Chapter 7 Iterative Techniques in Matrix Algebra

Baodong LIU
baodong@sdu.edu.cn

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 高斯消元法，部分列主元消去法，比列列主元消去法;
- 矩阵的$\mathbf{LU}$ 分解：Gaussian 方法，直接分解方法
- 借助原矩阵的$\mathbf{LU}$ 分解，

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LUx}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b},\\ \mathbf{Ux}=\mathbf{y},\end{cases}$$

- 严格对角占优矩阵和对称正定矩阵可直接进行高斯消元
- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LL}^T$, 相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LDL}^T$，相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LDL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Lz}=\mathbf{b};\\ \mathbf{Dy}=\mathbf{z};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

回顾线性方程组$\mathbf{Ax = b}$的直接解法

- 高斯消元法，部分列主元消去法，比列列主元消去法;
- 矩阵的$\mathbf{LU}$ 分解：Gaussian 方法，直接分解方法
- 借助原矩阵的$\mathbf{LU}$ 分解，

$$\mathbf{Ax = b} \Rightarrow \mathbf{LUx = b} \Rightarrow \begin{cases} \mathbf{Ly = b}, \\ \mathbf{Ux = y}, \end{cases}$$

- 严格对角占优矩阵和对称正定矩阵可直接进行高斯消元
- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LL}^T$, 相应的

$$\mathbf{Ax = b} \Rightarrow \mathbf{LL}^T\mathbf{x = b} \Rightarrow \begin{cases} \mathbf{Ly = b}; \\ \mathbf{L}^T\mathbf{x = y}, \end{cases}$$

- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LDL}^T$，相应的

$$\mathbf{Ax = b} \Rightarrow \mathbf{LDL}^T\mathbf{x = b} \Rightarrow \begin{cases} \mathbf{Lz = b}; \\ \mathbf{Dy = z}; \\ \mathbf{L}^T\mathbf{x = y}, \end{cases}$$

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 高斯消元法，部分列主元消去法，比列列主元消去法;
- 矩阵的$\mathbf{LU}$ 分解：Gaussian 方法，直接分解方法
- 借助原矩阵的$\mathbf{LU}$ 分解，

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LUx}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b},\\ \mathbf{Ux}=\mathbf{y},\end{cases}$$

- 严格对角占优矩阵和对称正定矩阵可直接进行高斯消元
- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LL}^T$, 相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LDL}^T$，相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LDL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Lz}=\mathbf{b};\\ \mathbf{Dy}=\mathbf{z};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 高斯消元法，部分列主元消去法，比列列主元消去法;
- 矩阵的$\mathbf{LU}$ 分解：Gaussian 方法，直接分解方法
- 借助原矩阵的$\mathbf{LU}$ 分解，

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LUx}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b},\\ \mathbf{Ux}=\mathbf{y},\end{cases}$$

- 严格对角占优矩阵和对称正定矩阵可直接进行高斯消元
- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LL}^T$, 相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LDL}^T$，相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LDL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Lz}=\mathbf{b};\\ \mathbf{Dy}=\mathbf{z};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 高斯消元法，部分列主元消去法，比列列主元消去法;
- 矩阵的$\mathbf{LU}$ 分解：Gaussian 方法，直接分解方法
- 借助原矩阵的$\mathbf{LU}$ 分解，

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LUx}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b},\\ \mathbf{Ux}=\mathbf{y},\end{cases}$$

- 严格对角占优矩阵和对称正定矩阵可直接进行高斯消元
- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LL}^T$, 相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LDL}^T$，相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LDL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Lz}=\mathbf{b};\\ \mathbf{Dy}=\mathbf{z};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 高斯消元法，部分列主元消去法，比列列主元消去法;
- 矩阵的$\mathbf{LU}$ 分解：Gaussian 方法，直接分解方法
- 借助原矩阵的$\mathbf{LU}$ 分解，

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LUx}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b},\\ \mathbf{Ux}=\mathbf{y},\end{cases}$$

- 严格对角占优矩阵和对称正定矩阵可直接进行高斯消元
- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LL}^T$, 相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Ly}=\mathbf{b};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

- 若$\mathbf{A}$ 为对称正定矩阵，则可分解为$\mathbf{LDL}^T$，相应的

$$\mathbf{Ax}=\mathbf{b}\Rightarrow\mathbf{LDL}^T\mathbf{x}=\mathbf{b}\Rightarrow\begin{cases}\mathbf{Lz}=\mathbf{b};\\ \mathbf{Dy}=\mathbf{z};\\ \mathbf{L}^T\mathbf{x}=\mathbf{y},\end{cases}$$

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 若$\mathbf{A}$ 为带状矩阵，尤其是三对角对称矩阵，则此为上述情况的特例。前述分解形式同样适用.
- 若把$\mathbf{Ax}=\mathbf{b}\rightarrow\mathbf{Ax}-\mathbf{b}=\mathbf{0}$ 则化为n 维线性方程组的求根问题。考虑到一元情形：$f(x)=0$ 可等价为不动点$x=g(x)$ 形式，从而借助于不动点迭代格式

$$x_{k+1}=g(x_k),k=0,1,\cdots$$

进行迭代求解.

- **问题：**对n 维形式是否也可以做类似不动点迭代形式进行求解？如果可以，如何设计迭代格式？如何确定迭代解序列$\{\mathbf{x}_k\}_0^\infty$ 收敛？收敛准则如何确定？

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 若$\mathbf{A}$ 为带状矩阵，尤其是三对角对称矩阵，则此为上述情况的特例。前述分解形式同样适用.
- 若把$\mathbf{Ax}=\mathbf{b}\rightarrow\mathbf{Ax}-\mathbf{b}=\mathbf{0}$ 则化为n 维线性方程组的求根问题。考虑到一元情形：$f(x)=0$ 可等价为不动点$x=g(x)$ 形式，从而借助于不动点迭代格式

$$x_{k+1}=g(x_k),k=0,1,\cdots$$

进行迭代求解.

- **问题：**对n 维形式是否也可以做类似不动点迭代形式进行求解？如果可以，如何设计迭代格式？如何确定迭代解序列$\{\mathbf{x}_k\}_0^\infty$ 收敛？收敛准则如何确定？

回顾线性方程组$\mathbf{Ax}=\mathbf{b}$的直接解法

- 若$\mathbf{A}$ 为带状矩阵，尤其是三对角对称矩阵，则此为上述情况的特例。前述分解形式同样适用.
- 若把$\mathbf{Ax}=\mathbf{b}\rightarrow\mathbf{Ax}-\mathbf{b}=\mathbf{0}$ 则化为n 维线性方程组的求根问题。考虑到一元情形：$f(x)=0$ 可等价为不动点$x=g(x)$ 形式，从而借助于不动点迭代格式

$$x_{k+1}=g(x_k),k=0,1,\cdots$$

进行迭代求解.

- **问题：**对n 维形式是否也可以做类似不动点迭代形式进行求解？如果可以，如何设计迭代格式？如何确定迭代解序列$\{\mathbf{x}_k\}_0^\infty$ 收敛？收敛准则如何确定？

7.1 Norms of Vectors and Matrices

- Let $\mathbb{R}^n$ denote the set of all n-dimensional column vectors with real-number components.
- To define a distance in $\mathbb{R}^n$, we use the notion of a norm.

