

数值分析
Numerical Analysis

Dait

目 录

第一章 数学基础知识	1
1.1 线性空间	1
1.1.1 线性空间	1
1.1.2 线性空间的基和维度	1
1.2 范数	2
1.2.1 度量空间	2
1.2.2 赋范空间与范数	3
1.2.3 范数的等价性	5
1.3 内积	5
1.4 矩阵空间	8
1.4.1 矩阵范数	8
1.4.2 算子范数	9
1.4.3 扰动定理	12
第二章 函数插值和重构	13
2.1 一维多项式插值	13
2.1.1 Lagrange 插值	13
2.1.2 Lagrange 插值的收敛性	15
2.1.3 Newton 插值公式	16
2.1.4 Hermite 插值	17
2.2 分段插值	18
2.2.1 分段线性插值	18
2.2.2 分段三次 Hermite 插值	19
2.3 Fourier 插值	20
第三章 函数逼近	21
3.1 最佳平方逼近	21
3.1.1 法方程	21
3.1.2 正交多项式	22
3.2 最小二乘法	23
3.3 最佳一致逼近	24

第四章 数值积分	26
4.1 Newton-Cotes 公式	27
4.1.1 闭型 Newton-Cotes 公式	27
4.1.2 开型 Newton-Cotes 公式	28
4.1.3 积分法的一致稳定性	28
4.2 复合求积公式	29
4.3 Romberg 求积方法	30
4.3.1 Euler-Maclaurin 求积公式	30
4.3.2 Richardson 外推方法	31
4.3.3 Romberg 求积方法	31
4.4 Gauss 求积公式	32
第五章 线性方程组的直接解法	34
5.1 矩阵操作	34
5.2 Gauss 消元法	38
5.2.1 Gauss 消元法	38
5.2.2 LU 分解	39
5.2.3 Cholesky 分解	42
5.2.4 Thomas 方法	42
5.3 稳定性分析	43
第六章 线性方程组的迭代解法	47
6.1 迭代法基本概念	47
6.1.1 向量序列和矩阵序列的极限	47
6.1.2 迭代公式的构造	48
6.2 (单步) 定常线性迭代	50
6.2.1 Richardson 迭代法	50
6.2.2 Jacobi 迭代法和 Gauss-Seidel 迭代法	51
6.2.3 超松弛迭代法	52
6.3 非线性迭代方法	54
6.3.1 子空间方法	55
6.3.2 最速下降法	55
6.3.3 共轭梯度法	56
6.3.4 共轭梯度法的预处理	59
第七章 非线性方程和方程组的数值解法	60
7.1 非线性方程的不动点迭代法	60
7.1.1 迭代法的收敛性	61
7.1.2 Newton 法	62
7.1.3 割线法	64
7.1.4 Aitken 加速方法	65
7.1.5 Steffensen 迭代法	65

目 录	iii
7.2 非线性方程组的不动点迭代法	66
第八章 矩阵特征值问题的数值方法	69
第九章 常微分方程初值问题的数值解法	70
9.1 常微分方程初值问题	70
9.2 Euler 方法	71

第一章 数学基础知识

1.1 线性空间

1.1.1 线性空间

定义 1.1.1: 线性空间

给定一个数域 \mathbb{F} (本笔记只涉及实数域 \mathbb{R} 和复数域 \mathbb{C}) 和一个集合 V , 定义加法 $+: V \times V \rightarrow V$ 满足:

- 结合律: $(a + b) + c = a + (b + c)$,
- 零元: $a + 0 = a$,
- 逆元: $a + (-a) = 0$,
- 交换律: $a + b = b + a$;

数乘 $\cdot: \mathbb{F} \times V \rightarrow V$ 满足:

- 单位元: $1a = a$,
- 结合律: $(\lambda\mu)a = \lambda(\mu a)$,
- 分配率 1: $(\lambda + \mu)a = \lambda a + \mu a$,
- 分配率 2: $\lambda(a + b) = \lambda a + \lambda b$.

则称 V 在 \mathbb{F} 上构成一个线性空间 (linear space).

例 1.1.1: 线性空间的例子

- $C^n[a, b]$: 全体在 $[a, b]$ 上 n 次导数连续 (continuous) 的函数构成的集合.
- $\mathcal{P}_n = \{a_0 + a_1x + \cdots + a_nx^n \mid a_i \in \mathbb{F}\}$: n 次多项式函数空间.

1.1.2 线性空间的基和维度

定义 1.1.2: 线性无关

设 V 是线性空间, 给定其中 n 个元素 $x_1, \dots, x_n \in V$, 若

$$a_1x_1 + \cdots + a_nx_n = 0, \iff a_1 = \cdots = a_n = 0, \quad (1.1)$$

则称 x_1, \dots, x_n 线性无关 (linear independent). 反之, 称为线性相关.

例 1.1.2: 线性无关的例子

- \mathcal{P}_n 中 $\{1, x, \dots, x^n\}$ 是线性无关的;
- 所有 $[-\pi, \pi]$ 上的周期函数构成的函数空间是线性空间, 其中

$$\{1, \cos x, \sin x, \dots, \cos nx, \sin nx\}$$

线性无关.

定义 1.1.3: 线性空间的基和维度

给定线性空间 V 中的一组元素 $\{x_1, \dots, x_n\}$, 若 $\forall x \in V$ 都可以被唯一表示为其线性组合

$$x = a_1 x_1 + \dots + a_n x_n, \quad (1.2)$$

则称 $\{x_1, \dots, x_n\}$ 构成一组基 (base), 且 V 的维度 $\dim(V) = n$.

定理 1.1.1

线性空间的维度与基的选取没有关系.

例 1.1.3: 典型线性空间的维度

- $\dim(\mathcal{P}_n) = n$;
- $\dim(\mathcal{C}[a, b]) = +\infty$.

1.2 范数

1.2.1 度量空间

定义 1.2.1: 度量空间

给定集合 M , 若映射 $d: M \times M \rightarrow \mathbb{R}$ 满足: $\forall x, y \in M$

- 正定性: $d(x, y) \geq 0$, 且 $d(x, y) = 0 \iff x = y$;
- 交换性: $d(x, y) = d(y, x)$;
- 三角不等式: $d(x, y) + d(y, z) \geq d(x, z)$.

则称 M 是一个度量空间 (metric space) 或距离空间, d 为度量函数或距离函数.

定义 1.2.2: Cauchy 序列

对于序列 $\{x_n\}$, 若 $\forall \varepsilon > 0, \exists N \in \mathbb{N}$ 满足:

$$d(x_m, x_n) < \varepsilon, \quad \forall m, n > N, \quad (1.3)$$

则称 $\{x_n\}$ 为 Cauchy 序列 (Cauchy sequence).

定义 1.2.3: 度量空间的完备性

若 \forall Cauchy 序列 $\{x_n\} \subset M$, $\exists x \in M$ 满足:

$$\lim_{n \rightarrow \infty} d(x_n, x) = 0, \iff \lim_{n \rightarrow \infty} x_n = x,$$

则称度量空间 M 是完备的 (complete).

注. 不严谨地说, 完备性要求: 所有在 M 中收敛的序列都会收敛到 M 中.

例 1.2.1: 实数公理

实数集 \mathbb{R} 中的度量函数 $d(a, b) = |a - b|$ 是完备的.

注. 有理数域 \mathbb{Q} 中的度量函数 $d(a, b) = |a - b|$ 不是完备的, 比如

$$x_0 = 1, \quad x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n} \rightarrow \sqrt{2} \notin \mathbb{Q}.$$

定理 1.2.1: 完备化定理

若 (M, d) 是一个度量空间, 则存在唯一等距同构的完备化空间.

证明. 构造性证明: 令 \tilde{M} 是 M 中所有 Cauchy 序列 $\{x_n\}$ 的集合. 在 \tilde{M} 中定义等价关系:

$$\{x_n\} \sim \{y_n\} \iff \lim_{n \rightarrow \infty} d(x_n, y_n) = 0.$$

令 $[\{x_n\}] = \{\{y_n\} \in \tilde{M} \mid \{x_n\} \sim \{y_n\}\}$ 表示 $\{x_n\}$ 的等价类, $\hat{M} = \{[\{x_n\}] \mid \{x_n\} \in \tilde{M}\}$ 是 M 中所有 Cauchy 序列的等价类构成的集合. 定义 \hat{M} 上的度量为

$$\hat{d}([\{x_n\}], [\{y_n\}]) := \lim_{n \rightarrow \infty} d(x_n, y_n),$$

可证 (\hat{M}, \hat{d}) 是完备的度量空间. 且存在等距嵌入

$$i: M \rightarrow \hat{M}, \quad x \mapsto [\{x, x, \dots\}].$$

即 i 是映射到对应常数序列的等价类. □

1.2.2 赋范空间与范数**定义 1.2.4: 赋范空间与范数**

给定数域 \mathbb{F} 上的线性空间 V , 若映射 $\|\cdot\|: V \rightarrow \mathbb{R}$ 满足: $\forall a, b \in V, \lambda \in \mathbb{F}$

- 正定性: $\|a\| \geq 0$, 且 $\|a\| = 0 \iff a = 0$;
- 齐次性: $\|\lambda a\| = |\lambda| \|a\|$;
- 三角不等式: $\|a + b\| \leq \|a\| + \|b\|$.

