quad \frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} = \frac{0.2001}{1} = \mathbf{0.2001}.$$

- Also, the residual bounds in Thm 7.27 are computed as

$$\|\tilde{x} - x\|_\infty \leq K_\infty(A) \frac{\|r\|_\infty}{\|A\|_\infty} = \frac{(15999)(0.27413)}{15934} = \mathbf{0.27525},$$

$$\frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} \leq K_\infty(A) \frac{\|r\|_\infty}{\|b\|_\infty} = \frac{(15999)(0.27413)}{15913} = \mathbf{0.27561}.$$

- **This example illustrates the sharpness (or tightness) of the error bounds in Thm 7.27!**

Iterative Refinement (or Improvement)

- In above derivation, $x \approx \tilde{x} + \tilde{y}$ is **more accurate than** $\tilde{x}$ as an approximation to the sol. of $Ax = b$, where $\tilde{y}$ is the computed sol. to $Ay = r$.
- **Basic Idea:** Let $x^{(1)} = \tilde{x}$. For $k = 1, 2, \ldots$

$$r^{(k)} = b - Ax^{(k)}, \quad Ay^{(k)} = r^{(k)}, \quad x^{(k+1)} = x^{(k)} + y^{(k)}.$$

- All steps, except Step 3 below for computing the residual vector $r^{(k)}$, of Iterative Refinement are performed in the **t-digit arithmetic**.

Algorithm 7.4: Iterative Refinement

INPUT matrices $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$; tolerance TOL; max. no. of iter. N; no. of precision t.

OUTPUT approx. sol. xx to $Ax = b$; $K_\infty(A) \approx COND$.

Step 0 Solve $Ax = b$ for x using GE.

Step 1 Set $k = 1$.

Step 2 While $(k \leq N)$ do **Steps 3–9**

- Step 3 Set residual vector $r = b - Ax$ using **2t-digit arithmetic**.
- Step 4 Solve $Ay = r$ for y using the LU fact. generated from Step 0.
- Step 5 Set $xx = x + y$.
- Step 6 If $k = 1$ then set $COND = \frac{\|y\|_\infty}{\|xx\|_\infty} 10^t$.
- Step 7 If $\|xx - x\|_\infty < TOL$ then OUTPUT(xx and COND); **STOP**.
- Step 8 Set $k = k + 1$.
- Step 9 Set $x = xx$. (Update x.)

Step 10 OUTPUT('Max. no. of iter. exceeded' and COND); **STOP**.

Comments on Algorithm 7.4

- The double-precision (or $2t$-digit) arithmetic is required in Step 3 in order to avoid the loss of significance for two nearly equal numbers.
- If t-digit arithmetic is used and $K_\infty(A) \approx 10^q$ $(0 \le q \le t)$, then after k iterations of Iterative Refinement, the sol. xx has approximately $\min\{\mathbf{t}, \mathbf{k}(\mathbf{t}-\mathbf{q})\}$ **correct digits**.
- If A (or the linear system) is well-conditioned, usually **only one or two iterations** of Iterative Refinement are required for obtaining highly accurate approximation to the linear system.
- If A is ill-conditioned with $K_\infty(A) > 10^t$, then extended precision should be used in the calculations.

Example

The 3×3 linear system $Ax = b$ with

$$A = \begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$$

has the unique solution $\mathbf{x} = [\mathbf{1}, \mathbf{1}, \mathbf{1}]^{\mathbf{T}} \in \mathbb{R}^3$. Perform 2 iterations of Iterative Refinement using **5-digit rounding** arithmetic.

Sol: With the computed approx. $\tilde{x}$ and its residual

$$x^{(1)} = \tilde{x} = [1.2001, 0.99991, 0.92538]^T$$

$$r^{(1)} = r = [-0.00518, 0.27412914, -0.186160367],$$

we solve $Ay = r^{(1)}$ for $y^{(1)} \Longrightarrow$

$$y^{(1)} = \tilde{y} = [-0.20008, 8.9987 \times 10^{-5}, 0.074607]^T.$$

- After fist iteration of Iterative Refinement, we have

$$x^{(2)} = x^{(1)} + y^{(1)} = [\mathbf{1.0000}, \mathbf{1.0000}, \mathbf{0.99999}]^{\mathbf{T}}$$

with absolute (or relative) error $\|x^{(2)} - x\|_\infty = \mathbf{1} \times \mathbf{10^{-5}}$.