Definition 7.1

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

1. $\|\mathbf{x}\| \geq 0$ for all $x \in \mathbb{R}^n$
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = (0, 0, \cdots, 0)^T \equiv \mathbf{0}$
3. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for all $\alpha \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$
4. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

Two classes of useful norms: l_2 norm and l_∞ norm:

Definition7.2

The l_2 and l_∞ norm for vectors $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ are defined by

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

In general, the l_2 norm is called **Euclidean Norm**.

To show that the definition of l_∞ norm satisfies the properties of definition 7.1.

- It is easy to see that, l_∞ norm follows from the similar results for absolute values, and satisfy the properties (1)-(3) obviously.
- For property (4), we can see that, if

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$$

and

$$\mathbf{y} = (y_1, y_2, \cdots, y_n)^T,$$

then

$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|_\infty &= \max_{1\le i\le n}\{|x_i+y_i|\} \\ &\le \max_{1\le i\le n}\{|x_i|+|y_i|\} \\ &= \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty. \end{aligned}$$

To show that the definition of l_2 norm satisfies the properties of definition 7.1.

- For l_2 norm, the properties of (1)-(3) of definition 7.1 are satisfied obviously.
- Next we prove it satisfies the property (4) also.
- To prove this, we need a famous inequality.

Theorem 7.3 (Cauchy-Buniakowsky-Schwarz Inequality)

For each $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ in $\mathbb{R}^n$, there has

$$\sum_{i=1}^{n} |x_i y_i| \leq \left\{ \sum_{i=1}^{n} x_i^2 \right\}^{1/2} \left\{ \sum_{i=1}^{n} y_i^2 \right\}^{1/2}$$

Proof of theorem 7.3

- Suppose $\lambda \in \mathbb{R}$, and

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^T \neq \mathbf{0}$$

and

$$\mathbf{y} = (y_1, y_2, \cdots, y_n)^T \neq \mathbf{0}$$

in $\mathbb{R}^n$.

- Thus according to the definition of l_2 norm, we obtain

$$\begin{aligned} 0 &\leq \|\mathbf{x} - \lambda \mathbf{y}\|_2^2 \\ &= \sum_{i=1}^{n} (x_i - \lambda y_i)^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\lambda \sum_{i=1}^{n} x_i y_i + \lambda^2 \sum_{i=1}^{n} y_i^2. \end{aligned}$$

- Thus

$$2\lambda \sum_{i=1}^{n} x_i y_i \leq \sum_{i=1}^{n} x_i^2 + \lambda^2 \sum_{i=1}^{n} y_i^2$$

Let

$$\lambda = \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} \neq 0$$

and substitutes into previous inequality.

$$\begin{aligned} 2\frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2}\sum_{i=1}^{n} x_i y_i &\leq \sum_{i=1}^{n} x_i^2 + \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{y}\|_2^2}\sum_{i=1}^{n} y_i^2 \\ &= \|\mathbf{x}\|_2^2 + \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{y}\|_2^2}\|\mathbf{y}\|_2^2 = 2\|\mathbf{x}\|_2^2 \end{aligned}$$

and divided by λ to produce

$$\sum_{i=1}^{n} x_i y_i \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2 = \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2}\left\{\sum_{i=1}^{n} y_i^2\right\}^{1/2}$$

- If $x_i y_i < 0$, replace x_i by $-x_i$, and call the new vector $\tilde{\mathbf{x}} = (\tilde{x}_i) = (-x_i)$.
- Then $\|\tilde{\mathbf{x}}\|_2 = \|\mathbf{x}\|_2$ and

$$\begin{aligned}\sum_{i=1}^{n} |x_i y_i| &= \sum_{i=1}^{n} \tilde{x} y_i \leq \|\tilde{\mathbf{x}}\|_2 \|\mathbf{y}\|_2 \\ &= \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2} \left\{\sum_{i=1}^{n} y_i^2\right\}^{1/2}, \blacksquare\blacksquare\blacksquare.\end{aligned}$$

Definition 7.4

If $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ are vectors in $\mathbb{R}^n$, the l_2 and l_∞ distances between $\mathbf{x}$ and $\mathbf{y}$ are defined by

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left\{ \sum_{i=1}^{n} (x_i - y_i)^2 \right\}^{1/2}$$

and

$$\|\mathbf{x} - \mathbf{y}\|_\infty = \max_{1 \le i \le n} |x_i - y_i|.$$

The Limit of a Sequence of Vectors in $\mathbb{R}^n$

Definition7.5

A sequence $\{\mathbf{x}^{(k)}\}_0^\infty$ of vectors in $\mathbb{R}^n$ is said to converge to $\mathbf{x}$ with respect to the norm$\|\cdot\|$ if, given any $\epsilon > 0$, there exists an integer $N(\epsilon)$ such that

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \leq \epsilon \quad \text{for all} \quad k \geq N(\epsilon)$$

Theorem 7.6

The sequence $\{\mathbf{x}^{(k)}\}_0^\infty$ of vectors converges to $\mathbf{x}$ in $\mathbb{R}^n$ with respect to the norm $\|\cdot\|_\infty$ if and only if

$$\lim_{k\to\infty} x_i^{(k)} = x_i \quad \text{for each} \quad i = 1, 2, \cdots, n.$$

Proof:

- Suppose that the sequence $\{\mathbf{x}^{(k)}\}_0^\infty$ of vectors converges to $\mathbf{x}$ with respect to the norm $\|\cdot\|_\infty$
- Thus given $\epsilon > 0$, there exists an integer $N(\epsilon)$, such that for all $k \geq N(\epsilon)$,

$$\max_{1 \leq i \leq n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \epsilon$$

  holds.
- This inequality implies that

$$|x_i^{(k)} - x_i| < \epsilon \text{ for each } i = 1, 2, \cdots, n.$$

  So $\lim_{k \to \infty} x_i^{(k)} = x_i$, for each $i = 1, 2, \cdots, n$.

- Conversely, suppose that $\lim_{k\to\infty} x_i^{(k)} = x_i$, for each $i = 1, 2, \cdots, n$.
- Thus for a given positive number $\epsilon > 0$, there exists an integer $N_i(\epsilon)$, such that

$$|x_i^{(k)} - x_i| < \epsilon$$

  for i, $i = 1, 2, \cdots, n$.
- Let $N(\epsilon) = \max_{1\le i\le n} N_i(\epsilon)$, then for all $k \ge N(\epsilon)$, $|x_i^{(k)} - x_i| < \epsilon, i = 1, 2, \cdots, n$ hold.
- Thus we have

$$\max_{1\le i\le n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \epsilon.$$

- This implies that the sequence $\{\mathbf{x}^{(k)}\}_0^\infty$ of vectors converges to $\mathbf{x}$.■

向量范数之间的等价性

Theorem 7.7

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

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

Proof:

- Let x_j be a coordinate of $\mathbf{x}$ such that

$$\|\mathbf{x}\|_\infty = \max_{1\le i\le n} |x_i| = |x_j|.$$

- Then

$$\|\mathbf{x}\|_\infty^2 = |x_j|^2 = x_j^2 \le \sum_{i=1}^n x_i^2 \le \sum_{i=1}^n x_j^2 = nx_j^2 = n\|\mathbf{x}\|_\infty^2.$$

- Thus

$$\|\mathbf{x}\|_\infty \le \{\sum_{i=1}^n x_i^2\}^{1/2} = \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty.\blacksquare\blacksquare\blacksquare$$

- It can be shown that all norms on $\mathbb{R}^n$ are equivalent with respect to convergence;
- that is, if $\|\cdot\|$ and $\|\cdot\|'$ are any two norms on $\mathbb{R}^n$ and $\{\mathbf{x}^k\}_{k=1}^{\infty}$ has the limit $\mathbf{x}$ with respect to $\|\cdot\|$, then $\{\mathbf{x}^k\}_{k=1}^{\infty}$ also has the limit $\mathbf{x}$ with respect to $\|\cdot\|'$.

