矩阵范数

Eliauk

2024 年 7 月 27 日

目录

1 向量空间的范数

本文的 $\mathbb{F}$ 均指代域 $\mathbb{R}$ 或者 $\mathbb{C}$. 对于一个 $\mathbb{F}$ 上的向量空间 V, 如果函数 $\|\cdot\|: V \to \mathbb{R}$ 满足:

(1) 任取 $v \in V$, 有 $\|v\| \geq 0$, 且 $\|v\| = 0$ 当且仅当 $v = 0$;

(2) 任取 $c \in \mathbb{F}$, $v \in V$, 有 $\|cv\| = |c| \, \|v\|$;

(3) 任取 $u, v \in V$, 有 $\|u + v\| \leq \|u\| + \|v\|$,

那么 $\|\cdot\|$ 被称为 V 上的**范数**. 一个配备了范数的向量空间被称为**赋范向量空间**. Cauchy-Schwarz 不等式告诉我们 V 上的范数可以自然地来源于内积.

定理 1.1 (Cauchy-Schwarz 不等式). 设 V 是内积空间, 那么

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle \langle v, v\rangle \quad \forall u, v \in V.$$

等号成立当且仅当 u, v 线性相关, 即存在 $c \in \mathbb{F}$ 使得 $u = cv$ 或者 $v = cu$.

Proof. 任取 $u, v \in V$, 如果 $v = 0$, 显然不等式成立. 设 $v \neq 0$, 我们将 u 向 v 做投影, 即

$$u = \frac{\langle u, v\rangle}{\langle v, v\rangle} v + w,$$

此时 $\langle w, v\rangle = 0$. 故

$$\langle u, u\rangle = \frac{|\langle u, v\rangle|^2}{\langle v, v\rangle^2} \langle v, v\rangle + \langle w, w\rangle \geq \frac{|\langle u, v\rangle|^2}{\langle v, v\rangle},$$

即

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle \langle v, v\rangle .$$

等号成立当且仅当 $\langle w, w\rangle = 0$, 当且仅当 $w = 0$, 当且仅当 u, v 线性相关. □

推论 1.2. 设 V 是内积空间，定义 $\|\cdot\| : V \to \mathbb{R}$ 为

$$\|v\| = \sqrt{\langle v, v\rangle},$$

那么这是一个范数.

Proof. 任取 $u, v \in V$, 由 Cauchy-Schwarz 不等式, 有

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v\rangle = \langle u, u\rangle + \langle v, v\rangle + 2\operatorname{Re}\langle u, v\rangle \\ &\leq \langle u, u\rangle + \langle v, v\rangle + 2|\langle u, v\rangle| \\ &\leq \langle u, u\rangle + \langle v, v\rangle + 2\|u\|\|v\| \\ &= (\|u\| + \|v\|)^2 . \end{aligned}$$

□

上述范数 $\|v\| = \sqrt{\langle v, v\rangle}$ 被称为**内积诱导的范数**, 在不另外说明的情况下, 内积空间的范数均指代内积诱导的范数. 那么是否范数都能由某一个内积导出? 我们有下面的结论, 这个结论的证明比较繁琐, 这里不予说明.

定理 1.3. 对于 $\mathbb{F}$ 上的赋范向量空间 V，范数 $\|\cdot\|$ 是由某个内积诱导的当且仅当其满足平行四边形恒等式：

$$\frac{1}{2}\left(\|u + v\|^2 + \|u - v\|^2\right) = \|u\|^2 + \|v\|^2 \quad \forall u, v \in V.$$

在 $\mathbb{C}^n$ 上最常见的范数是 l_p-范数, 即对于 $x = (x_1, \ldots, x_n) \in \mathbb{C}^n$, 定义

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

其中 $p \geq 1$, 该范数的三角不等式由 Minkowski 不等式保证. 令 $p \to \infty$, 可以定义 l_∞-范数:

$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\},$$

不难验证其三角不等式成立. 值得一提的是, 当 $n \geq 2$ 的时候 ($n = 1$ 时所有的 l_p-范数都是相同的), 考虑 $x = (1, 0, 0, \ldots, 0)$, $y = (0, 1, 0, \ldots, 0)$, 那么

$$\frac{1}{2}\left(\|x+y\|_p^2 + \|x-y\|_p^2\right) = 2^{2/p}, \quad \|x\|_p^2 + \|y\|_p^2 = 2,$$

所以只有 $p = 2$ 时, l_p-范数才可能满足平行四边形恒等式, 根据定理 1.3, l_p-范数只有在 $p = 2$ 时才能从内积诱导出来.

对于一个赋范向量空间 V, 可以定义度量 $d : V \times V \to \mathbb{R}$ 为

$$d(u, v) = \|u - v\|,$$

这使得 V 成为一个度量空间, 从而拥有自然地拓扑结构. 那么不同的范数诱导出的拓扑结构是否相同? 这就引出了范数等价的概念. 对于复向量空间 V 上的两个范数 $\|\cdot\|_a$ 和 $\|\cdot\|_b$, 如果存在正实数 $C_1 \leq C_2$ 使得

$$C_1 \|v\|_b \leq \|v\|_a \leq C_2 \|v\|_b \quad \forall v \in V,$$

那么我们说范数 $\|\cdot\|_a$ 和 $\|\cdot\|_b$ 是**等价的**. 不难验证这个关系满足自反性、对称性和传递性, 所以这是一个等价关系.

命题 1.4. 对于复向量空间 V 上的两个范数 $\|\cdot\|_a$ 和 $\|\cdot\|_b$，这两个范数给出相同的拓扑结构当且仅当它们是等价的.

