1 引论

1.1 向量和矩阵范数

1.1.1 向量范数的定义

称 $\mathbb{R}^n$ 上的一个函数 $\|\cdot\|$ 为范数，若满足：

1. **非负性** $\|\mathbf{x}\| \geq 0, \quad \|\mathbf{x}\| = 0$ 当且仅当 $\mathbf{x} = \mathbf{0}$
2. **齐次性** $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|,\ \alpha \in \mathbb{R}$
3. **三角不等式** $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

常见的向量范数有：

- **1-范数** $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
- **2-范数** $\|\mathbf{x}\|_2 = (\sum_{i=1}^n |x_i|^2)^{\frac{1}{2}}$
- **∞-范数** $\|\mathbf{x}\|_\infty = \max\limits_{1\leq i\leq n} |x_i|$

1.1.2 向量范数的性质

向量范数的等价性 对于 $\mathbb{R}^n$ 中的任意两种范数 $\|\mathbf{x}\|_\alpha$ 和 $\|\mathbf{x}\|_\beta$，都存在两个正数 c_1, c_2，使得对任意 $\mathbf{x} \in \mathbb{R}^n$ 都有

$$c_1\|\mathbf{x}\|_\alpha \leq \|\mathbf{x}\|_\beta \leq c_2\|\mathbf{x}\|_\alpha$$

证明. 令 $f(\mathbf{x}) = \|\mathbf{x}\|_\beta$，$\mathbf{S} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_\alpha = 1\}$ 为一个有界闭集，则连续函数 $f(\mathbf{x})$ 在 $\mathbf{S}$ 上存在最小值 $c_1 = \min\limits_{\mathbf{x}\in\mathbf{S}} f(\mathbf{x})$ 和最大值 $c_2 = \max\limits_{\mathbf{x}\in\mathbf{S}} f(\mathbf{x})$。对于 $\mathbf{x} \neq \mathbf{0}$ 有 $\frac{\mathbf{x}}{\|\mathbf{x}\|_\alpha} \in \mathbf{S}$，则有

$$c_1 \leq f(\frac{\mathbf{x}}{\|\mathbf{x}\|_\alpha}) = \frac{\|\mathbf{x}\|_\beta}{\|\mathbf{x}\|_\alpha} \leq c_2$$

即

$$c_1\|\mathbf{x}\|_\alpha \leq \|\mathbf{x}\|_\beta \leq c_2\|\mathbf{x}\|_\alpha$$

□

对于常用的向量范数，有如下关系：

$$\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$$

$$\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_1 \le n\|\mathbf{x}\|_\infty$$

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

向量序列的收敛性　在空间 $\mathbb{R}^n$ 中，向量序列 $\{\mathbf{x}^{(k)}\}$ 收敛于向量 $\mathbf{x}^*$ 的充要条件是存在范数 $\|\cdot\|$ 使得

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0$$

压缩映射　设有非空集合 $\mathbf{D} \subset \mathbb{R}^n$，对于映射 $\mathbf{f}: \mathbf{D} \to \mathbf{D}$，若存在范数 $\|\cdot\|$ 和常数 $q \in [0,1)$ 使得对任意 $\mathbf{x}, \mathbf{y} \in \mathbf{D}$ 都有

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \le q\|\mathbf{x} - \mathbf{y}\|$$

则称 $\mathbf{f}$ 为 $\mathbf{D}$ 上的压缩映射。

Banach 压缩映射原理　设 $\mathbf{D} \subset \mathbb{R}^n$ 为闭集，映射 $\mathbf{f}$ 为 $\mathbf{D}$ 上的压缩映射，则 $\mathbf{f}$ 在 $\mathbf{D}$ 上有唯一不动点 $\mathbf{x}$，使得 $\mathbf{f}(\mathbf{x}) = \mathbf{x}$。

证明. 设 $\mathbf{x}^{(0)} \in \mathbf{D}$，构造序列 $\{\mathbf{x}^{(k)}\}$ 如下：

$$\mathbf{x}^{(k+1)} = \mathbf{f}(\mathbf{x}^{(k)}), \quad k = 0, 1, 2, \ldots$$

则对任意 $k \ge 0$，有

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| = \|\mathbf{f}(\mathbf{x}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k-1)})\| \le q\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

由此可得

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le q^k\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$$