- After second iteration, we obtain

$$y^{(2)} = [\mathbf{1.5002} \times \mathbf{10^{-9}}, \mathbf{2.0951} \times \mathbf{10^{-10}}, \mathbf{1.0000} \times \mathbf{10^{-5}}]$$

and hence the next approx. sol. is given by

$$x^{(3)} = x^{(2)} + y^{(2)} = [\mathbf{1.0000}, \mathbf{1.0000}, \mathbf{1.0000}]^{\mathbf{T}},$$

which is the exact sol. x to the given linear system.

Question

For any linear system $Ax = b$ with nonsingular A, the computed sol. $\tilde{x}$ is obtained by solving the **perturbed linear system**

$$(A + \delta A)\tilde{x} = b + \delta b \quad \text{with } \|\delta A\| = O(10^{-t}), \|\delta b\| = O(10^{-t}).$$

Is it *always* true that $\|\tilde{x} - x\| = O(10^{-t})$?

Thm 7.29 (線性系統的擾動上界)

If $A \in \mathbb{R}^{n \times n}$ and $\|\delta A\| \cdot \|A^{-1}\| < 1$ for any natural norm $\|\cdot\|$, then

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

Section 7.6
The Conjugate Gradient Method (共軛梯度法; 簡稱 CG Method)

Goal: To develop an iterative method for solving **large-scale** linear systems with **positive definite** coefficient matrices.

Thm 7.30 (內積的基本性質)

The inner product (or dot product) of $x, y \in \mathbb{R}^n$ is defined by

$$\langle x, y\rangle = x^T y = \sum_{i=1}^{n} x_i y_i.$$

For any $x, y, z \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, we have

(a) $\langle x, y\rangle = \langle y, x\rangle$;

(b) $\langle \alpha x, y\rangle = \langle x, \alpha y\rangle = \alpha\langle x, y\rangle$;

(c) $\langle x, y+z\rangle = \langle x, y\rangle + \langle x, z\rangle$;

(d) $\langle x, x\rangle \geq 0$;

(e) $\langle x, x\rangle = 0 \iff x = 0$;

(f) $\langle x, Ay\rangle = \langle Ax, y\rangle$ if $A = A^T$.

Thm 7.31 (正定線性系統與優化問題的關聯性)

x^* is the sol. to the **positive definite** linear system $Ax = b \Longleftrightarrow$ x^* produces the minimal value of $g(x) = \langle x, Ax\rangle - 2\langle x, b\rangle$.

Note: Since $A^T = A$, for $x, 0 \neq v \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we have

$$\begin{aligned} g(x+tv) =& \langle x+tv, Ax+tAv\rangle - 2\langle x+tv, b\rangle \\ =& \langle x, Ax\rangle + t\langle x, Av\rangle + t\langle v, Ax\rangle + t^2\langle v, Av\rangle \\ & - 2\langle x, b\rangle - 2t\langle v, b\rangle \\ =& \langle x, Ax\rangle - 2\langle x, b\rangle + 2t\langle v, Ax\rangle - 2t\langle v, b\rangle + t^2\langle v, Av\rangle \\ =& g(x) - 2t\langle v, b-Ax\rangle + t^2\langle v, Av\rangle. \end{aligned} \tag{8}$$

Proof of Thm 7.31 (1/2)

From (8), for x and $v \neq 0$, define a quadratic function h in t by

$$h(t) = g(x + tv) = g(x) - 2t\langle v, b - Ax\rangle + t^2\langle v, Av\rangle.$$

Since $\langle v, Av\rangle > 0$, h has a minimal value at some $\hat{t}$ and hence

$$0 = h'(\hat{t}) = -2\langle v, b - Ax\rangle + 2\hat{t}\langle v, Av\rangle.$$

So, we obtain

$$\hat{t} = \frac{\langle v, b - Ax\rangle}{\langle v, Av\rangle} \quad \text{and} \quad g(x + \hat{t}v) = g(x) - \frac{\langle v, b - Ax\rangle^2}{\langle v, Av\rangle} < g(x)$$

unless $\langle v, b - Ax\rangle = 0$.

