线性方程组的敏度分析与消去法的舍入误差分析

邓建松

2022 年 9 月 8 日

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素

对解的影响

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素
  - 数据不准确

  对解的影响

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素
  - 数据不准确
  - 机器的有限精度

  对解的影响

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素
    - 数据不准确
    - 机器的有限精度

  对解的影响
- 为此，我们需要用范数来描述向量与矩阵的扰动大小

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足
  - 正定性：$\forall x \in \mathbb{R}^n$, $\|x\| \geqslant 0$; 且$\|x\| = 0$当且仅当$x = 0$

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足
  - 正定性：$\forall x \in \mathbb{R}^n$, $\|x\| \geqslant 0$; 且$\|x\| = 0$当且仅当$x = 0$
  - 齐次性：$\forall x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $\|\alpha x\| = |\alpha|\,\|x\|$

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足
  - 正定性：$\forall x \in \mathbb{R}^n$, $\|x\| \geqslant 0$; 且$\|x\| = 0$当且仅当$x = 0$
  - 齐次性：$\forall x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $\|\alpha x\| = |\alpha|\,\|x\|$
  - 三角不等式：$\forall x, y \in \mathbb{R}^n$, $\|x + y\| \leqslant \|x\| + \|y\|$

向量范数的性质

由范数的定义，易得

- $\forall x, y \in \mathbb{R}^n$,

$$\big|\|x\| - \|y\|\big| \leqslant \|x - y\| \leqslant \max_{1\leqslant i\leqslant n} \|e_i\| \sum_{i=1}^{n} |x_i - y_i|$$

向量范数的性质

由范数的定义，易得

- $\forall x, y \in \mathbb{R}^n$,

$$\big|\|x\| - \|y\|\big| \leqslant \|x - y\| \leqslant \max_{1 \leqslant i \leqslant n} \|e_i\| \sum_{i=1}^{n} |x_i - y_i|$$

- 所以$\|\cdot\|$作为$\mathbb{R}^n$上的实函数是连续的

p范数

最典型的一类范数是p范数

- 也称为Hölder范数

p范数

最典型的一类范数是p范数

- 也称为Hölder范数
- $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}, \ p \geqslant 1$

p范数

最典型的一类范数是p范数

- 也称为Hölder范数
- $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$, $p \geqslant 1$
- $p = 1, 2, \infty$是最常见的，也是最重要的

常用的p范数

- 1范数：$\|x\|_1 = |x_1| + \cdots + |x_n|$

常用的p范数

- 1范数：$\|x\|_1 = |x_1| + \cdots + |x_n|$
- 2范数：

$$\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2} = \sqrt{x^T x}$$

常用的p范数

- 1范数：$\|x\|_1 = |x_1| + \cdots + |x_n|$
- 2范数：
  $$\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2} = \sqrt{x^T x}$$
- ∞范数：$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$

范数的等价性

定理

设$\|\cdot\|_\alpha$和$\|\cdot\|_\beta$是$\mathbb{R}^n$上任意两个范数，则存在正常数c_1，c_2，使得对一切$x \in \mathbb{R}^n$，

$$c_1\|x\|_\alpha \leqslant \|x\|_\beta \leqslant c_2\|x\|_\alpha$$

例

- (数分练习8.1 第6题)

$$\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2$$

例

- (数分练习8.1 第6题)

$$\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2$$

- (数分练习8.1 第7题)

$$\|x\|_\infty \leqslant \|x\|_2 \leqslant \sqrt{n}\|x\|_\infty$$

例

- (数分练习8.1 第6题)

$$\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2$$

- (数分练习8.1 第7题)

$$\|x\|_\infty \leqslant \|x\|_2 \leqslant \sqrt{n}\|x\|_\infty$$

- 显然有

$$\|x\|_\infty \leqslant \|x\|_1 \leqslant n\|x\|_\infty$$

收敛定理

定理

设$x_k \in \mathbb{R}^n$，则$\lim_{k\to\infty} \|x_k - x\| = 0$当且仅当

$$\lim_{k\to\infty} \left|x_i^{(k)} - x_i\right| = 0, \quad i = 1, \ldots, n$$

即向量序列的范数收敛等价于其分量收敛

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数
- 这种范数称为广义矩阵范数

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数
- 这种范数称为广义矩阵范数
- 矩阵的尺寸是任意的，并不需要是方阵

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数
- 这种范数称为广义矩阵范数
- 矩阵的尺寸是任意的，并不需要是方阵
- 如此定义没有考虑矩阵的乘法运算

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A \in \mathbb{R}^{n\times n}$, $\|A\| \geqslant 0$; 且$\|A\| = 0$当且仅当$A = 0$

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A \in \mathbb{R}^{n\times n}$, $\|A\| \geqslant 0$; 且$\|A\| = 0$当且仅当$A = 0$
- 齐次性：$\forall A \in \mathbb{R}^{n\times n}$, $\alpha \in \mathbb{R}$, $\|\alpha A\| = |\alpha|\,\|A\|$

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A \in \mathbb{R}^{n\times n}$, $\|A\| \geqslant 0$; 且$\|A\| = 0$当且仅当$A = 0$
- 齐次性：$\forall A \in \mathbb{R}^{n\times n}$, $\alpha \in \mathbb{R}$, $\|\alpha A\| = |\alpha| \, \|A\|$
- 三角不等式：$\forall A, B \in \mathbb{R}^{n\times n}$, $\|A + B\| \leqslant \|A\| + \|B\|$

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A \in \mathbb{R}^{n\times n}$, $\|A\| \geqslant 0$; 且$\|A\| = 0$当且仅当$A = 0$
- 齐次性：$\forall A \in \mathbb{R}^{n\times n}$, $\alpha \in \mathbb{R}$, $\|\alpha A\| = |\alpha|\,\|A\|$
- 三角不等式：$\forall A, B \in \mathbb{R}^{n\times n}$, $\|A+B\| \leqslant \|A\| + \|B\|$
- 相容性：$\forall A, B \in \mathbb{R}^{n\times n}$, $\|AB\| \leqslant \|A\|\,\|B\|$

矩阵范数的性质

- $\mathbb{R}^{n\times n}$上的矩阵范数可以看作$\mathbb{R}^{n^2}$上的向量范数

矩阵范数的性质

- $\mathbb{R}^{n\times n}$上的矩阵范数可以看作$\mathbb{R}^{n^2}$上的向量范数
- $\mathbb{R}^{n\times n}$上的任意两个矩阵范数是等价的

矩阵范数的性质

- $\mathbb{R}^{n\times n}$上的矩阵范数可以看作$\mathbb{R}^{n^2}$上的向量范数
- $\mathbb{R}^{n\times n}$上的任意两个矩阵范数是等价的
- 矩阵序列的范数收敛等价于逐元素收敛

矩阵范数与向量范数的协调性

- 矩阵与向量的乘积是矩阵计算中经常用到的运算

矩阵范数与向量范数的协调性

- 矩阵与向量的乘积是矩阵计算中经常用到的运算
- 自然期望矩阵范数与向量范数之间具有某种协调性：把向量看作特殊的矩阵，那么定义中的相容性就是所期望的性质

相容性

- 若矩阵范数$\|\cdot\|_M$和向量范数$\|\cdot\|_v$满足

  $$\|Ax\|_v \leqslant \|A\|_M\|x\|_v, \quad \forall A \in \mathbb{R}^{n\times n}, x \in \mathbb{R}^n$$

  则称矩阵范数$\|\cdot\|_M$与$\|\cdot\|_v$是**相容的**

相容性

- 若矩阵范数$\|\cdot\|_M$和向量范数$\|\cdot\|_v$满足
$$\|Ax\|_v \leqslant \|A\|_M\|x\|_v, \quad \forall A \in \mathbb{R}^{n\times n}, x \in \mathbb{R}^n$$
则称矩阵范数$\|\cdot\|_M$与$\|\cdot\|_v$是相容的
- 今后凡同时涉及到向量范数与矩阵范数，都假定它们是相容的

相容性

- 若矩阵范数$\|\cdot\|_M$和向量范数$\|\cdot\|_v$满足

$$\|Ax\|_v \leqslant \|A\|_M \|x\|_v, \quad \forall A \in \mathbb{R}^{n\times n}, x \in \mathbb{R}^n$$

则称矩阵范数$\|\cdot\|_M$与$\|\cdot\|_v$是**相容的**
- 今后凡同时涉及到向量范数与矩阵范数，都假定它们是相容的
- 对任意给定的向量范数，我们都可以定义一个矩阵范数，使之与向量范数是相容的

从属于向量范数的矩阵范数

定理

设$\|\cdot\|$是$\mathbb{R}^n$上的一个向量范数。若定义

$$\|A\| = \max_{\|x\|=1} \|Ax\|, \quad A \in \mathbb{R}^{n\times n}$$

则$\|A\|$是$\mathbb{R}^{n\times n}$上的一个矩阵范数

证明

- 最大值可达到：$\|x\| = 1$是有界闭集，$\|\cdot\|$是连续的

证明

- 最大值可达到：$\|x\| = 1$是有界闭集，$\|\cdot\|$是连续的
- 矩阵与向量乘法的相容性：

$$\frac{\|Ax\|}{\|x\|} = \left\|A\frac{x}{\|x\|}\right\| \leqslant \|A\|$$

$$\Longrightarrow \|Ax\| \leqslant \|A\|\|x\|, \forall x \in \mathbb{R}^n$$

证明

- 正定性：取$A \neq 0$, 不妨设A的第i列是非零向量，即$Ae_i \neq 0$. 则

$$0 < \|Ae_i\| \leqslant \|A\| \|e_i\| \Longrightarrow \|A\| > 0$$

证明

- 正定性：取$A \neq 0$, 不妨设A的第i列是非零向量，即$Ae_i \neq 0$. 则

$$0 < \|Ae_i\| \leqslant \|A\|\|e_i\| \implies \|A\| > 0$$

- 齐次性：显然

证明

- 三角不等式：取x, $\|x\| = 1$使得$\|(A+B)x\| = \|A+B\|$, 则

$$\begin{aligned}\|A+B\| &= \|(A+B)x\| \leqslant \|Ax\| + \|Bx\| \\ &\leqslant \|A\|\|x\| + \|B\|\|x\| = \|A\| + \|B\|\end{aligned}$$