则称 $(V, \|\cdot\|)$ 构成一个赋范空间 (normed space), $\|\cdot\|$ 是 V 的范数 (norm).

推论. 显然, 赋范空间也是度量空间, 只需定义 $d(a, b) := \|a - b\|$ 即可.

注. 赋范空间不一定是完备的, 完备的赋范空间称为 Banach 空间.

例 1.2.2: \mathbb{R}^n 的 p -范数

向量 $x = (x_1, \dots, x_n)^\top \in \mathbb{F}^n$ 的 p -范数为

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad p > 0. \quad (1.4)$$

特别地, $\|x\|_\infty = \lim_{p \rightarrow \infty} \|x\|_p = \max_i |x_i|$.

证明. 记 $k = \arg \max_i |x_i|$, 对下式两边开 p 次方并令 $p \rightarrow \infty$ 即得

$$|x_k|^p \leq \sum_{i=1}^n |x_i|^p \leq n |x_k|^p. \quad \square$$

二维位置向量 $r = (x, y) \in \mathbb{R}^2$, 方程 $\|r\|_p = 1$ 在直角坐标中的图像为

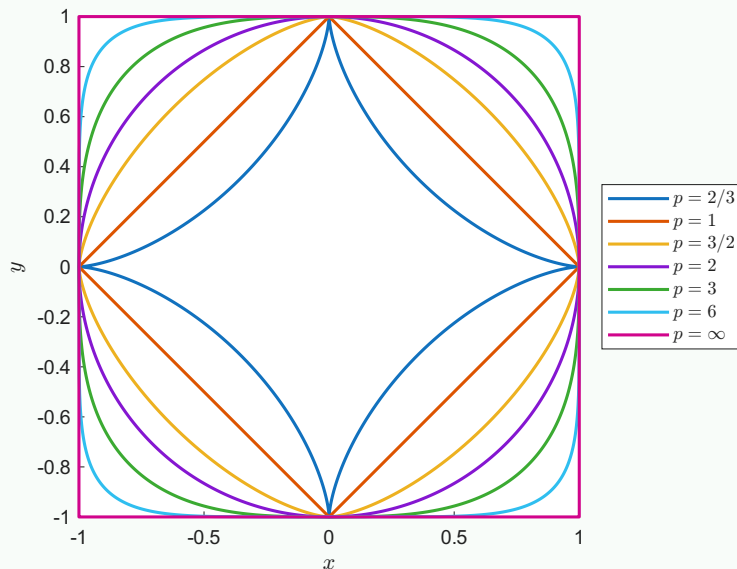


图 1.1: 不同的 p -范数

特别地, $p = 2/3$ 的曲线叫星形线 (astroid).

例 1.2.3: $C[a, b]$ 的 p -范数

函数 $f \in C[a, b]$ 的 p -范数为

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{1/p}, \quad (1.5)$$

当 $1 < p < \infty$ 时, $\|\cdot\|_p$ 是不完备的: 连续函数的 Cauchy 序列可以按 p -范数收敛于一个不连续的函数. 但 $\|\cdot\|_\infty$ 是完备的.

1.2.3 范数的等价性

定义 1.2.5: 范数的等价性

给定线性空间 S 上的两个范数 $\|\cdot\|_\alpha, \|\cdot\|_\beta$, 若 $\exists C_1, C_2 > 0$ 满足:

$$C_1 \|x\|_\alpha \leq \|x\|_\beta \leq C_2 \|x\|_\alpha,$$

则称 $\|\cdot\|_\alpha, \|\cdot\|_\beta$ 等价 (equivalent).

推论. 向量的 p -范数都是等价的, 因为由 Hölder 不等式, $p < q$ 时

$$\|x\|_q \leq \|x\|_p \leq n^{1/p-1/q} \|x\|_p. \quad (1.6)$$

定理 1.2.2

有限维线性空间中, 任意两个范数都是等价的.

1.3 内积

定义 1.3.1: 内积

给定线性空间 S , 内积 (inner product) 是一个映射 $\langle \cdot, \cdot \rangle : S \times S \rightarrow \mathbb{F}$, 满足:

- 正定性: $\langle x, x \rangle \geq 0$, 且 $\langle x, x \rangle = 0 \iff x = 0$;
- 交换共轭: $\langle x, y \rangle = \overline{\langle y, x \rangle}$;
- 对第一个变量线性: $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$.

若 $\langle x, y \rangle = 0$, 则称 x, y 是正交的 (orthogonal).

注. 对第二个变量不一定线性, 会出现复共轭:

$$\langle x, ay + bz \rangle = \bar{a} \langle x, y \rangle + \bar{b} \langle x, z \rangle.$$

除非内积的值域 $\mathbb{F} \subset \mathbb{R}$, 此时内积满足对称性: $\langle x, y \rangle = \langle y, x \rangle$.

推论. 根据内积可以自然定义出一个范数, 称为内积诱导的范数:

$$\|x\| := \sqrt{\langle x, x \rangle}, \quad (1.7)$$

因此内积空间也是赋范空间, 完备的内积空间称为 Hilbert 空间.

例 1.3.1: 内积实例

- \mathbb{C}^n 中的标准内积为

$$\langle x, y \rangle = \sum_{i=1}^n x_i \bar{y}_i,$$

标准内积诱导出来的范数就是 2-范数.

- $\mathcal{C}[a, b]$ 的一个内积为

$$\langle f, g \rangle = \int_a^b f(x) \overline{g(x)} dx.$$

定义 1.3.2: Gram 矩阵

给定一组向量 x_1, \dots, x_n , Gram 矩阵的第 i, j 个元素由内积 $\langle x_i, x_j \rangle$ 给出:

$$G := \begin{bmatrix} \langle x_1, x_1 \rangle & \cdots & \langle x_1, x_n \rangle \\ \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \cdots & \langle x_n, x_n \rangle \end{bmatrix}. \quad (1.8)$$

由于内积交换共轭 $\langle x_i, x_j \rangle = \overline{\langle x_j, x_i \rangle}$, Gram 矩阵是 Hermite 的: $G^\dagger = G$.

定理 1.3.1: 内积空间线性无关的判定

给定内积空间 S , $x_1, \dots, x_n \in S$ 是线性无关的 \iff Gram 矩阵可逆.

证明. 设 $a = (a_1, \dots, a_n)$ 满足 $aG = 0$, 则 $\forall k = 1, \dots, n$

$$(aG)_k = a_1 \langle x_1, x_k \rangle + \cdots + a_n \langle x_n, x_k \rangle = \langle a_1 x_1 + \cdots, a_n x_n, x_k \rangle = 0,$$

特别地,

$$\langle a_1 x_1 + \cdots + a_n x_n, a_1 x_1 + \cdots + a_n x_n \rangle = 0, \iff a_1 x_1 + \cdots, a_n x_n = 0,$$

则 G 可逆 $\iff N(G^\top) = \{0\} \iff a$ 只有零解. □

例 1.3.2

给定内积空间 S 的一组基 $\{x_1, \dots, x_n\}$, 则 $\forall x \in S$ 均可以写成

$$x = a_1 x_1 + \cdots + a_n x_n,$$

下面计算 a_1, \dots, a_n . 两边分别与 x_i 做内积:

$$\langle x, x_i \rangle = a_1 \langle x_1, x_i \rangle + \cdots + a_n \langle x_n, x_i \rangle,$$

即

$$[\langle x, x_1 \rangle \cdots \langle x, x_n \rangle] = [a_1 \cdots a_n] G,$$

若基是正交的, 即 $\forall i \neq j, \langle x_i, x_j \rangle = 0$, 则

$$a_i = \frac{\langle x, x_i \rangle}{\langle x_i, x_i \rangle}.$$

定理 1.3.2: Schmidt 正交化

设 x_1, \dots, x_n 是一组线性无关的基, 为得到一组正交基, 定义

$$y_i = x_i - \sum_{j=1}^{i-1} \frac{\langle x_i, y_j \rangle}{\langle y_j, y_j \rangle} y_j. \quad (1.9)$$

则 y_1, \dots, y_n 是正交的.

定义 1.3.3: 带权内积

设 $\rho \in C(a, b)$ 是一个几乎处处为正¹的函数, 且

$$\int_a^b \rho(x) dx < +\infty,$$

定义内积

$$\langle f, g \rangle = \int_a^b f(x)g(x)\rho(x) dx. \quad (1.10)$$

¹即 $\{x | \rho(x) \leq 0\}$ 的 Lebesgue 测度为 0.

定义 1.3.4: 正交多项式

已知 $\{1, x, \dots, x^n\}$ 是线性无关的. 考虑 $C[a, b]$ 上的带权内积, Schmidt 正交化得到一组多项式函数

$$\psi_0(x), \psi_1(x), \dots, \psi_n(x),$$

显然, $\deg(\psi_i) = i$.