Proof. ($\Rightarrow$) 若 $\|\cdot\|_a$ 和 $\|\cdot\|_b$ 给出相同的拓扑结构, 记这两个范数对应的度量分别为 d_a 和 d_b, 开球

$$B_r^{(a)}(x_0) = \{x \in V \mid d_a(x, x_0) < r\}, \quad B_r^{(b)}(x_0) = \{x \in V \mid d_b(x - x_0) < r\}.$$

(V, d_a) 和 (V, d_b) 是相同的拓扑空间表明 $B_1^{(a)}(0)$ 是 (V, d_b) 中的开集, $B_1^{(b)}(0)$ 是 (V, d_a) 中的开集, 所以存在 $r_1 > 0$ 和 $r_2 > 0$ 使得

$$B_{r_1}^{(b)}(0) \subseteq B_1^{(a)}(0), \quad B_{r_2}^{(a)}(0) \subseteq B_1^{(b)}(0),$$

任取 $0 \neq x \in V$, 那么 $y = r_1 x / (2\|x\|_b)$ 满足 $\|y\|_b = r_1/2 < r_1$, 所以 $y \in B_{r_1}^{(b)}(0) \subseteq B_1^{(a)}(0)$, 即

$$\frac{r_1}{2}\frac{\|x\|_a}{\|x\|_b} = \|y\|_a \leq 1,$$

即 $\|x\|_a \leq 2/r_1 \|x\|_b$. 类似地, 可以证明

$$\frac{r_2}{2}\|x\|_b \leq \|x\|_a \leq \frac{2}{r_1}\|x\|_b,$$

总可以令 $r_1, r_2 < 1$, 所以 $\|\cdot\|_a$ 和 $\|\cdot\|_b$ 是等价的.

($\Leftarrow$) 若 $\|\cdot\|_a$ 和 $\|\cdot\|_b$ 是等价的. 即存在正实数 $C_1 \le C_2$ 使得

$$C_1 \|v\|_b \le \|v\|_a \le C_2 \|v\|_b \quad \forall v \in V.$$

若 X 是 (V, d_a) 中的开集, 即任取 $x \in X$, 都存在 $r > 0$ 使得

$$x \in B_r^{(a)}(x) \subseteq X,$$

令 $r' = r/C_2$, 那么 $\|y-x\|_b \le r'$ 就能表明 $\|y-x\|_a \le C_2 \|y-x\|_b \le r$, 所以

$$x \in B_{r'}^{(b)}(x) \subseteq B_r^{(a)}(x) \subseteq X,$$

这表明 X 是 (V, d_b) 中的开集. 同理, 可证 (V, d_b) 中的开集也是 (V, d_a) 中的开集, 故二者的拓扑结构相同. □

命题 1.4 告诉我们等价的范数给出相同的拓扑, 从而在这样的赋范向量空间中, 序列的收敛性、函数的连续性等相关概念都是和所用的范数无关的. 下面的定理告诉我们, 有限维向量空间上的任意两个范数都是等价的.

定理 1.5. V 是有限维实或者复向量空间, $\|\cdot\|_a$ 和 $\|\cdot\|_b$ 是 V 上任意两个范数, 那么 $\|\cdot\|_a$ 和 $\|\cdot\|_b$ 是等价的.

Proof. 设 V 的一组基为 $e_1, \ldots, e_n$. 任取 $x = x_1 e_1 + \cdots + x_n e_n$, 定义 l_1-范数为

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

容易证明这是一个范数. 根据范数等价的传递性, 我们只需要证明任意范数 $\|\cdot\|_a$ 都和 $\|\cdot\|_1$ 等价即可. 即存在正数 $C_1 \le C_2$ 使得

$$C_1 \|x\|_1 \le \|x\|_a \le C_2 \|x\|_1 .$$

显然 $x = 0$ 时成立, 下面假设 $x \ne 0$. 此时等价于证明

$$C_1 \le \left\| \frac{x}{\|x\|_1} \right\|_a \le C_2,$$

即证明: 对于任意的满足 $\|u\|_1 = 1$ 的 $u \in V$, 存在正数 $C_1 \le C_2$ 使得

$$C_1 \le \|u\|_a \le C_2.$$

现在考虑 V 上的度量 $d_1(x, y) = \|x - y\|_1$, 此时 (V, d_1) 成为度量空间. 将 $\|\cdot\|_a$ 视为 $V \to \mathbb{R}$ 的函数, 记 S 是 (V, d_1) 中的单位圆:

$$S = \{x \in V \mid \|x\|_1 = 1\}.$$

此时 S 是紧集, 如果我们能说明 $\|\cdot\|_a$ 是连续函数, 那么就表明 $\|\cdot\|_a$ 限制在 S 上是有界的, 从而完成证明. 所以下面我们说明 $\|\cdot\|_a$ 是连续函数.

给定 $x = x_1e_1 + \cdots + x_ne_n \in V$, 对于任意的 $\epsilon > 0$, 取 $\delta = \epsilon$, 令 $y = y_1e_1 + \cdots + y_ne_n \in V$, 只要 $\|y - x\|_1 \le \delta$, 就有

$$\left|\|y\|_a - \|x\|_a\right| \le \|y - x\|_a \le \sum_{i=1}^{n} |y_i - x_i| \, \|e_i\|_a \le \left(\max_i \|e_i\|_a\right) \|y - x\|_1 \le \left(\max_i \|e_i\|_a\right) \varepsilon,$$

这就证明了 $\|\cdot\|_a$ 是连续函数. □

这告诉我们对于有限维赋范向量空间而言, 无论其采用何种范数, 导出的拓扑结构都是相同的, 这一点使得我们在处理一些问题的时候可以选取更容易处理的范数. 下面就是一个例子.

命题 1.6. 令 V, W 是两个有限维赋范向量空间, 令 $\varphi : V \to W$ 是线性映射, 那么 φ 一定连续.