证明

- 三角不等式：取x, $\|x\| = 1$使得$\|(A+B)x\| = \|A+B\|$, 则

$$\begin{aligned}\|A+B\| = \|(A+B)x\| &\leqslant \|Ax\| + \|Bx\| \\ &\leqslant \|A\|\|x\| + \|B\|\|x\| = \|A\| + \|B\|\end{aligned}$$

- 相容性：取x, $\|x\| = 1$使得$\|ABx\| = \|AB\|$,

$$\begin{aligned}\|AB\| = \|ABx\| &\leqslant \|A\|\|Bx\| \\ &\leqslant \|A\|\|B\|\|x\| = \|A\|\|B\|\end{aligned}$$

算子范数

- 称由

$$\|A\| = \max_{\|x\|=1} \|Ax\|, \quad A \in \mathbb{R}^{n\times n}$$

所定义的范数$\|\cdot\|$为从属于向量范数$\|\cdot\|$的矩阵范数，也称为由向量范数$\|\cdot\|$诱导出的算子范数

算子范数

- 称由

$$\|A\| = \max_{\|x\|=1} \|Ax\|, \quad A \in \mathbb{R}^{n\times n}$$

所定义的范数$\|\cdot\|$为从属于向量范数$\|\cdot\|$的矩阵范数，也称为由向量范数$\|\cdot\|$诱导出的算子范数

- 今后仍把$\|\cdot\|$记作$\|\cdot\|$

单位阵的范数

- 对广义矩阵范数，$\|I\| > 0$

单位阵的范数

- 对广义矩阵范数，$\|I\| > 0$
- 对矩阵范数，$\|I\| \geqslant 1$

单位阵的范数

- 对广义矩阵范数，$\|I\| > 0$
- 对矩阵范数，$\|I\| \geqslant 1$
- 对算子范数，$\|I\| = 1$

算子范数$\|\cdot\|_p$

- 前面定义了向量的p范数

算子范数$\|\cdot\|_p$

- 前面定义了向量的p范数
- 由其诱导出矩阵的p范数如下：

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

算子范数$\|\cdot\|_p$

- 前面定义了向量的p范数
- 由其诱导出矩阵的p范数如下：

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

- $p = 1, 2, \infty$时对应的算子范数具有简捷的表示

“列和”范数

对于矩阵$A = (a_{ij}) \in \mathbb{R}^{n\times n}$, 我们有

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|$$

- 记A的第j列为向量a_j, 不妨设$\delta = \|a_1\| = \max\limits_{1\leqslant j\leqslant n} \|a_j\|_1$

“列和”范数

对于矩阵$A = (a_{ij}) \in \mathbb{R}^{n\times n}$, 我们有

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|$$

- 记A的第j列为向量a_j, 不妨设$\delta = \|a_1\| = \max\limits_{1\leqslant j\leqslant n} \|a_j\|_1$
- 取x, $\|x\|_1 = 1$, 则

$$\|Ax\|_1 = \left\|\sum_{j=1}^{n} x_j a_j\right\|_1 \leqslant \sum_{j=1}^{n} |x_j| \|a_j\|_1 \leqslant \delta$$

- 注意到$\|e_1\|_1 = 1$

- 注意到$\|e_1\|_1 = 1$
- $\|Ae_1\|_1 = \|a_1\|_1 = \delta$

- 注意到$\|e_1\|_1 = 1$
- $\|Ae_1\|_1 = \|a_1\|_1 = \delta$
- 所以$\|A\|_1 = \delta = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|$

“行和”范数

对于矩阵$A = (a_{ij}) \in \mathbb{R}^{n\times n}$, 我们有

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

- 对于任意x, $\|x\|_\infty = 1$, 我们有

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

- 不妨设A中第1行对应向量的1范数最大，取

$$\bar{x} = (\operatorname{sgn} a_{11}, \ldots, \operatorname{sgn} a_{1n})^T$$

- 不妨设A中第1行对应向量的1范数最大，取

$$\bar{x} = (\operatorname{sgn} a_{11}, \ldots, \operatorname{sgn} a_{1n})^T$$

- $A \neq 0 \Longrightarrow \|\bar{x}\|_\infty = 1$, 而且

$$\|A\bar{x}\|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|$$

谱范数

对于矩阵$A = (a_{ij}) \in \mathbb{R}^{n\times n}$, 我们有

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$$

这里$\lambda_{\max}$表示参数矩阵的最大特征值

- 根据定义

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \sqrt{x^T A^T A x}$$

- 设对称半正定矩阵A^TA的特征值为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_n \geqslant 0$$

对应规范正交特征向量为$v_1, \cdots, v_n$

- 设对称半正定矩阵A^TA的特征值为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_n \geqslant 0$$

对应规范正交特征向量为$v_1, \cdots, v_n$

- 取x, $\|x\|_2 = 1$, 则

$$x = \sum_{i=1}^{n} \alpha_i v_i, \quad \sum_{i=1}^{n} \alpha_i^2 = 1$$

- $$x^T A^T A x = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$

- $$x^TA^TAx = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$
- 取$x = v_1$, 则
$$x^TA^TAx = v_1^TA^TAv_1 = v_1^T\lambda_1 v_1 = \lambda_1$$

- $$x^T A^T A x = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$
- 取$x = v_1$, 则
$$x^T A^T A x = v_1^T A^T A v_1 = v_1^T \lambda_1 v_1 = \lambda_1$$
- 从而
$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$$

- $$x^TA^TAx = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$
- 取$x = v_1$, 则
$$x^TA^TAx = v_1^TA^TAv_1 = v_1^T\lambda_1 v_1 = \lambda_1$$
- 从而
$$\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$$
- A^TA的特征值的开方称为A的奇异值

算子范数的性质

- $\|A\|_2 = \max\{|y^T Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$

算子范数的性质

- $\|A\|_2 = \max\{|y^T Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$
- $\|A^T\|_2 = \|A\|_2 = \sqrt{\|A^T A\|_2}$

算子范数的性质

- $\|A\|_2 = \max\{|y^T Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$
- $\|A^T\|_2 = \|A\|_2 = \sqrt{\|A^T A\|_2}$
- 对任意n阶正交阵U, V, 我们有$\|UA\|_2 = \|AV\|_2 = \|A\|_2$

算子范数的性质

- $\|A\|_2 = \max\{|y^TAx| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$
- $\|A^T\|_2 = \|A\|_2 = \sqrt{\|A^TA\|_2}$
- 对任意n阶正交阵U, V, 我们有$\|UA\|_2 = \|AV\|_2 = \|A\|_2$
- $\|A\|_\infty = \|A^T\|_1$

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

- 这是向量2范数的自然推广，也称为Hilbert-Schmidt范数

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

- 这是向量2范数的自然推广，也称为Hilbert-Schmidt范数
- 该范数满足对矩阵乘法的相容性

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

- 这是向量2范数的自然推广，也称为Hilbert-Schmidt范数
- 该范数满足对矩阵乘法的相容性
- 该范数与向量的2范数是相容的

Frobenius范数的性质

- $\|A\|_F = \sqrt{\mathrm{trace}(A^T A)}$

Frobenius范数的性质

- $\|A\|_F = \sqrt{\mathrm{trace}(A^T A)}$
- $\|AB\|_F \leqslant \|A\|_2 \|B\|_F$, $\|AB\|_F \leqslant \|A\|_F \|B\|_2$

Frobenius范数的性质

- $\|A\|_F = \sqrt{\mathrm{trace}(A^T A)}$
- $\|AB\|_F \leqslant \|A\|_2 \|B\|_F$, $\|AB\|_F \leqslant \|A\|_F \|B\|_2$
- $\|\cdot\|_F$是矩阵范数，但它不是算子范数（为什么？）

Frobenius范数的性质

- $\|A\|_F = \sqrt{\mathrm{trace}(A^TA)}$
- $\|AB\|_F \leqslant \|A\|_2\|B\|_F$, $\|AB\|_F \leqslant \|A\|_F\|B\|_2$
- $\|\cdot\|_F$是矩阵范数，但它不是算子范数（为什么？）
    - $\|I\|_F = \sqrt{n}$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max\limits_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max\limits_{i,j} |a_{ij}|$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max\limits_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max\limits_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max\limits_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max\limits_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$
- $\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$
- $\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$
- 利用$\|A\|_F$的trace表达式证明性质1; 利用$\|A\|_2$的定义，取特殊的x证明2,3,4,5等性质

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max\limits_{i,j}|a_{ij}| \leqslant \|A\|_2 \leqslant n\max\limits_{i,j}|a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$
- $\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$
- 利用$\|A\|_F$的trace表达式证明性质1; 利用$\|A\|_2$的定义，取特殊的x证明2,3,4,5等性质
- 各系数是最佳的（为什么？）

证明示例：$\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$

- 取$x \neq 0$, 使得$A^TAx = \lambda^2 x$, 其中$\lambda = \|A\|_2$

证明示例：$\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$

- 取$x \neq 0$，使得$A^TAx = \lambda^2 x$，其中$\lambda = \|A\|_2$
- 那么

$$\begin{aligned}\lambda^2\|x\|_1 &= \|A^TAx\|_1 \\ &\leqslant \|A^T\|_1\|A\|_1\|x\|_1 \\ &= \|A\|_\infty\|A\|_1\|x\|_1\end{aligned}$$

谱半径

- 设$A \in \mathbb{C}^{n\times n}$, 则称

$$\rho(A) = \max\{|\lambda| : \lambda \in \lambda(A)\}$$

为A的谱半径，这里$\lambda(A)$表示A的特征值的全体

谱半径与矩阵范数

定理

设$A \in \mathbb{C}^{n\times n}$，则

- 对任意矩阵范数$\|\cdot\|$，$\rho(A) \leqslant \|A\|$；
- 对$\forall \varepsilon > 0$，存在算子范数$\|\cdot\|$，使得

$$\|A\| \leqslant \rho(A) + \varepsilon$$

$\rho(A) \leqslant \|A\|$的证明

- 取$x \neq 0$, $Ax = \lambda x$, $|\lambda| = \rho(A)$

$\rho(A) \leqslant \|A\|$的证明

- 取$x \neq 0$, $Ax = \lambda x$, $|\lambda| = \rho(A)$
- 则有(xe_1^T表示以x为第一列的增广矩阵)

$$\rho(A)\|xe_1^T\| = \|\lambda xe_1^T\| = \|Axe_1^T\| \leqslant \|A\|\|xe_1^T\|$$