例 1.3.3: Legendre 多项式

权函数 $\rho = 1$, 区间 $[-1, 1]$, 得到 Legendre 多项式

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n. \quad (1.11)$$

性质:

- 内积

$$\langle P_n, P_m \rangle = \frac{2}{2n+1} \delta_{nm}. \quad (1.12)$$

- 递归关系

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x). \quad (1.13)$$

- 奇偶性

$$P_n(-x) = (-1)^n P_n(x). \quad (1.14)$$

例 1.3.4: Chebyshev 多项式

权函数 $\rho(x) = (1 - x^2)^{-1/2}$, 区间 $[-1, 1]$, 得到 Chebyshev 多项式

$$T_n(x) = \cos(n \arccos x). \quad (1.15)$$

- 内积

$$\langle T_n, T_n \rangle = \begin{cases} \pi, & n = 0 \\ \pi/2, & n \geq 1 \end{cases} \quad (1.16)$$

- 递推关系

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \quad (1.17)$$

- 奇偶性

$$T_n(-x) = (-1)^n T_n(x); \quad (1.18)$$

- T_n 的 n 个实单根为 $\cos\left(\frac{2k-1}{2n}\pi\right)$, $(n+1)$ 个极值点为 $\cos\left(\frac{k}{n}\pi\right)$
- 当 $|x| \geq 1$ 时,

$$T_n(x) = \frac{1}{2} \left[\left(x + \sqrt{x^2 - 1} \right)^k + \left(x - \sqrt{x^2 - 1} \right)^k \right]. \quad (1.19)$$

1.4 矩阵空间

1.4.1 矩阵范数

定义 1.4.1: 矩阵范数

矩阵空间上的范数 $\|\cdot\| : \mathbb{F}^{n \times n} \rightarrow \mathbb{R}$ 若满足

$$\|AB\| \leq \|A\| \|B\|, \quad (1.20)$$

则 $\|\cdot\|$ 称为矩阵范数 (matrix norm).

例 1.4.1: Frobenius 范数

定义 Frobenius 范数

$$\|A\|_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^n |A_{ij}|^2}. \quad (1.21)$$

是一个矩阵范数.

证明. 由 Cauchy-Schwarz 不等式

$$\begin{aligned}\|AB\|_F &= \sqrt{\sum_{i=1}^n \sum_{j=1}^n \left| \sum_{k=1}^n A_{ik} B_{kj} \right|^2} \leq \sqrt{\sum_{i=1}^n \sum_{j=1}^n \left(\sum_{k=1}^n |A_{ik}|^2 \sum_{k=1}^n |B_{kj}|^2 \right)} \\ &= \sqrt{\sum_{i=1}^n \sum_{k=1}^n |A_{ik}|^2 \sum_{j=1}^n \sum_{k=1}^n |B_{kj}|^2} = \|A\|_F \|B\|_F. \quad \square\end{aligned}$$

注意到

$$\operatorname{tr}(A^\dagger A) = \sum_{i=1}^n (A^\dagger A)_{ii} = \sum_{i=1}^n \sum_{j=1}^n A_{ij}^\dagger A_{ji} = \sum_{i=1}^n \sum_{j=1}^n |A_{ji}|^2.$$

故

$$\|A\|_F = \sqrt{\operatorname{tr}(A^\dagger A)} = \sqrt{\operatorname{tr}(AA^\dagger)}. \quad (1.22)$$

定义 1.4.2: 矩阵范数与向量范数的相容

给定矩阵范数 $\|\cdot\|_M$ 和向量范数 $\|\cdot\|_V$, 若 $\forall A \in \mathbb{F}^{n \times n}, x \in \mathbb{F}^n$

$$\|Ax\|_V \leq \|A\|_M \|x\|_V, \quad (1.23)$$

则称他们是相容的. 在不引起混淆的情况下, 可以略去下标.

定理 1.4.1

Frobenius 范数与向量 2 - 范数相容.

证明.

$$\|Ax\|_2 = \sqrt{\sum_{i=1}^n \left| \sum_{j=1}^n A_{ij} x_j \right|^2} \leq \sqrt{\sum_{i=1}^n \left(\sum_{j=1}^n |A_{ij}|^2 \sum_{j=1}^n |x_j|^2 \right)} = \|A\|_F \|x\|_2. \quad \square$$

1.4.2 算子范数

定义 1.4.3: 算子范数

给定向量范数 $\|\cdot\|$, 定义矩阵的算子范数 (operate norm)

$$N : \mathbb{F}^{n \times n} \rightarrow \mathbb{R}, \quad A \mapsto \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}. \quad (1.24)$$

称算子范数 N 是该向量范数 $\|\cdot\|$ 诱导出来的矩阵范数.

注. 注意 $N(I) \equiv 1$. 故 Frobenius 范数不是算子范数.¹

¹不考虑 $n = 1$ 的平凡情况.

定理 1.4.2

$N(\cdot)$ 是一个矩阵范数, 并与向量范数 $\|\cdot\|$ 相容.

证明. 易得 $\forall x \neq 0, \|Ax\| \leq N(A) \|x\|$.

$$N(AB) = \sup_{x \neq 0} \frac{\|ABx\|}{\|x\|} \leq \sup_{x \neq 0} \frac{N(A) \|Bx\|}{\|x\|} = N(A)N(B). \quad \square$$

定义 1.4.4: 谱半径

矩阵 A 全体特征值的集合称为 A 的谱 (spectrum), 记作 $\sigma(A)$, 特征值模的最大值称为谱半径, 记作 $\rho(A)$.

注. 对于一般的矩阵 $A \in \mathbb{F}^{m \times n}$, 有 $\sigma(A^\dagger A) \setminus \{0\} = \sigma(AA^\dagger) \setminus \{0\}$, 即非 0 特征值相同, 称为 A 的奇异值 (singular value).

例 1.4.2

设 $A \in \mathbb{F}^{n \times n}$, 则 p -向量范数诱导出来的矩阵范数为:

$$\|A\|_\infty = \max_i \sum_{j=1}^n |A_{ij}|, \quad (1.25a)$$

$$\|A\|_1 = \max_j \sum_{i=1}^n |A_{ij}|, \quad (1.25b)$$

$$\|A\|_2 = \sqrt{\rho(A^\dagger A)}, \quad (1.25c)$$

证明. 先证明 (1.25b), 将 A 写作列向量的形式 (A_1, \dots, A_n) , 令 $k = \arg \max_j \|A_j\|_1$,

则 $\forall x \in \mathbb{F}^n$ 且 $\|x\|_1 = \sum_{j=1}^n |x_j| = 1$, 有

$$\|Ax\|_1 = \left\| \sum_{j=1}^n A_j x_j \right\|_1 \leq \sum_{j=1}^n |x_j| \|A_j\|_1 \leq \|A_k\|_1 \sum_{j=1}^n |x_j| = \|A_k\|_1 \|x\|_1 = \|A_k\|_1,$$

特别地, 取 $x = e_k$ 可使等号成立, 故

$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 = \|A_k\|_1 = \max_j \sum_{i=1}^n |A_{ij}|;$$

然后证明 (1.25a), 令 $k = \arg \max_i \sum_{j=1}^n |A_{ij}|$, $\forall x \in \mathbb{F}^n$ 且 $\|x\|_\infty = \max_j |x_j| = 1$, 有

$$\|Ax\|_\infty = \max_i \left| \sum_{j=1}^n A_{ij} x_j \right| \leq \max_i \sum_{j=1}^n |A_{ij}| |x_j| \leq \max_i \sum_{j=1}^n |A_{ij}| \max_j |x_j| = \sum_{j=1}^n |A_{kj}|.$$

特别地, 取 $x_j = \text{sgn}(A_{kj})$ 可使等号成立, 故

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \max_i \sum_{j=1}^n |A_{ij}|;$$

最后证明 (1.25c), 由 2 - 范数的性质:

$$\|A\|_2^2 = \sup_{\|x\|_2=1} \langle Ax, Ax \rangle = \sup_{\|x\|_2=1} \langle A^\dagger Ax, x \rangle = \rho(A^\dagger A). \quad \square$$

定理 1.4.3: 谱半径小于矩阵范数

谱半径 $\rho(A)$ 和矩阵范数的关系:

$$\rho(A) \leq \|A\|. \quad (1.26)$$

证明. 考虑 A 的一个特征值 λ 和特征向量 x , 则

$$|\lambda| \|xx'\| = \|Axx'\| \leq \|A\| \|xx'\|.$$

于是 $\|A\| \geq |\lambda|$. \square

定理 1.4.4

给定 $A \in \mathbb{F}^{n \times n}$, $\epsilon > 0$, 存在算子范数 $\|\cdot\|$ 满足:

$$\|A\| \leq \rho(A) + \epsilon. \quad (1.27)$$

引理. 若 $\|\cdot\|_\alpha$ 是 \mathbb{F}^n 中的向量范数, $P \in \mathbb{F}^{n \times n}$ 非奇异, 则

$$\|\cdot\|_{P,\alpha} : x \mapsto \|Px\|_\alpha,$$

构成另一个向量范数, 诱导的算子范数为

$$\|A\|_{P,\alpha} = \|PAP^{-1}\|_\alpha.$$

证明. 令 P 将 A 相似变换为 Jordan 型, 即

$$PAP^{-1} = J = \text{diag}(J_1, \dots, J_r), \quad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}$$

令 $D_\epsilon = \text{diag}(1, \epsilon, \dots, \epsilon^{n-1})$, 则

$$\hat{J} = D_\epsilon^{-1} J D_\epsilon = \text{diag}(\hat{J}_1, \dots, \hat{J}_r), \quad \hat{J}_i = \begin{bmatrix} \lambda_i & \epsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \epsilon \\ & & & \lambda_i \end{bmatrix}$$

则

$$\|A\|_{D_\epsilon^{-1}P,\infty} = \|D_\epsilon^{-1}PAP^{-1}D_\epsilon\|_\infty = \|\hat{J}\|_\infty \leq \rho(A) + \epsilon. \quad \square$$

注. 对任意满足 $|\lambda| = \rho(A)$ 的特征值 λ , 对应 Jordan 块是对角的, 则存在一个算子范数满足

$$\|A\| = \rho(A).$$

1.4.3 扰动定理

一个矩阵不可逆的概率是很低的,¹¹那如何度量矩阵的奇异性?