Proof. 设 $e_1, \ldots, e_n$ 是 V 的一组基, 对于 $x = x_1e_1 + \cdots + x_ne_n$, 定义 V 上的范数 $\|x\| = |x_1| + \cdots + |x_n|$. 根据定理 1.5 和命题 1.4, 我们说明 φ 在这个意义下连续即可. 任取 $y = y_1e_1 + \cdots + y_ne_n$, 那么

$$\|\varphi(x) - \varphi(y)\| = \left\| \sum_{i=1}^{n} (x_i - y_i)\varphi(e_i) \right\| \le \sum_{i=1}^{n} |x_i - y_i| \, \|\varphi(e_i)\| \le \|x - y\| \sum_{i=1}^{n} \|\varphi(e_i)\|,$$

这就表明 φ 是连续映射. □

2 矩阵范数

令 $M_n(\mathbb{F})$ 表示 $\mathbb{F}$ 上的 n 阶矩阵构成的向量空间, 此时其同构于向量空间 $\mathbb{F}^{n^2}$, 所以可以自然地采用 $\mathbb{F}^{n^2}$ 上的范数, 但是 $M_n(\mathbb{F})$ 有重要的乘法结构, 所以对于矩阵范数, 我们定义一个额外的条件.

如果函数 $\|\cdot\| : M_n(\mathbb{F}) \to \mathbb{R}$ 满足:

(1) 任取 $A \in M_n(\mathbb{F})$, 有 $\|A\| \ge 0$, 且 $\|A\| = 0$ 当且仅当 $A = 0$;

(2) 任取 $c \in \mathbb{F}$, $A \in M_n(\mathbb{F})$, 有 $\|cA\| = |c| \, \|A\|$;

(3) 任取 $A, B \in M_n(\mathbb{F})$, 有 $\|A + B\| \le \|A\| + \|B\|$;

(4) 任取 $A, B \in M_n(\mathbb{F})$, 有 $\|AB\| \le \|A\| \, \|B\|$,

那么 $\|\cdot\|$ 被称为**矩阵范数**. 根据定义, 总是有 $\|A^2\| \le \|A\|^2$, 所以对于非零幂等矩阵 $A^2 = A$ 来说, 总是有 $\|A\| \ge 1$. 特别地, 单位矩阵的范数 $\|I\| \ge 1$. 若 A 是可逆矩阵, 那么 $\|I\| \le \|A\| \, \|A^{-1}\|$, 所以我们有逆矩阵范数的一个下界:

$$\left\|A^{-1}\right\| \ge \frac{\|I\|}{\|A\|}.$$

我们将矩阵视为 n^2 维向量, 自然可以继承向量的 l_p-范数, 此时我们只需要验证矩阵范数的条件 (4) 即可.

例 2.1. 矩阵 $A \in M_n(\mathbb{F})$ 的 l_1-范数定义为

$$\|A\|_{[1]} = \sum_{i,j=1}^{n} |a_{ij}|,$$

此时

$$\begin{aligned}
\|AB\|_{[1]} &= \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right| \le \sum_{i,j=1}^{n} \sum_{k=1}^{n} |a_{ik} b_{kj}| \\
&\le \sum_{i,j=1}^{n} \sum_{k,\ell=1}^{n} |a_{ik} b_{\ell j}| = \left(\sum_{i,k=1}^{n} |a_{ik}| \right) \left(\sum_{j,\ell=1}^{n} |b_{\ell j}| \right) \\
&= \|A\|_{[1]} \|B\|_{[1]},
\end{aligned}$$

所以 l_1-范数是矩阵范数. 这里我们使用 $\|A\|_{[1]}$ 而不是 $\|A\|_1$ 来表示 l_1-范数是因为将 $\|A\|_1$ 的记号留给后面的算子范数, 算子范数更加常用.

例 2.2. 矩阵 $A \in M_n(\mathbb{F})$ 的 l_2-范数 (也被称为 Frobenius 范数) 定义为

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

此时

$$\begin{aligned}
\|AB\|_F &= \left(\sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \right)^{1/2} \le \left(\sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{\ell=1}^{n} |b_{\ell j}|^2 \right) \right)^{1/2} \\
&= \left(\sum_{i,k=1}^{n} |a_{ik}|^2 \right)^{1/2} \left(\sum_{j,\ell=1}^{n} |b_{\ell j}|^2 \right)^{1/2} = \|A\|_F \|B\|_F,
\end{aligned}$$

所以 l_2-范数是矩阵范数. 注意到 A^*A 的迹是 A^*A 的特征值之和, 所以 $\|A\|_F$ 是 A 的奇异值的平方和的平方根. 此外, 容易证明

$$\|A\|_F = \|A^*\|_F, \quad \|A\|_F = \|PAQ\|_F,$$

其中 P, Q 是酉 (正交) 矩阵.

例 2.3. 向量的 l_∞-范数推广到矩阵为

$$\|A\|_{[\infty]} = \max_{i,j} |a_{ij}|,$$

但是这不是一个矩阵范数, 考虑

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}, \quad A^2 = nA,$$

但是 $\left\|A^2\right\|_{[\infty]} = n \geq 1 = \|A\|_{[\infty]}^2$, 不满足矩阵范数的条件 (4).

下面我们定义一个非常重要的矩阵范数, 通常被称为**算子范数**. 令 $\|\cdot\|$ 为 $\mathbb{F}^n$ 上的向量范数, 我们定义 $M_n(\mathbb{F})$ 上的矩阵范数为

$$\|A\| = \sup\left\{\|Ax\| \mid x \in \mathbb{F}^n, \|x\| \leq 1\right\}.$$

由于 $\{x \in \mathbb{F}^n \mid \|x\| \leq 1\}$ 是度量空间 $\mathbb{F}^n$ 中的有界闭集, 从而是紧集, 所以连续函数 $x \mapsto \|Ax\|$ 是有界的, 并且能取到最大值, 所以定义中的 sup 可以改为 max.