Proof of Thm 7.31 (2/2)

Therefore, we conclude that

$$
\begin{aligned}
& x^* \text{ is the unique sol. to } Ax = b \\
\Longleftrightarrow\ & Ax^* = b \quad \text{or} \quad b - Ax^* = 0 \\
\Longleftrightarrow\ & \langle v, b - Ax^* \rangle = 0 \quad \forall\, v \neq 0 \\
\Longleftrightarrow\ & g(x^* + \hat{t}v) = g(x^*) \quad \forall\, v \neq 0 \\
\Longleftrightarrow\ & g \text{ has a minimal value at } x^*.
\end{aligned}
$$

Basic Idea of CG Method

INPUT $x^{(0)}$ is an initial approximation to x^*; $v^{(1)} \neq 0$ is an initial **search direction** s.t. $\langle \mathbf{v^{(1)}}, \mathbf{b} - \mathbf{Ax^{(0)}} \rangle \neq \mathbf{0}$.

OUTPUT an approx. sol. to the linear system $Ax = b$.

- For $k = 1, 2, \ldots$ until convergence

$$t_k = \frac{\langle v^{(k)}, b - Ax^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}, \quad x^{(k)} = x^{(k-1)} + t_k v^{(k)}.$$

- But, what is the next search direction $v^{(2)}$ in above iteration?

Question

How to find suitable and feasible search directions $v^{(k)}$ for $k \geq 2$?

Two Choices of Search Directions

- **Steepest Descent Method:** (最速下降法)
  - Note that $\nabla g(x) = 2(Ax - b) = -2r$. (Check!)
  - The direction where $g(x)$ decreases most rapidly is $-\nabla g(x) = r$, which is the residual vector.
  - Select $v^{(k)} = r^{(k)} = b - Ax^{(k-1)}$ for $k = 1, 2, \ldots$.
  - But this method is not used for solving linear systems because of its **slow convergence**.

- **Conjugate Gradient Method:** (共軛梯度法)
  - Select A-orthogonal set of nonzero vectors $\{v^{(1)}, v^{(2)}, \cdots\}$.
  - **A-orthogonality:** Two vectors $v^{(i)}$ and $v^{(j)}$ are called **A-orthogonal** (A-垂直向量) if $\langle v^{(i)}, Av^{(j)}\rangle = 0$ for $i \neq j$.
  - The CG method of Hestenes and Steifel [HS, 1952] was originally developed as **a direct method** for solving an $n \times n$ positive definite linear system.

Reference

[HS] M. R. Hestenes and E. Steifel, *Conjugate gradient methods in optimization*, Journal of Research of the National Bureau of Standards, Vol. 49, pp. 409–436, 1952.

Thm 7.32 (CG 法的收斂性)

Let $\{v^{(1)}, v^{(2)}, \cdots, v^{(n)}\}$ be A-orthogonal associated with positive definite A and $x^{(0)}$ be arbitrary. Define

$$t_k = \frac{\langle v^{(k)}, b - Ax^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}, \quad x^{(k)} = x^{(k-1)} + t_k v^{(k)}$$

for $k = 1, 2, \ldots, n$. Then, **assuming exact arithmetic**, $Ax^{(n)} = b$. (在無捨入誤差的精確算術環境中，CG 算法只要 n 步迭代即可求解線性系統!)

Exercise 13, p. 494

Let $S = \{v^{(1)}, v^{(2)}, \cdots, v^{(n)}\}$ be an A-**orthogonal** set of nonzero vectors associated with **positive definite** $A \in \mathbb{R}^{n\times n}$. Then

(a) S is **linearly independent** and hence forms a basis for $\mathbb{R}^n$.

(b) $\langle z, v^{(k)} \rangle = 0$ for $k = 1, 2, \ldots, n \iff z = 0$.