定理 1.4.5: 扰动定理

给定扰动 B , 若 $\|B\| < 1$, 则 $I + B$ 可逆且

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \quad (1.28)$$

证明. 若 $I + B$ 不可逆, 则 $\rho(B) \geq 1 > \|B\|$ 矛盾. 记 $D := (I + B)^{-1}$, 则

$$(I + B)D = I, \iff D = I - BD, \implies \|D\| \leq 1 + \|B\| \|D\|. \quad \square$$

定理 1.4.6: 扰动定理 · 二

给定 A, C , 若 A 非奇异且

$$\|C - A\| \leq \|A^{-1}\|^{-1},$$

则 C 也非奇异, 且

$$\|C^{-1}\| \leq \frac{1}{\|A^{-1}\|^{-1} - \|C - A\|}. \quad (1.29)$$

证明. 令 $B = I - A^{-1}C$ 即可. \square

¹¹因为多了一个 $\det(A) = 0$ 的限制条件.

第二章 函数插值和重构

基本问题 已知关于某函数 f 的一组信息，如何重构 f ? 事实上，由于信息缺失，无法准确重构.

定义 2.0.1: 重构

给定函数空间 X 上的一组线性无关泛函 ϕ_1, \dots, ϕ_n , $f \in X$ 且 $\phi_i(f)$ 已知，希望确定 $f^* \in Y \subset X$ 满足：

$$\phi_i(f^*) = \phi_i(f), \quad i = 1, \dots, n. \quad (2.1)$$

Y 称为插值空间或重构空间， $\{\phi_i\}$ 为信息泛函.

例 2.0.1: 采样空间的选择

- 多项式函数空间

$$\mathcal{P}_n = \{a_0 + a_1x + \dots + a_nx^n\},$$

- 样条函数 (spline) 空间：分段多项式函数；
- 三角多项式函数空间

$$\mathcal{Y}_n = \{a_0 + a_1 \cos x + b_1 \sin x + \dots + a_n \cos nx + b_n \sin nx\}.$$

2.1 一维多项式插值

2.1.1 Lagrange 插值

定义 2.1.1: Lagrange 插值

插值空间 Y 由 $n+1$ 个参数 a_0, \dots, a_n 标定，即

$$y = y(x; a_0, \dots, a_n).$$

给定一组插值节点 (采样点) x_i 和采样值 $f_i = f(x_i)$ ，希望确定参数满足

$$y(x_i) = f(x_i), \quad \forall i \in I. \quad (2.2)$$

定理 2.1.1: 多项式插值定理

给定 $n+1$ 个不同插值点 x_0, \dots, x_n , 存在唯一的多项式函数 $p_n \in \mathcal{P}_n$ 满足插值条件.

证明. 存在性: 采取直接构造的方法. 定义插值基函数 $\ell_i \in \mathcal{P}_n$:

$$\ell_i(x) := \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}. \quad (2.3)$$

易验证, ℓ_i 满足:

$$\ell_i(x_j) = \delta_{ij}. \quad (2.4)$$

故插值多项式为

$$p_n(x) = \sum_{i=0}^n f(x_i) \ell_i(x). \quad (2.5)$$

唯一性: 若还存在 $q_n \in \mathcal{P}_n$ 满足插值条件, 则 $p_n - q_n \in \mathcal{P}_n$ 且有 x_0, \dots, x_n 共 $n+1$ 个零点, 故 $p_n - q_n \equiv 0$. \square

定义 2.1.2: 余项

定义插值函数 $p_n(x)$ 与被插值函数 $f(x)$ 之间的差为余项 (remainder)

$$R_n(x) := f(x) - p_n(x). \quad (2.6)$$

定理 2.1.2: 中值定理与余项

若 $f \in \mathcal{C}^{n+1}[a, b]$, 则 $\forall x \in [a, b], \exists \xi(x) \in (a, b)$ 使得

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{i=0}^n (x - x_i). \quad (2.7)$$

证明. 当 $x = x_i$ 时, $R_n(x_i) = 0$ 显然成立; 给定 $x \in [a, b]$ 且 $x \neq x_i$, 定义

$$g(t) := R_n(t) - \frac{\prod_{i=0}^n (t - x_i)}{\prod_{i=0}^n (x - x_i)} R_n(x),$$

则 $g \in \mathcal{C}^{n+1}[a, b]$ 且在 $[a, b]$ 上有 x_0, \dots, x_n, x 共 $n+2$ 个零点, 在其划分的 $n+1$ 个区间中应用 Rolle 定理: 对应存在 $\xi_1^{(1)}, \dots, \xi_{n+1}^{(1)}$ 使得

$$g'(\xi_i^{(1)}) = 0, \quad i = 0, 1, \dots, n+1,$$

继续对 $\xi_0^{(1)}, \dots, \xi_{n+1}^{(1)}$ 划分的 n 个区间上应用 Rolle 定理, 直到 $g^{(n+1)}(\xi_0^{(n+1)}) = 0$:

$$g^{(n+1)}(\xi_0^{(n+1)}) = f^{(n+1)}(\xi_0^{(n+1)}) - \frac{(n+1)!}{\prod_{i=0}^n (x - x_i)} R_n(x) = 0.$$

取 $\xi = \xi_0^{(n+1)}$ 即证. \square

推论. 若 $f \in \mathcal{C}^{n+1}[a, b]$, $h = \max(x_{i+1} - x_i)$ 则

$$\|R_n\|_\infty \leq \frac{h^{n+1}}{4(n+1)} \|f^{(n+1)}\|_\infty. \quad (2.8)$$

2.1.2 Lagrange 插值的收敛性

当什么条件下, $n \rightarrow \infty$, 误差 $\|R_n\|_\infty \rightarrow 0$?

定理 2.1.3: Lagrange 插值收敛性的一个充分条件

若被插值函数的任意阶导数一致有界, 则误差收敛到 0.

例 2.1.1

给定 $f(x) = \sin x$, $x \in [0, \pi]$, 由于 $\forall x \in [0, \pi]$

$$|f^{(n+1)}(x)| \leq 1, \quad \left| \prod_{i=0}^n (x - x_i) \right| \leq \pi^{n+1},$$

故

$$|R_n(x)| \leq \frac{\pi^{n+1}}{(n+1)!} \rightarrow 0,$$

说明 Lagrange 插值多项式 p_n 在 $[0, \pi]$ 上一致收敛到 f .

例 2.1.2: Runge 现象

Runge 函数 $f(x) = \frac{1}{1+25x^2}$ 在 $[-1, 1]$ 上等距插值.

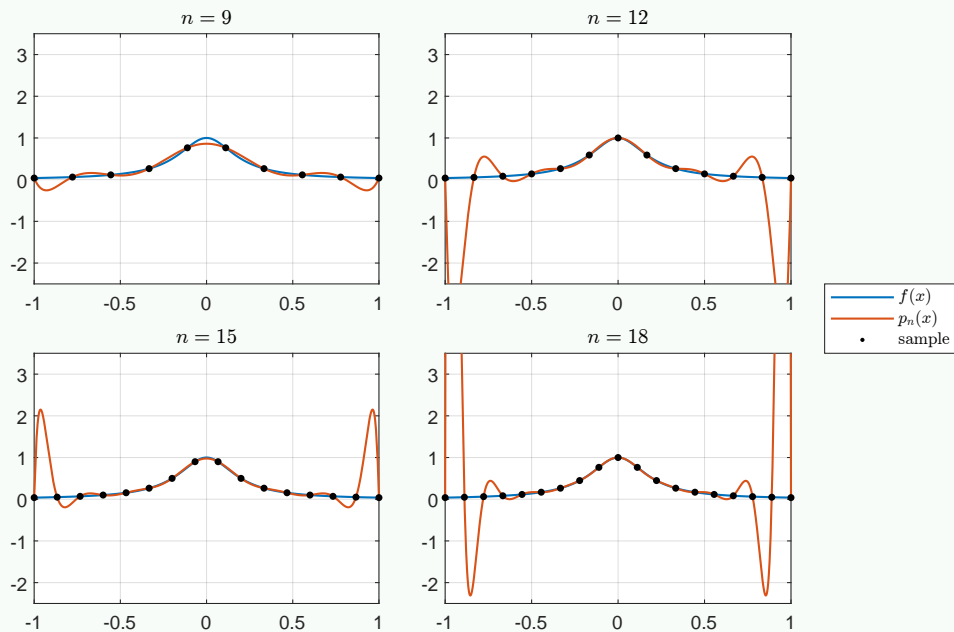


图 2.1: 等距插值 Runge 现象

显然 $f(x)$ 在 \mathbb{R} 上是解析的, 但在 \mathbb{C} 上存在奇点 $\pm i/5$.