$\rho(A) \leqslant \|A\|$的证明

- 取$x \neq 0$, $Ax = \lambda x$, $|\lambda| = \rho(A)$
- 则有(xe_1^T表示以x为第一列的增广矩阵)

$$\rho(A)\|xe_1^T\| = \|\lambda xe_1^T\| = \|Axe_1^T\| \leqslant \|A\|\|xe_1^T\|$$

- 所以

$$\rho(A) \leqslant \|A\|$$

$\|A\| \leqslant \rho(A)+\varepsilon$的证明

- Jordan分解定理：

$$X^{-1}AX = \begin{pmatrix} \lambda_1 & \delta_1 & & & \\ & \lambda_2 & \delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \delta_{n-1} \\ & & & & \lambda_n \end{pmatrix}$$

其中$\delta_i = 1$或0

$\|A\| \leqslant \rho(A) + \varepsilon$的证明

- 对$\varepsilon > 0$取$D_\varepsilon = \mathrm{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1})$

$$D_\varepsilon^{-1} X^{-1} A X D_\varepsilon = \begin{pmatrix} \lambda_1 & \varepsilon\delta_1 & & & \\ & \lambda_2 & \varepsilon\delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \varepsilon\delta_{n-1} \\ & & & & \lambda_n \end{pmatrix}$$

$\|A\| \leqslant \rho(A) + \varepsilon$的证明

- 定义

$$\|G\|_\varepsilon = \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)\|_\infty$$

这是由下述向量范数诱导出的算子范数：

$$\|x\|_{XD_\varepsilon} = \|(XD_\varepsilon)^{-1}x\|_\infty$$

诱导性的证明

- 由定义：

$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

诱导性的证明

- 由定义：

$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

-

$$\begin{aligned}\|Gx\|_{XD_\varepsilon} &= \|(XD_\varepsilon)^{-1}Gx\|_\infty \\ &= \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)(XD_\varepsilon)^{-1}x\|_\infty \\ &\leqslant \|G\|_\varepsilon \|x\|_{XD_\varepsilon} = \|G\|_\varepsilon\end{aligned}$$

诱导性的证明

- 由定义：

$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

-

$$\begin{aligned}\|Gx\|_{XD_\varepsilon} &= \|(XD_\varepsilon)^{-1}Gx\|_\infty \\ &= \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)(XD_\varepsilon)^{-1}x\|_\infty \\ &\leqslant \|G\|_\varepsilon \|x\|_{XD_\varepsilon} = \|G\|_\varepsilon\end{aligned}$$

- 所以$\|G\|_{XD_\varepsilon} \leqslant \|G\|_\varepsilon$.

诱导性的证明

- 由定义：
$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

-
$$\begin{aligned}\|Gx\|_{XD_\varepsilon} &= \|(XD_\varepsilon)^{-1}Gx\|_\infty \\ &= \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)(XD_\varepsilon)^{-1}x\|_\infty \\ &\leqslant \|G\|_\varepsilon \|x\|_{XD_\varepsilon} = \|G\|_\varepsilon\end{aligned}$$

- 所以$\|G\|_{XD_\varepsilon} \leqslant \|G\|_\varepsilon$.
- 由矩阵∞范数的性质，可知等号成立

$\|A\| \leqslant \rho(A) + \varepsilon$的证明

- 从而有

$$\|A\|_\varepsilon = \max_{1\leqslant i\leqslant n}(|\lambda_i| + |\varepsilon\delta_i|) \leqslant \rho(A) + \varepsilon$$

谱半径与收敛

定理

设$A \in \mathbb{C}^{n\times n}$,

$$\lim_{k\to\infty} A^k = 0 \iff \rho(A) < 1$$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$
- $\forall k,\ \lambda^k \in \lambda(A^k)$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$
- $\forall k,\ \lambda^k \in \lambda(A^k)$
- $(\rho(A))^k = |\lambda|^k \leqslant \rho(A^k) \leqslant \|A^k\|_2$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$
- $\forall k,\ \lambda^k \in \lambda(A^k)$
- $(\rho(A))^k = |\lambda|^k \leqslant \rho(A^k) \leqslant \|A^k\|_2$
- 所以有$\rho(A) < 1$

充分性

- 设$\rho(A) < 1$

充分性

- 设$\rho(A) < 1$
- 存在算子范数$\|\cdot\|$，使得$\|A\| < 1$

充分性

- 设$\rho(A) < 1$
- 存在算子范数$\|\cdot\|$, 使得$\|A\| < 1$
- $0 \leqslant \|A^k\| \leqslant \|A\|^k \to 0, \quad k \to \infty$

充分性

- 设$\rho(A) < 1$
- 存在算子范数$\|\cdot\|$, 使得$\|A\| < 1$
- $0 \leqslant \|A^k\| \leqslant \|A\|^k \to 0, \quad k \to \infty$
- 所以$\lim_{k\to\infty} A^k = 0.$

矩阵级数的收敛性

定理

- $\sum_{k=0}^{\infty} A^k$收敛当且仅当$\rho(A) < 1$
- 当$\sum_{k=0}^{\infty} A^k$收敛时我们有$\sum_{k=0}^{\infty} A^k = (I - A)^{-1}$ 且存在算子范数$\|\cdot\|$使得对$\forall m \geqslant 0$

$$\left\| (I - A)^{-1} - \sum_{k=0}^{m} A^k \right\| \leqslant \frac{\|A\|^{m+1}}{1 - \|A\|}$$

推论

- 设$\|\cdot\|$是$\mathbb{C}^{n\times n}$上的矩阵范数，并假设$A\in\mathbb{C}^{n\times n}$满足$\|A\|<1$, 则$I-A$可逆

推论

- 设$\|\cdot\|$是$\mathbb{C}^{n\times n}$上的矩阵范数，并假设$A\in\mathbb{C}^{n\times n}$满足$\|A\|<1$, 则$I-A$可逆
- 进一步，若对此范数有$\|I\|=1$, 则

$$\|(I-A)^{-1}\|\leqslant\frac{1}{1-\|A\|}$$

敏度分析的问题描述

- 一个线性方程组$Ax = b$是由系数矩阵A和右端项b所确定的

敏度分析的问题描述

- 一个线性方程组$Ax = b$是由系数矩阵A和右端项b所确定的
- A或b通常是带有误差的，这些误差相对于精确值是微小的

敏度分析的问题描述

- 一个线性方程组$Ax = b$是由系数矩阵A和右端项b所确定的
- A或b通常是带有误差的，这些误差相对于精确值是微小的
- 这种微小误差对解有何影响？这就是线性方程组的敏感性问题

微小扰动

- 确实有方程组，系数的小扰动使解产生巨大变化。Mathematica2.2.1.nb

微小扰动

- 确实有方程组，系数的小扰动使解产生巨大变化。Mathematica2.2.1.nb
- 对于$Ax = b$, 经扰动后变为

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

微小扰动

- 确实有方程组，系数的小扰动使解产生巨大变化。Mathematica2.2.1.nb
- 对于$Ax = b$, 经扰动后变为

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

- 利用$Ax = b$简化后

$$(A + \delta A)\delta x = \delta b - (\delta A)x$$

- A非奇异，由前述推论，当$\|A^{-1}\|\|\delta A\| < 1$时，$I + A^{-1}\delta A$可逆，而且若$\|I\| = 1$则

$$\|(I + A^{-1}\delta A)^{-1}\| \leqslant \frac{1}{1 - \|A^{-1}\|\|\delta A\|}$$

- A非奇异，由前述推论，当$\|A^{-1}\|\|\delta A\| < 1$时，$I + A^{-1}\delta A$可逆，而且若$\|I\| = 1$则

$$\|(I + A^{-1}\delta A)^{-1}\| \leqslant \frac{1}{1 - \|A^{-1}\|\|\delta A\|}$$

- 从而$A + \delta A = A(I + A^{-1}\delta A)$非奇异，

$$\begin{aligned}\delta x &= (A + \delta A)^{-1}(\delta b - (\delta A)x) \\ &= (I + A^{-1}\delta A)^{-1}A^{-1}(\delta b - (\delta A)x)\end{aligned}$$

- 两边取范数

$$
\begin{aligned}
\|\delta x\| &\leqslant \|(I + A^{-1}\delta A)^{-1}\|\|A^{-1}\|(\|\delta b\| + \|\delta A\|\|x\|) \\
&\leqslant \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\delta A\|}(\|\delta b\| + \|\delta A\|\|x\|)
\end{aligned}
$$

- 两边取范数

$$
\begin{aligned}
\|\delta x\| &\leqslant \|(I+A^{-1}\delta A)^{-1}\|\|A^{-1}\|(\|\delta b\|+\|\delta A\|\|x\|)\\
&\leqslant \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta b\|+\|\delta A\|\|x\|)
\end{aligned}
$$

- 当$x\neq 0$, 注意到$\|b\|\leqslant\|A\|\|x\|$, 上式两边同除以$\|x\|$我们有

$$
\frac{\|\delta x\|}{\|x\|}\leqslant\frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)
$$

定理

定理

$\|\cdot\|$是一个满足$\|I\|=1$的矩阵范数，A非奇异，$b\neq 0$, $\|A^{-1}\|\|\delta A\|<1$. 若x和$x+\delta x$分别是$Ax=b$和$(A+\delta x)(x+\delta x)=b+\delta b$的解，则$(\kappa(A)=\|A\|\|A^{-1}\|)$

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

$\kappa(A)$

- 当$\|\delta A\|/\|A\|$较小时，

$$\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}} \approx \kappa(A)$$

$\kappa(A)$

- 当$\|\delta A\|/\|A\|$较小时，

$$\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}} \approx \kappa(A)$$

- 从而有

$$\frac{\|\delta x\|}{\|x\|} \lessapprox \kappa(A)\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)$$

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关
  - 若$\kappa(A)$不大，则扰动对解的影响也不会太大

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关
  - 若$\kappa(A)$不大，则扰动对解的影响也不会太大
  - 若$\kappa(A)$很大，则扰动对解的影响可能就很大