命题 2.4. 矩阵 A 的算子范数的等价形式:

$$\|A\| = \sup_{\|x\| \leq 1} \|Ax\| = \sup_{\|x\|=1} \|Ax\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \inf\left\{c \mid \|Ax\| \leq c\,\|x\|, \forall x \in \mathbb{F}^n\right\}.$$

Proof. 记右端四个值分别为 S_1, S_2, S_3, S_4, 显然 $S_2 \leq S_1$. 因为

$$\frac{\|Ax\|}{\|x\|} = \left\|A\left(\frac{x}{\|x\|}\right)\right\| \leq S_2,$$

所以 $S_3 \leq S_2$. 当 $\|x\| \leq 1$ 的时候, 有

$$\|Ax\| \leq \frac{\|Ax\|}{\|x\|} \leq S_3,$$

所以 $S_1 \leq S_3$. 这就表明 $S_1 = S_2 = S_3$. 注意到

$$\|Ax\| \leq S_3\,\|x\| \quad \forall x \in \mathbb{F}^n,$$

所以 $S_4 \leq S_3$. 根据 sup 的定义, 对于任意的正整数 n, 都存在 $x_n \in \mathbb{F}^n$ 使得

$$S_4 \geq \|Ax_n\| / \|x_n\| \geq S_3 - 1/n,$$

这就表明 $S_3 = S_4$. □

定理 2.5. 上述算子范数满足下面的性质：

(1) $\|I\| = 1$；

(2) 对于任意的 $A \in M_n(\mathbb{F})$ 和 $v \in \mathbb{F}^n$，有 $\|Av\| \le \|A\|\,\|v\|$；

(3) 算子范数是一种矩阵范数.

Proof. (1) 由于

$$\sup_{\|x\|=1} \|Ix\| = \sup_{\|x\|=1} \|x\| = 1,$$

所以 $\|I\| = 1$.

(2) $v = 0$ 时显然成立. 假设 $v \neq 0$, 那么

$$\frac{1}{\|v\|}\|Av\| = \left\| A\frac{v}{\|v\|} \right\| \le \max_{\|x\|=1} \|Ax\| = \|A\|,$$

所以 $\|Av\| \le \|A\|\,\|v\|$.

(3) 矩阵范数显然满足正定性和齐次性. 任取单位向量 $x \in \mathbb{F}^n$, 那么

$$\|(A+B)x\| = \|Ax + Bx\| \le \|Ax\| + \|Bx\| \le \|A\| + \|B\|,$$

所以 $\|A+B\| \le \|A\| + \|B\|$. 此外, 根据 (2), 还有

$$\|(AB)x\| = \|A(Bx)\| \le \|A\|\,\|Bx\| \le \|A\|\,\|B\|,$$

所以 $\|AB\| \le \|A\|\,\|B\|$. 这就表明算子范数是一种矩阵范数. □

所以算子范数 $\|A\|$ 通常被称为向量空间 $\mathbb{F}^n$ 上的范数 $\|\cdot\|$ 诱导的矩阵范数. 定理 2.5 表明向量空间上任意范数都可以诱导出一个矩阵范数. 如果矩阵范数和向量范数满足 定理 2.5 的 (2), 那么我们说这个矩阵范数和向量范数是相容的. 对于任意向量范数, 都存在与之相容的矩阵范数（即诱导的矩阵范数）.

例 2.6. 我们考察 $\mathbb{F}^n$ 上的 l_1-范数诱导的矩阵范数. 对于矩阵 $A \in M_n(\mathbb{F})$, 设其列向量为 $A = [\alpha_1, \ldots, \alpha_n]$, 任取 $x = (x_1, \ldots, x_n)$, 那么

$$\|Ax\|_1 = \left\| \sum_{i=1}^n x_i\alpha_i \right\|_1 \le \sum_{i=1}^n |x_i|\,\|\alpha_i\|_1 \le \left(\max_{1\le i\le n} \|\alpha_i\|_1 \right) \sum_{i=1}^n |x_i| = \|x\|_1 \left(\max_{1\le i\le n} \|\alpha_i\|_1 \right),$$

另一方面, 我们取 $x = e_i$ 是标准基, 那么

$$\sup_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} \ge \|Ae_i\|_1 = \|\alpha_i\|_1,$$

这表明

$$\sup_{x\neq 0}\frac{\|Ax\|_1}{\|x\|_1}\geq \max_{1\leq i\leq n}\|\alpha_i\|_1 .$$

故 A 的 1-范数为:

$$\|A\|_1=\sup_{x\neq 0}\frac{\|Ax\|_1}{\|x\|_1}=\max_{1\leq i\leq n}\|\alpha_i\|_1=\max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ji}|.$$

即 $\|A\|_1$ 为 A 的列向量 l_1-范数的最大值.

例 2.7. 考察 $\mathbb{F}^n$ 上的 l_∞-范数诱导的矩阵范数. 任取 $x=(x_1,\ldots,x_n)$, 那么

$$\|Ax\|_\infty=\max_{1\leq i\leq n}\left|\sum_{j=1}^{n}a_{ij}x_j\right|\leq \max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}||x_j|\leq \|x\|_\infty\left(\max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}|\right),$$

这表明

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

现在假设 $A\neq 0$, 那么设 $a_{ij}\neq 0$. 令 $y=(y_1,\ldots,y_n)$ 满足

$$y_k=\begin{cases}\bar{a}_{ik}/|a_{ik}| & a_{ik}\neq 0,\\ 1 & a_{ik}=0.\end{cases}$$

那么 $\|y\|_\infty=1$, 并且 $a_{ik}y_k=|a_{ik}|$, 所以

$$\|A\|_\infty=\sup_{\|x\|_\infty=1}\|Ax\|_\infty\geq \|Ay\|_\infty=\max_{1\leq i\leq n}\left|\sum_{k=1}^{n}a_{ik}y_k\right|\geq \left|\sum_{k=1}^{n}a_{ik}y_k\right|\geq \sum_{k=1}^{n}|a_{ik}|,$$

所以

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

故 A 的 ∞-范数为

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

即 $\|A\|_\infty$ 为 A 的行向量 l_1-范数的最大值.

