第二章 线性方程组的敏度分析与消去法的舍入误差分析

—范数, 敏度分析, 舍入误差分析等

杜磊

dulei@dlut.edu.cn

大连理工大学 数学科学学院

创新园大厦 B1207

2019 年 9 月 24 日

内容提要

向量范数

定义 (向量范数)

一个从 $\mathbb{R}^n$ 到 $\mathbb{R}$ 的非负函数 $\|\cdot\|$ 叫做 $\mathbb{R}^n$ 上的向量范数, 如果它满足:

1. 正定性: 对所有的 $x \in \mathbb{R}^n$, 有 $\|x\| \geq 0$, 而且 $\|x\| = 0$ 当且仅当 $x = 0$;
2. 齐次性: 对所有的 $x \in \mathbb{R}^n$ 和 $\alpha \in \mathbb{R}$, 有 $\|\alpha x\| = |\alpha| \, \|x\|$;
3. 三角不等式: 对所有的 $x, y \in \mathbb{R}^n$, 有 $\|x + y\| \leq \|x\| + \|y\|$.

向量范数

定义 (向量范数)

一个从 $\mathbb{R}^n$ 到 $\mathbb{R}$ 的非负函数 $\|\cdot\|$ 叫做 $\mathbb{R}^n$ 上的向量范数, 如果它满足:

1. 正定性: 对所有的 $x \in \mathbb{R}^n$, 有 $\|x\| \geq 0$, 而且 $\|x\| = 0$ 当且仅当 $x = 0$;
2. 齐次性: 对所有的 $x \in \mathbb{R}^n$ 和 $\alpha \in \mathbb{R}$, 有 $\|\alpha x\| = |\alpha| \, \|x\|$;
3. 三角不等式: 对所有的 $x, y \in \mathbb{R}^n$, 有 $\|x + y\| \leq \|x\| + \|y\|$.

范数的连续性:

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

常见的向量范数

最常见的向量范数是 p 范数:

$$\|x\|_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}}, \quad p \geq 1,$$

常见的向量范数

最常见的向量范数是 p 范数:

$$\|x\|_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}}, \quad p \geq 1,$$

其中 $p = 1, 2, \infty$ 是最重要的, 即

- 1-范数: $\|x\|_1 = |x_1| + \cdots + |x_n|$,
- 2-范数: $\|x\|_2 = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{\frac{1}{2}}$,
- ∞-范数: $\|x\|_\infty = \max\{|x_i| : i = 1, \cdots, n\}$.

常见的向量范数

最常见的向量范数是 p 范数:

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}, \quad p \geq 1,$$

其中 $p = 1, 2, \infty$ 是最重要的, 即

- 1-范数: $\|x\|_1 = |x_1| + \cdots + |x_n|$,
- 2-范数: $\|x\|_2 = (|x_1|^2 + \cdots + |x_n|^2)^{\frac{1}{2}}$,
- ∞-范数: $\|x\|_\infty = \max\{|x_i| : i = 1, \cdots, n\}$.