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关
  - 若$\kappa(A)$不大，则扰动对解的影响也不会太大
  - 若$\kappa(A)$很大，则扰动对解的影响可能就很大
- $\kappa(A)$称为线性方程组$Ax = b$的条件数

病态与良态方程组

- 条件数在一定程度上刻画了扰动对方程组解的影响程度

病态与良态方程组

- 条件数在一定程度上刻画了扰动对方程组解的影响程度
- 若条件数很大，就称该线性方程组的求解问题是病态的，有时也直接称A是病态的

病态与良态方程组

- 条件数在一定程度上刻画了扰动对方程组解的影响程度
- 若条件数很大，就称该线性方程组的求解问题是病态的，有时也直接称A是病态的
- 若条件数很小，就称该线性方程组的求解问题是良态的，有时也直接称A是良态的

条件数与范数

- 条件数与所采用的范数有关

条件数与范数

- 条件数与所采用的范数有关
- 可以通过在条件数上加下标表明所采用的范数：$\kappa_2(A)$

条件数与范数

- 条件数与所采用的范数有关
- 可以通过在条件数上加下标表明所采用的范数：$\kappa_2(A)$
- 病态、良态性是否与范数有关呢？

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的
- $\frac{1}{n}\kappa_2(A) \leqslant \kappa_1(A) \leqslant n\kappa_2(A)$

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的
- $\frac{1}{n}\kappa_2(A) \leqslant \kappa_1(A) \leqslant n\kappa_2(A)$
- $\frac{1}{n}\kappa_\infty(A) \leqslant \kappa_2(A) \leqslant n\kappa_\infty(A)$

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的
- $\frac{1}{n}\kappa_2(A) \leqslant \kappa_1(A) \leqslant n\kappa_2(A)$
- $\frac{1}{n}\kappa_\infty(A) \leqslant \kappa_2(A) \leqslant n\kappa_\infty(A)$
- $\frac{1}{n^2}\kappa_1(A) \leqslant \kappa_\infty(A) \leqslant n^2\kappa_1(A)$

矩阵求逆

- 设$\|\cdot\|$是满足$\|I\|=1$的矩阵范数，A非奇异，且$\|A^{-1}\|\|\delta A\|<1$，则$A+\delta A$也是非奇异的，而且有

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\|A^{-1}\|}\leqslant\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}$$

矩阵求逆

- 设$\|\cdot\|$是满足$\|I\|=1$的矩阵范数，A非奇异，且$\|A^{-1}\|\|\delta A\|<1$, 则$A+\delta A$也是非奇异的，而且有

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\|A^{-1}\|}\leqslant \frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}$$

- $(A+\delta A)^{-1}-A^{-1}=(A+\delta A)^{-1}(\delta A)A^{-1}$

矩阵求逆

- 设$\|\cdot\|$是满足$\|I\|=1$的矩阵范数，A非奇异，且$\|A^{-1}\|\|\delta A\|<1$, 则$A+\delta A$也是非奇异的，而且有

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\|A^{-1}\|}\leqslant\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}$$

  - $(A+\delta A)^{-1}-A^{-1}=(A+\delta A)^{-1}(\delta A)A^{-1}$

- 这表明$\kappa(A)$也可以看作矩阵求逆问题的条件数

条件数的几何意义

定理

设A非奇异,

$$\min\left\{\frac{\|\delta A\|_2}{\|A\|_2} : A+\delta A\text{奇异}\right\} = \frac{1}{\kappa_2(A)}$$

- 即在谱范数下，矩阵的条件数的倒数恰好等于该矩阵与全体奇异矩阵所成集合的相对距离

条件数的几何意义

定理

设A非奇异，

$$\min\left\{\frac{\|\delta A\|_2}{\|A\|_2}: A+\delta A\text{奇异}\right\}=\frac{1}{\kappa_2(A)}$$

- 即在谱范数下，矩阵的条件数的倒数恰好等于该矩阵与全体奇异矩阵所成集合的相对距离
- 当A十分病态时，它与一个奇异矩阵十分靠近

证明

只需证明：

$$\min\left\{\|\delta A\|_2 : A+\delta A\text{奇异}\right\} = \frac{1}{\|A^{-1}\|_2}$$

- 由推论，当$\|A^{-1}\|_2\|\delta A\|_2 < 1$时，$A+\delta A$非奇异，从而有

$$\min\left\{\|\delta A\|_2 : A+\delta A\text{奇异}\right\} \geqslant \frac{1}{\|A^{-1}\|_2}$$

证明

- 由于谱范数是向量2范数诱导出的算子范数，所以存在x, $\|x\|_2 = 1$, 使得$\|A^{-1}x\|_2 = \|A^{-1}\|_2$

证明

- 由于谱范数是向量2范数诱导出的算子范数，所以存在x, $\|x\|_2 = 1$, 使得$\|A^{-1}x\|_2 = \|A^{-1}\|_2$
- 令$y = \dfrac{A^{-1}x}{\|A^{-1}x\|_2}$, $\delta A = -\dfrac{xy^T}{\|A^{-1}\|_2}$ 则$\|y\|_2 = 1$而且

$$\begin{aligned}(A+\delta A)y &= Ay + (\delta A)y\\ &= \frac{x}{\|A^{-1}x\|_2} - \frac{x}{\|A^{-1}\|_2}\\ &= 0\end{aligned}$$

证明

- 由于谱范数是向量2范数诱导出的算子范数，所以存在x, $\|x\|_2 = 1$, 使得$\|A^{-1}x\|_2 = \|A^{-1}\|_2$
- 令$y = \dfrac{A^{-1}x}{\|A^{-1}x\|_2}$, $\delta A = -\dfrac{xy^T}{\|A^{-1}\|_2}$ 则$\|y\|_2 = 1$而且
$$
\begin{aligned}
(A+\delta A)y &= Ay + (\delta A)y \\
&= \frac{x}{\|A^{-1}x\|_2} - \frac{x}{\|A^{-1}\|_2} \\
&= 0
\end{aligned}
$$
- 这说明A的秩1校正$A+\delta A$奇异

证明

- 注意到

$$
\begin{aligned}
\|\delta A\|_2 &= \max_{\|z\|_2=1} \left\| \frac{xy^T}{\|A^{-1}\|_2} z \right\|_2 \\
&= \frac{\|x\|_2}{\|A^{-1}\|_2} \max_{\|z\|_2=1} |y^T z| \\
&= \frac{1}{\|A^{-1}\|_2}
\end{aligned}
$$

证明

- 注意到

$$\begin{aligned}\|\delta A\|_2 &= \max_{\|z\|_2=1}\left\|\frac{xy^T}{\|A^{-1}\|_2}z\right\|_2\\ &= \frac{\|x\|_2}{\|A^{-1}\|_2}\max_{\|z\|_2=1}|y^Tz|\\ &= \frac{1}{\|A^{-1}\|_2}\end{aligned}$$

- 这说明$\|\delta A\|_2 = \|A^{-1}\|_2^{-1}$

浮点数

- 计算机中浮点数f可表示为

$$f = \pm w \times \beta^J, \quad L \leqslant J \leqslant U$$

其中β是机器所用浮点数的基底，J是阶，w是尾数

浮点数

- 计算机中浮点数f可表示为

  $$f = \pm w \times \beta^J, \quad L \leqslant J \leqslant U$$

  其中β是机器所用浮点数的基底，J是阶， w是尾数
- 尾数的一般形式为$w = 0.d_1 d_2 \cdots d_t$, 其中t是尾数位数，称为字长，$0 \leqslant d_i < \beta$.

浮点数

- 计算机中浮点数f可表示为
$$f = \pm w \times \beta^J, \quad L \leqslant J \leqslant U$$
其中β是机器所用浮点数的基底，J是阶，w是尾数
- 尾数的一般形式为$w = 0.d_1 d_2 \cdots d_t$, 其中t是尾数位数，称为字长，$0 \leqslant d_i < \beta$.
- 若$d_1 \neq 0$, 称该浮点数为规格化的

浮点数全体

- 计算机系统的浮点数全体是

$$\mathcal{F} = \{0\} \cup \{f : f = \pm 0.d_1 d_2 \cdots d_t \times \beta^J, \\ 0 \leqslant d_i < \beta, \\ d_1 \neq 0, L \leqslant J \leqslant U\}$$

浮点数全体

- 计算机系统的浮点数全体是

$$\mathcal{F} = \{0\} \cup \{f : f = \pm 0.d_1 d_2 \cdots d_t \times \beta^J, \quad 0 \leqslant d_i < \beta, \quad d_1 \neq 0, L \leqslant J \leqslant U\}$$

- 可以用四元数组(β, t, L, U)描述$\mathcal{F}$. 典型取值是$(2, 56, -63, 64)$

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m, M]$和$[-M, -m]$中，这里$m=\beta^{L-1}$, $M=\beta^U(1-\beta^{-t})$

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m, M]$和$[-M, -m]$中，这里$m=\beta^{L-1}$, $M=\beta^U(1-\beta^{-t})$
  - 最小正数为$0.1\times\beta^L=\beta^{L-1}$, 其中0.1为$\beta$进制数

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m,M]$和$[-M,-m]$中，这里$m=\beta^{L-1}$, $M=\beta^U(1-\beta^{-t})$
  - 最小正数为$0.1\times\beta^L=\beta^{L-1}$, 其中0.1为$\beta$进制数
  - 最大正数为$0.\underbrace{(\beta-1)\cdots(\beta-1)}_{t位}\times\beta^U$

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m, M]$和$[-M, -m]$中，这里$m=\beta^{L-1}$, $M=\beta^U(1-\beta^{-t})$
  - 最小正数为$0.1\times\beta^L=\beta^{L-1}$, 其中0.1为$\beta$进制数
  - 最大正数为$0.\underbrace{(\beta-1)\cdots(\beta-1)}_{t\text{位}}\times\beta^U$
- 这些数在区间中非均匀分布

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$, $[-M, -m]$中所有实数

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$, $[-M, -m]$中所有实数
  - 0.584635无法用4位10进制浮点数表示

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$, $[-M, -m]$中所有实数
  - 0.584635无法用4位10进制浮点数表示
  - 两个t位浮点数的乘积一般需要$2t$位浮点数