- It can be shown that all norms on $\mathbb{R}^n$ are equivalent with respect to convergence;
- that is, if $\|\cdot\|$ and $\|\cdot\|'$ are any two norms on $\mathbb{R}^n$ and $\{\mathbf{x}^k\}_{k=1}^{\infty}$ has the limit $\mathbf{x}$ with respect to $\|\cdot\|$, then $\{\mathbf{x}^k\}_{k=1}^{\infty}$ also has the limit $\mathbf{x}$ with respect to $\|\cdot\|'$.

Definition 7.8

A matrix norm on the set of all $n \times n$ matrices is a real-valued function, $\|\cdot\|$, defined on this set, satisfying for all $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ and all real numbers α:

- $\|\mathbf{A}\| \geq 0$.
- $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A}$ is $\mathbf{0}$.
- $\|\alpha\mathbf{A}\| = \alpha\|\mathbf{A}\|$.
- $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.
- $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$.

Distance

A **distance** between $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ with respect to this matrix norm is

$$\|\mathbf{A} - \mathbf{B}\|$$

Theorem 7.9

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

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

is a matrix norm.

- This is called the **natural**, or **induced, matrix norm** associated with the vector norm.
- The following corollary is often used to bound a value of $\|\mathbf{Ax}\|$.

Corollary 7.10

For any vector $\mathbf{x} \neq 0$, matrix $\mathbf{A}$, and any natural norm $\|\cdot\|$, we have

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|.$$

Proof:

- First note that for $\mathbf{x} \neq 0$, the vector $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ has length 1.
- So by theorem 7.9 we have

$$\left\| \mathbf{A}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|} \right) \right\| \leq \|\mathbf{A}\|.$$

- Since $\|\mathbf{x}\|$ is a nonzero real number, which implies that

$$\mathbf{A}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|} \right) = \frac{1}{\|\mathbf{x}\|}\mathbf{A}\mathbf{x}$$

- Hence

$$\frac{1}{\|\mathbf{x}\|}\|\mathbf{A}\mathbf{x}\| = \left\| \frac{1}{\|\mathbf{x}\|}\mathbf{A}\mathbf{x} \right\| = \left\| \mathbf{A}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|} \right) \right\| \leq \|\mathbf{A}\|,$$

  which implies that

$$\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\| \blacksquare$$

The matrix norms we will consider have the forms

$$\|\mathbf{A}\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{Ax}\|_\infty, \texttt{the} \quad l_\infty \quad \texttt{norm}$$

and

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2, \texttt{the} \quad l_2 \quad \texttt{norm}$$

Theorem 7.11

If $\mathbf{A} = (a_{ij})$ is an $n \times n$ matrix, then

$$\|\mathbf{A}\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Proof:

First we show that

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

Let $\mathbf{x}$ be an n-dimensional column vector with

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

Since $\mathbf{Ax}$ is also an n-dimensional column vector,

$$\begin{aligned}
\|\mathbf{Ax}\|_\infty &= \max_{1\leq i\leq n} |(\mathbf{Ax})_i| \\
&= \max_{1\leq i\leq n} |\sum_{j=1}^{n} a_{ij}\, x_j| \\
&\leq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \max_{1\leq j\leq n} |x_j|
\end{aligned}$$

Since $\|\mathbf{x}\|_\infty = 1$, we have

$$\|\mathbf{Ax}\|_\infty \leq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Consequently,

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

- Now we need to show the opposite inequality, that

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

- Let p be an integer with

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

and $\mathbf{x}$ be the vector with components

$$x_j = \begin{cases} 1, \texttt{if} & a_{pj} \geq 0 \\ -1, \texttt{if} & a_{pj} < 0 \end{cases}$$

- Then $\|\mathbf{x}\|_\infty = 1$ and $a_{pj}\, x_j = |a_{pj}|$, for all $j = 1, 2, \cdots, n$, so

$$\begin{aligned} \|\mathbf{Ax}\|_\infty &= \max_{1\leq i\leq n} |\sum_{j=1}^{n} a_{ij}\, x_j| \geq |\sum_{j=1}^{n} a_{pj}\, x_j| \\ &= \left|\sum_{j=1}^{n} |a_{pj}|\right| = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \end{aligned}$$

This result implies that

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

which together with inequality previous, gives

$$\|\mathbf{A}\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.\blacksquare$$

Definition 7.12

If $\mathbf{A}$ is a square matrix, the polynomial defined by

$$p(\mathbf{A}) = \det(\mathbf{A} - \lambda \mathbf{I})$$

is called the **characteristic polynomial** of $\mathbf{A}$.

- It is not difficult to show that p is an nth-degree polynomial and, consequently, has at most n distinct zeros, some of which may be complex.
- If λ is a zero of p, then, since

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0,$$

this implies that the linear system defined by

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

has a solution other than $\mathbf{x} = \mathbf{0}$.

- We wish to study the zeros of p and the nonzero solutions corresponding to these systems.

Definition 7.13

- If p is the **characteristic polynomial** of the matrix $\mathbf{A}$, the zeros of p are called **eigenvalues**, or **characteristic values**, of the matrix $\mathbf{A}$.
- If λ is a eigenvalue of $\mathbf{A}$ and $\mathbf{x} \neq \mathbf{0}$ has the property that

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0},$$

then $\mathbf{x}$ is called an **eigenvector**, or **characteristic vector**, of $\mathbf{A}$ corresponding to the eigenvalue λ.

- If $\mathbf{x}$ is an eigenvector associated with the eigenvalue $\mathbf{A}$, then

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$$

- the matrix $\mathbf{A}$ takes the vector $\mathbf{x}$ into a scalar multiple of itself.

Definition 7.14

The **spectral radius** $\rho(\mathbf{A})$ of a matrix $\mathbf{A}$ is defined by

$$\rho(\mathbf{A}) = \max |\lambda|,$$

where λ is an eigenvalue of $\mathbf{A}$.
(**Note that** if λ is a complex number and $\lambda = \alpha + \beta i$, then we have $|\lambda| = (\alpha^2 + \beta^2)^{1/2}$).

Theorem 7.15

If $\mathbf{A}$ is an $n \times n$ matrix, then
(i) $\|\mathbf{A}\|_2 = [\rho(\mathbf{A}^T\mathbf{A})]^{1/2}$,
(ii) $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$, for any natural norm $\|\cdot\|$.