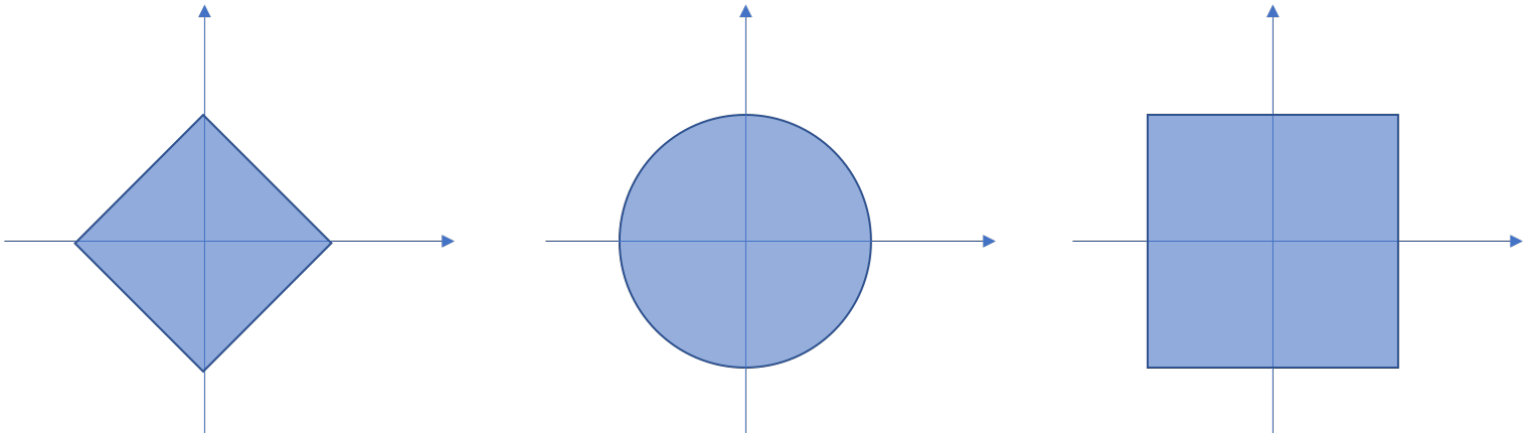

图: 闭单位球 $\{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$, 取 $n = 2, p = 1, 2, \infty$.

向量范数的等价性

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 和 c_2 使得对一切 $x \in \mathbb{R}^n$, 有

$$c_1 \|x\|_\alpha \leq \|x\|_\beta \leq c_2 \|x\|_\alpha .$$

向量范数的等价性

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 和 c_2 使得对一切 $x \in \mathbb{R}^n$, 有

$$c_1 \|x\|_\alpha \leq \|x\|_\beta \leq c_2 \|x\|_\alpha .$$

特别的, 对于 $\|\cdot\|_1 , \|\cdot\|_2 , \|\cdot\|_\infty$ 有

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

向量范数的等价性

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 和 c_2 使得对一切 $x \in \mathbb{R}^n$, 有

$$c_1 \|x\|_\alpha \leq \|x\|_\beta \leq c_2 \|x\|_\alpha .$$

特别的, 对于 $\|\cdot\|_1 , \|\cdot\|_2 , \|\cdot\|_\infty$ 有

$$\begin{aligned} \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n} \|x\|_2 , \\ \|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty , \\ \|x\|_\infty &\leq \|x\|_1 \leq n \|x\|_\infty . \end{aligned}$$

定理

设 $x_k \in \mathbb{R}^n$, 则 $\lim\limits_{k\to\infty} \|x_k - x\| = 0$ 的充分必要条件是

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

即向量序列的范数收敛等价于其分量收敛

带权重的向量范数

给定任意向量范数 $\|\cdot\|$, 对任意非奇异矩阵 W, 定义

$$\|x\|_W := \|Wx\|.$$

可证 $\|\cdot\|_W$ 也为向量范数.

矩阵范数

定义 (矩阵范数)

一个从 $\mathbb{R}^{n\times n}$ 到 $\mathbb{R}$ 的非负函数 $\|\cdot\|$ 叫做 $\mathbb{R}^{n\times n}$ 上的矩阵范数, 如果它满足:

1. 正定性: 对所有的 $A \in \mathbb{R}^{n\times n}$, 有 $\|A\| \geq 0$, 而且 $\|A\| = 0$ 当且仅当 $A = 0$;
2. 齐次性: 对所有的 $A \in \mathbb{R}^{n\times n}$ 和 $\alpha \in \mathbb{R}$, 有 $\|\alpha A\| = |\alpha| \, \|A\|$;
3. 三角不等式: 对所有的 $A, B \in \mathbb{R}^{n\times n}$, 有 $\|A + B\| \leq \|A\| + \|B\|$;
4. 相容性: 对所有的 $A, B \in \mathbb{R}^{n\times n}$, 有 $\|AB\| \leq \|A\| \, \|B\|$.

矩阵范数

定义 (矩阵范数)

一个从 $\mathbb{R}^{n\times n}$ 到 $\mathbb{R}$ 的非负函数 $\|\cdot\|$ 叫做 $\mathbb{R}^{n\times n}$ 上的矩阵范数, 如果它满足:

1. 正定性: 对所有的 $A\in\mathbb{R}^{n\times n}$, 有 $\|A\|\geq 0$, 而且 $\|A\|=0$ 当且仅当 $A=0$;
2. 齐次性: 对所有的 $A\in\mathbb{R}^{n\times n}$ 和 $\alpha\in\mathbb{R}$, 有 $\|\alpha A\|=|\alpha|\,\|A\|$;
3. 三角不等式: 对所有的 $A,B\in\mathbb{R}^{n\times n}$, 有 $\|A+B\|\leq\|A\|+\|B\|$;
4. 相容性: 对所有的 $A,B\in\mathbb{R}^{n\times n}$, 有 $\|AB\|\leq\|A\|\,\|B\|$.