改用 $n+1$ 阶 Chebyshev 多项式零点

$$\cos\left(\frac{\pi}{2(n+1)}\right), \cos\left(\frac{3\pi}{2(n+1)}\right), \dots, \cos\left(\frac{(2n+1)\pi}{2(n+1)}\right),$$

作为插值节点, 可以消去 Runge 现象.

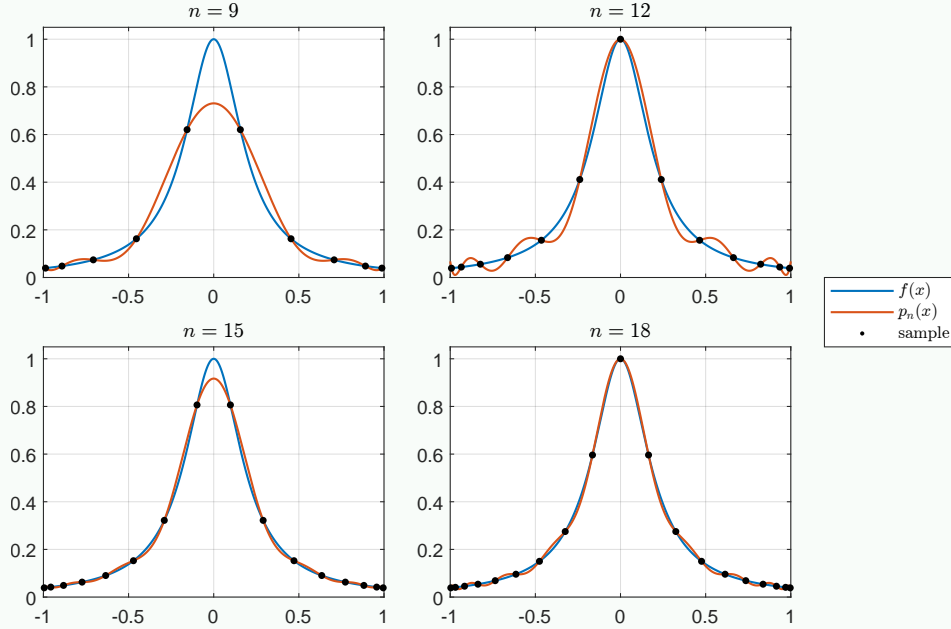


图 2.2: Chebyshev 多项式零点插值

2.1.3 Newton 插值公式

定义 2.1.3: 均差

定义 f 在节点集 i_0, i_1, \dots, i_k 上的 k 阶均差 (divided differences) 递归地定义为:

$$f[x_{i_0}, \dots, x_{i_k}] := \frac{f[x_{i_1}, \dots, x_{i_k}] - f[x_{i_0}, \dots, x_{i_{k-1}}]}{x_{i_k} - x_{i_0}}. \quad (2.9)$$

特别地, f 在 x_i 上的零阶均差 $f[x_i] := f(x_i)$.

推论. k 阶均差 $f[x_0, \dots, x_k]$ 是 $f(x_0), \dots, f(x_k)$ 的线性组合:

$$f[x_0, \dots, x_k] = \sum_{i=0}^k \frac{f(x_i)}{\prod_{j \neq i} (x_i - x_j)}. \quad (2.10)$$

因此, 均差对于节点是对称的, 即任意改变节点的顺序, 均差的值不变.

定理 2.1.4: Newton 插值公式

利用均差迭代得到 n 次 Newton 插值多项式:

$$p_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + \dots + f[x_0, \dots, x_n](x - x_0) \cdots (x - x_{n-1}). \quad (2.11)$$

其余项为

$$R_n(x) = f[x, x_0, \dots, x_n](x - x_0) \cdots (x - x_n). \quad (2.12)$$

对均差形式的余项应用定理 2.1.2 得到:

定理 2.1.5: 中值定理与均差

如果 $f \in C[a, b]$, $x_0, \dots, x_n \in [a, b]$, 则 $\exists \xi \in I(x_0, \dots, x_n)$,

$$f[x_0, \dots, x_n] = \frac{f^{(n)}(\xi)}{n!}. \quad (2.13)$$

推论. 特别地, n 阶均差:

$$f[\underbrace{x, \dots, x}_{n+1}] = \frac{f^{(n)}(x)}{n!} \quad (2.14)$$

推论. 均差的导数:

$$\frac{d}{dx} f[x_0, \dots, x_n, x] = f[x_0, \dots, x_n, x, x]. \quad (2.15)$$

定理 2.1.6: Neville 算法

给定插值节点 x_0, \dots, x_n , 定义 $p_{i,j} \in \mathcal{P}^{j-i}$ 为满足节点 x_i, \dots, x_j 插值条件的多项式插值函数, 则递归地, 有 $p_{i,i}(x) \equiv f(x_i)$,

$$p_{i,j}(x) = \frac{(x - x_i)p_{i+1,j}(x) + (x_j - x)p_{i,j-1}(x)}{x_j - x_i}, \quad (2.16)$$

2.1.4 Hermite 插值

现推广 Lagrange 插值的概念: 除了要求插值函数在节点上相等外, 还要求在节点上的导数值相等.

定义 2.1.4: Hermite 插值问题

给定插值节点 $x_0 < x_1 < \dots < x_m$ 及插值条件

$$(x_i, f^{(k)}(x_i)), \quad i = 0, 1, \dots, m, \quad k = 0, 1, \dots, n_i - 1$$

确定次数为 $n = n_0 + \dots + n_m - 1$ 的多项式函数 $p \in \mathcal{P}_n$ 满足插值条件.

定理 2.1.7

Hermite 插值问题的解存在且唯一.

证明. 定义拓展均差:

$$f[x_0, x_1, \dots, x_n] := \int_0^{t_0} \int_0^{t_1} \cdots \int_0^{t_{n-1}} f^{(n)}(t_n(x_n - x_{n-1}) + \cdots + t_1(x_1 - x_0) + t_0 x_0) dt_n \cdots dt_2 dt_1. \quad (2.17)$$

有递推式:

$$f[x_0, \dots, x_n] = \frac{f[x_0, \dots, x_{n-2}, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_{n-1}}. \quad (2.18)$$

即证. \square

例 2.1.3: 均差表

给定 $f(a), f'(a), f(b), f'(b)$, 则插值函数

$$p_3(x) = f(a) + f[a, a](x - a) + f[a, a, b](x - a)^2 + f[a, a, b, b](x - a)^2(x - b).$$

计算到 $f[a, a, b, b]$, 给出均差表:

0	1	2	3
$f(a)$	$f[a, a] = f'(a)$	$f[a, a, b] = \frac{f[a, b] - f[a, a]}{b - a}$	$f[a, a, b, b] = \cdots$
$f(a)$	$f[a, b] = \frac{f(b) - f(a)}{b - a}$	$f[a, b, b] = \frac{f[b, b] - f[a, b]}{b - a}$	
$f(b)$	$f[b, b] = f'(b)$		

2.2 分段插值

2.2.1 分段线性插值

定义 2.2.1: 分段线性插值

给定节点 $a = x_0 < x_1 < \cdots < x_n = b$ 及 $f(x_i)$, 求插值函数 φ 满足:

- $\varphi \in \mathcal{C}[a, b]$ 是连续函数;
- $[x_i, x_{i+1}]$ 上 $\varphi \in \mathcal{P}_1$ 是线性函数;
- $\varphi(x_i) = f(x_i)$.

满足前两个性质的函数组成插值空间 Φ , 且 $\dim(\Phi) = n + 1$.

定理 2.2.1

分段线性插值函数是存在且唯一的.

证明. 定义插值基函数

$$I_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}] \\ 0, & \text{otherwise} \end{cases} \quad (2.19)$$

则

$$\varphi(x) = \sum_{i=0}^n f(x_i) I_i(x). \quad (2.20)$$

□

定理 2.2.2: 误差收敛性

定义 $h := \max_i (x_i - x_{i-1})$,

- 如果 $f \in \mathcal{C}[a, b]$, 则 $\lim_{h \rightarrow 0} \|f - \varphi\|_{\infty} = 0$;
- 如果 $f \in \mathcal{C}^1[a, b]$, 则 $\|f - \varphi\|_{\infty} \leq 2h \|f'\|_{\infty}$;
- 如果 $f \in \mathcal{C}^2[a, b]$, 则 $\|f - \varphi\|_{\infty} \leq h^2/8 \cdot \|f''\|_{\infty}$.

2.2.2 分段三次 Hermite 插值

定义 2.2.2: 分段三次 Hermite 插值

给定节点 $a = x_0 < x_1 < \cdots < x_n = b$ 及 $f(x_i), f'(x_i)$, 求插值函数 φ 满足:

- $\varphi \in \mathcal{C}^1[a, b]$ 连续可导 (即 φ' 也连续);
- $[x_i, x_{i+1}]$ 上 $\varphi \in \mathcal{P}_3$ 是三次多项式函数;
- $\varphi(x_i) = f(x_i), \varphi'(x_i) = f'(x_i)$.