Proof:

- The proof of part (i) requires more information concerning eigenvalues than We presently have available.
- To prove part (ii), suppose λ is an eigenvalue of A with eigenvector $\mathbf{x}$ where $\|\mathbf{x}\| = 1$.
- Since $\mathbf{Ax} = \lambda\mathbf{x}$, for any natural norm

$$|\lambda| = |\lambda| \cdot \|\mathbf{x}\| = \|\lambda\mathbf{x}\| = \|\mathbf{Ax}\| \leq \|\mathbf{A}\|\|\mathbf{x}\| = \|\mathbf{A}\|.$$

- Thus,

$$\rho(\mathbf{A}) = \max|\lambda| \leq \|\mathbf{A}\|.\blacksquare$$

Example:

If

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}$$

then

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{pmatrix}$$

To calculate $\rho(\mathbf{A}^T\mathbf{A})$ we need the eigenvalues of $\mathbf{A}^T\mathbf{A}$. If

$$\begin{aligned} 0 &= \det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}) \\ &= \det\begin{pmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{pmatrix} \\ &= -\lambda^3 + 14\lambda^2 - 42\lambda = -\lambda(\lambda^2 - 14\lambda + 42) \end{aligned}$$

then

$$\lambda = 0, \texttt{or}\ \ \lambda = 7 \pm \sqrt{7},$$

so

$$\begin{aligned} \|\mathbf{A}\|_2 &= \sqrt{\rho(A^T A)} \\ &= \sqrt{\max\{0, 7+\sqrt{7}, 7-\sqrt{7}\}} = \sqrt{7+\sqrt{7}} \\ &\approx 3.106. \end{aligned}$$

Definition 7.16

We call an $n \times n$ matrix $\mathbf{A}$ convergent if

$$\lim_{k \to \infty} (\mathbf{A}^k)_{ij} = 0,$$

for each $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, n$.

Example:

Let

$$\mathbf{A} = \begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

Computing the power of A, we obtain

$$\mathbf{A}^2 = \begin{pmatrix} \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}, \mathbf{A}^3 = \begin{pmatrix} \frac{1}{8} & 0 \\ \frac{3}{16} & \frac{1}{8} \end{pmatrix}, \mathbf{A}^4 = \begin{pmatrix} \frac{1}{16} & 0 \\ \frac{1}{8} & \frac{1}{16} \end{pmatrix}, \cdots$$

and in general,

$$\mathbf{A}^k = \begin{pmatrix} (\frac{1}{2})^k & 0 \\ \frac{k}{2^{k+1}} & (\frac{1}{2})^k \end{pmatrix}.$$

Since

$$\lim_{k\to\infty} \left(\frac{1}{2}\right)^k = 0, \lim_{k\to\infty} \frac{k}{2^{k+1}} = 0,$$

so $\mathbf{A}$ is a convergent matrix.

Theorem 7.17

The following statements are equivalent.

(1) $\mathbf{A}$ is a convergent matrix.

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

(3) $\lim_{n\to\infty} \|\mathbf{A}^n\| = 0$, for all natural norm.

(4) $\rho(\mathbf{A}) < 1$.

(5) $\lim_{n\to\infty} \mathbf{A}^n\mathbf{x} = \mathbf{0}$, for every $\mathbf{x}$.

7.3 Iterative Techniques for Solving Linear Systems

- An iterative technique to solve the linear system

$$\mathbf{Ax} = \mathbf{b}$$

starts with an initial approximation $\mathbf{x}^{(0)}$ to the solution $\mathbf{x}$ and generates a sequence of vectors $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ that converges to $\mathbf{x}$.

- Iterative techniques involve a process that converts the system $\mathbf{Ax} = \mathbf{b}$ into an equivalent system of the form

$$\mathbf{x} = \mathbf{Tx} + \mathbf{c}$$

for some fixed matrix T and vector $\mathbf{c}$.

- After the initial vector $\mathbf{x}^{(0)}$ is selected, the sequence of approximate solution vectors is generated by computing

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}$$

  for each $k = 1, 2, 3, \cdots$,
- Iterative techniques are seldom used for solving linear systems of small dimension since the time required for sufficient accuracy exceeds that required for direct techniques such as the Gaussian elimination method.
- For large systems with a high percentage of zero entries, however, these techniques are efficient in terms of both computer storage and computational time.
- Systems of this type arise frequently in circuit analysis and in the numerical solution of boundary-value problems and partial differential equations.

Jacobi Iterative Method

To solve the linear system of equation in the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1i}x_i + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2i}x_i + \cdots + a_{2n}x_n = b_2 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ii}x_i + \cdots + a_{in}x_n = b_i \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{ni}x_i + \cdots + a_{nn}x_n = b_n \end{cases}$$

If $a_{ii} \neq 0$, for each $i = 1, 2, \cdots, n$, we can rewrite above equation as the follows

$$\begin{cases} x_1 = \frac{1}{a_{11}}[b_1 \quad - a_{12}x_2 - \cdots - a_{1i}x_i - \cdots - a_{1n}x_n] \\ x_2 = \frac{1}{a_{22}}[b_2 - a_{21}x_1 \quad - a_{23}x_3 - \cdots - a_{2i}x_i - \cdots - a_{2n}x_n] \\ \vdots \\ x_i = \frac{1}{a_{ii}}[b_i - a_{i1}x_1 - \cdots - a_{i,i-1}x_{i-1} \quad - a_{i,i+1}x_{i+1} - \cdots - \\ \vdots \\ x_n = \frac{1}{a_{nn}}[b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{ni}x_i - \cdots - a_{n,n-1}x_{n-1} \end{cases}$$

- Jacobi Iterative Method consists of solving the ith equation in $\mathbf{Ax} = \mathbf{b}$ for x_i to obtain (provided $a_{ii} \neq 0$)

$$x_i = \sum_{j=1, j\neq i}^{n} \left(-\frac{a_{ij} x_j}{a_{ii}} \right) + \frac{b_i}{a_{ii}}, \quad \texttt{for} \quad i = 1, 2, \cdots, n$$

- generating each $x_i^{(k)}$ from components of $\mathbf{x}^{(k-1)}$ for $k \geq 1$ by

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

for $i = 1, 2, \cdots, n$.

$$
\begin{cases}
x_1^{(k)} = \frac{1}{a_{11}}[-a_{12}x_2^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1] \\
x_2^{(k)} = \frac{1}{a_{22}}[-a_{21}x_1^{(k-1)} - a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2] \\
\vdots \\
x_i^{(k)} = \frac{1}{a_{ii}}[-a_{i1}x_1^{(k-1)} - \cdots - a_{i,i-1}x_{i-1}^{(k-1)} - a_{i,i+1}x_{i+1}^{(k-1)} - \cdots - a_{in}x_n^{(k-1)} + b_i] \\
\vdots \\
x_n^{(k)} = \frac{1}{a_{nn}}[-a_{n1}x_1^{(k-1)} - a_{n2}x_2^{(k-1)} - \cdots - a_{n,n-1}x_{n-1}^{(k-1)} + b_n]
\end{cases}
$$