- Since $x^{(k)} = x^{(k-1)} + t_k v^{(k)}$ for $k \geq 1$, we have

$$\begin{aligned} Ax^{(n)} &= Ax^{(n-1)} + t_n Av^{(n)} \\ &= Ax^{(n-2)} + t_{n-1} Av^{(n-1)} + t_n Av^{(n)} \\ &\vdots \\ &\quad Ax^{(0)} + t_1 Av^{(1)} + \cdots + t_n Av^{(n)}. \end{aligned}$$

- So, $Ax^{(n)} - b = Ax^{(0)} - b + t_1 Av^{(1)} + \cdots + t_n Av^{(n)}$.
- Because $\langle v^{(k)}, Av^{(i)} \rangle = 0$ for $k \neq i$, we see that

$$\langle v^{(k)}, Ax^{(n)} - b \rangle = \langle v^{(k)}, Ax^{(0)} - b \rangle + t_k \langle v^{(k)}, Av^{(k)} \rangle.$$

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

- Again, by induction and A-orthogonality, notice that

$$t_k\langle v^{(k)}, Av^{(k)}\rangle = \langle v^{(k)}, b - Ax^{(k-1)}\rangle$$
$$= \langle v^{(k)}, b - Ax^{(0)} - t_1 Av^{(1)} - \cdots - t_{k-1}Av^{(k-1)}\rangle$$
$$= \langle v^{(k)}, b - Ax^{(0)}\rangle.$$

- Thus, we conclude that for $k = 1, 2, \ldots, n$,

$$\langle v^{(k)}, Ax^{(n)} - b\rangle = \langle v^{(k)}, Ax^{(0)} - b\rangle + \langle v^{(k)}, b - Ax^{(0)}\rangle = 0.$$

- From Exercise 13(b), we know that

$$Ax^{(n)} - b = 0 \quad \text{or} \quad Ax^{(n)} = b,$$

i.e., $x^{(n)}$ is the **exact solution** to $Ax = b$!

Example 1, p. 483

Apply Thm 7.32 to solve the positive definite system $Ax = b$ with

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}, b = \begin{bmatrix} 24 \\ 30 \\ -24 \end{bmatrix}$$

using $\mathbf{x}^{(0)} = [\mathbf{0}, \mathbf{0}, \mathbf{0}]^{\mathbf{T}}$ and A-orthogonal vectors $\mathbf{v}^{(1)} = [\mathbf{1}, \mathbf{0}, \mathbf{0}]^{\mathbf{T}}$, $\mathbf{v}^{(2)} = [-\mathbf{3}/\mathbf{4}, \mathbf{1}, \mathbf{0}]^{\mathbf{T}}$, $\mathbf{v}^{(3)} = [-\mathbf{3}/\mathbf{7}, \mathbf{4}/\mathbf{7}, \mathbf{1}]^{\mathbf{T}}$.

Note: See the textbook for showing the A-orthogonality conditions:

$$\langle v^{(1)}, Av^{(2)} \rangle = 0, \quad \langle v^{(1)}, Av^{(3)} \rangle = 0, \quad \langle v^{(2)}, Av^{(3)} \rangle = 0.$$

Solution of Example 1

k=1: $r^{(0)} = b - Ax^{(0)} = b = [24, 30, -24]^T$ and

$$t_1 = \frac{\langle v^{(1)}, r^{(0)}\rangle}{\langle v^{(1)}, Av^{(1)}\rangle} = \frac{24}{4} = \mathbf{6}, \quad x^{(1)} = x^{(0)} + t_1 v^{(1)} = [\mathbf{6}, \mathbf{0}, \mathbf{0}]^{\mathbf{T}}.$$

k=2: $r^{(1)} = b - Ax^{(1)} = [0, 12, -24]^T$ and

$$t_2 = \frac{\langle v^{(2)}, r^{(1)}\rangle}{\langle v^{(2)}, Av^{(2)}\rangle} = \frac{12}{7/4} = \mathbf{\frac{48}{7}}, \quad x^{(2)} = x^{(1)} + t_2 v^{(2)} = [\mathbf{\frac{6}{7}}, \mathbf{\frac{48}{7}}, \mathbf{0}]^{\mathbf{T}}.$$

k=3: $r^{(2)} = b - Ax^{(2)} = [0, 0, \frac{-120}{7}]^T$ and

$$t_3 = \frac{\langle v^{(3)}, r^{(2)}\rangle}{\langle v^{(3)}, Av^{(3)}\rangle} = -\mathbf{5}, \quad x^{(3)} = x^{(2)} + t_3 v^{(3)} = [\mathbf{3}, \mathbf{4}, -\mathbf{5}]^{\mathbf{T}},$$

which is the **exact solution** to $Ax = b$ after **3** iterations!