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$, $[-M, -m]$中所有实数
  - 0.584635无法用4位10进制浮点数表示
  - 两个t位浮点数的乘积一般需要$2t$位浮点数
- 因此舍入误差不可避免

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x=0$, 则$\mathrm{fl}(x)=0$

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$, 则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$,

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$，则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$，
  - 舍入法：取$\mathrm{fl}(x)$为$\mathcal{F}$中最接近于x的数

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$, 则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$,
    - 舍入法：取$\mathrm{fl}(x)$为$\mathcal{F}$中最接近于x的数
    - 截断法：取$\mathrm{fl}(x)$为$\mathcal{F}$中满足$|f| \leqslant |x|$且最接近于x的数

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$，则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$，
  - 舍入法：取$\mathrm{fl}(x)$为$\mathcal{F}$中最接近于x的数
  - 截断法：取$\mathrm{fl}(x)$为$\mathcal{F}$中满足$|f| \leqslant |x|$且最接近于x的数
  - 例：$(\beta, t, L, U) = (10, 3, 0, 2)$，$x = 5.45627$，舍入法时$\mathrm{fl}(x) = 0.546 \times 10$，截断法时$\mathrm{fl}(x) = 0.545 \times 10$

机器精度与相对误差

- 定义机器精度**u**为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

机器精度与相对误差

- 定义机器精度$\mathbf{u}$为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

- 由于$t \geqslant 1$, 我们有$\mathbf{u} \in (0, 1]$

机器精度与相对误差

- 定义机器精度$\mathbf{u}$为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

- 由于$t \geqslant 1$, 我们有$\mathbf{u} \in (0, 1]$

机器精度与相对误差

- 定义机器精度$\mathbf{u}$为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

- 由于$t \geqslant 1$, 我们有$\mathbf{u} \in (0, 1]$

定理

设$m \leqslant |x| \leqslant M$, 则

$$\mathrm{fl}(x) = x(1+\delta), \quad |\delta| \leqslant \mathbf{u}$$

证明

- 不妨设$x > 0$, 取α使得

$$\beta^{\alpha-1} \leqslant x < \beta^{\alpha}$$

证明

- 不妨设$x > 0$, 取α使得

$$\beta^{\alpha-1} \leqslant x < \beta^{\alpha}$$

- 在$[\beta^{\alpha-1}, \beta^{\alpha})$中浮点数的阶为$\alpha$, 所以在这个区间中所有$t$位的浮点数以间距$\beta^{\alpha-t}$均匀分布

证明

- 对于舍入法，

$$|\,\mathrm{fl}(x) - x| \leqslant \frac{1}{2}\beta^{\alpha-t} = \frac{1}{2}\beta^{\alpha-1}\beta^{1-t} \leqslant \frac{1}{2}x\beta^{1-t}$$

从而有

$$\frac{|\,\mathrm{fl}(x) - x|}{x} \leqslant \frac{1}{2}\beta^{1-t}$$

证明

- 对于截断法,

$$|\mathrm{fl}(x) - x| \leqslant \beta^{\alpha-t} = \beta^{\alpha-1}\beta^{1-t} \leqslant x\beta^{1-t}$$

从而有

$$\frac{|\mathrm{fl}(x) - x|}{x} \leqslant \beta^{1-t}$$

注解

- 有时为了使用方便，我们也将上述结果表示为

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \quad |\delta| \leqslant \mathbf{u}$$

注解

- 有时为了使用方便，我们也将上述结果表示为

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \quad |\delta| \leqslant \mathbf{u}$$

- 为此，只需要在证明中考虑$|\mathrm{fl}(x) - x|$相对于$\mathrm{fl}(x)$的界即可

- 对于舍入法，
  $$\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) \leqslant \beta^{\alpha}$$

- 对于舍入法，
  $\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) \leqslant \beta^{\alpha}$
- 从而

$$|\,\mathrm{fl}(x) - x| \leqslant \frac{1}{2}\beta^{\alpha-t} = \frac{1}{2}\beta^{\alpha-1}\beta^{1-t} \leqslant \frac{1}{2}\,\mathrm{fl}(x)\beta^{1-t}$$

从而有

$$\frac{|\,\mathrm{fl}(x) - x|}{\mathrm{fl}(x)} \leqslant \frac{1}{2}\beta^{1-t}$$

- 对于截断法，

$$\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) < \beta^{\alpha}$$

- 对于截断法，
  $\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) < \beta^{\alpha}$
- 从而

$$|\,\mathrm{fl}(x) - x| \leqslant \beta^{\alpha-t} = \beta^{\alpha-1}\beta^{1-t} \leqslant \mathrm{fl}(x)\beta^{1-t}$$

从而有

$$\frac{|\,\mathrm{fl}(x) - x|}{\mathrm{fl}(x)} \leqslant \beta^{1-t}$$

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$
- 用$\circ$表示$+, -, \times, \div$中任意一种运算

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$
- 用$\circ$表示$+, -, \times, \div$中任意一种运算
- $\mathrm{fl}(a \circ b)$表示先把a, b看作实数进行运算，然后再按舍入规则把结果转化为浮点数。运算时若$|a \circ b| > M$或$0 < |a \circ b| < m$, 则发生了上溢或下溢

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$
- 用$\circ$表示$+, -, \times, \div$中任意一种运算
- $\mathrm{fl}(a \circ b)$表示先把a, b看作实数进行运算，然后再按舍入规则把结果转化为浮点数。运算时若$|a \circ b| > M$或$0 < |a \circ b| < m$, 则发生了上溢或下溢
- 在不发生溢出时，由前述定理可有

$$\mathrm{fl}(a \circ b) = (a \circ b)(1 + \delta), \quad |\delta| \leqslant \mathbf{u}$$

例

设x, y是两个由浮点数构成的n维向量。试估计$|\,\mathrm{fl}(x^Ty) - x^Ty|$的上界

- 令$S_k = \mathrm{fl}\left(\sum_{i=1}^{k} x_i y_i\right)$

$$
\begin{aligned}
S_1 &= x_1y_1(1+\gamma_1), \quad |\gamma_1| \leqslant \mathbf{u} \\
S_k &= \mathrm{fl}(S_{k-1} + \mathrm{fl}(x_k y_k)) \\
&= (S_{k-1} + x_k y_k(1+\gamma_k))(1+\delta_k), \\
&\qquad |\delta_k|, |\gamma_k| \leqslant \mathbf{u}
\end{aligned}
$$

- 于是有

$$\mathrm{fl}(x^T y) = S_n = \sum_{i=1}^{n} x_i y_i (1+\gamma_i) \prod_{j=i}^{n} (1+\delta_j)$$

$$= \sum_{i=1}^{n} (1+\varepsilon_i) x_i y_i,$$

其中

$$1+\varepsilon_i = (1+\gamma_i) \prod_{j=i}^{n} (1+\delta_j), \delta_1 = 0$$

- 所以我们得到

$$
\begin{aligned}
|\,\mathrm{fl}(x^T y) - x^T y| &\leqslant \sum_{i=1}^{n} |\varepsilon_i||x_i y_i| \\
&\leqslant 1.01n\mathbf{u} \sum_{i=1}^{n} |x_i y_i|
\end{aligned}
$$

这里我们用到了当给定条件$n\mathbf{u} \leqslant 0.01$时，后面马上要证明的结论

- 所以我们得到

$$\begin{aligned} |\operatorname{fl}(x^Ty) - x^Ty| &\leqslant \sum_{i=1}^{n} |\varepsilon_i||x_iy_i| \\ &\leqslant 1.01n\mathbf{u} \sum_{i=1}^{n} |x_iy_i| \end{aligned}$$

这里我们用到了当给定条件$n\mathbf{u} \leqslant 0.01$时，后面马上要证明的结论

- 上式表明若$|x^Ty| \ll \sum_{i=1}^{n} |x_iy_i|$, 则$\operatorname{fl}(x^Ty)$的相对误差可能会很大

定理

若$|\delta_i| \leqslant \mathbf{u}$且$n\mathbf{u} \leqslant 0.01$, 那么

$$1 - 1.01n\mathbf{u} \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant 1 + 1.01n\mathbf{u}$$

或写成

$$\prod_{i=1}^{n}(1+\delta_i) = 1+\delta, \quad |\delta| \leqslant 1.01n\mathbf{u}$$

证明

- 由于$|\delta_i| \leqslant \mathbf{u}$, 所以

$$(1-\mathbf{u})^n \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant (1+\mathbf{u})^n$$

证明

- 由于$|\delta_i| \leqslant \mathbf{u}$, 所以

$$(1-\mathbf{u})^n \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant (1+\mathbf{u})^n$$

- 由于当$x \in (0,1)$时$1-nx \leqslant (1-x)^n$, 所以

$$1-1.01n\mathbf{u} \leqslant 1-n\mathbf{u} \leqslant (1-\mathbf{u})^n$$

- 注意到e^x的Taylor展开式

$$
\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\
&= 1 + x + \frac{x}{2} \cdot x \cdot \left(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots\right)
\end{aligned}
$$

- 注意到e^x的Taylor展开式

$$
\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\
&= 1 + x + \frac{x}{2} \cdot x \cdot \left(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots\right)
\end{aligned}
$$

- 所以当$x \in [0, 0.01]$时，(注：$e^{0.01} \approx 1.01$)

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

- 在左端不等式中取$x = \mathbf{u}$, 得到

$$(1+u)^n \leqslant e^{n\mathbf{u}}$$

- 在左端不等式中取$x = \mathbf{u}$, 得到

$$(1+u)^n \leqslant e^{n\mathbf{u}}$$

- 在右端不等式中取$x = n\mathbf{u}$,

$$e^{n\mathbf{u}} \leqslant 1 + 1.01n\mathbf{u}$$

- 在左端不等式中取$x = \mathbf{u}$, 得到

$$(1+u)^n \leqslant e^{n\mathbf{u}}$$

- 在右端不等式中取$x = n\mathbf{u}$,

$$e^{n\mathbf{u}} \leqslant 1 + 1.01n\mathbf{u}$$

- 综合得到$(1+\mathbf{u})^n \leqslant 1 + 1.01n\mathbf{u}$

矩阵计算

- 取n阶方阵E, 定义记号$|E| = (|e_{ij}|)$

矩阵计算

- 取n阶方阵E, 定义记号$|E| = (|e_{ij}|)$
- 规定$|E| \leqslant |F| \iff |e_{ij}| \leqslant |f_{ij}|$, $i, j = 1, \ldots, n$