满足前两个性质的函数组成插值空间 Φ , 且 $\dim(\Phi) = 2n + 2$.

定义插值基函数, 在 $[x_i, x_{i+1}]$ 上

$$\alpha_i(x) = \left(1 + 2 \frac{x - x_i}{x_{i+1} - x_i}\right) \left(\frac{x - x_{i+1}}{x_i - x_{i+1}}\right)^2, \quad (2.21a)$$

$$\alpha_{i+1}(x) = \left(1 + 2 \frac{x - x_{i+1}}{x_i - x_{i+1}}\right) \left(\frac{x - x_i}{x_{i+1} - x_i}\right)^2, \quad (2.21b)$$

$$\beta_i(x) = (x - x_i) \left(\frac{x - x_{i+1}}{x_i - x_{i+1}}\right)^2, \quad (2.21c)$$

$$\beta_{i+1}(x) = (x - x_{i+1}) \left(\frac{x - x_i}{x_{i+1} - x_i}\right)^2, \quad (2.21d)$$

满足

$$\alpha_i(x_j) = \delta_{ij}, \quad \alpha'_i(x_j) = 0, \quad (2.22a)$$

$$\beta_i(x_j) = 0, \quad \beta'_i(x_j) = \delta_{ij}. \quad (2.22b)$$

则

$$\varphi(x) = \sum_{i=1}^n f(x_i) \alpha_i(x) + f'(x) \beta_i(x). \quad (2.23)$$

2.3 Fourier 插值

定义 2.3.1: Fourier 级数

周期函数展开成 Fourier 级数

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)]. \quad (2.24)$$

其中

$$a_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(nx) dx, \quad (2.25a)$$

$$b_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(nx) dx. \quad (2.25b)$$

如果 $f \in \mathcal{C}^M$, 则 $a_n = \mathcal{O}(n^{-M})$, $b_n = \mathcal{O}(n^{-M})$, 且

$$\left\| f(x) - \left[\frac{a_0}{2} + \sum_{n=1}^N [a_n \cos(nx) + b_n \sin(nx)] \right] \right\|_{\infty} = \mathcal{O}(N^{-M}).$$

定义 2.3.2: 三角多项式插值

给定周期为 2π 的函数 f 在节点 $x_i = 2\pi i/N$ 的值, 希望重构函数 ψ 满足 $\psi(x_i) = f(x_i)$.

寻找相多项式

$$p(x) = \beta_0 + \beta_1 e^{ix} + \cdots + \beta_{N-1} e^{i(N-1)x}. \quad (2.26)$$

定理 2.3.1: Fourier 插值

存在唯一的相多项式满足 Lagrange 插值条件且

$$\beta_j = \frac{1}{N} \sum_{k=1}^{N-1} f(x_k) \omega^{-kj}, \quad \omega = e^{2\pi i/N}. \quad (2.27)$$

对应三角多项式为

$$A_j = \frac{2}{N} \sum_{k=0}^{N-1} f(x_k) \cos\left(\frac{2\pi kj}{N}\right), \quad (2.28a)$$

$$B_j = \frac{2}{N} \sum_{k=0}^{N-1} f(x_k) \sin\left(\frac{2\pi kj}{N}\right), \quad (2.28b)$$

第三章 函数逼近

定义 3.0.1: 函数逼近

给定函数 $f \in C[a, b]$ 和子集 $\Phi \subset C[a, b]$ (如多项式函数), 寻找最佳逼近:

$$\varphi^* = \arg \min_{\varphi \in \Phi} \|f - \varphi\|. \quad (3.1)$$

注.

- Φ 一般是简单函数集合, 如多项式函数. 但 Φ 未必是线性空间.
- $f \notin \Phi$, 且关于 f 的信息可能有误差;
- 近似程度的度量: 平方 (squares) 逼近 $\|\cdot\|_2$ 和一致 (uniform) 逼近 $\|\cdot\|_\infty$.

3.1 最佳平方逼近

3.1.1 法方程

最佳平方逼近中, $\|\cdot\|_2$ 是由 $C[a, b]$ 的某个内积 $\langle \cdot, \cdot \rangle$ 诱导出的范数:

$$\|f\|_2 := \sqrt{\langle f, f \rangle}.$$

给定 Φ 的一组基 $\varphi_0, \dots, \varphi_n$, 则

$$\varphi = \sum_{i=0}^n a_i \varphi_i,$$

实函数情况, 简单计算得到

$$\|f - \varphi\|_2^2 = \sum_{i,j=0}^n \langle \varphi_i, \varphi_j \rangle a_i a_j - 2 \sum_{i=0}^n \langle \varphi_i, f \rangle a_i + \langle f, f \rangle.$$

是一个关于 a_0, \dots, a_n 的二次函数, 故 Hess 矩阵

$$\nabla^2 \|f - \varphi\|_2^2 = 2 \begin{bmatrix} \langle \varphi_0, \varphi_0 \rangle & \cdots & \langle \varphi_0, \varphi_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \varphi_n, \varphi_0 \rangle & \cdots & \langle \varphi_n, \varphi_n \rangle \end{bmatrix} \quad (3.2)$$

为对称正定矩阵, 因而 $\|f - \varphi\|_2^2$ 有最小值. 由

$$\frac{\partial}{\partial a_i} \|f - \varphi\|_2^2 = 2 \left[\sum_{j=0}^n \langle \varphi_i, \varphi_j \rangle a_j - \langle \varphi_i, f \rangle \right] = 0,$$

得到

定理 3.1.1: 法方程

$$\begin{bmatrix} \langle \varphi_0, \varphi_0 \rangle & \cdots & \langle \varphi_0, \varphi_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \varphi_n, \varphi_0 \rangle & \cdots & \langle \varphi_n, \varphi_n \rangle \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \langle \varphi_0, f \rangle \\ \vdots \\ \langle \varphi_n, f \rangle \end{bmatrix}. \quad (3.3)$$

注. 最佳平方逼近 φ^* 满足 $f - \varphi^* \perp \Phi$, 有

$$\|f\|_2^2 = \|\varphi^*\|_2^2 + \|f - \varphi^*\|_2^2. \quad (3.4)$$

例 3.1.1

给定 $\Phi = \text{span}(1, x^2)$, 求 $f = x$ 在 $[0, 1]$ 上的最佳平方逼近.

$$\begin{bmatrix} \langle 1, 1 \rangle & \langle 1, x^2 \rangle \\ \langle x^2, 1 \rangle & \langle x^2, x^2 \rangle \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \langle 1, x \rangle \\ \langle x^2, x \rangle \end{bmatrix}$$

解得 $\varphi^*(x) = a + bx^2$ 的系数

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 & 1/3 \\ 1/3 & 1/5 \end{bmatrix}^{-1} \begin{bmatrix} 1/2 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 3/16 \\ 15/16 \end{bmatrix}.$$

3.1.2 正交多项式

定义 3.1.1: Hilbert 矩阵

考虑 $f \in C[0, 1]$ 在 \mathcal{P}_n 上的最佳平方逼近, $1, x, \dots, x^n$ 作为基, 法方程的系数矩阵为 Hilbert 矩阵

$$H_n := \begin{bmatrix} 1 & 1/2 & \cdots & 1/(n+1) \\ 1/2 & 1/3 & \cdots & 1/(n+2) \\ \vdots & \vdots & \ddots & \vdots \\ 1/(n+1) & 1/(n+2) & \cdots & 1/(2n+1) \end{bmatrix} \quad (3.5)$$

这个矩阵是严重病态的.

注. 解决方法: 将 \mathcal{P}_n 的基改进为正交基 $\varphi_0, \dots, \varphi_n$, 即

$$\langle \varphi_i, \varphi_j \rangle = \|\varphi_i\|_2^2 \delta_{ij},$$

则系数矩阵是对角的,

$$\varphi^* = \sum_{i=0}^n \frac{\langle f, \varphi_i \rangle}{\|\varphi_i\|_2^2} \varphi_i. \quad (3.6)$$

定义 3.1.2: 广义 Fourier 级数

给定 $C[a, b]$ 上的一组正交归一函数 $\varphi_0, \varphi_1, \dots$, f 在 $\Phi_n := \text{span}(\varphi_0, \dots, \varphi_n)$ 中的最佳平方逼近为

$$f_n = \sum_{i=0}^n a_i \varphi_i, \quad a_i := \langle f, \varphi_i \rangle,$$

由于

$$\sum_{i=0}^n |a_i|^2 = \|f_n\|_2^2 \leq \|f\|_2^2, \implies \sum_{i=0}^{\infty} |a_i|^2 < \infty,$$

故 f_∞ 收敛且 $f_\infty \in \overline{\langle C[a, b], \|\cdot\|_2 \rangle}$, 称为广义 Fourier 级数.