对于 $m > n \ge 0$，有

$$\|\mathbf{x}^{(m)} - \mathbf{x}^{(n)}\| \le \sum_{k=n}^{m-1} \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \sum_{k=n}^{m-1} q^k\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| = \frac{q^n - q^m}{1-q}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$$

由于 $q \in [0,1)$，当 $n \to \infty$ 时

$$\lim_{n\to\infty} \|\mathbf{x}^{(m)} - \mathbf{x}^{(n)}\| = 0$$

即序列 $\{\mathbf{x}^{(k)}\}$ 为 Cauchy 序列，故在 $\mathbb{R}^n$ 中收敛，设其极限为 $\mathbf{x}^* \in \mathbf{D}$，则由映射的连续性可得

$$\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^{(k+1)} = \lim_{k\to\infty} \mathbf{f}(\mathbf{x}^{(k)}) = \mathbf{f}(\lim_{k\to\infty} \mathbf{x}^{(k)}) = \mathbf{f}(\mathbf{x}^*)$$

即 $\mathbf{x}^*$ 为 $\mathbf{f}$ 的不动点。设存在不动点 $\mathbf{x}_1, \mathbf{x}_2$，则有

$$\|\mathbf{x}_1 - \mathbf{x}_2\| = \|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\| \le q\|\mathbf{x}_1 - \mathbf{x}_2\|$$

由于 $q \in [0, 1)$，上式只在 $\mathbf{x}_1 = \mathbf{x}_2$ 时成立，故不动点唯一。

$\square$

1.1.3 矩阵范数的定义

称 $\mathbb{R}^{n\times n}$ 上的一个函数 $\|\cdot\|$ 为范数，若满足：

1. **非负性** $\|\mathbf{A}\| \ge 0, \quad \|\mathbf{A}\| = 0$ 当且仅当 $\mathbf{A} = \mathbf{0}$
2. **齐次性** $\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|,\ \alpha \in \mathbb{R}$
3. **三角不等式** $\|\mathbf{A} + \mathbf{B}\| \le \|\mathbf{A}\| + \|\mathbf{B}\|$
4. **矩阵乘法不等式** $\|\mathbf{AB}\| \le \|\mathbf{A}\| \cdot \|\mathbf{B}\|$

常见的矩阵范数有：

- **列范数** $\|\mathbf{A}\|_1 = \max\limits_{1\le j\le n} \sum_{i=1}^n |a_{ij}|$
- **行范数** $\|\mathbf{A}\|_\infty = \max\limits_{1\le i\le n} \sum_{j=1}^n |a_{ij}|$
- **谱范数** $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$
- **F-范数** $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2}$

1.1.4 矩阵范数的性质

算子范数 称向量范数导出的矩阵范数为算子范数，定义如下：

$$\|\mathbf{A}\| = \max_{\mathbf{x}\ne\mathbf{0}} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|$$

由定义可知算子范数是矩阵范数，且与向量范数相容：

$$\|\mathbf{Ax}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\|$$

矩阵范数的等价性　对于 $\mathbb{R}^{n\times n}$ 中的任意两种范数 $\|\mathbf{A}\|_\alpha$ 和 $\|\mathbf{A}\|_\beta$，都存在两个正数 c_1, c_2，使得对任意 $\mathbf{A}\in\mathbb{R}^{n\times n}$ 都有

$$c_1\|\mathbf{A}\|_\alpha \le \|\mathbf{A}\|_\beta \le c_2\|\mathbf{A}\|_\alpha$$

对于常用的矩阵范数，有如下关系：

1. $$\frac{1}{n}\|\mathbf{A}\|_\infty \le \|\mathbf{A}\|_1 \le n\|\mathbf{A}\|_\infty$$

   证明. 对于任意 $a_{ij}\in\mathbf{A}$，有

   $$|a_{ij}| \le \|\mathbf{A}\|_\infty$$

   则有

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

   取最大值可得

   $$\|\mathbf{A}\|_1 = \max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}| \le n\|\mathbf{A}\|_\infty$$

   同理可得

   $$\|\mathbf{A}\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}| \le n\|\mathbf{A}\|_1$$

   □

2. $$\frac{1}{\sqrt{n}}\|\mathbf{A}\|_\infty \le \|\mathbf{A}\|_2 \le \sqrt{n}\|\mathbf{A}\|_\infty$$

   证明. 先证明 $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty$：

   对于任意 $x_i\in\mathbf{x}$，有 $|x_i| \le \|\mathbf{x}\|_2$，取最大值可得 $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2$