例 2.8. 考察 $\mathbb{F}^n$ 上的 l_2-范数诱导的矩阵范数, 这个范数被称为**谱范数**. 设 $s_1,\ldots,s_k$ 是 A 的非零奇异值 (从大到小排列), 根据奇异值分解, 那么存在 $\mathbb{F}^n$ 的正交向量组 $e_1,\ldots,e_k$ 和 $f_1,\ldots,f_k$ 使得

$$Ax=s_1\langle x,e_1\rangle f_1+\cdots+s_k\langle x,e_k\rangle f_k,$$

所以

$$\begin{aligned}\|Ax\|_2^2 &= s_1^2\,|\langle x, e_1\rangle|^2 + \cdots + s_k^2\,|\langle x, e_k\rangle|^2 \\ &\le s_1^2\left(|\langle x, e_1\rangle|^2 + \cdots + |\langle x, e_k\rangle|^2\right) \le s_1^2\,\|x\|_2^2,\end{aligned}$$

这表明 $\|A\|_2 \le s_1$. 另一方面, $\|Ae_1\|_2 = \|s_1 f_1\|_2 = s_1$, 所以 $\|A\|_2 \ge s_1$, 故

$$\|A\|_2 = s_1.$$

即 $\|A\|_2$ 为 A 的奇异值的最大值.

3 谱半径和矩阵级数

矩阵范数的第一个重要应用在于提供了矩阵谱半径的界. 矩阵 A 的谱半径为其特征值模长的最大值, 记为 $\rho(A)$.

命题 3.1. 若 $A \in M_n(\mathbb{F})$ 是正规矩阵, 那么

$$\rho(A) = \|A\|_2 = s_1,$$

其中 s_1 是 A 的最大的奇异值.

Proof. 将 A 始终看作复正规矩阵, 根据谱定理, 存在正交矩阵 P 使得

$$P^{-1}AP = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

其中 $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$ 是 A 的所有特征值(注意正规矩阵的特征值都是实数), 于是 $\rho(A) = |\lambda_1|$. A 和 $P^{-1}AP$ 的奇异值相同,而 $P^{-1}AP$ 的奇异值显然为 $|\lambda_1|, \ldots, |\lambda_n|$,所以 $\rho(A) = \|A\|_2$. □

设 λ 是 A 的一个特征值, x 是 λ 的特征向量, 考虑矩阵 $X = (x, \ldots, x) \in M_n(\mathbb{F})$, 此时有 $AX = \lambda X$. 如果 $\|\cdot\|$ 是一个矩阵范数, 那么

$$|\lambda|\,\|X\| = \|\lambda X\| = \|AX\| \le \|A\|\,\|X\|,$$

因此 $|\lambda| \le \|A\|$, 这就表明 $\rho(A) \le \|A\|$. 若 A 是可逆矩阵, 那么 λ^{-1} 是 A^{-1} 的特征值, 所以 $|\lambda^{-1}| \le \left\|A^{-1}\right\|$, 即 $|\lambda| \ge 1/\left\|A^{-1}\right\|$. 于是我们得到了下面的结论.

定理 3.2. 令 $\|\cdot\|$ 是矩阵范数, $A \in M_n(\mathbb{F})$, λ 是 A 的特征值, 那么

$$|\lambda| \le \rho(A) \le \|A\|,$$

如果 A 可逆, 那么

$$\rho(A) \ge |\lambda| \ge 1/\left\|A^{-1}\right\|.$$

定理 3.3. 给定 $A \in M_n(\mathbb{F})$ 和 $\epsilon > 0$，存在一个矩阵范数 $\|\cdot\|$ 使得

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

Proof. 设 A 的 Jordan 标准型为

$$P^{-1}AP = \begin{pmatrix} J_{n_1}(\lambda_1) & & & \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{n_k}(\lambda_k) \end{pmatrix} = J,$$

其中 $\lambda_1, \dots, \lambda_k$ 是 A 的所有特征值，$n_1 + \cdots + n_k = n$. 令

$$D_{n_i}(\epsilon) = \operatorname{diag}(\epsilon, \epsilon^2, \dots, \epsilon^{n_i}), \quad D = \operatorname{diag}\big(D_{n_1}(\epsilon), \dots, D_{n_k}(\epsilon)\big),$$

定义 $\mathbb{F}^n$ 上的向量范数

$$\|x\|_D = \left\|D^{-1}P^{-1}x\right\|_1,$$

不难验证这确实是一个范数. 考虑其诱导的矩阵范数，有

$$\begin{aligned} \|A\|_D &= \max_{x \ne 0} \frac{\|Ax\|_D}{\|x\|_D} = \max_{x \ne 0} \frac{\left\|D^{-1}P^{-1}Ax\right\|_1}{\|D^{-1}P^{-1}x\|_1} \\ &= \max_{y \ne 0} \frac{\left\|D^{-1}P^{-1}APDy\right\|_1}{\|y\|_1} = \left\|D^{-1}JD\right\|_1. \end{aligned}$$

计算得

$$D^{-1}JD = \operatorname{diag}\big(B_{n_1}(\lambda_1, \epsilon), \dots, B_{n_k}(\lambda_k, \epsilon)\big),$$

其中

$$B_{n_i}(\lambda_i, \epsilon) = \begin{pmatrix} \lambda_i & \epsilon & & & \\ & \lambda_i & \epsilon & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & \epsilon \\ & & & & \lambda_i \end{pmatrix},$$

这就表明 $\|A\|_D \le \rho(A) + \epsilon$. □

定理 3.2 和 定理 3.3 共同表明

$$\rho(A) = \inf\{\|A\| \mid \|\cdot\| \text{ is a matrix norm}\}. \tag{1}$$

这给出了谱半径和矩阵范数的一个重要关系，而谱半径是研究矩阵级数的一个重要工具.

引理 3.4. 令 $A \in M_n(\mathbb{F})$，如果存在矩阵范数 $\|\cdot\|$ 使得 $\|A\| < 1$，那么 $\lim_{k\to\infty} A^k = 0$，这表明 A^k 的每个元素都趋于 0.