矩阵计算

- 取n阶方阵E, 定义记号$|E| = (|e_{ij}|)$
- 规定$|E| \leqslant |F| \Longleftrightarrow |e_{ij}| \leqslant |f_{ij}|$, $i,j = 1,\ldots,n$
- 设$A, B \in \mathcal{F}^{n\times n}$, $\alpha \in \mathcal{F}$, 则

$$\mathrm{fl}(\alpha A) = \alpha A + E, |E| \leqslant \mathbf{u}|\alpha A|,$$

$$\mathrm{fl}(A+B) = (A+B) + E, |E| \leqslant \mathbf{u}|A+B|$$

矩阵乘积

- 基于前面的例子，我们可有

$$\mathrm{fl}(AB) = AB + E, |E| \leqslant 1.01n\mathbf{u}|A||B|$$

矩阵乘积

- 基于前面的例子，我们可有

$$\mathrm{fl}(AB) = AB + E, |E| \leqslant 1.01n\mathbf{u}|A||B|$$

- 注意$|AB|$可能比$|A||B|$小得多

矩阵乘积

- 基于前面的例子，我们可有

$$\mathrm{fl}(AB) = AB + E, |E| \leqslant 1.01n\mathbf{u}|A||B|$$

- 注意$|AB|$可能比$|A||B|$小得多
- 计算内积时，我们通常会先用双精度进行计算，然后再舍入到单精度数

向前误差分析

- 上述舍入误差界，是通过估计计算解与精确解之间的误差得到

向前误差分析

- 上述舍入误差界，是通过估计计算解与精确解之间的误差得到
- 界与精确解有关

向前误差分析

- 上述舍入误差界，是通过估计计算解与精确解之间的误差得到
- 界与精确解有关
- 这种误差分析的方法称为向前误差分析法

再议矩阵乘法

- 假设A, B是两个2×2的上三角阵，则

$$\mathrm{fl}(AB) = \begin{pmatrix} a_{11}b_{11}(1+\varepsilon_1) & \tilde{a}_{12} \\ 0 & a_{22}b_{22}(1+\varepsilon_5) \end{pmatrix}$$

其中

$$\tilde{a}_{12} = (a_{11}b_{12}(1+\varepsilon_2) + a_{12}b_{22}(1+\varepsilon_3))(1+\varepsilon_4),$$
$$|\varepsilon_i| \leqslant \mathbf{u}, \quad i = 1, 2, 3, 4, 5$$

- 令

$$\tilde{A} = \begin{pmatrix} a_{11} & a_{12}(1+\varepsilon_3)(1+\varepsilon_4) \\ 0 & a_{22}(1+\varepsilon_5) \end{pmatrix}$$

$$\tilde{B} = \begin{pmatrix} b_{11}(1+\varepsilon_1) & b_{12}(1+\varepsilon_2)(1+\varepsilon_4) \\ 0 & b_{22} \end{pmatrix}$$

- 令

$$\tilde{A} = \begin{pmatrix} a_{11} & a_{12}(1+\varepsilon_3)(1+\varepsilon_4) \\ 0 & a_{22}(1+\varepsilon_5) \end{pmatrix}$$

$$\tilde{B} = \begin{pmatrix} b_{11}(1+\varepsilon_1) & b_{12}(1+\varepsilon_2)(1+\varepsilon_4) \\ 0 & b_{22} \end{pmatrix}$$

- 则可证$\mathrm{fl}(AB) = \tilde{A}\tilde{B}$, 其中$\tilde{A} = A + E$, $|E| \leqslant 3\mathbf{u}|A|$; $\tilde{B} = B + F$, $|F| \leqslant 3\mathbf{u}|B|$

向后误差分析

- $\mathrm{fl}(AB)$是有了微小扰动的两个矩阵$\tilde{A}$, $\tilde{B}$的精确乘积

向后误差分析

- $\mathrm{fl}(AB)$是有了微小扰动的两个矩阵$\tilde{A}$, $\tilde{B}$的精确乘积
- 这种把计算过程产生的误差归结为具有误差的原始数据的精确运算的分析方法称为向后误差分析法

向后误差分析

- $\mathrm{fl}(AB)$是有了微小扰动的两个矩阵$\tilde{A}$, $\tilde{B}$的精确乘积
- 这种把计算过程产生的误差归结为具有误差的原始数据的精确运算的分析方法称为**向后误差分析法**
- 这种方法把浮点数的运算转化为实数的精确运算，从而可以在分析过程中便利地应用实数的代数运算法则

列主元Gauss消去法的舍入误差分析

本节我们利用浮点数基本运算的舍入误差理论对求解线性方程组的列主元Gauss消去法进行向后误差分析

- 将证明用列主元Gauss消去法求解$Ax = b$时，它的计算解$\tilde{x}$满足

$$(A + E)\tilde{x} = b$$

列主元Gauss消去法的舍入误差分析

本节我们利用浮点数基本运算的舍入误差理论对求解线性方程组的列主元Gauss消去法进行向后误差分析

- 将证明用列主元Gauss消去法求解$Ax = b$时，它的计算解$\tilde{x}$满足

$$(A + E)\tilde{x} = b$$

- 并且给出误差矩阵E的上界估计

三角分解时的舍入误差

定理

设$n \times n$浮点数矩阵$A = (a_{ij})$有三角分解，且$1.01n\mathbf{u} \leqslant 0.01$，则用Gauss消去法计算得到的单位下三角阵$\tilde{L}$和上三角阵$\tilde{U}$满足

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

其中$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$

证明之$\widetilde{U}$的计算

- 设$\tilde{U} = (\tilde{u}_{ij})$, $\tilde{L} = (\tilde{\ell}_{ij})$, 则$a_{ij} = \sum_{k=1}^{\min(i,j)} \ell_{ik} u_{kj}$

证明之$\widetilde{U}$的计算

- 设$\tilde{U} = (\tilde{u}_{ij})$, $\tilde{L} = (\tilde{\ell}_{ij})$, 则$a_{ij} = \sum_{k=1}^{\min(i,j)} \ell_{ik} u_{kj}$
- 由Gauss消去法可知$\tilde{u}_{ij}(i \leqslant j)$是从$a_{ij}$中依次减去$\tilde{\ell}_{ik}\tilde{u}_{kj}$而得到的，$k = 1, \ldots, i-1$,即

$$
\begin{aligned}
a_{ij}^{(0)} &= a_{ij}, \\
a_{ij}^{(k)} &= \mathrm{fl}(a_{ij}^{(k-1)} - \mathrm{fl}(\tilde{\ell}_{ik}\tilde{u}_{kj})), k = 1, \ldots, i-2, \\
\tilde{u}_{ij} = a_{ij}^{(i-1)} &= \mathrm{fl}(a_{ij}^{(i-2)} - \mathrm{fl}(\tilde{\ell}_{i,i-1}\tilde{u}_{i-1,i}))
\end{aligned}
$$

证明之$\widetilde{U}$的计算

- 由基本运算舍入误差分析的结果可知

$$a_{ij}^{(k)} = (a_{ij}^{(k-1)} - (\tilde{\ell}_{i,k}\tilde{u}_{k,j})(1+\gamma_k))(1+\varepsilon_k)$$

其中$|\gamma_k|, |\varepsilon_k| \leqslant \mathbf{u}$

证明之$\widetilde{U}$的计算

- 由基本运算舍入误差分析的结果可知

$$a_{ij}^{(k)} = (a_{ij}^{(k-1)} - (\tilde{\ell}_{i,k}\tilde{u}_{k,j})(1+\gamma_k))(1+\varepsilon_k)$$

其中$|\gamma_k|, |\varepsilon_k| \leqslant \mathbf{u}$

- 从而仿照计算内积的例题所示，我们有

$$\tilde{u}_{ij} = a_{ij}(1+\delta_i) - \sum_{k=1}^{i-1}(\tilde{\ell}_{ik}\tilde{u}_{kj})(1+\delta_k)$$

其中$|\delta_k| \leqslant 1.01n\mathbf{u}$, $k = 1, \ldots, i$

证明之$\widetilde{U}$的计算

- 由前式解出a_{ij}：

$$a_{ij} = \frac{\tilde{u}_{ij}}{1+\delta_i} + \sum_{k=1}^{i-1}(\tilde{\ell}_{ik}\tilde{u}_{kj})\frac{1+\delta_k}{1+\delta_i}$$

$$= \sum_{k=1}^{i}\tilde{\ell}_{ik}\tilde{u}_{kj} - e_{ij}$$

其中$\tilde{\ell}_{ii} = 1$,

$$e_{ij} = (\tilde{\ell}_{ii}\tilde{u}_{ij})\frac{\delta_i}{1+\delta_i} + \sum_{k=1}^{i-1}(\tilde{\ell}_{ik}\tilde{u}_{kj})\frac{\delta_i-\delta_k}{1+\delta_i}$$

证明之$\widetilde{U}$的计算

- 注意到$|\delta_k| \leqslant 1.01n\mathbf{u} < 0.01$, 我们有

$$
\begin{aligned}
|e_{ij}| &\leqslant \frac{2.02n\mathbf{u}}{1-0.01}\sum_{k=1}^{i}|\tilde{\ell}_{ik}||\tilde{u}_{kj}| \\
&\leqslant 2.05n\mathbf{u}\sum_{k=1}^{i}|\tilde{\ell}_{ik}||\tilde{u}_{kj}|
\end{aligned}
$$

证明之$\widetilde{L}$的计算

- 由Gauss消去法的具体实现可知，$\tilde{\ell}_{ij}(i > j)$的计算过程为

$$
\begin{aligned}
a_{ij}^{(0)} &= a_{ij}, \\
a_{ij}^{(k)} &= \mathrm{fl}(a_{ij}^{(k-1)} - \mathrm{fl}(\tilde{\ell}_{ik}\tilde{u}_{kj})), k = 1, \ldots, j-1 \\
\tilde{\ell}_{ij} &= \mathrm{fl}(a_{ij}^{(j-1)}/\tilde{u}_{jj})
\end{aligned}
$$