   对于任意 $x_i\in\mathbf{x}$，有 $|x_i| \le \|\mathbf{x}\|_\infty$，则有 $|x_i|^2 \le \|\mathbf{x}\|_\infty^2$，对所有 i 求和可得 $\|\mathbf{x}\|_2^2 \le n\|\mathbf{x}\|_\infty^2$，即 $\|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty$

   由算子范数定义可得

   $$\|\mathbf{A}\|_2 = \max_{\mathbf{x}\ne\mathbf{0}}\frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2},\ \|\mathbf{A}\|_\infty = \max_{\mathbf{x}\ne\mathbf{0}}\frac{\|\mathbf{A}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}$$

对于任意 $\mathbf{x} \neq \mathbf{0}$，有

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

取最大值可得

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

对于任意 $\mathbf{x} \neq \mathbf{0}$，有

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

取最大值可得

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

□

1.2 误差

1.2.1 误差的类型

误差描述了数值计算中近似解的精确程度，可分为以下几类：

- **截断误差** 在数值运算中运用近似方法表示准确数值运算或数量而引起的，也叫方法误差。
- **舍入误差** 由于计算机字长的限制而产生的误差。
- 不与数值方法相关的误差，如测量误差等。

1.2.2 误差的度量

绝对误差 设 x^* 为精确值，$\tilde{x}$ 为近似值，则称 $\tilde{x}$ 的绝对误差为

$$E(\tilde{x}) = x^* - \tilde{x}$$

绝对误差具有量纲，反映了近似值与精确值之间的差距，但不能很好地反映近似值的精度。

绝对误差极限 若 $\exists\, \delta > 0$，使得 $|E(\tilde{x})| = |x^* - \tilde{x}| \leq \delta$，则称 δ 为 $\tilde{x}$ 的绝对误差极限。

相对误差　设 x^* 为精确值，$\tilde{x}$ 为近似值，则称 $\tilde{x}$ 的相对误差为

$$E_r(\tilde{x}) = \frac{x^* - \tilde{x}}{\tilde{x}} \times 100 \quad (x^* \neq 0)$$

相对误差不具有量纲，能够较好地反映误差的特性及近似值的精度。

相对误差极限　若 $\exists\, \delta_r > 0$，使得 $|E_r(\tilde{x})| = |\frac{x^*-\tilde{x}}{\tilde{x}}| \leq \delta_r$，则称 δ_r 为 $\tilde{x}$ 的相对误差极限。

1.2.3　有效数字

如果近似值 $\tilde{x}$ 的误差不超过某位的半个单位，该位数字到 $\tilde{x}$ 的第一位非零数字共有 n 位，那么这 n 位数字称为 $\tilde{x}$ 的有效数字。

$$\tilde{x} = \pm 10^k \times 0.a_1 a_2 \cdots a_n$$

$|x^* - \tilde{x}| \leq \frac{1}{2} \times 10^{k-n}$ 时称 $\tilde{x}$ 是 x^* 的 n 位有效数字。

误差和有效数字的关系　由相对误差的定义 $E_r(\tilde{x}) = \frac{E(\tilde{x})}{|\tilde{x}|}$ 可知 $\delta_r(\tilde{x}) = \frac{\delta(\tilde{x})}{|\tilde{x}|}$。有效数字和相对误差限的关系由以下定理给出：

- 若 $\tilde{x}$ 有 n 位有效数字，则 $\left|\frac{x^*-\tilde{x}}{\tilde{x}}\right| \leq \frac{1}{2a_1} \times 10^{1-n}$.
- 若 $\left|\frac{x^*-\tilde{x}}{\tilde{x}}\right| \leq \frac{1}{2(a_1+1)} \times 10^{1-n}$，则 $\tilde{x}$ 至少具有 n 位有效数字。

1.2.4　误差的传播

函数误差的传播　若 $f(x)$ 在 $\tilde{x}$ 的邻域上可微，由其在 $x = \tilde{x}$ 的泰勒展开式 $f(\tilde{x}) \approx f(x^*) + f'(x^*)(\tilde{x} - x^*)$ 近似可得

$$|f(\tilde{x}) - f(x^*)| \leq |f'(\tilde{x})| \cdot |\tilde{x} - x^*|$$

由此对近似函数 $f(\tilde{x})$ 的误差限和相对误差限分别有如下估计式：

$$\begin{cases} \delta f(\tilde{x}) \leq |f'(\tilde{x})| \cdot \delta(\tilde{x}) \\ \delta_r f(\tilde{x}) \leq \left|\frac{f'(\tilde{x})}{f(\tilde{x})}\right| \cdot \delta(\tilde{x}) \end{cases}$$