Definition

The numerical method is called a **conjugate direction method** if it uses an A-orthogonal set $\{v^{(1)}, v^{(2)}, \cdots, v^{(n)}\}$ of direction vectors.

Thm 7.33 (殘量與共軛方向向量的關係)

For a conjugate direction method, its residual vectors $r^{(k)}$, where $k = 1, 2, \ldots, n$, satisfy

$$\langle r^{(k)}, v^{(j)} \rangle = 0, \quad j = 1, 2, \ldots, k.$$

(第 k 步的殘餘向量與前 k 個共軛方向向量均垂直)

Proof of Thm 7.33

k=1:

$$\langle v^{(1)}, r^{(1)} \rangle = \langle v^{(1)}, b - Ax^{(1)} \rangle$$
$$= \langle v^{(1)}, b - Ax^{(0)} \rangle - t_1 \langle v^{(1)}, Av^{(1)} \rangle = 0.$$

k=2:

$$\langle v^{(1)}, r^{(2)} \rangle = \langle v^{(1)}, b - Ax^{(2)} \rangle$$
$$= \langle v^{(1)}, b - Ax^{(1)} \rangle - t_2 \langle v^{(1)}, Av^{(2)} \rangle$$
$$= \langle v^{(1)}, r^{(1)} \rangle - 0 = 0.$$

and

$$\langle v^{(2)}, r^{(2)} \rangle = \langle v^{(2)}, b - Ax^{(2)} \rangle$$
$$= \langle v^{(2)}, b - Ax^{(1)} \rangle - t_2 \langle v^{(2)}, Av^{(2)} \rangle = 0.$$

$k \geq 3$ By mathematical induction! See also **Exercise 14**.

Let $x^{(0)}$ be an initial approx. with residual $r^{(0)} = b - Ax^{(0)} \neq 0$.

- Firstly, choose $v^{(1)} = r^{(0)}$. (the **steepest descent direction**)
- Assume the conjugate directions $v^{(1)}, \cdots, v^{(k-1)}$ and approx. $x^{(1)}, \cdots, x^{(k-1)}$ are computed with

$$\langle v^{(i)}, Av^{(j)} \rangle = 0 \quad \text{and} \quad \langle r^{(i)}, r^{(j)} \rangle = 0 \quad \text{for } i \neq j.$$

- If $r^{(k-1)} = b - Ax^{(k-1)} = 0$, we are done. Otherwise, we have

$$\langle r^{(k-1)}, v^{(i)} \rangle = 0, \quad i = 1, 2, \ldots, k-1$$

by Thm 7.33.

Define the kth Conjugate Direction

$$v^{(k)} = r^{(k-1)} + s_{k-1} v^{(k-1)} \quad \text{for some } s_{k-1} \in \mathbb{R}.$$

- Since we want $\langle v^{(k-1)}, Av^{(k)}\rangle = 0$, it follows that

$$0 = \langle v^{(k-1)}, Av^{(k)}\rangle = \langle v^{(k-1)}, Ar^{(k-1)}\rangle + s_{k-1}\langle v^{(k-1)}, Av^{(k-1)}\rangle,$$

and hence the scalar s_{k-1} is given by

$$s_{k-1} = \frac{-\langle v^{(k-1)}, Ar^{(k-1)}\rangle}{\langle v^{(k-1)}, Av^{(k-1)}\rangle} = \frac{-\langle r^{(k-1)}, Av^{(k-1)}\rangle}{\langle v^{(k-1)}, Av^{(k-1)}\rangle}. \qquad (9)$$

- From p. 245 of [Lu], it can be shown that $\{v^{(1)}, v^{(2)}, \cdots, v^{(k)}\}$ is an A-**orthogonal** set.

Reference

[Lu] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading MA, 1984.