证明之$\widetilde{L}$的计算

- 由基本运算舍入误差分析的结果可知

$$
\begin{aligned}
a_{ij}^{(k)} &= (a_{ij}^{(k-1)} - (\tilde{\ell}_{ik}\tilde{u}_{kj})(1+\gamma_k))(1+\varepsilon_k) \\
\tilde{\ell}_{ij} &= a_{ij}^{(j-1)}/(\tilde{u}_{jj}(1+\delta))
\end{aligned}
$$

其中$|\delta|, |\gamma_k|, |\varepsilon_k| \leqslant \mathbf{u}$

证明之$\widetilde{L}$的计算

- 由此可类似证明

$$a_{ij} = \sum_{k=1}^{j} \tilde{\ell}_{ik}\tilde{u}_{kj} - e_{ij},$$

其中

$$|e_{ij}| \leqslant 2.05nu \sum_{k=1}^{j} |\tilde{\ell}_{ik}||\tilde{u}_{kj}|$$

有行列交换时的情形

定理

设$n \times n$浮点数矩阵$A = (a_{ij})$非奇异，且$1.01n\mathbf{u} \leqslant 0.01$，则用列主元Gauss消去法计算得到的单位下三角阵$\tilde{L}$，上三角阵$\tilde{U}$以及排列方阵$\tilde{P}$满足$\tilde{L}\tilde{U} = \tilde{P}A + E$，其中$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$

注意：行列交换并不引入舍入误差

求解三角形方程组的舍入误差

引理

设$n \times n$浮点数三角阵S非奇异，
且$1.01n\mathbf{u} \leqslant 0.01$，则用前一章方法求
解$Sx = b$得到的计算解$\tilde{x}$满足

$$(S + H)\tilde{x} = b$$

其中$|H| \leqslant 1.01n\mathbf{u}|S|$

证明

- 不妨设S是下三角阵

证明

- 不妨设S是下三角阵
- 采用归纳法：对n进行归纳

证明

- 不妨设S是下三角阵
- 采用归纳法：对n进行归纳
- $n = 1$时成立

证明

- 不妨设S是下三角阵
- 采用归纳法：对n进行归纳
- $n = 1$时成立
- 假设$n - 1$阶时结论成立

证明

- 采用前代法求解$Lx = b$，得到计算解$\tilde{x}$

证明

- 采用前代法求解$Lx = b$，得到计算解$\tilde{x}$
- 对b, $\tilde{x}$进行分块：$b = (b_1, c^T)^T$, $\tilde{x} = (\tilde{x}_1, \tilde{y}^T)^T$, 其中$b_1$, $\tilde{x}_1 \in \mathcal{F}$, $c, \tilde{y}$为列向量;

证明

- 采用前代法求解$Lx = b$，得到计算解$\tilde{x}$
- 对b, $\tilde{x}$进行分块：$b = (b_1, c^T)^T$, $\tilde{x} = (\tilde{x}_1, \tilde{y}^T)^T$, 其中$b_1$, $\tilde{x}_1 \in \mathcal{F}$, $c, \tilde{y}$为列向量;
- 对L进行相应分块：

$$L = \begin{pmatrix} \ell_{11} & 0 \\ \ell_1 & L_1 \end{pmatrix}$$

- 我们有

$$\tilde{x}_1 = \mathrm{fl}(b_1/\ell_{11}) = \frac{b_1}{\ell_{11}(1+\delta_1)}, |\delta_1| \leqslant \mathbf{u}$$

- 我们有

$$\tilde{x}_1 = \mathrm{fl}(b_1/\ell_{11}) = \frac{b_1}{\ell_{11}(1+\delta_1)}, |\delta_1| \leqslant \mathbf{u}$$

- 而$\tilde{y}$是用前代法求解下述$n-1$阶线性方程组所得到的计算解

$$L_1 y = \mathrm{fl}(c - \tilde{x}_1 \ell_1)$$

- 我们有

$$\tilde{x}_1 = \mathrm{fl}(b_1/\ell_{11}) = \frac{b_1}{\ell_{11}(1+\delta_1)}, |\delta_1| \leqslant \mathbf{u}$$

- 而$\tilde{y}$是用前代法求解下述$n-1$阶线性方程组所得到的计算解

$$L_1 y = \mathrm{fl}(c - \tilde{x}_1 \ell_1)$$

- 由归纳假设，$(L_1 + H_1)\tilde{y} = \mathrm{fl}(c - \tilde{x}_1 \ell_1)$，这里$|H_1| \leqslant 1.01(n-1)\mathbf{u}|L_1|$

- 根据基本运算的舍入误差结果，我们有

$$\begin{aligned}\mathrm{fl}(c - \tilde{x}_1\ell_1) &= \mathrm{fl}(c - \mathrm{fl}(\tilde{x}_1\ell_1)) \\ &= (I + D_\gamma)^{-1}(c - \tilde{x}_1\ell_1 - \tilde{x}_1 D_\delta \ell_1),\end{aligned}$$

  其中

$$\begin{aligned}D_\gamma &= \mathrm{diag}(\gamma_2, \ldots, \gamma_n), \\ D_\delta &= \mathrm{diag}(\delta_2, \ldots, \delta_n), \\ |\gamma_i|, |\delta_i| &\leqslant \mathbf{u}, \quad i = 2, 3, \ldots, n\end{aligned}$$

- 于是

$$\tilde{x}_1\ell_1+\tilde{x}_1D_\delta\ell_1+(I+D_\gamma)(L_1+H_1)\tilde{y}=c$$

- 于是

$$\tilde{x}_1\ell_1 + \tilde{x}_1 D_\delta \ell_1 + (I + D_\gamma)(L_1 + H_1)\tilde{y} = c$$

- 从而有

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

其中

$$H = \begin{pmatrix} \delta_1\ell_{11} & 0 \\ D_\delta\ell_1 & H_1 + D_\gamma(L_1 + H_1) \end{pmatrix}$$

$$
\begin{aligned}
|H| &\leqslant \begin{pmatrix} |\delta_1||\ell_{11}| & 0 \\ |D_\delta||\ell_1| & |H_1| + |D_\gamma|(|L_1| + |H_1|) \end{pmatrix} \\
&\leqslant \begin{pmatrix} \mathbf{u}|\ell_{11}| & 0 \\ \mathbf{u}|\ell_1| & |H_1| + \mathbf{u}(|L_1| + |H_1|) \end{pmatrix} \\
&\leqslant \mathbf{u} \begin{pmatrix} |\ell_{11}| & 0 \\ |\ell_1| & (1.01(n-1) + 1 + 1.01(n-1)\mathbf{u})|L_1| \end{pmatrix} \\
&\leqslant 1.01 n \mathbf{u} |L|
\end{aligned}
$$

最后一个不等式中用到了条件$1.01n\mathbf{u} \leqslant 0.01$

完整求解

- 把上述引理应用于三角形方程组

$$\tilde{L}y = \tilde{P}b, \quad \tilde{U}x = y$$

可知最后得到的计算解$\tilde{x}$应满足

$$(\tilde{L} + F)(\tilde{U} + G)\tilde{x} = \tilde{P}b$$

即

$$(\tilde{L}\tilde{U} + F\tilde{U} + \tilde{L}G + FG)\tilde{x} = \tilde{P}b$$

其中$|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$, $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$

舍入误差

- 把$\tilde{L}\tilde{U} = \tilde{P}A + E$代入得

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

其中

$$\delta A = \tilde{P}^T(E + F\tilde{U} + \tilde{L}G + FG)$$

舍入误差

- 把$\tilde{L}\tilde{U} = \tilde{P}A + E$代入得

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

其中

$$\delta A = \tilde{P}^T(E + F\tilde{U} + \tilde{L}G + FG)$$

- 由已有结论，我们有

$$|\delta A| \leqslant 4.09n\mathbf{u}\tilde{P}^T|\tilde{L}||\tilde{U}|$$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$
- $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$
- $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$
- $|FG| \leqslant 0.01(1.01n\mathbf{u})|\tilde{L}||\tilde{U}| \leqslant 0.02n\mathbf{u}|\tilde{L}||\tilde{U}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$
- $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$
- $|FG| \leqslant 0.01(1.01n\mathbf{u})|\tilde{L}||\tilde{U}| \leqslant 0.02n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $2.05 + 1.01 + 1.01 + 0.02 = 4.09$

增长因子

- 由于$\tilde{L}$的元素的绝对值不超过1，所以$\|\tilde{L}\|_\infty \leqslant n$

增长因子

- 由于$\tilde{L}$的元素的绝对值不超过1，所以$\|\tilde{L}\|_\infty \leqslant n$
- 为了估计$\|\tilde{U}\|_\infty$, 定义

$$\rho = \max_{i,j} |\tilde{u}_{ij}| / \max_{i,j} |a_{ij}|$$

称之为列主元Gauss消去法的增长因子

增长因子

- 由于$\tilde{L}$的元素的绝对值不超过1，所以$\|\tilde{L}\|_\infty \leqslant n$
- 为了估计$\|\tilde{U}\|_\infty$, 定义

$$\rho = \max_{i,j} |\tilde{u}_{ij}| / \max_{i,j} |a_{ij}|$$

称之为列主元Gauss消去法的增长因子

- 于是我们有

$$\|\tilde{U}\|_\infty \leqslant n \max_{i,j} |\tilde{u}_{ij}| = n\rho \max_{i,j} |a_{ij}| \leqslant n\rho \|A\|_\infty$$

Gauss消去法的舍入误差定理

定理

设$n \times n$浮点数矩阵A非奇异，且$1.01n\mathbf{u} \leqslant 0.01$，则用列主元Gauss消去法求解线性方程组$Ax = b$所得到的计算解$\tilde{x}$满足

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

其中

$$\|\delta A\|_\infty / \|A\|_\infty \leqslant 4.09 n^3 \rho \mathbf{u}$$

注解

- 求解过程中引入舍入误差产生的计算解相当于系数矩阵作某些扰动而得到的扰动方程组的精确解

注解

- 求解过程中引入舍入误差产生的计算解相当于系数矩阵作某些扰动而得到的扰动方程组的精确解
- 通常δA的元素比A的元素的初始误差小得多，从这个意义上说，列主元Gauss消去法是数值稳定的