由于 $\mathbb{R}^{n\times n}$ 上矩阵范数可以看作是 $\mathbb{R}^{n^2}$ 上的向量范数的推广, 所以矩阵范数具有向量范数的一切性质, 例如:

1. $\mathbb{R}^{n\times n}$ 上的任意两个矩阵范数是等价的;
2. 矩阵序列的范数收敛等价于其元素收敛, 即
$$\lim_{k\to\infty}\|A_k-A\|=0\Leftrightarrow\lim_{k\to\infty}a_{ij}^{(k)}=a_{ij},\ i,j=1,\cdots,n.$$

矩阵范数

定义 (矩阵向量范数的相容性)

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

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

则称矩阵范数 $\|\cdot\|_M$ 和向量范数 $\|\cdot\|_v$ 是相容的.

矩阵范数

定义 (矩阵向量范数的相容性)

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

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

则称矩阵范数 $\|\cdot\|_M$ 和向量范数 $\|\cdot\|_v$ 是相容的.

Frobenius 范数 (F-范数)

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2}.$$

算子范数

定理

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

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

则 $\|\cdot\|$ 是 $\mathbb{R}^{n\times n}$ 上的一个矩阵范数. 该矩阵范数称为从属于向量范数 $\|\cdot\|$ 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

定理

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

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

则 $\|\cdot\|$ 是 $\mathbb{R}^{n\times n}$ 上的一个矩阵范数. 该矩阵范数称为从属于向量范数 $\|\cdot\|$ 的矩阵范数, 也称为由向量范数 $\|\cdot\|$ 诱导出的算子范数.

矩阵 p-范数

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

矩阵的 1-范数, 2-范数和 ∞-范数

定理

设 $A \in \mathbb{R}^{n\times n}$, 则有

$$1\text{-范数: } \|A\|_1 = \max_{1\le j\le n}\left(\sum_{i=1}^{n}|a_{ij}|\right);$$

$$\infty\text{-范数: } \|A\|_\infty = \max_{1\le i\le n}\left(\sum_{j=1}^{n}|a_{ij}|\right);$$

$$2\text{-范数: } \|A\|_2 = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}.$$

矩阵的 1-范数, 2-范数和 ∞-范数

定理

设 $A \in \mathbb{R}^{n\times n}$, 则有

$$1\text{-范数: } \|A\|_1 = \max_{1\le j\le n}\left(\sum_{i=1}^{n}|a_{ij}|\right);$$

$$\infty\text{-范数: } \|A\|_\infty = \max_{1\le i\le n}\left(\sum_{j=1}^{n}|a_{ij}|\right);$$

$$2\text{-范数: } \|A\|_2 = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)}.$$

定理 (矩阵范数的等价性)

$\mathbb{R}^{n\times n}$ 空间上所有范数都是等价的, 特别的, 有

$$\frac{1}{\sqrt{n}}\|A\|_1 \le \|A\|_2 \le \sqrt{n}\|A\|_1, \quad \frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{n}\|A\|_\infty,$$

$$\frac{1}{n}\|A\|_\infty \le \|A\|_1 \le n\|A\|_\infty, \quad \frac{1}{\sqrt{n}}\|A\|_1 \le \|A\|_F \le \sqrt{n}\|A\|_2.$$

矩阵 2-范数, 谱半径

定理

设 $A \in \mathbb{R}^{n\times n}$, 则

1. $\|A\|_2 = \max\{|y^{\mathrm{T}}Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$;
2. $\left\|A^{\mathrm{T}}\right\|_2 = \|A\|_2 = \sqrt{\|A^{\mathrm{T}}A\|_2}$;
3. 对任意的 n 阶正交矩阵 U 和 V, 有 $\|UA\|_2 = \|AV\|_2 = \|A\|_2$.
4. 对任意的 n 阶正交矩阵 U 和 V, 有 $\|UA\|_F = \|AV\|_F = \|A\|_F$.

矩阵 2-范数, 谱半径

定理

设 $A \in \mathbb{R}^{n\times n}$, 则

1. $\|A\|_2 = \max\{|y^{\mathrm{T}}Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$;
2. $\left\|A^{\mathrm{T}}\right\|_2 = \|A\|_2 = \sqrt{\|A^{\mathrm{T}}A\|_2}$;
3. 对任意的 n 阶正交矩阵 U 和 V, 有 $\|UA\|_2 = \|AV\|_2 = \|A\|_2$.
4. 对任意的 n 阶正交矩阵 U 和 V, 有 $\|UA\|_F = \|AV\|_F = \|A\|_F$.