Proof. 我们有 $\|A^k\| \le \|A\|^k \to 0$, 所以 A^k 在范数 $\|\cdot\|$ 的意义下收敛到 0, 定理 1.5 表明任意两个矩阵范数都是等价的, 所以 A^k 在任意范数意义下收敛到 0, 特别地, $A^k \to 0$ 在无穷范数 $\|\cdot\|_\infty$ 意义下成立, 即 A^k 的行向量模长的最大值趋于零, 即 A^k 的每个元素都趋于 0. □

定理 3.5. 令 $A \in M_n(\mathbb{F})$，那么 $\lim_{k\to\infty} A^k = 0$ 当且仅当 $\rho(A) < 1$.

Proof. $(\Rightarrow)$ 若 $A^k \to 0$, 设 x 是 A 的特征向量, 即 $Ax = \lambda x$, 那么 $\lambda^k x = A^k x \to 0$, 故 $\lambda^k \to 0$, 这表明 $|\lambda| < 1$, 故 $\rho(A) < 1$.

$(\Leftarrow)$ 若 $\rho(A) < 1$, (1) 式表明存在某个矩阵范数 $\|\cdot\|$ 使得 $\|A\| < 1$, 引理 3.4 表明 $A^k \to 0$. □

推论 3.6 (Gelfand). 令 $A \in M_n(\mathbb{F})$，$\|\cdot\|$ 是一个矩阵范数，那么

$$\rho(A) = \lim_{k\to\infty} \left\|A^k\right\|^{1/k}.$$

Proof. 因为 $\rho(A)^k = \rho(A^k) \le \|A^k\|$, 所以 $\rho(A) \le \|A^k\|^{1/k}$. 任取 $\epsilon > 0$, 记 $B = (\rho(A)+\epsilon)^{-1}A$, 那么 $\rho(B) = \rho(A)/(\rho(A)+\epsilon) < 1$, 根据定理 3.5, 所以 $B^k \to 0$, 故 $\|B^k\| \to 0$, 所以存在 N 使得 $k \ge N$ 时 $\|B^k\| < 1$, 此时

$$\left\|A^k\right\|^{1/k} = \left\|(\rho(A)+\epsilon)^k B^k\right\|^{1/k} = (\rho(A)+\epsilon)\left\|B^k\right\|^{1/k} < \rho(A)+\epsilon,$$

所以 $k \ge N$ 时有 $\rho(A) \le \|A^k\|^{1/k} < \rho(A)+\epsilon$, 故

$$\rho(A) = \lim_{k\to\infty} \left\|A^k\right\|^{1/k}.$$ □

回顾幂级数, 对于一个复幂级数 $\sum_{k=0}^\infty a_k z^k$, 我们知道其有一个收敛半径 R, 当 $|z| < R$ 时该幂级数是绝对收敛的, $|z| > R$ 时幂级数发散. 下面我们考虑对应的矩阵级数 $\sum_{k=0}^\infty a_k A^k$, 注意到

$$\left\|\sum_{k=0}^n a_k A^k\right\| \le \sum_{k=0}^n |a_k| \left\|A^k\right\| \le \sum_{k=0}^n |a_k| \|A\|^k,$$

所以当 $\|A\| < R$ 时, $\sum_{k=0}^\infty a_k A^k$ 是绝对收敛的, 从而收敛. 注意正这个矩阵范数是可以任意选取的, (1) 式告诉我们存在使得 $\|A\| < R$ 的矩阵范数当且仅当 $\rho(A) < R$. 于是我们得到了下面的定理.

定理 3.7. 设 R 是幂级数 $\sum_{k=0}^\infty a_k z^k$ 的收敛半径，矩阵 $A \in M_n(\mathbb{C})$，那么 $\rho(A) < R$ 时矩阵级数 $\sum_{k=0}^\infty a_k A^k$ 收敛. 等价地说，如果存在矩阵范数 $\|\cdot\|$ 使得 $\|A\| < R$，那么 $\sum_{k=0}^\infty a_k A^k$ 收敛.

由于 $e^z = \sum_{k=0}^{\infty} z^k/k!$ 的收敛半径是无穷大, 所以对于任意的复矩阵 A, $e^A = \sum_{k=0}^{\infty} A^k/k!$ 总是有意义的. 类似地, $\sin A, \cos A$ 也都是良好定义的. $\log(1+z) = \sum_{k=1}^{\infty}(-1)^{k-1}z^k/k$ 在 $|z| < 1$ 时收敛, 故 $\rho(A) < 1$ 时, $\log(I + A)$ 是良好定义的.

推论 3.8. 令 $A \in M_n(\mathbb{F})$，如果存在矩阵范数 $\|\cdot\|$ 使得 $\|I - A\| < 1$，那么 A 可逆，此时

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

Proof. 如果 $\|I - A\| < 1$, 那么 $\sum_{k=0}^{\infty}(I - A)^k$ 收敛, 并且

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

令 $n \to \infty$, 便有

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

□

4 条件数和线性方程组

给定一个可逆矩阵 $A \in M_n(\mathbb{F})$, 计算机计算 A^{-1} 时存在舍入误差和截断误差, 此外 A 的元素也可能是实验得到的, 存在不确定性. 所以我们要研究这些误差如何影响逆矩阵的计算.

给定一个矩阵范数 $\|\cdot\|$. 我们希望计算 A 的逆矩阵, 但是实际情况中我们需要处理的可能是矩阵 $B = A + \Delta A$. 当然, 我们需要保证 B 仍然是可逆的, 注意到 $B = A(I + A^{-1}\Delta A)$, 所以我们要求

$$\left\|A^{-1}\Delta A\right\| < 1, \tag{2}$$

此时 $\rho(A^{-1}\Delta A) \le \left\|A^{-1}\Delta A\right\| < 1$, 这表明 -1 不是 $A^{-1}\Delta A$ 的特征值, 所以 0 不是 $I + A^{-1}\Delta A$ 的特征值, 从而 $I + A^{-1}\Delta A$ 可逆, 从而保证 B 是可逆的. 从直观上来说, 这要求扰动 ΔA 不能太大从而造成 B 不可逆. 从拓扑视角来看, 由于可逆矩阵的集合是矩阵空间的开集, 所以只要 ΔA 足够小, 总能使得 B 可逆.