注. 若 $\|\cdot\|_2$ 是权系数 ρ 的内积 $\langle \cdot, \cdot \rangle_\rho$ 诱导的范数, 则

$$\overline{\langle C[a, b], \|\cdot\|_2 \rangle} = \mathcal{L}_\rho^2[a, b] \quad (3.7)$$

定理 3.1.2

给定有界闭区间 $[a, b]$, 若 $f \in C[a, b]$ 在 \mathcal{P}_n 上的最佳平方逼近为 f_n , 则

$$\lim_{n \rightarrow \infty} \|f - f_n\|_2 = 0. \quad (3.8)$$

这说明多项式函数是完备的.

定理 3.1.3: Legendre 多项式作最佳平方逼近

若内积为 $[-1, 1]$ 上的 $\rho \equiv 1$ 内积, 若 $f \in C^2[-1, 1]$, 则 $\forall \epsilon > 0$, $\exists N > 0$ 使得 $\forall n \geq N$

$$\|f - f_n\|_\infty \leq \frac{\epsilon}{\sqrt{n}}. \quad (3.9)$$

定理 3.1.4: 零平方误差最小

在所有首项系数为 1 的 n 次多项式中, Legendre 多项式 $P_n(x)$ 在 $[-1, 1]$ 与 0 的平方误差最小.

证明. $\forall f \in \mathcal{P}_n$ 且首项系数为 1, 有

$$f = P_n + a_{n-1}P_{n-1} + \dots + a_0P_0,$$

则

$$\|f\|_2^2 = \|P_n\|_2^2 + |a_{n-1}|^2 \|P_{n-1}\|_2^2 + \dots + |a_0|^2 \|P_0\|_2^2 \geq \|P_n\|_2^2. \quad \square$$

3.2 最小二乘法

给定 $\Phi = \text{span}\{\varphi_0, \dots, \varphi_n\}$

定义 3.2.1: Haar 条件

$\forall \varphi_i \in \Phi$ 且 $\varphi_i \neq 0$ 在 x_0, \dots, x_m 上不同时为 0, 则称 Φ 满足 Haar 条件.

例 3.2.1

若 $m \geq n$, 则 \mathcal{P}_n 在 x_0, \dots, x_m 上满足 Haar 条件.

3.3 最佳一致逼近

考虑用多项式函数进行最佳一致逼近:

$$\varphi^* = \arg \inf_{\varphi \in \mathcal{P}_n} \|f - \varphi\|_{\infty}. \quad (3.10)$$

定义 3.3.1

称 $\|f - p_n\|_{\infty}$ 为 p_n 关于 f 的偏差. 若 $\exists x$ 使得

$$f(x) - p_n(x) = \pm \|f - p_n\|_{\infty}. \quad (3.11)$$

则称 x 为 p_n 关于 f 的正 (负) 偏差点.

定理 3.3.1: 最佳一致逼近的存在性

\mathcal{P}_n 关于 $f \in C[a, b]$ 的最小偏差可以达到.

证明. 记 $\varphi(f, p_n) = \|f - p_n\|_{\infty}$, 则 $\varphi(f, \cdot)$ 是关于 p_n 系数 a_0, \dots, a_n 的连续函数, 且 $\varphi(0; p_n)$ 在单位球面 $a_0^2 + \dots + a_n^2 = 1$ 上达到正最小值 μ .

$$\|f - p_n\|_{\infty} \geq \|p_n\|_{\infty} - \|f\|_{\infty} \geq \mu \sqrt{\sum_{i=0}^n a_i^2} - \|f\|_{\infty}.$$

当

$$\sum_{i=0}^n a_i^2 \geq \frac{4}{\mu^2} \|f\|_{\infty}^2$$

时, $\|f - p_n\|_{\infty} \geq \|f\|_{\infty}$, 故最佳一致逼近存在. \square

引理. 最佳一致逼近多项式 p_n^* 关于 f 的正负偏差点同时存在.

定义 3.3.2: Chebyshev 交错点

若 x_1, \dots, x_m 是 p_n 关于 f 的偏差点且轮流为正负, 则称其为一组 Chebyshev 交错点.

定理 3.3.2

p_n^* 是 f 的最佳一致逼近多项式 \iff 存在 $n+2$ 个 Chebyshev 交错点.

推论. 最佳一致逼近多项式是一个 Lagrange 插值多项式, 其插值点在 (a, b) 内.

定理 3.3.3: 最佳一致逼近的唯一性

最佳一致逼近多项式是唯一的.

定理 3.3.4: 零偏差最小

在所有首项系数为 1 的 n 次多项式中, Chebyshev 多项式 $T_n(x)$ 在 $[-1, 1]$ 与 0 的偏差最小.

例 3.3.1: 最佳一次一致逼近多项式的求法

给定 $f \in C^2[a, b]$ 且 $f''(x) \neq 0$, 求 f 的最佳一次一致逼近多项式 $g = a_0 + a_1x$.
由 f'' 恒正或恒负可知 $f - g$ 在 (a, b) 上的极值点是唯一的, 记作 x^* ,

$$f'(x^*) - g'(x^*) = 0,$$

且 $\{a, x^*, b\}$ 是一组 Chebyshev 交错点, 故

$$f(a) - g(a) = -[f(x^*) - g(x^*)] = f(b) - g(b)$$

可确定 a_1, x^*, a_0 :

$$\begin{cases} a_1 = \frac{f(b) - f(a)}{b - a}, \\ x^* = f'^{-1}(a_0), \\ a_0 = \frac{1}{2}[f(a) + f(x^*) - a_1(a + x^*)] \end{cases}$$

第四章 数值积分

给定函数 $f \in \mathcal{C}[a, b]$ 和权函数 $\rho(x)$, 计算积分

$$I(f) \equiv \int_a^b f(x)\rho(x) dx, \quad (4.1)$$

但很多情况下, 我们只知道关于 f 的部分信息, 不能用 Newton-Leibniz 公式. 因此我们希望构造一种不依赖于 f 具体表达式的近似积分方法.

定义 4.0.1: 数值积分

给定系列积分节点 x_0, \dots, x_n , 数值积分 $\tilde{I}(f)$ 一般具有如下形式:

$$\tilde{I}(f) = \sum_{k=0}^n A_k f(x_k). \quad (4.2)$$

其中 A_k 称为求积系数. 积分方法 \tilde{I} 的误差为

$$E(f) := I(f) - \tilde{I}(f). \quad (4.3)$$

给定一种积分方法 \tilde{I} , 如何评价其优劣?

定义 4.0.2: 代数精确度

给定积分法 \tilde{I} , 若 $\forall f \in \mathcal{P}_n$ 均有 $E(f) = 0$; 但 $\exists f \in \mathcal{P}_{n+1}$ 使得 $E(f) \neq 0$, 则 \tilde{I} 的代数精确度为 n .

定义 4.0.3: 插值型求积公式

给定系列插值节点 $x_0 < x_1 < \dots < x_n$, 由式 (2.5), 函数 f 的 n 次插值多项式为

$$p_n(x) = \sum_{i=0}^n f(x_i) \ell_i(x) \approx f(x),$$

选定

$$I_n(f) \equiv I(p_n) = \sum_{i=0}^n I(\ell_i) f(x_i) \equiv \sum_{i=0}^n A_i^{(n)} f(x_i), \quad (4.4)$$

其中 $A_k^{(n)}$ 与 f 无关, 称为求积系数.

4.1 Newton-Cotes 公式

考虑 $\rho \equiv 1$ 的情形.

4.1.1 闭型 Newton-Cotes 公式

定理 4.1.1: (闭型) Newton-Cotes 公式

取等距节点 $a = x_0 < x_1 < \cdots < x_n = b$, 节点间距 $h := (b - a)/n$, 有

$$I_n(f) = (b - a) \sum_{k=0}^n C_k^{(n)} f(x_k), \quad (4.5)$$

其中 $C_k^{(n)}$ 称为求积 Cotes 系数, 由式 (2.3)

$$C_k^{(n)} = \frac{(-1)^{n-k}}{k!(n-k)!} \frac{1}{n} \int_0^n \prod_{j \neq k} (t - j) dt. \quad (4.6)$$

注. Cotes 系数不但与 f 无关, 也与积分区间 $[a, b]$ 无关. 且 $C_k^{(n)} = C_{n-k}^{(n)}$.

例 4.1.1: 前几个 Cotes 系数

n	$C_k^{(n)}$								
1	1 1 / 2								
2	1 4 1 / 6								
3	1 3 3 1 / 8								
4	7 32 12 12 7 / 90								
5	19 75 50 50 75 19 / 288								
6	41 216 27 572 27 216 41 / 840								
7	751 3577 1323 2989 2989 1323 3577 751 / 17280								
8	989 5888 -928 10496 -4540 10496 -928 5888 989 / 28350								

每行斜线 / 的左边表示 $C_0^{(n)}, \dots, C_n^{(n)}$ 的分子, 右边表示他们共同的分母.

例 4.1.2: $n = 1$: 梯形公式

$n = 1$ 时, 只有 a, b 两个求积节点, 得到梯形公式 (trapezoidal)

$$I_1(f) = \frac{b-a}{2} [f(a) + f(b)]. \quad (4.7)$$

若 $f \in \mathcal{C}^2[a, b]$, 则 $\exists \xi \in [a, b]$, 积分误差

$$E_1(f) = -\frac{h^3}{12} f''(\xi). \quad (4.8)$$

梯形公式的代数精确度为 1.