定义

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.

矩阵范数与谱半径关系

定理

设 $A \in \mathbb{C}^{n\times n}$, 则有

1. 对 $\mathbb{C}^{n\times n}$ 上的任意矩阵范数 $\|\cdot\|$, 有

$$\rho(A) \leq \|A\|;$$

2. 对任给的 $\epsilon > 0$, 存在 $\mathbb{C}^{n\times n}$ 上的算子范数 $\|\cdot\|$, 使得

$$\|A\| \leq \rho(A) + \epsilon.$$

矩阵范数与谱半径关系

定理

设 $A \in \mathbb{C}^{n\times n}$, 则有

1. 对 $\mathbb{C}^{n\times n}$ 上的任意矩阵范数 $\|\cdot\|$, 有

$$\rho(A) \leq \|A\|;$$

2. 对任给的 $\epsilon > 0$, 存在 $\mathbb{C}^{n\times n}$ 上的算子范数 $\|\cdot\|$, 使得

$$\|A\| \leq \rho(A) + \epsilon.$$

定理

设 $A \in \mathbb{C}^{n\times n}$, 则有

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

矩阵级数, 矩阵逆的范数估计

定理

设 $A \in \mathbb{C}^{n\times n}$, 则有

1. $\sum\limits_{k=0}^{\infty} A^k$ 收敛的充分必要条件是 $\rho(A) < 1$;
2. 当 $\sum\limits_{k=0}^{\infty} A^k$ 收敛时, 有

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

而且存在 $\mathbb{C}^{n\times n}$ 上的算子范数 $\|\cdot\|$, 使得

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

对一切的自然数 m 成立.

矩阵级数, 矩阵逆的范数估计

推论

设 $\|\cdot\|$ 是 $\mathbb{C}^{n\times n}$ 上的一个满足条件 $\|I\| = 1$ 的矩阵范数, 并假定 $A \in \mathbb{C}^{n\times n}$ 满足 $\|A\| < 1$, 则 $I - A$ 可逆, 且有

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

线性方程组的敏度分析

定理

设 $\|\cdot\|$ 是 $\mathbb{R}^{n\times n}$ 上的一个满足条件 $\|I\| = 1$ 的矩阵范数, 并假定 $A \in \mathbb{R}^{n\times n}$ 非奇异, $b \in \mathbb{R}^n$ 非零; 再假定 $\delta A \in \mathbb{R}^{n\times n}$ 满足 $\|A\|^{-1}\|\delta A\| < 1$. 若 x 和 $x+\delta x$ 分别是线性方程组

$$Ax = b \quad \text{和} \quad (A+\delta A)(x+\delta x) = b+\delta b$$

的解, 则

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

其中 $\kappa(A) = \|A\|\left\|A^{-1}\right\|$.

条件数及其几何意义

定义

称数 $\kappa(A) = \|A\| \left\|A^{-1}\right\|$ 为线性方程组 $Ax = b$ 的条件数.

通常, 矩阵 A 的条件数 $\kappa(A)$ 很大, 则称 A 是病态的; 反之, 若 A 的条件数 $\kappa(A)$ 很小, 则称 A 是良态的.

条件数及其几何意义

定义

称数 $\kappa(A) = \|A\| \left\|A^{-1}\right\|$ 为线性方程组 $Ax = b$ 的条件数.

通常, 矩阵 A 的条件数 $\kappa(A)$ 很大, 则称 A 是病态的; 反之, 若 A 的条件数 $\kappa(A)$ 很小, 则称 A 是良态的.
由矩阵范数的等价性可推出, $\mathbb{R}^{n\times n}$ 上任意两个范数下的条件数 $\kappa_\alpha(A)$ 和 $\kappa_\beta(A)$ 都是等价的, 即存在常数 c_1 和 c_2, 使得

$$c_1\kappa_\alpha(A) \leq \kappa_\beta(A) \leq c_2\kappa_\alpha(A).$$

条件数及其几何意义

定义

称数 $\kappa(A) = \|A\| \left\|A^{-1}\right\|$ 为线性方程组 $Ax = b$ 的条件数.