我们有 $A^{-1}(\Delta A)B^{-1} = A^{-1}(B - A)B^{-1} = A^{-1} - B^{-1}$, 所以

$$\left\|A^{-1} - B^{-1}\right\| = \left\|A^{-1}(\Delta A)B^{-1}\right\| \le \left\|A^{-1}\Delta A\right\|\left\|B^{-1}\right\|. \tag{3}$$

因为 $B^{-1} = A^{-1} - A^{-1}(\Delta A)B^{-1}$, 所以

$$\left\|B^{-1}\right\| \le \left\|A^{-1}\right\| + \left\|A^{-1}(\Delta A)B^{-1}\right\| \le \left\|A^{-1}\right\| + \left\|A^{-1}\Delta A\right\|\left\|B^{-1}\right\|,$$

这表明
$$\|B^{-1}\| = \|(A+\Delta A)^{-1}\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}\Delta A\|}, \tag{4}$$
代入 (3) 式, 得到
$$\|A^{-1}-B^{-1}\| \le \frac{\|A^{-1}\|\,\|A^{-1}\Delta A\|}{1-\|A^{-1}\Delta A\|} \le \frac{\|A^{-1}\|\,\|A^{-1}\|\,\|\Delta A\|}{1-\|A^{-1}\Delta A\|},$$
这表明所求逆矩阵 B^{-1} 和真实逆矩阵 A^{-1} 之间的相对误差满足
$$\frac{\|A^{-1}-B^{-1}\|}{\|A^{-1}\|} \le \frac{\|A^{-1}\|\,\|A\|}{1-\|A^{-1}\Delta A\|}\frac{\|\Delta A\|}{\|A\|}. \tag{5}$$

对于矩阵 A, 定义**相对于矩阵范数 $\|\cdot\|$ 的条件数**为
$$\kappa(A) = \begin{cases} \|A^{-1}\|\,\|A\| & \text{if } A \text{ is nonsingular,} \\ \infty & \text{if } A \text{ is singular.} \end{cases}$$
注意到 $\kappa(A) = \|A^{-1}\|\,\|A\| \ge \|A^{-1}A\| = \|I\| \ge 1$.

如果我们把 (2) 的要求加强一些, 即要求
$$\|A^{-1}\|\,\|\Delta A\| < 1, \tag{6}$$
此时相对误差 (5) 可以进一步变为
$$\frac{\|A^{-1}-B^{-1}\|}{\|A^{-1}\|} \le \frac{\|A^{-1}\|\,\|A\|}{1-\|A^{-1}\|\,\|\Delta A\|}\frac{\|\Delta A\|}{\|A\|} = \frac{\kappa(A)}{1-\kappa(A)\frac{\|\Delta A\|}{\|A\|}}\frac{\|\Delta A\|}{\|A\|},$$
这能进一步看出条件数的意义. 可以发现, 当条件数不大的时候, 逆矩阵 B^{-1} 的相对误差和 $\|\Delta A\|\,/\,\|A\|$ 的阶是差不多的, 也就是说此时只要扰动 ΔA 不大, 逆矩阵 B^{-1} 也不会距离真实的 A^{-1} “太远”. 如果条件数很大, 那么这个误差的上界就会很大, 此时即使扰动 ΔA 很小, 也可能造成 B^{-1} 与 A^{-1} 有很大的误差. 所以当条件数 $\kappa(A)$ 越大的时候, 我们说矩阵 A 越**病态**.

命题 4.1. 在采用谱范数 (例 2.8) 的情况下，设可逆矩阵 $A \in M_n(\mathbb{F})$ 的奇异值为 $s_1 \ge \cdots \ge s_n$，那么
$$\kappa(A) = \frac{s_1}{s_n},$$
即 A 的相对于谱范数的条件数为最大奇异值与最小奇异值之比. 特别地，根据 命题 3.1，如果 A 是正规矩阵，那么 $\kappa(A) = \rho(A)\rho(A^{-1})$.

Proof. A^{-1} 的奇异值为 $1/s_n \ge \cdots \ge 1/s_1$, 所以
$$\kappa(A) = \|A^{-1}\|_2\,\|A\|_2 = \frac{s_1}{s_n}.$$
□

命题 4.2. 对于任意矩阵 A, B, 任取矩阵范数 $\|\cdot\|$, 都有

$$\kappa(AB) \le \kappa(A)\kappa(B).$$

Proof. 若 A, B 至少有一个不可逆, 结论显然成立. 假设 A, B 都可逆, 那么

$$\kappa(AB) = \left\|B^{-1}A^{-1}\right\| \|AB\| \le \left\|B^{-1}\right\| \left\|A^{-1}\right\| \|A\| \|B\| = \kappa(A)\kappa(B). \qquad \square$$

利用 $\rho(A) \le \|A\|$, 所以矩阵的条件数有下届

$$\kappa(A) \ge \rho(A)\rho(A^{-1}),$$

这表明如果 A 特征值模长的最大值和最小值之比如果很大, 那么其条件数也会很大.

例 4.3 (Hilbert 矩阵)**.** Hilbert 矩阵被定义为 $H_n = [(i+j-1)^{-1}] \in M_n(\mathbb{F})$, 即 H_n 的 (i,j)-元为 $(i+j-1)^{-1}$, 例如

$$H_5 = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{pmatrix}.$$

显然 H_n 是对称阵. 其行列式为

$$\det H_n = \frac{(1!2!\cdots(n-1)!)^4}{1!2!\cdots(2n-1)!},$$

故 Hilbert 矩阵总是可逆. Hilbert 矩阵是一个经典的病态矩阵的例子. 下面是 $n = 2, \ldots, 8$ 的时候 H_n 相对于谱范数的条件数列表:

$n =$	2	3	4	5	6	7	8
$\kappa(H_n) \approx$	19	524	15513	4.8×10^5	1.5×10^7	4.8×10^8	1.5×10^{10}