算法的矩阵-向量描述

- The method is written in the form

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}$$

  by splitting $\mathbf{A}$ into its diagonal and off-diagonal parts.
- To see this, let
  - $\mathbf{D}$ be the diagonal matrix whose diagonal is the same as $\mathbf{A}$
  - $-\mathbf{L}$ be the strictly lower-triangular part of $\mathbf{A}$
  - $-\mathbf{U}$ be the strictly upper-triangular part of $\mathbf{A}$

With this notation, $\mathbf{A}$ is split into

$$
\begin{aligned}
\mathbf{A} &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \\
&= \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -a_{21} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix} \\
&\quad - \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -a_{n-1,n} \\ 0 & \cdots & 0 & 0 \end{bmatrix} \\
&= \mathbf{D} - \mathbf{L} - \mathbf{U}
\end{aligned}
$$

- The equation $\mathbf{Ax} = \mathbf{b}$, or $(\mathbf{D} - \mathbf{L} - \mathbf{U})\mathbf{x} = \mathbf{b}$, is then transformed into

$$\mathbf{Dx} = (\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}$$

  and, finally, $\mathbf{x} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}$.
- This results in the **matrix form of the Jacobi iterative technique**:

$$\mathbf{x}^{(k)} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k-1)} + \mathbf{D}^{-1}\mathbf{b}, k = 1, 2, \cdots \quad (2)$$

- Introducing the notation $\mathbf{T} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ and $\mathbf{c} = \mathbf{D}^{-1}\mathbf{b}$, the Jacobi technique has the form

$$\mathbf{x}^{(k)} = \mathbf{Tx}^{(k-1)} + \mathbf{c}. \quad (3)$$

- In practice, Eq. (1) is used in computation and Eq. (3) for theoretical purposes.

I Jacobi Iterative Algorithm

To solve $\mathbf{Ax} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$:

INPUT:
- the number of equations and unknowns n;
- the entries $a_{ij}, 1 \leq i, j \leq n$ of the matrix $\mathbf{A}$;
- the entries $b_i, 1 \leq i \leq n$ of $\mathbf{b}$;
- the entries $XO_i, 1 \leq i \leq n$ of $\mathbf{XO} = \mathbf{x}^{(0)}$;
- tolerance TOL;
- maximum number of iterations N.

OUTPUT: the approximate solution $x_1, x_2, \cdots, x_n$ or a message that the number of iterations was exceeded.

Step 1 Set $k = 1$.

Step 2 While $(k \leq N)$ do Steps 3-6.

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

$$x_i = \frac{-\sum_{j=1, j\neq i}^{n} (a_{ij} XO_j) + b_i}{a_{ii}}$$

Step 4 If $\|\mathbf{x} - \mathbf{XO}\| < TOL$ then OUTPUT $(x_1, x_2, \cdots, x_n)$; (Procedure completed successfully.) STOP.

Step 5 Set $k = k + 1$.

Step 6 For $i = 1, \cdots, n$ set $XO_i = x_i$.

Step 7 OUTPUT ('Maximum number of iterations exceeded'); (Procedure completed unsuccessfully.) STOP.

Remarks:

1. Step 3 of the algorithm requires that $a_{ii} \neq 0$ for each $i = 1, 2, \cdots, n$.
2. If one of the a_{ii} entries is zero and the system is nonsingular, a reordering of the equations (row interchange) can be performed so that no $a_{ii} = 0$.
3. To speed convergence, the equations should be arranged so that a_{ii} is as large as possible.
4. Another possible stopping criterion in Step 4 is to iterate until
$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|} < \varepsilon,$$
where $\varepsilon > 0$ is the tolerance.
5. For this purpose, any convenient norm can be used, the usual being the l_∞ norm.

Improvement in Jacobi Iterative Technique —- Gauss-Seidel iterative technique

- To compute $x_i^{(k)}$, the components of $\mathbf{x}^{(k-1)}$ are used.
- Since, for

$$i > 1, x_1^{(k)}, x_2^{(k)}, \cdots, x_{i-1}^{(k)}$$

have already been computed and are likely to be better approximations to the actual solutions

$$x_1, x_2, \cdots, x_{i-1}$$

than

$$x_1^{(k-1)}, x_2^{(k-1)}, \cdots, x_{i-1}^{(k-1)},$$

- It seems more reasonable to compute $x_i^{(k)}$ using these most recently calculated values; that is,

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

  for each $i = 1, 2, \cdots, n$, instead of Eq. (1).

This modification is called the **Gauss-Seidel iterative technique** .

In details:

$$
\begin{cases}
x_1^{(k)} = \frac{1}{a_{11}}[\qquad -a_{12}x_2^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1] \\
x_2^{(k)} = \frac{1}{a_{22}}[-a_{21}x_1^{(k)} \qquad - a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2] \\
\vdots \\
x_i^{(k)} = \frac{1}{a_{ii}}[-a_{i1}x_1^{(k)} - \cdots - a_{i,i-1}x_{i-1}^{(k)} \qquad - a_{i,i+1}x_{i+1}^{(k-1)} - \cdots - a_{in}x_n^{(k-1)} + b_i] \\
\vdots \\
x_n^{(k)} = \frac{1}{a_{nn}}[-a_{n1}x_1^{(k)} - a_{n2}x_2^{(k)} - \cdots - a_{n,n-1}x_{n-1}^{(k)} \qquad + b_n]
\end{cases}
$$

Gauss-Seidel iterative technique

- To write the Gauss-Seidel method in matrix form, multiply both sides of Eq. (4) by a_{ii} and collect all kth iterate terms to give

$$\begin{aligned} a_{i1}x_1^{(k)} + a_{i2}x_2^{(k)} + \cdots + a_{ii}x_i^{(k)} &= -a_{i,i+1}x_{i+1}^{(k-1)} - a_{i,i+2}x_{i+2}^{(k-1)} - \cdots - a_{i,n}x_n^{(k-1)} + b_i \end{aligned}$$

  for each $i = 1, 2, \cdots, n$.
- Writing all n equations gives

$$\begin{aligned} a_{11}x_1^{(k)} &= -a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} \cdots - a_{1n}x_n^{(k-1)} + b_1, \\ a_{21}x_1^{(k)} + a_{22}x_2^{(k)} &= -a_{23}x_3^{(k-1)} \cdots - a_{2n}x_n^{(k-1)} + b_2, \\ &\vdots = \vdots \\ a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} \cdots + a_{nn}x_n^{(k)} &= b_n; \end{aligned}$$

Gauss-Seidel 方法的矩阵-向量描述

- It follows that the matrix form of the Gauss-Seidel method is

$$(\mathbf{D}-\mathbf{L})\mathbf{x}^{(k)} = \mathbf{U}\mathbf{x}^{(k-1)} + \mathbf{b}$$

or

$$\mathbf{x}^{(k)} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(k-1)} + (\mathbf{D}-\mathbf{L})^{-1}\mathbf{b}, k = 1, 2, \cdots \quad (5)$$

- Letting $\mathbf{T}_g = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}$ and $\mathbf{c}_g = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{b}$, the Gauss-Seidel technique has the form

$$\mathbf{x}^{(k)} = \mathbf{T}_g\mathbf{x}^{(k-1)} + \mathbf{c}_g.$$

- For the lower-triangular matrix $\mathbf{D}-\mathbf{L}$ to be nonsingular, it is necessary and sufficient that $a_{ii} \neq 0$ for each $i = 1, 2, \cdots, n$.