通常, 矩阵 A 的条件数 $\kappa(A)$ 很大, 则称 A 是病态的; 反之, 若 A 的条件数 $\kappa(A)$ 很小, 则称 A 是良态的.
由矩阵范数的等价性可推出, $\mathbb{R}^{n\times n}$ 上任意两个范数下的条件数 $\kappa_\alpha(A)$ 和 $\kappa_\beta(A)$ 都是等价的, 即存在常数 c_1 和 c_2, 使得

$$c_1\kappa_\alpha(A) \le \kappa_\beta(A) \le c_2\kappa_\alpha(A).$$

例如, 有

$$\frac{1}{n}\kappa_2(A) \le \kappa_1(A) \le n\kappa_2(A),$$
$$\frac{1}{n}\kappa_\infty(A) \le \kappa_2(A) \le n\kappa_\infty(A),$$
$$\frac{1}{n^2}\kappa_1(A) \le \kappa_\infty(A) \le n^2\kappa_1(A).$$

条件数及其几何意义

推论

设 $\|\cdot\|$ 是 $\mathbb{C}^{n\times n}$ 上的一个满足条件 $\|I\| = 1$ 的矩阵范数, 并假定 $A \in \mathbb{C}^{n\times n}$ 非奇异, 且 $\delta A \in \mathbb{R}^{n\times n}$ 满足 $\left\|A^{-1}\right\| \|\delta A\| < 1$, 则 $A + \delta A$ 也是非奇异的, 且有

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

条件数及其几何意义

推论

设 $\|\cdot\|$ 是 $\mathbb{C}^{n\times n}$ 上的一个满足条件 $\|I\| = 1$ 的矩阵范数, 并假定 $A \in \mathbb{C}^{n\times n}$ 非奇异, 且 $\delta A \in \mathbb{R}^{n\times n}$ 满足 $\left\|A^{-1}\right\| \|\delta A\| < 1$, 则 $A + \delta A$ 也是非奇异的, 且有

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

定理

设 $A \in \mathbb{R}^{n\times n}$ 非奇异, 则

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

(注意: 定理是在谱范数意义下成立.)

浮点数表示模型

计算机中的浮点数 f 通常可表示为

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

浮点数表示模型

计算机中的浮点数 f 通常可表示为

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

其中:

- β 表示浮点数的基底.
- J 是阶或指数, 指数的位数确定所表示的数的范围.
- w 是尾数, 一般表示为 $w = 0.d_1\, d_2 \cdots d_t$, t 是尾数位数, $0 \leq d_i < \beta$. 若 $d_1 \neq 0$, 则称该浮点数为规格化浮点数. 尾数的位数确定表示数的精度.

浮点数表示模型

计算机中的浮点数 f 通常可表示为

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

其中:

- β 表示浮点数的基底.
- J 是阶或指数, 指数的位数确定所表示的数的范围.
- w 是尾数, 一般表示为 $w = 0.d_1\, d_2 \cdots d_t$, t 是尾数位数, $0 \leq d_i < \beta$. 若 $d_1 \neq 0$, 则称该浮点数为规格化浮点数. 尾数的位数确定表示数的精度.

若用 $\mathcal{F}$ 表示一个系统的浮点数全体所构成的集合, 则

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

浮点数表示模型

计算机中的浮点数 f 通常可表示为

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

其中:

- β 表示浮点数的基底.
- J 是阶或指数, 指数的位数确定所表示的数的范围.
- w 是尾数, 一般表示为 $w = 0.d_1\, d_2 \cdots d_t$, t 是尾数位数, $0 \leq d_i < \beta$. 若 $d_1 \neq 0$, 则称该浮点数为规格化浮点数. 尾数的位数确定表示数的精度.

若用 $\mathcal{F}$ 表示一个系统的浮点数全体所构成的集合, 则

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

- 集合 $\mathcal{F}$ 是一个有限集, 包含 $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$ 个数.

浮点数表示模型

计算机中的浮点数 f 通常可表示为

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

其中:

- β 表示浮点数的基底.
- J 是阶或指数, 指数的位数确定所表示的数的范围.
- w 是尾数, 一般表示为 $w = 0.d_1\, d_2 \cdots d_t$, t 是尾数位数, $0 \leq d_i < \beta$. 若 $d_1 \neq 0$, 则称该浮点数为规格化浮点数. 尾数的位数确定表示数的精度.

若用 $\mathcal{F}$ 表示一个系统的浮点数全体所构成的集合, 则

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

- 集合 $\mathcal{F}$ 是一个有限集, 包含 $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$ 个数.
- 浮点数对称分布在区间 $[m, M]$ 和 $[-M, -m]$ 中, 其中 $m = \beta^{L-1}, M = \beta^U(1 - \beta^{-t})$.