对于二元函数，若 $f(x,y)$ 在 $(\tilde{x},\tilde{y})$ 的邻域上可微，由其泰勒展开式 $f(x^*,y^*)\approx f(\tilde{x},\tilde{y})+\frac{\partial f(\tilde{x},\tilde{y})}{\partial x}(x^*-\tilde{x})+\frac{\partial f(\tilde{x},\tilde{y})}{\partial y}(y^*-\tilde{y})$ 近似可得

$$|f(x^*,y^*)-f(\tilde{x},\tilde{y})|\leq\left|\frac{\partial f(\tilde{x},\tilde{y})}{\partial x}\right|\cdot|x^*-\tilde{x}|+\left|\frac{\partial f(\tilde{x},\tilde{y})}{\partial y}\right|\cdot|y^*-\tilde{y}|$$

由此对近似函数 $f(\tilde{x},\tilde{y})$ 的误差限和相对误差限分别有如下估计式：

$$\begin{cases}\delta f(\tilde{x},\tilde{y})\leq\left|\frac{\partial f(\tilde{x},\tilde{y})}{\partial x}\right|\cdot\delta(\tilde{x})+\left|\frac{\partial f(\tilde{x},\tilde{y})}{\partial y}\right|\cdot\delta(\tilde{y})\\ \delta_r f(\tilde{x},\tilde{y})=\left|\frac{\delta f(\tilde{x},\tilde{y})}{f(\tilde{x},\tilde{y})}\right|\end{cases}$$

算术误差的传播 将算术运算视为二元函数，可以算出加减乘除运算的误差传播公式：

$$\begin{cases}\delta(\tilde{x}\pm\tilde{y})\leq\delta(\tilde{x})+\delta(\tilde{y})\\ \delta_r(\tilde{x}\pm\tilde{y})\leq\frac{\delta(\tilde{x})+\delta(\tilde{y})}{|\tilde{x}\pm\tilde{y}|}\end{cases}$$

$$\begin{cases}\delta(\tilde{x}\tilde{y})\leq|\tilde{y}|\delta(\tilde{x})+|\tilde{x}|\delta(\tilde{y})\\ \delta_r(\tilde{x}\tilde{y})\leq\frac{\delta(\tilde{x})}{|\tilde{x}|}+\frac{\delta(\tilde{y})}{|\tilde{y}|}=\delta_r(\tilde{x})+\delta_r(\tilde{y})\end{cases}$$

$$\begin{cases}\delta(\frac{\tilde{x}}{\tilde{y}})\leq\frac{1}{|\tilde{y}|}\delta(\tilde{x})+|\frac{\tilde{x}}{\tilde{y}^2}|\delta(\tilde{y})\\ \delta_r(\frac{\tilde{x}}{\tilde{y}})\leq\frac{\delta(\tilde{x})}{|\tilde{x}|}+\frac{\delta(\tilde{y})}{|\tilde{y}|}=\delta_r(\tilde{x})+\delta_r(\tilde{y})\end{cases}$$

1.3 数值计算原则

1.3.1 适定问题

称一个数学问题是适定的，如果它满足以下三个条件：

- 存在解
- 解是唯一的
- 解连续的取决于初边值条件

即适定问题的解满足存在性、唯一性和稳定性三个条件。否则称其为不适定问题。

1.3.2 数值稳定性

对于某个数值算法，其稳定性可分为以下几类：

- 数值不稳定：输入数据的误差在计算过程中不断扩大
- 条件稳定（相对稳定）：算法在一定条件下数值稳定
- 无条件稳定（绝对稳定）：算法在任何条件下都数值稳定

1.3.3 数值计算原则

在进行数值计算时，应遵循以下原则：

1. **避免两个相近数相减**
2. **避免用绝对值过小的数做除数**
3. **防止大数吃掉小数**（避免对数量级差异过大的数作加减法）

除了具体运算中的误差规避，还可以从整体算法设计上控制误差：

- 简化计算步骤，提高计算效率：**减少计算量和减少误差积累**
- 使用数值稳定的算法：**控制误差的传播**