II Gauss-Seidel Iterative Algorithm

To solve $\mathbf{Ax} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$:

INPUT: the number of equations and unknowns n; the entries $a_{ij}, 1 \leq i, j \leq n$ of the matrix A; the entries $b_i, 1 \leq i \leq n$ of $\mathbf{b}$; the entries $XO_i, 1 \leq i \leq n$ of $\mathbf{XO} = \mathbf{x}^{(0)}$; tolerance TOL; maximum number of iterations N.

OUTPUT the approximate solution $x_1, x_2, \cdots, x_n$ or a message that the number of iterations was exceeded.

Continued

Step 1 Set $k = 1$.

Step 2 While $(k \leq N)$ do Steps 3-6.

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

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

Step 4 If $\|\mathbf{x} - \mathbf{XO}\| < TOL$ then OUTPUT $(x_1, x_2, \cdots, x_n)$; (Procedure completed successfully.) STOP.

Step 5 Set $k = k + 1$

Step 6 For $i = 1, \cdots, n$ set $XO_i = x_i$.

Step 7 OUTPUT ('Maximum number of iterations exceeded'); (Procedure completed unsuccessfully.) STOP.

Remarks:

1. Step 3 of the algorithm requires that $a_{ii} \neq 0$ for each $i = 1, 2, \cdots, n$. If one of the a_{ii} entries is zero and the system is nonsingular, a reordering of the equations can be performed so that no $a_{ii} = 0$.

2. To speed convergence, the equations should be rearranged so that a_{ii} is as large as possible.

3. Another possible stopping criterion in Step 4 is to iterate until

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

where $\varepsilon > 0$ is the tolerance. For this purpose, any convenient norm can be used, the usual being the l_∞ norm.

III. Convergence Analysis for Two Iterative Techniques

To study the convergence of general iteration techniques, we consider the formula

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}.$$

where $\mathbf{x}^{(0)}$ is arbitrary.

Lemma 7.18

If the spectral radius $\rho(\mathbf{T})$ satisfies $\rho(\mathbf{T}) < 1$, then $(\mathbf{I} - \mathbf{T})^{-1}$ exists, and

$$(\mathbf{I} - \mathbf{T})^{-1} = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots .$$

Proof:

- Since $\mathbf{Tx} = \lambda\mathbf{x}$ is true precisely when

$$(\mathbf{I} - \mathbf{T})\mathbf{x} = (1 - \lambda)\mathbf{x},$$

  we have λ as an eigenvalue of $\mathbf{T}$ precisely when $1 - \lambda$ is an eigenvalue of $\mathbf{I} - \mathbf{T}$.
- But $|\lambda| \leq \rho(\mathbf{T}) < 1$, So $\lambda = 1$ is not an eigenvalue of $\mathbf{T}$, and 0 cannot be an eigenvalue of $\mathbf{I} - \mathbf{T}$.
- Hence $\mathbf{I} - \mathbf{T}$ is nonsingular.
- Let

$$\mathbf{S}_m = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots + \mathbf{T}^m.$$

- Then

$$\begin{aligned}(\mathbf{I}-\mathbf{T})\mathbf{S}_m &= (\mathbf{I}+\mathbf{T}+\mathbf{T}^2+\ldots+\mathbf{T}^m)\\ &\quad -(\mathbf{T}+\mathbf{T}^2+\cdots+\mathbf{T}^{m+1})\\ &= \mathbf{I}-\mathbf{T}^{m+1}.\end{aligned}$$

- Since $\rho(\mathbf{T})<1$, the result at the end of Section 7.2 implies that $\mathbf{T}$ is convergent and

$$\begin{aligned}\lim_{m\to\infty}(\mathbf{I}-\mathbf{T})\mathbf{S}_m &= \lim_{m\to\infty}(\mathbf{I}-\mathbf{T}^{m+1})\\ &= \mathbf{I}.\end{aligned}$$

- Thus,

$$(\mathbf{I}-\mathbf{T})^{-1}=\lim_{m\to\infty}\mathbf{S}_m=\mathbf{I}+\mathbf{T}+\mathbf{T}^2+\cdots\ .\blacksquare\blacksquare\blacksquare.$$

Theorem 7.19

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

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}, \quad \texttt{for each} \quad k \geq 1. \tag{6}$$

converges to the unique solution of $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$ if and only if $\rho(\mathbf{T}) < 1$.

Proof:

First assume that $\rho(\mathbf{T}) < 1$. From Eq. (6),

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

Since $\rho(\mathbf{T}) < 1$, the matrix $\mathbf{T}$ is convergent and

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

- Lemma 7.18 implies that

$$\begin{aligned}\lim_{k\to\infty}\mathbf{x}^{(k)} &= \lim_{k\to\infty}\mathbf{T}^k\mathbf{x}^{(0)} + \lim_{k\to\infty}\Big(\sum_{j=0}^{k-1}\mathbf{T}^j\Big)\mathbf{c} \\ &= \mathbf{0} + (\mathbf{I}-\mathbf{T})^{-1}\mathbf{c} = (\mathbf{I}-\mathbf{T})^{-1}\mathbf{c}.\end{aligned}$$

- Since $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$ implies that $(\mathbf{I}-\mathbf{T})\mathbf{x} = \mathbf{c}$, the sequence $\{\mathbf{x}^{(k)}\}$ converges to the unique solution to the equation, the vector $\mathbf{x} = (\mathbf{I}-\mathbf{T})^{-1}\mathbf{c}$.

- To prove the converse, we show that for any $\mathbf{z} \in \mathbb{R}^n$ we have

$$\lim_{k\to\infty} \mathbf{T}^k \mathbf{z} = 0$$

- Let $\mathbf{x}$ be the unique solution to the equation $\mathbf{x} = \mathbf{T}\mathbf{x}$, that is, Eq. (6) with $\mathbf{c} = \mathbf{0}$.
- For $\mathbf{x}^{(0)} = \mathbf{x} - \mathbf{z}$, we have

$$\begin{aligned}
\lim_{k\to\infty} \mathbf{T}^k \mathbf{z} &= \lim_{k\to\infty} \mathbf{T}^k (\mathbf{x} - \mathbf{x}^{(0)}) \\
&= \lim_{k\to\infty} \mathbf{T}^{k-1} (\mathbf{T}\mathbf{x} - \mathbf{T}\mathbf{x}^{(0)}) \\
&= \lim_{k\to\infty} \mathbf{T}^{k-1} (\mathbf{x} - \mathbf{x}^{(1)}).
\end{aligned}$$