浮点数表示模型

计算机中的浮点数 f 通常可表示为

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

其中:

- β 表示浮点数的基底.
- J 是阶或指数, 指数的位数确定所表示的数的范围.
- w 是尾数, 一般表示为 $w = 0.d_1\, d_2 \cdots d_t$, t 是尾数位数, $0 \leq d_i < \beta$. 若 $d_1 \neq 0$, 则称该浮点数为规格化浮点数. 尾数的位数确定表示数的精度.

若用 $\mathcal{F}$ 表示一个系统的浮点数全体所构成的集合, 则

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

- 集合 $\mathcal{F}$ 是一个有限集, 包含 $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$ 个数.
- 浮点数对称分布在区间 $[m, M]$ 和 $[-M, -m]$ 中, 其中 $m = \beta^{L-1}, M = \beta^U(1 - \beta^{-t})$.
- 浮点数在区间 $[m, M]$ 和 $[-M, -m]$ 中的分布是不等距的, 但在任意 $[\beta^{J-1}, \beta^J)$ 内是等间距分布的.

浮点数图例

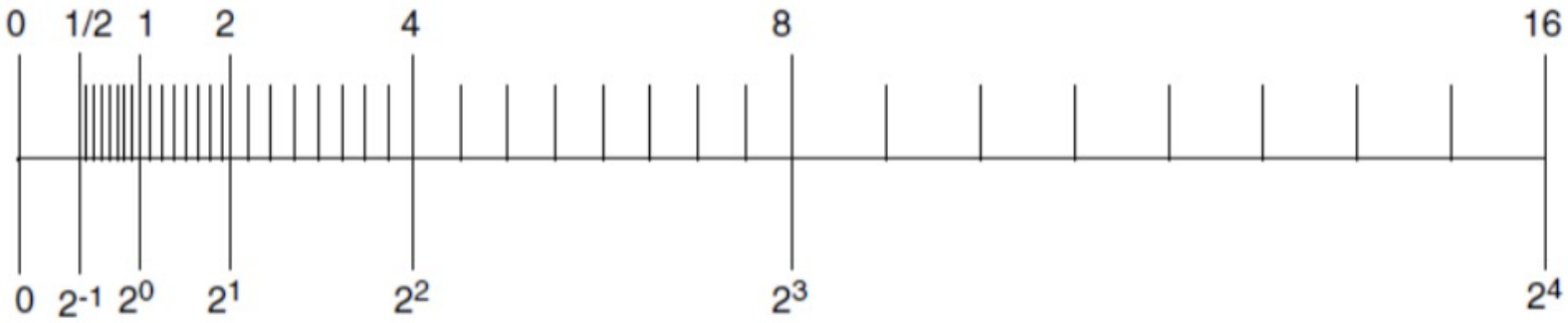

图: $\beta = 2,\ t = 4,\ L = -1,\ U = 4$ 对应的有限个数的分布图.

IEEE 754 标准

IEEE 的浮点数专业小组于七十年代末期开始制定浮点数的标准。在 1980 年, Intel 推出了单片的 8087 浮点数处理器. 由于其浮点数表示法及运算的合理性, 被 IEEE 采用作为浮点数的标准, 并于 1985 年发布.

图: William Kahan, 数学家与计算机科学家, 1989 年获图灵奖, 被称为浮点数之父.

参考文献

G. David

What every computer scientist should know about floating-point arithmetic

ACM Computing Surveys 23:1(1991), 5-48.

M. Overton

Numerical computing with IEEE floating point arithmetic

Vol. 76. SIAM, 2001.

J.-M. Muller, B. Nicolas, F. Dinechin, C.-P. Jeannerod, V. Lefèvre, G. Melquiond, N. Revol, D. Stehlè, and S. Torres

Handbook of floating-point arithmetic

2nd edition, Birkhäuser 2010.

R. Brent and Z. Paul

Modern computer arithmetic

Vol. 18. Cambridge University Press, 2010.

舍入模式与相对误差

由于 $\mathcal{F}$ 只是一个有限集, 无法表示 $[m, M]$ 和 $[-M, -m]$ 中任意实数. 因此在计算机中浮点数的运算是无法精确进行的. 实数表示成浮点数通常采用以下四种舍入模式:

- 向最接近的可表示的值, 当有两个最接近的可表示的值时首选 “偶数” 值;
- 向负无穷大方向 (向下);
- 向正无穷大方向 (向上);
- 向 0 方向 (截断).