可以发现 $\kappa(H_n)$ 随着 n 的增大增长非常快, 实际上, 我们有

$$\kappa(H_n) \sim e^{cn} \quad n \to \infty,$$

其中常数 c 大约为 3.5. 根据命题 4.1, 我们有 $\kappa(H_n) = \rho(H_n)\rho(H_n^{-1})$. 意外的是, 当 $n \to \infty$ 时, 其谱半径有渐近 $\rho(H_n) = \pi + \mathcal{O}(1/\log n)$, 所以其谱半径趋于一个不大的常数, 但是其条件数却

增长非常迅速, 这意味着 $\rho(H_n^{-1}) = \kappa(H_n)/\rho(H_n)$ 增长非常迅速, 而 H_n^{-1} 的特征值是 H_n 特征值的倒数, 所以这表明 H_n 的特征值的最小值会随着 n 增大迅速减小. 以 H_5 为例, 我们有

$$H_5^{-1} = \begin{pmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{pmatrix}.$$

此时我们添加微小的扰动 $\Delta H_5 = 0.01 I_5$, 逆矩阵变为

$$(H_5 + \Delta H_5)^{-1} \approx \begin{pmatrix} 4.39 & -9.84 & 1.14 & 2.59 & 2.25 \\ -9.84 & 44.49 & -29.22 & -10.67 & 0.50 \\ 1.14 & -29.22 & 70.39 & -24.55 & -19.37 \\ 2.59 & -10.67 & -24.55 & 69.52 & -32.38 \\ 2.25 & 0.50 & -19.37 & -32.38 & 60.1125 \end{pmatrix},$$

其与真实的 H_5^{-1} 已经差别非常大了, 这意味着计算 Hilbert 矩阵的逆矩阵的时候, 即使微小的误差也会带来结果上巨大的差异, 这也是这类矩阵被称为**病态矩阵**的原因.

下面我们考虑条件数对线性方程组的影响. 对于线性方程组

$$Ax = b,$$

其中 A 可逆, $b \in \mathbb{F}^n$ 非零. 添加扰动 ΔA 和 Δb, 此时实际求解方程组

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

设 $\tilde{x} = x + \Delta x$, 我们分析 Δx 的大小. 取一个与向量范数相容的矩阵范数 $\|\cdot\|$, 即其满足 $\|Ax\| \leq \|A\|\,\|x\|$. 那么

$$\begin{aligned} (A + \Delta A)\tilde{x} &= (A + \Delta A)(x + \Delta x) = Ax + (\Delta A)x + (A + \Delta A)\Delta x \\ &= b + (\Delta A)x + (A + \Delta A)\Delta x = b + \Delta b, \end{aligned}$$

所以

$$(\Delta A)x + (A + \Delta A)\Delta x = \Delta b.$$

这表明 $\Delta x = (A + \Delta A)^{-1}(\Delta b - (\Delta A)x)$, 所以

$$\|\Delta x\| \leq \left\|(A + \Delta A)^{-1}\right\| \left(\|\Delta b\| + \|(\Delta A)x\|\right),$$

根据 (4) 式以及范数的相容性, 得到

$$\|\Delta x\| \leq \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\Delta A\|} \left(\|\Delta b\| + \|\Delta A\|\,\|x\|\right),$$

所以

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

再利用 $\|b\| = \|Ax\| \le \|A\| \|x\|$, 所以

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\kappa(A)}{1 - \|A^{-1}\Delta A\|}\left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right). \tag{7}$$

同样, 如果我们采用更严格的要求 $\left\|A^{-1}\right\| \|\Delta A\| < 1$, 那么

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right). \tag{8}$$

这表明对于一个系数矩阵条件数不大的线性方程组而言, 其解的相对误差和系数矩阵与数据 b 的相对误差大致处于同一个水平上.

更进一步的，如果我们已经计算出了 $Ax = b$ 的一个近似解 $\hat{x}$，我们可以估计这个近似解和精确解 x 的相对误差. 定义残差向量为 $r = b - A\hat{x}$. 此时 $A^{-1}r = A^{-1}b - \hat{x} = x - \hat{x}$，所以 $\|x - \hat{x}\| \le \left\|A^{-1}\right\| \|r\|$, 进一步考虑 $\|b\| = \|Ax\| \le \|A\| \|x\|$, 那么

$$\begin{aligned} \|x - \hat{x}\| &\le \left\|A^{-1}\right\| \|r\| = \left\|A^{-1}\right\| \|A\| \frac{\|x\| \|r\|}{\|A\| \|x\|} \\ &\le \left\|A^{-1}\right\| \|A\| \frac{\|x\| \|r\|}{\|b\|}, \end{aligned}$$

所以相对误差满足

$$\frac{\|x - \hat{x}\|}{\|x\|} \le \kappa(A)\frac{\|r\|}{\|b\|}. \tag{9}$$

这表明如果系数矩阵条件数不大, 那么计算误差大致和残差是一个水平, 否则即使计算解的残差很小, 也可能与精确解相差很远.

例 4.4. 考虑 Hilbert 矩阵为系数矩阵的线性方程组

$$H_5 x = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}^T = b,$$

其显然有精确解 $x = (1, 0, 0, 0, 0)^T$. 添加扰动 $\Delta b = 0.01(1, 1, 1, 1, 1)^T$, 解变为了

$$x + \Delta x = H_5^{-1}(b + \Delta b) = (1.05, -1.2, 6.3, -11.2, 6.3)^T,$$

此时相对误差为

$$\frac{\|\Delta x\|_2}{\|x\|_2} \approx 14.36, \quad \frac{\|\Delta b\|_2}{\|b\|_2} = 0.018,$$

二者相差了接近 1000 倍! 所以病态矩阵作为系数矩阵的线性方程组，其解的数值稳定性是非常差的, 即使一个微小的误差也可能造成解的巨大差异.

参考文献

[1] Horn RA, Johnson CR. Matrix Analysis. Cambridge university press; 2012 Oct 22.

[2] Axler S. Linear Algebra Done Right. Springer Nature; 2024.