- Continuing in this manner, we have

$$\begin{aligned}
\lim_{k\to\infty} \mathbf{T}^k\mathbf{z} &= \lim_{k\to\infty} \mathbf{T}^{k-1}(\mathbf{x}-\mathbf{x}^{(1)}) \\
&= \lim_{k\to\infty} \mathbf{T}^{k-2}(\mathbf{x}-\mathbf{x}^{(2)}) \\
&\vdots \\
&= \lim_{k\to\infty} (\mathbf{x}-\mathbf{x}^{(k)}) = \mathbf{0}.
\end{aligned}$$

- Since $\mathbf{z} \in \mathbb{R}^n$ was arbitrary, Theorem 7.17 implies that $\mathbf{T}$ is a convergent matrix and that $\rho(\mathbf{T}) < 1$.■■■

Corollary 7.20

If $\|\mathbf{T}\| < 1$ for any natural matrix norm and $\mathbf{c}$ is a given vector, then the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ defined by

$$\mathbf{x}^{(k)} = \mathbf{T}\mathbf{x}^{(k-1)} + \mathbf{c}$$

converges, for any $\mathbf{x}^{(0)} \in \mathbb{R}^n$, to a vector $\mathbf{x} \in \mathbb{R}^n$, and the following error bounds hold:

1. $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \|\mathbf{T}\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\|$.
2. $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \frac{\|\mathbf{T}\|^k}{1-\|\mathbf{T}\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$.

The convergence analysis on the techniques: Jacobi and Gauss-seidel Iterative methods

- For Jacobi Iterative method, it can be written as

$$\mathbf{x}^{(k)} = \mathbf{T}_j \mathbf{x}^{(k-1)} + \mathbf{c}_j$$

  with $\mathbf{T}_j = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}), \mathbf{c}_j = \mathbf{D}^{-1}\mathbf{b}$
- For Gauss-Seidel Iterative method, it can be written as

$$\mathbf{x}^{(k)} = \mathbf{T}_g \mathbf{x}^{(k-1)} + \mathbf{c}_g$$

  where $\mathbf{T}_g = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}, \mathbf{c}_g = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{b}$.
- If $\rho(\mathbf{T}_j)$ or $\rho(\mathbf{T}_g)$ is less than 1, then the corresponding sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ will converge to the solution $\mathbf{x}$ of $\mathbf{Ax} = \mathbf{b}$.

For example, the Jacobi scheme has

$$\mathbf{x}^{(k)} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k-1)} + \mathbf{D}^{-1}\mathbf{b},$$

and, if $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converges to $\mathbf{x}$, then

$$\mathbf{x} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b},$$

this implies that

$$\mathbf{D}\mathbf{x} = (\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b},$$

and

$$(\mathbf{D} - \mathbf{L} - \mathbf{U})\mathbf{x} = \mathbf{b},$$

that is

$$\mathbf{A}\mathbf{x} = \mathbf{b}.$$

The Sufficiency Conditions for Convergence of the Jacobi and Gauss-Seidel Methods

Theorem 7.21

If $\mathbf{A}$ is strictly diagonally dominant, then for any choice of $\mathbf{x}^{(0)}$, both the Jacobi and Gauss- Seidel methods give sequences $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ that converge to the unique solution of $\mathbf{Ax} = \mathbf{b}$.

Theorem 7.22 (Stein-Rosenberg)

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

- a. $0 \leq \rho(\mathbf{T}_g) < \rho(\mathbf{T}_j) < 1$.
- b. $1 < \rho(\mathbf{T}_j) < \rho(\mathbf{T}_g)$.
- c. $\rho(\mathbf{T}_j) = \rho(\mathbf{T}_g) = 0$
- d. $\rho(\mathbf{T}_j) = \rho(\mathbf{T}_g) = 1$.

For the special case described in Theorem 7.22, we see:

- Part (a) shows that when one method gives convergence, then both give convergence, and the Gauss-Seidel method converges faster than the Jacobi method.
- Part (b) indicates that when one method diverges then both diverge, and the divergence is more pronounced for the Gauss-Seidel method.

Definition 7.23

- Suppose $\tilde{\mathbf{x}} \in \mathbb{R}^n$ is an approximation to the solution of the linear system defined by $\mathbf{Ax} = \mathbf{b}$.
- The **residual vector** for $\tilde{\mathbf{x}}$ with respect to this system is $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$.

In procedures such as the Jacobi or Gauss-Seidel methods, a residual vector is associated with each calculation of an approximation component to the solution vector.

- The object of the method is to generate a sequence of approximations that will cause the associated residual vectors to converge rapidly to zero.
- Suppose we let

$$\mathbf{r}_i^{(k)} = \left(r_{i1}^{(k)}, r_{i2}^{(k)}, \cdots, r_{in}^{(k)}\right)^T$$

denote the residual vector for the Gauss-Seidel method corresponding to the approximate solution vector $\mathbf{x}_i^{(k)}$ defined by

$$\mathbf{x}_i^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \cdots, x_{i-1}^{(k)}, x_i^{(k-1)}, \cdots, x_n^{(k-1)}\right)^T$$

The mth component of $\mathbf{r}_i^{(k)}$ is

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^{n} a_{mj} x_j^{(k-1)}, \tag{7}$$

or, equivalently,

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)},$$

for each $m = 1, 2, \cdots, n$.

In particular, the ith component of $\mathbf{r}_i^{(k)}$ is

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

so

$$a_{ii} x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)}, \qquad (8)$$

Recall, however, that in the Gauss-Seidel method, $x_i^{(k)}$ is chosen to be

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right]. \qquad (9)$$

so Eq. (8) can be rewritten as

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}x_i^{(k)}. \tag{10}$$

Consequently, the Gauss-Seidel method can be characterized as choosing $x_i^{(k)}$ to satisfy

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}. \tag{11}$$

We can derive another connection between the residual vectors and the Gauss- Seidel technique.

Consider the residual vector $\mathbf{r}_{i+1}^{(k)}$ associated with the vector

$$\mathbf{x}_{i+1}^{(k)} = (x_1^{(k)}, \cdots, x_i^{(k)}, x_{i+1}^{(k-1)}, \cdots, x_n^{(k-1)})^T,$$

by (7), the ith component of $\mathbf{r}_{i+1}^{(k)}$ is

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

Equation (9) implies that $r_{i,i+1}^{(k)} = 0$.

In a sense, then, the Gauss-Seidel technique is also characterized by choosing $x_i^{(k)}$ in such a way that the ith component of $\mathbf{r}_{i+1}^{(k)}$ is zero.

III. Relaxation Methods

- Reducing one coordinate of the residual vector to zero, however, is not generally the most efficient way to reduce the overall size of the vector $\mathbf{r}_{i+1}^{(k)}$.
- Instead, we need to choose $x_i^{(k)}$ so that $\|\mathbf{r}_{i+1}^{(k)}\|$ is small.
- Modifying the Gauss-Seidel procedure as given by Eq. (11) to

$$x_i^{(k)} = x_i^{(k-1)} + \omega \frac{r_{ii}^{(k)}}{a_{ii}}. \tag{12}$$

for certain choices of positive ω reduces the norm of the residual vector and leads to significantly faster convergence.