舍入模式与相对误差

由于 $\mathcal{F}$ 只是一个有限集, 无法表示 $[m, M]$ 和 $[-M, -m]$ 中任意实数. 因此在计算机中浮点数的运算是无法精确进行的. 实数表示成浮点数通常采用以下四种舍入模式:

- 向最接近的可表示的值, 当有两个最接近的可表示的值时首选 "偶数" 值;
- 向负无穷大方向 (向下);
- 向正无穷大方向 (向上);
- 向 0 方向 (截断).

定理

设 $m \le |x| \le M$, 其中 $m = \beta^{L-1}, M = \beta^U(1-\beta^{-t})$, 则

$$\mathrm{fl}(x) = x(1+\delta), \quad |\delta| \le u, \quad 或 \quad \left|\frac{\mathrm{fl}(x)-x}{x}\right| \le u,$$

其中 u 为机器精度, 即

$$u = \begin{cases} \frac{1}{2}\beta^{1-t}, & 用舍入法, \\ \beta^{1-t}, & 用截断法. \end{cases}$$

基本运算的舍入误差

设 $a, b \in \mathcal{F}$ 是给定的浮点数, $\circ$ 表示 $+, -, \times, /$ 中任一运算, $\mathrm{fl}(a \circ b)$ 表示先计算精确实数, 再按舍入规则表示成浮点数. 若出现 $|a \circ b| > M$ 或 $0 < |a \circ b| < m$, 则称为发生了上溢或下溢.

基本运算的舍入误差

设 $a, b \in \mathcal{F}$ 是给定的浮点数, $\circ$ 表示 $+, -, \times, /$ 中任一运算, $\mathrm{fl}(a \circ b)$ 表示先计算精确实数, 再按舍入规则表示成浮点数. 若出现 $|a \circ b| > M$ 或 $0 < |a \circ b| < m$, 则称为发生了上溢或下溢.

定理

设 $a, b \in \mathcal{F}$, 则

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

基本运算的舍入误差

设 $a, b \in \mathcal{F}$ 是给定的浮点数, $\circ$ 表示 $+, -, \times, /$ 中任一运算, $\mathrm{fl}(a \circ b)$ 表示先计算精确实数, 再按舍入规则表示成浮点数. 若出现 $|a \circ b| > M$ 或 $0 < |a \circ b| < m$, 则称为发生了上溢或下溢.

定理

设 $a, b \in \mathcal{F}$, 则

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

定理

设 $|\delta_i| \leq u$ 且 $nu \leq 0.01$, 那么

$$1 - nu \leq \prod_{i=1}^{n}(1 + \delta_i) \leq 1 + 1.01nu.$$

或写成

$$\prod_{i=1}^{n}(1 + \delta_i) = 1 + \delta, \ |\delta| \leq 1.01nu.$$

向前误差分析法: 例一

通过估计计算解与精确解之间的误差得到舍入误差的界, 舍入误差的界与精确解有关, 这种误差分析方法称为向前误差分析法.

向前误差分析法: 例一

通过估计计算解与精确解之间的误差得到舍入误差的界, 舍入误差的界与精确解有关, 这种误差分析方法称为向前误差分析法.

例

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

向前误差分析法: 例二

例

设 A, B 是两个由浮点数构成的 $n \times n$ 矩阵, α 为浮点数. 记 $|A| = [|a_{ij}|]$, 且规定

$$|A| \leq |B| \text{ 当且仅当 } |a_{ij}| \leq |b_{ij}|, \ i, j = 1, \cdots, n.$$

则下列结论成立:

1. $\mathrm{fl}(\alpha A) = \alpha A + E, \ |E| \leq u|\alpha A|$
2. $\mathrm{fl}(A + B) = (A + B) + E, \ |E| \leq u|A + B|$
3. $\mathrm{fl}(AB) = AB + E, \ |E| \leq 1.01nu|A||B|$[a]

[a]注意: 有可能 $|AB| \ll |A||B|$.

向后误差分析法: 例三

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误差分析方法称为向后误差分析法.

向后误差分析法: 例三

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误差分析方法称为向后误差分析法.