Reformulation for t_k, $r^{(k)}$ and s_k

- The scalar t_k can be rewritten as

$$
\begin{aligned}
t_k &= \frac{\langle v^{(k)}, r^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} = \frac{\langle r^{(k-1)} + s_{k-1} v^{(k-1)}, r^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} \\
&= \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} + s_{k-1} \frac{\langle v^{(k-1)}, r^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} \\
&= \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}. \qquad (10)
\end{aligned}
$$

- The kth residual vector can be obtained by

$$
\begin{aligned}
r^{(k)} &= b - Ax^{(k)} = b - Ax^{(k-1)} - t_k A v^{(k)} \\
&= r^{(k-1)} - t_k A v^{(k)}. \qquad (11)
\end{aligned}
$$

- From (10) and (11), we notice that

$$\langle r^{(k-1)}, r^{(k-1)} \rangle = t_k \langle v^{(k)}, Av^{(k)} \rangle,$$
$$\langle r^{(k)}, r^{(k)} \rangle = \langle r^{(k)}, r^{(k-1)} \rangle - t_k \langle r^{(k)}, Av^{(k)} \rangle$$
$$= -t_k \langle r^{(k)}, Av^{(k)} \rangle.$$

- From Eq. (9) for s_k, it can be rewritten as

$$s_k = \frac{-\langle r^{(k)}, Av^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} = \frac{-t_k \langle r^{(k)}, Av^{(k)} \rangle}{t_k \langle v^{(k)}, Av^{(k)} \rangle}$$
$$= \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k-1)}, r^{(k-1)} \rangle}. \tag{12}$$

Put Eqs. (10), (11) and (12) together to get

Basic Algorithm of CG Method

Let $x^{(0)}$ be an initial guess with residual $r^{(0)} = b - Ax^{(0)} \neq 0$.

Step 1 Set $v^{(1)} = r^{(0)}$.

Step 2 For $k = 1, 2, \ldots, n$, set

$$t_k = \langle r^{(k-1)}, r^{(k-1)} \rangle / \langle v^{(k)}, Av^{(k)} \rangle;$$
$$x^{(k)} = x^{(k-1)} + t_k v^{(k)};$$
$$r^{(k)} = r^{(k-1)} - t_k A v^{(k)};$$

If $k < n$, set

$$s_k = \langle r^{(k)}, r^{(k)} \rangle / \langle r^{(k-1)}, r^{(k-1)} \rangle;$$
$$v^{(k+1)} = r^{(k)} + s_k v^{(k)}.$$

Step 3 OUTPUT($x^{(n)}$); **STOP**.

Example 2, p. 488

Apply basic algorithm of CG method to solve the positive definite system $Ax = b$ with

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}, b = \begin{bmatrix} 24 \\ 30 \\ -24 \end{bmatrix}$$

using $\mathbf{x}^{(0)} = [\mathbf{0}, \mathbf{0}, \mathbf{0}]^{\mathbf{T}}$. The actual solution is $x = [3, 4, -5]^T$.

$k = 1$: we obtain $x^{(1)}$ and $r^{(1)}$ as

$$x^{(1)} = [3.525773196, 4.407216495, -3.525773196]^T$$
$$r^{(1)} = [-3.32474227, -1.73195876, -5.48969072]^T.$$

The relative error and relative residual with respect to b are

$$\frac{\|x^{(1)} - x\|_\infty}{\|x\|_\infty} \approx \mathbf{2.95 \times 10^{-1}}, \quad \frac{\|r^{(1)}\|_\infty}{\|b\|_\infty} \approx \mathbf{1.83 \times 10^{-1}}.$$

$k = 2$: we obtain $x^{(2)}$ and $r^{(2)}$ as

$$x^{(2)} = [2.858011121, 4.148971939, -4.954222164]^T$$
$$r^{(2)} = [0.121039698, -0.124143281, -0.034139402]^T.$$

The relative error and relative residual with respect to b are

$$\frac{\|x^{(2)} - x\|_\infty}{\|x\|_\infty} \approx \mathbf{2.98 \times 10^{-2}}, \quad \frac{\|r^{(2)}\|_\infty}{\|b\|_\infty} \approx \mathbf{4.14 \times 10^{-3}}.$$

- Note that **SOR method** with $\omega = 1.25$ requires **14** iterations for obtaining **7** significant digits of the approximate solution.