注解

- 理论上可以证明$\rho \leqslant 2^{n-1}$, 而且上界可以达到；实际中常遇到的情形是ρ不会超过n

注解

- 理论上可以证明$\rho \leqslant 2^{n-1}$, 而且上界可以达到；实际中常遇到的情形是ρ不会超过n
- 定理中的上界比真正的$\|\delta A\|_\infty/\|A\|_\infty$大很多，而实际计算中几乎都有$\|\delta A\|_\infty/\|A\|_\infty \approx \mathbf{u}$

精度估计

- 设采用某种方法求解$Ax = b$得到的计算解为$\tilde{x}$

精度估计

- 设采用某种方法求解$Ax = b$得到的计算解为$\tilde{x}$
- 令$r = b - A\tilde{x}$, 则有

$$r = Ax - A\tilde{x} = A(x - \tilde{x})$$

精度估计

- 设采用某种方法求解$Ax = b$得到的计算解为$\tilde{x}$
- 令$r = b - A\tilde{x}$, 则有

$$r = Ax - A\tilde{x} = A(x - \tilde{x})$$

- 从而

$$\|x - \tilde{x}\| = \|A^{-1}r\| \leqslant \|A^{-1}\|\|r\|$$

精度估计

- 注意到$\|b\| \leqslant \|A\|\|x\|$, 所以我们有

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leqslant \|A^{-1}\|\|A\|\frac{\|r\|}{\|b\|}$$

精度估计

- 注意到$\|b\| \leqslant \|A\|\|x\|$, 所以我们有

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leqslant \|A^{-1}\|\|A\|\frac{\|r\|}{\|b\|}$$

- 在上式中取∞范数，有

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} \leqslant \kappa_\infty(A)\frac{\|r\|_\infty}{\|b\|_\infty}$$

精度估计公式

- 通过计算$\kappa_\infty(A)\dfrac{\|r\|_\infty}{\|b\|_\infty}$估计计算解的精度

精度估计公式

- 通过计算$\kappa_\infty(A)\dfrac{\|r\|_\infty}{\|b\|_\infty}$估计计算解的精度
- 这个公式中除$\kappa_\infty(A)$外，其余量都容易计算

精度估计公式

- 通过计算$\kappa_\infty(A)\dfrac{\|r\|_\infty}{\|b\|_\infty}$估计计算解的精度
- 这个公式中除$\kappa_\infty(A)$外，其余量都容易计算
- 而$\|A\|_\infty$是容易计算的，问题就余下了如何计算$\|A^{-1}\|_\infty$，或者如何计算$\|A^{-T}\|_1$

多元函数极值问题

- 为了计算$\|B\|_1$, 我们在区域$D = \{x \in \mathbb{R}^n\colon \|x\|_1 \leqslant 1\}$上考虑多元函数极值问题

$$\max f(x) = \|Bx\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} b_{ij} x_j \right|$$

多元函数极值问题

- 为了计算$\|B\|_1$, 我们在区域$\mathcal{D} = \{x \in \mathbb{R}^n : \|x\|_1 \leqslant 1\}$上考虑多元函数极值问题

$$\max f(x) = \|Bx\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} b_{ij} x_j \right|$$

- 易证f是凸函数，$\mathcal{D}$是凸集。从而最大值一定在边界上达到，但可能是局部极大值

多元函数极值问题

- 为了计算$\|B\|_1$, 我们在区域$\mathcal{D} = \{x \in \mathbb{R}^n\colon \|x\|_1 \leqslant 1\}$上考虑多元函数极值问题

$$\max f(x) = \|Bx\|_1 = \sum_{i=1}^n \left| \sum_{j=1}^n b_{ij} x_j \right|$$

- 易证f是凸函数，$\mathcal{D}$是凸集。从而最大值一定在边界上达到，但可能是局部极大值
- 可以应用著名的“盲人爬山法”求解这个优化问题

梯度

- 对于$x_0 = (x_j^{(0)}) \in \mathbf{R}^n$, $\|x_0\|_1 = 1$满足
$$\sum_{j=1}^{n} b_{ij} x_j^{(0)} \neq 0, \quad i = 1, \ldots, n$$
则f在$x = x_0$点的梯度$\nabla f(x_0)$存在

梯度

- 对于$x_0 = (x_j^{(0)}) \in \mathbf{R}^n$, $\|x_0\|_1 = 1$满足
$$\sum_{j=1}^{n} b_{ij} x_j^{(0)} \neq 0, \quad i = 1, \ldots, n$$
则f在$x = x_0$点的梯度$\nabla f(x_0)$存在
- 由凸函数的性质，
$$f(y) \geqslant f(x_0) + \nabla f(x_0)(y - x), y \in \mathbf{R}^n$$

梯度计算

- 取$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$,
  $v = (\xi_1, \ldots, \xi_n)^T$

梯度计算

- 取$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$,
  $v = (\xi_1, \ldots, \xi_n)^T$
- 在x_0附近$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i b_{ij} x_j$

梯度计算

- 取$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$,
  $v = (\xi_1, \ldots, \xi_n)^T$
- 在x_0附近$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i b_{ij} x_j$
- $\nabla f(x_0) = v^T B = (B^T v)^T$

梯度计算

- 取$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$, $v = (\xi_1, \ldots, \xi_n)^T$
- 在x_0附近$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i b_{ij} x_j$
- $\nabla f(x_0) = v^T B = (B^T v)^T$
- 定义$w = Bx_0$, $z = B^T v$

定理

定理

B, x_0, v, w, z如前所述。则有

- 若$\|z\|_\infty \leqslant z^T x_0$, 则$\|w\|_1 = \|Bx_0\|_1$是$f(x)$在$\mathcal{D}$中的局部极大值
- 若$\|z\|_\infty > z^T x_0$, 则$\|Be_j\|_1 > \|Bx_0\|_1$, 其中j满足$|z_j| = \|z\|_\infty$

证明：当$\|z\|_\infty \leqslant z^T x_0$时

- 由于在x_0附近f是线性函数，所以

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

证明：当$\|z\|_\infty \leqslant z^T x_0$时

- 由于在x_0附近f是线性函数，所以

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

- 因此我们只需证在x_0附近

$$\nabla f(x_0)(x - x_0) = z^T(x - x_0) \leqslant 0$$

证明：当$\|z\|_\infty \leqslant z^T x_0$时

- 由于在x_0附近f是线性函数，所以

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

- 因此我们只需证在x_0附近

$$\nabla f(x_0)(x - x_0) = z^T(x - x_0) \leqslant 0$$

- 实际上对于$\|x\|_1 \leqslant 1$我们有

$$z^T(x - x_0) = z^T x - z^T x_0 \leqslant \|z\|_\infty \|x\|_1 - z^T x_0$$
$$\leqslant \|z\|_\infty - z^T x_0 \leqslant 0$$

证明：当$\|z\|_\infty > z^T x_0$时

- 取$\tilde{x} = e_j \operatorname{sgn}(z_j)$，则有

$$\begin{aligned}
\|Be_j\|_1 = \|B\tilde{x}\|_1 &= f(\tilde{x}) \\
&\geqslant f(x_0) + \nabla f(x_0)(\tilde{x} - x_0) \\
&= \|Bx_0\|_1 + z^T\tilde{x} - z^T x_0 \\
&= \|Bx_0\|_1 + |z_j| - z^T x_0 \\
&= \|Bx_0\|_1 + \|z\|_\infty - z^T x_0 \\
&> \|Bx_0\|_1
\end{aligned}$$

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$
- 对$B = A^{-T}$应用上述定理，其中$w = Bx$和$z = B^Tv$相当于求解方程组$A^Tw = x$和$Az = v$

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$
- 对$B = A^{-T}$应用上述定理，其中$w = Bx$和$z = B^T v$相当于求解方程组$A^T w = x$和$Az = v$
- 如此仅用$O(n^2)$就可以完成估计

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$
- 对$B = A^{-T}$应用上述定理，其中$w = Bx$和$z = B^T v$相当于求解方程组$A^T w = x$和$Az = v$
- 如此仅用$O(n^2)$就可以完成估计
- 初值x可以任取，如$x_i = 1/n$

精度估算过程

- 估计$\|A^{-1}\|_\infty$的一个近似值$\tilde{\nu}$

精度估算过程

- 估计$\|A^{-1}\|_\infty$的一个近似值$\tilde{\nu}$
- 分别计算$\|r\|_\infty$, $\|b\|_\infty$, $\|A\|_\infty$的近似值$\tilde{\gamma}$, $\tilde{\beta}$, $\tilde{\mu}$

精度估算过程

- 估计$\|A^{-1}\|_\infty$的一个近似值$\tilde{\nu}$
- 分别计算$\|r\|_\infty$, $\|b\|_\infty$, $\|A\|_\infty$的近似值$\tilde{\gamma}$, $\tilde{\beta}$, $\tilde{\mu}$
- 计算$\tilde{\rho} = \tilde{\nu}\tilde{\mu}\tilde{\gamma}/\tilde{\beta}$得到计算解$\tilde{x}$的相对误差的一个估计

注解

- 大多数问题中这一方法可以得到相当好的估计

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为
  - 由于舍入误差的影响使得$\tilde{\gamma}$远远小于$\|r\|_\infty$的真值

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为
  - 由于舍入误差的影响使得$\tilde{\gamma}$远远小于$\|r\|_\infty$的真值
  - 当A十分病态时，三角分解已相当不准确，从而使得应用它估计出的$\tilde{\nu}$ 要比$\|A^{-1}\|_\infty$的真值小得多

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为
  - 由于舍入误差的影响使得$\tilde{\gamma}$远远小于$\|r\|_\infty$的真值
  - 当A十分病态时，三角分解已相当不准确，从而使得应用它估计出的$\tilde{\nu}$ 要比$\|A^{-1}\|_\infty$的真值小得多
- 如何估计得更好，仍是一个有待深入研究的问题

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
    - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
  - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$
  - 利用A的三角分解求解$Az = r$

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
  - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$
  - 利用A的三角分解求解$Az = r$
  - 计算$x = \tilde{x} + z$

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
  - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$
  - 利用A的三角分解求解$Az = r$
  - 计算$x = \tilde{x} + z$
  - 若$\dfrac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} \leqslant \varepsilon$ 则结束；否则令$\tilde{x} = x$ 转第一步