例

设 A, B 为 2×2 的上三角矩阵, 有前面例子可知

$$\mathrm{fl}(AB) = \begin{bmatrix} a_{11}b_{11}(1+\epsilon_1) & \tilde{a}_{12} \\ 0 & a_{22}b_{22}(1+\epsilon_5) \end{bmatrix},$$

其中 $\tilde{a}_{12} = \big(a_{11}b_{12}(1+\epsilon_2) + a_{12}b_{22}(1+\epsilon_3)\big)(1+\epsilon_4)$, $|\epsilon_i| \leq u, i = 1, \cdots, 5.$
若令

$$\tilde{A} = \begin{bmatrix} a_{11} & \tilde{a}_{12}(1+\epsilon_3)(1+\epsilon_4) \\ 0 & a_{22}(1+\epsilon_5) \end{bmatrix}, \ \tilde{B} = \begin{bmatrix} b_{11}(1+\epsilon_1) & \tilde{b}_{12}(1+\epsilon_2)(1+\epsilon_4) \\ 0 & b_{22} \end{bmatrix},$$

易证 $\mathrm{fl}(AB) = \tilde{A}\tilde{B}$, 且 $\tilde{A} = A + E$, $|E| \leq 3u|A|$ 及 $\tilde{B} = B + F$, $|F| \leq 3u|B|$.

列主元 Gauss 消去法的舍入误差分析

引理

设 $n \times n$ 浮点数矩阵 A 有三角分解且 $1.01nu \leq 0.01$, 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 $|E| \leq 2.05nu|\tilde{L}||\tilde{U}|$.

列主元 Gauss 消去法的舍入误差分析

引理

设 $n \times n$ 浮点数矩阵 A 有三角分解且 $1.01nu \leq 0.01$, 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 $|E| \leq 2.05nu|\tilde{L}||\tilde{U}|$.

推论

设 $n \times n$ 浮点数矩阵 A 是非奇异的, 且 $1.01nu \leq 0.01$, 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 $\tilde{U}$ 及排列方阵 $\tilde{P}$ 满足

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

其中 $|E| \leq 2.05nu|\tilde{L}||\tilde{U}|$.

列主元 Gauss 消去法的舍入误差分析

引理

设 $n \times n$ 浮点数三角阵 S 是非奇异的, 并且假定 $1.01nu \leq 0.01$, 则用向前/后代法解三角方程组的方法求解 $Sx = b$ 所得到的计算解 $\tilde{x}$ 满足

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

其中 $|H| \leq 1.01nu|S|$.

列主元 Gauss 消去法的舍入误差分析

引理

设 $n \times n$ 浮点数三角阵 S 是非奇异的, 并且假定 $1.01nu \leq 0.01$, 则用向前/后代法解三角方程组的方法求解 $Sx = b$ 所得到的计算解 $\tilde{x}$ 满足

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

其中 $|H| \leq 1.01nu|S|$.

定理

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

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

其中 $\|\delta A\|_\infty / \|A\|_\infty \leq 4.09n^3\rho u$.

精度估计

设 $\hat{x}$ 为用某种方法解 $Ax = b$ 得到的计算解. 令

$$r = b - A\hat{x},$$

则有

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

于是

$$\|x - \hat{x}\| = \left\|A^{-1}x\right\| \leq \left\|A^{-1}\right\| \|r\| .$$

再由于

$$\|b\| \leq \|A\| \, x,$$

即得

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \left\|A^{-1}\right\| \|A\| \frac{\|r\|}{\|b\|}.$$

特别地, 在上式中取 ∞ 范数便有

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \leq \kappa_\infty(A) \frac{\|r\|_\infty}{\|b\|_\infty}.$$

盲人爬山法

设 $B \in \mathbb{R}^{n\times n}$, 我们来估计 $\|B\|_1$. 定义

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

$$\mathcal{D} = \{x \in \mathbb{R}^n : \|x\|_1 \le 1\}.$$

易证 f 是凸函数, $\mathcal{D}$ 是凸集.

由于 $\left\|A^{-1}\right\|_\infty = \left\|A^{-T}\right\|_1$, 所以令 $B = A^{-T}$, 利用“盲人爬山法”便可得到 $\left\|A^{-1}\right\|_\infty$ 的一个估计值.

迭代改进

若计算解 $\hat{x}$ 得精度太低, 可将 $\hat{x}$ 做为初值, 应用 Newton 迭代法于函数 $f(x) = Ax - b$ 上, 来改进精度. 具体计算过程如下:

1. 计算 $r = b - A\hat{x}$ (用双精度和原始矩阵 A)
2. 求解 $Az = r$ (利用 A 的三角分解)
3. 计算 $x = \hat{x} + z$
4. 若 $\frac{\|x-\hat{x}\|_\infty}{\|x\|_\infty} \leq \epsilon$, 则结束; 否则, 令 $\hat{x} = x$, 转步 (1)