Solution (3/3)

$k = 3$: we obtain $x^{(3)}$ and $r^{(3)}$ as

$$x^{(3)} = [2.999999998, 4.000000002, -4.999999998]^T$$
$$r^{(3)} = [0.36 \times 10^{-8}, 0.39 \times 10^{-8}, -0.14 \times 10^{-8}]^T.$$

The relative error and relative residual with respect to b are

$$\frac{\|x^{(3)} - x\|_\infty}{\|x\|_\infty} \approx \mathbf{4.00 \times 10^{-10}}, \quad \frac{\|r^{(3)}\|_\infty}{\|b\|_\infty} \approx \mathbf{1.30 \times 10^{-10}}.$$

Preconditioning (預優處理)

- When A is **ill-conditioned**, CG method is **highly susceptible** (高度敏感的) to the rounding errors.
- The **preconditioning** strategy is to find a nonsingular matrix C so that the transformed coefficient matrix

$$\tilde{A} = C^{-1}AC^{-T}$$

is **better-conditioned**, where $C^{-T} \equiv (C^{-1})^T = (C^T)^{-1}$.
- The original linear system $Ax = b$ is transformed as

$$\tilde{A}\tilde{x} = \tilde{b} \Leftrightarrow (C^{-1}AC^{-T})(C^T x) = C^{-1}b \Leftrightarrow C^{-1}Ax = C^{-1}b,$$

where $\tilde{x} = C^T x$ and $\tilde{b} = C^{-1}b$.
- The inverse matrix of a **preconditioner** C should be cheaply obtained in practice!

- Let $\tilde{x}^{(k)} = C^T x^{(k)}$ for $k \geq 1$. Then

$$\begin{aligned}\tilde{r}^{(k)} &= \tilde{b} - \tilde{A}\tilde{x}^{(k)} = C^{-1}b - (C^{-1}AC^{-T})C^T x^{(k)} \\ &= C^{-1}(b - Ax^{(k)}) = C^{-1}r^{(k)}.\end{aligned}$$

- Let $\tilde{v}^{(k)} = C^T v^{(k)}$ and $w^{(k)} = C^{-1}r^{(k)}$ for $k \geq 1$. Then

$$\tilde{t}_k = \frac{\langle w^{(k-1)}, w^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}; \quad x^{(k)} = x^{(k-1)} + \tilde{t}_k v^{(k)};$$

$$r^{(k)} = r^{(k-1)} - \tilde{t}_k A v^{(k)}; \quad \tilde{s}_k = \frac{\langle w^{(k)}, w^{(k)} \rangle}{\langle w^{(k-1)}, w^{(k-1)} \rangle};$$

$$v^{(k+1)} = C^{-T} w^{(k)} + \tilde{s}_k v^{(k)}.$$

Algorithm 7.5: Preconditioned CG Method (1/2)

INPUT dimension n;
matrices $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$;
preconditioning matrix $C \in \mathbb{R}^{n\times n}$;
initial approximation $x^{(0)} \in \mathbb{R}^n$;
maximum number of iterations N;
tolerance TOL.

OUTPUT an approximate solution $x \in \mathbb{R}^n$ to $Ax = b$;
residual vector $r \in \mathbb{R}^n$.

Algorithm 7.5: Preconditioned CG Method (2/2)

Step 1 Set $x = x^{(0)}$; $r = b - Ax$; $w = C^{-1}r$;
$v = C^{-T}w$; $\alpha = \langle w, w \rangle$.

Step 2 Set $k = 1$.

Step 3 While $(k \leq N)$ do **Steps 4–7**

- Step 4 If $\|v\| < TOL$ then OUTPUT(x and r); **STOP**.
- Step 5 Set $u = Av$; $t = \alpha / \langle v, u \rangle$;
  $x = x + tv$; $r = r - tu$;
  $w = C^{-1}r$; $\beta = \langle w, w \rangle$.
- Step 6 If $|\beta| < TOL$ then
  if $\|r\| < TOL$ then OUTPUT(x and r) ; **STOP**.
- Step 7 Set $s = \beta/\alpha$; $v = C^{-T}w + sv$;
  $\alpha = \beta$; $k = k + 1$.

Step 8 OUTPUT('Maximum number of iterations exceeded'); **STOP**.

Example 3, p. 491

The 5×5 linear system $Ax = b$ with

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

has the solution

$$\begin{aligned} x^* =& [7.859713071, 0.4229264082, -0.07359223906, \\ & -0.5406430164, 0.01062616286]^T. \end{aligned}$$

The preconditioner used in PCG method is

$$C = \text{diag}(\sqrt{a_{11}}, \sqrt{a_{22}}, \sqrt{a_{33}}, \sqrt{a_{44}}, \sqrt{a_{55}}).$$

Numerical Results for Example 3

Thank you for your attention!