- Methods involving Eq. (12) are called **relaxation methods**.
- For choices of ω with $0 < \omega < 1$, the procedures are called **under-relaxation methods** and can be used to obtain convergence of some systems that are not convergent by the Gauss-Seidel method.
- For choices of ω with $1 < \omega$, the procedures are called **over-relaxation methods**, which are used to accelerate the convergence for systems that are convergent by the Gauss-Seidel technique.
- These methods are abbreviated **SOR**, for **Successive Over-Relaxation**, and are particularly useful for solving the linear systems that occur in the numerical solution of certain partial-differential equations.

Note that by using Eq. (7) with $m = i$, Eq. (12) can be reformulated for calculation purposes to

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

SOR 方法的矩阵描述

To determine the matrix of the SOR method, we rewrite this as

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

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

so

$$(\mathbf{D} - \omega\mathbf{L})\mathbf{x}^{(k)} = [(1-\omega)\mathbf{D} + \omega\mathbf{U}]\mathbf{x}^{(k-1)} + \omega\mathbf{b}.$$

or

$$\begin{aligned}\mathbf{x}^{(k)} &= (\mathbf{D} - \omega L)^{-1}[(1-\omega)\mathbf{D} + \omega\mathbf{U}]\mathbf{x}^{(k-1)} \\ &\quad +\omega(\mathbf{D} - \omega\mathbf{L})^{-1}\mathbf{b}.\end{aligned}$$

If we let

$$\mathbf{T}_\omega = (\mathbf{D} - \omega\mathbf{L})^{-1}[(1-\omega)\mathbf{D} + \omega\mathbf{U}]$$

and

$$\mathbf{c}_\omega = \omega(\mathbf{D} - \omega\mathbf{L})^{-1}\mathbf{b}$$

we can express the SOR technique in the form

$$\mathbf{x}^{(k)} = \mathbf{T}_\omega\mathbf{x}^{(k-1)} + \mathbf{c}_\omega.$$

Theorem 7.24 (Kahan)

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

$$\rho(\mathbf{T}_\omega) \geq |\omega - 1|.$$

- This implies that the SOR method can converge only if $0 < \omega < 2$.

Theorem 7.25 (Ostrowski-Reich)

If $\mathbf{A}$ is a positive definite matrix and

$$0 < \omega < 2,$$

then the SOR method converges for any choice of initial approximate vector $\mathbf{x}^{(0)}$.

Theorem 7.26

If $\mathbf{A}$ is positive definite and tridiagonal, then

$$\rho(\mathbf{T}_g) = [\rho(\mathbf{T}_j)]^2 < 1,$$

and the optimal choice of ω for the SOR method is

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

with this choice of ω, we have $\rho(\mathbf{T}_\omega) = \omega - 1$.

III. SOR Algorithm:

To solve $\mathbf{Ax} = \mathbf{b}$ given the parameter ω and an initial approximation $\mathbf{x}^{(0)}$:

INPUT the number of equations and unknowns n; the entries $a_{ij}, 1 \leq i, j \leq n$, of the matrix A; the entries $b_i, 1 \leq i \leq n$, of $\mathbf{b}$; the entries $XO_i, 1 \leq i \leq n$, of $\mathbf{XO} = \mathbf{x}^{(0)}$; the parameter ω; tolerance TOL; maximum number of iterations N.

OUTPUT the approximate solution $x_1, \cdots, x_n$ or a message that the number of iterations was exceeded.

III. SOR Algorithm:

Step 1 Set $k = 1$.

Step 2 While $(k \leq N)$ do Steps 3-6.

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

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

Step 4 If $\|\mathbf{x} - \mathbf{XO}\| < TOL$, then OUTPUT $(x_1, \cdots, x_n)$; (Procedure completed successfully.) STOP.

Step 5 Set $k = k + 1$.

Step 6 For $i = 1, \cdots, n$ set $XO_i = x_i$.

Step 7 OUTPUT ('Maximum number of iterations exceeded'); (Procedure completed unsuccessfully.) STOP.

7.4 Error Estimates and Iterative Refinement

- If $\widetilde{\mathbf{x}}$ is an approximation to the solution $\mathbf{x}$ of

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

and the residual vector

$$\mathbf{r} = \mathbf{b} - A\widetilde{\mathbf{x}}$$

has the property that $\|\mathbf{r}\|$ is small.

- then

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\|$$

would be small as well

- This is often the case, but certain systems, which occur frequently in practice, fail to have this property.

Theorem 7.27

Suppose that $\widetilde{\mathbf{x}}$ is an approximation to the solution of

$$\mathbf{Ax} = \mathbf{b},$$

$\mathbf{A}$ is a nonsingular matrix and $\mathbf{r}$ is the residual vector for $\widetilde{\mathbf{x}}$. Then for any norm,

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\| \leq \|\mathbf{r}\| \cdot \|\mathbf{A}^{-1}\|$$

and

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}, \tag{13}$$

provided $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$.

Proof

Since $\mathbf{r} = \mathbf{b} - \mathbf{A}\widetilde{\mathbf{x}} = \mathbf{A}\mathbf{x} - \mathbf{A}\widetilde{\mathbf{x}}$ and $\mathbf{A}$ is nonsingular,

$$\mathbf{x} - \widetilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{r}.$$

Corolllary 7.10 in section 7.1 implies that

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\| = \|\mathbf{A}^{-1}\mathbf{r}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{r}\|.$$

Moreover, since $\mathbf{b} = \mathbf{A}\mathbf{x}$, we have $\|\mathbf{b}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$, so

$$\frac{1}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|},$$

and

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

- The inequalities in Theorem 7.27 imply that the quantities $\|\mathbf{A}^{-1}\|$ and $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ provided an indication of the connection between the residual vector and the accuracy of the approximation.
- In general, the relative error

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\|/\|\mathbf{x}\|$$

  is of most interest
- by Inequality (7.18), this error is bounded by the product of $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ with the relative residual for this approximation, $\|\mathbf{r}\|/\|\mathbf{b}\|$.
- Any convenient norm can be used for this approximation; the only requirement is that it be used consistently throughout

Theorem 7.28

condition number of the nonsingular matrix $\mathbf{A}$ relative to a norm $\|\cdot\|$ is

$$K(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|.$$

- With this notation,the inequalities in Theorem 7.27 become

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\| \leq K(\mathbf{A})\|\mathbf{r}\|/\|\mathbf{A}\|$$

and

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq K(\mathbf{A})\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

- For any nonsingular matrix $\mathbf{A}$ and natural $\|\cdot\|$,

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

- A matrix $\mathbf{A}$ is **well-conditioned** if $K(\mathbf{A})$ is close to 1, and is **ill-conditioned** when $K(\mathbf{A})$ is significantly greater than 1.
- Condition in this context refers to the relative security that a small residual vector implies a correspondingly accurate approximate solution.

Theorem 7.29

- Suppose $\mathbf{A}$ is nonsingular, and

$$\|\delta\mathbf{b}\| < \frac{1}{\|\mathbf{A}^{-1}\|}.$$

- The solution $\widetilde{\mathbf{x}}$ to

$$(\mathbf{A} + \delta\mathbf{A})\widetilde{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$$

approximates to the solution $\mathbf{x}$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$ with error estimate

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{K(\mathbf{A})}{1 - K(\mathbf{A})(\|\delta\mathbf{A}\|/\|\mathbf{A}\|)}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\right).$$