例 4.1.3: $n = 2$: Simpson 公式

$n = 2$ 时, 有 $a, x_1 = (a + b)/2, b$ 三个求积节点, 得到 Simpson 公式:

$$I_2(f) = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]. \quad (4.9)$$

若 $f \in C^4[a, b]$, 则 $\exists \xi \in [a, b]$, 积分误差

$$E_2(f) = -\frac{h^5}{90} f^{(4)}(\xi). \quad (4.10)$$

Simpson 公式的代数精确度为 3.

注. $n = 3$ 时称为 Simpson 3/8 公式

$$E_3(f) = -\frac{3h^5}{80} f^{(4)}(\xi),$$

$n = 4$ 时称为 Boole 公式或 Cotes 公式

$$E_4(f) = -\frac{8h^7}{945} f^{(6)}(\xi).$$

定理 4.1.2: Newton-Cotes 公式的代数精确度

当 n 为奇数时, 代数精确度为 $n + 1$; 当 n 为偶数时, 代数精确度为 n .

4.1.2 开型 Newton-Cotes 公式**定理 4.1.3: 开型 Newton-Cotes 公式**

取等距节点 $a = x_{-1} < x_0 < \dots < x_n < x_{n+1} = b$, 以 x_0, \dots, x_n 为插值节点.

例 4.1.4: $n = 0$: 中点公式

$n = 0$ 的开型 Newton-Cotes 公式称为中点公式:

$$I_0 = (b-a) f\left(\frac{a+b}{2}\right). \quad (4.11)$$

若 $f \in C^2[a, b]$, 则 $\exists \xi \in [a, b]$, 积分误差

$$E_0(f) = \frac{h^3}{3} f''(\xi). \quad (4.12)$$

注. 一般来说, 闭型 Newton-Cotes 公式的结果比相应的开型 Newton-Cotes 公式的结果要好.

4.1.3 积分法的一致稳定性

由于积分节点 $f(x_i)$ 可能存在误差, 需要考虑随着渐进参数 n 的增大, 积分法的稳定性.

定义 4.1.1: 一致稳定

对于线性积分方法 \tilde{I}_n , 若 $\forall n \in \mathbb{N}$, $\exists M > 0$ 使得

$$\sum_{k=0}^n |A_k^{(n)}| < M, \quad (4.13)$$

则称 \tilde{I}_n 是一致稳定的.

推论. 若 $\|f - g\|_\infty \leq \delta$, 且 \tilde{I}_n 一致稳定, 则

$$|\tilde{I}_n(f) - \tilde{I}_n(g)| = |\tilde{I}_n(f - g)| \leq \sum_{k=0}^n |A_k^{(n)}| \|f - g\|_\infty \leq M\delta.$$

定理 4.1.4: 一致稳定的判定

若 $\forall n, k$ 均有 $A_k^{(n)} > 0$ 且 \tilde{I}_n 的代数精确度至少为 0, 则 \tilde{I}_n 一致稳定.

注. $n \geq 8$ 时, Newton-Catos 积分法存在 $A_k^{(n)} = (b - a)C_k^{(n)} < 0$, 故不是一致稳定的.

4.2 复合求积公式

对于一些函数 f 来说, 低阶 Newton-Catos 公式很不准确, 而高阶又存在数值不稳定问题. 为此需要新的求积方法, 一种思路是利用积分运算关于测度的可加性, 将积分区间分成若干小的区间进行分片求积.

定义 4.2.1: 复合求积公式

将积分区间 $[a, b]$ 划分成若干子区间, 再在每个子区间上采用低阶 Newton-Catos 公式.

例 4.2.1: 复合梯形公式

取节点 $a = x_0 < x_1 < \dots < x_n = b$, 子区间为 $[x_k, x_{k+1}]$, 区间间隔 $h_k = x_{k+1} - x_k$, 对每个子区间套用梯形积分公式 (4.7), 得到复合梯形公式为:

$$\begin{aligned} T_n(f) &= \sum_{k=0}^{n-1} \frac{h_k}{2} [f(x_k) + f(x_{k+1})] \\ &= \frac{h_0}{2} f(x_0) + \frac{h_0 + h_1}{2} f(x_1) + \dots + \frac{h_{n-2} + h_{n-1}}{2} f(x_{n-1}) + \frac{h_{n-1}}{2} f(x_n). \end{aligned} \quad (4.14)$$

由式 (4.8) 和中值定理知, $\exists \xi_k \in [x_k, x_{k+1}], \eta \in [a, b]$ 使得

$$E_n(f) = -\frac{1}{12} \sum_{k=0}^{n-1} h_k^3 f''(\xi_k) = -\frac{1}{12} \sum_{k=0}^{n-1} h_k^3 f''(\eta), \quad (4.15)$$

当等距均分时, $h_0 = h_1 = \dots = h_{n-1} = h = (b - a)/n$, 有

$$E_n(f) = -\frac{b-a}{12} h^2 f''(\eta). \quad (4.16)$$

例 4.2.2: 复合中点公式

对每个子区间套用中点积分公式 (4.11), 得到复合中点公式为:

$$H_n(f) = \sum_{k=0}^{n-1} h_k f(x_{k+1/2}), \quad x_{k+1/2} := \frac{x_k + x_{k+1}}{2}. \quad (4.17)$$

推论. 考虑每个子区间上增加一个节点 $x_{k+1/2}$, 得到递推公式:

$$T_{2n}(f) = \frac{T_n(f) + H_n(f)}{2}. \quad (4.18)$$

例 4.2.3: 复合 Simpson 公式

对每个子区间套用 Simpson 积分公式 (4.9), 得到复合 Simpson 公式为:

$$S_{2n}(f) = \sum_{k=0}^{n-1} \frac{h_k}{6} [f(x_k) + 4f(x_{k+1/2}) + f(x_{k+1})]. \quad (4.19)$$

若最大区间长度 $h := \max_k h_k$, 则 $\exists \eta \in [a, b]$ 使得

$$|E_{2n}(f)| \leq \frac{b-a}{2880} h^4 f^{(4)}(\eta). \quad (4.20)$$

注. 复合梯形公式、复合中点公式、复合 Simpson 公式都是一致稳定的.

推论. 递推关系:

$$S_{2n}(f) = \frac{T_n(f) + 2H_n(f)}{3} = \frac{4T_{2n}(f) - T_n(f)}{3}. \quad (4.21)$$

4.3 Romberg 求积方法

4.3.1 Euler-Maclaurin 求积公式

定义 4.3.1: Bernoulli 多项式

Bernoulli 多项式递归定义: $B_0(x) = 1$,

$$B_k(x) = k \int_0^x B_{k-1}(t) dt - k \int_0^1 \int_0^x B_{k-1}(t) dt dx \quad (4.22)$$

Bernoulli 数 $B_k := B_k(0)$.

例 4.3.1: 前几个 Bernoulli 多项式

$$\begin{aligned} B_1(x) &= x - \frac{1}{2}, & B_2(x) &= x^2 - x + \frac{1}{6}, \\ B_3(x) &= x^3 - \frac{3}{2}x^2 + \frac{1}{2}x, & B_4(x) &= x^4 - 2x^3 + x^2 - \frac{1}{30}, \\ B_5(x) &= x^5 - \frac{5}{2}x^4 + \frac{5}{3}x^3 - \frac{1}{6}x, & B_6(x) &= x^6 - 3x^5 + \frac{5}{2}x^4 - \frac{1}{2}x^2 + \frac{1}{42}. \end{aligned}$$

定理 4.3.1: Euler-Maclaurin 求积公式

若 $f \in C^{2m+2}[a, b]$, 则有

$$I(f) = T_n(f) + \sum_{\ell=1}^m \frac{B_{2\ell}}{(2\ell)!} [f^{(2\ell-1)}(a) - f^{(2\ell-1)}(b)] h^{2\ell} + r_{m+1}, \quad (4.23)$$

其中 T_n 是复合梯形公式,

$$r_{m+1} = -\frac{B_{2m+2}}{(2m+2)!} (b-a) f^{(2m+2)}(\eta) h^{2m+2}, \quad \exists \eta \in (a, b). \quad (4.24)$$

4.3.2 Richardson 外推方法

由 Euler-Maclaurin 求积公式, 考虑复合梯形公式关于 $h < 1$ 的渐进级数

$$T_n(f) = I(f) + \tau_1 h^2 + \tau_2 h^4 + \cdots + \tau_m h^{2m} + \mathcal{O}(h^{2m+2}).$$

当区间无限小时, $h \rightarrow 0$, $T_n(f) \rightarrow I(f)$ 是准确积分. 为了得到 $I(f)$ 的一个近似, 考虑

$$h_i = \frac{b-a}{n_i}, \quad i = 0, 1, \dots$$

且 $n_0 < n_1 < \cdots < n_m < \cdots$, 记关于 h^2 的插值多项式 $p_{i,j} \in \mathcal{P}^{j-i}$ 满足插值条件:

$$p_{i,j}(h_i) = T_{n_i}(f), \quad p_{i,j}(h_{i+1}) = T_{n_{i+1}}(f), \quad \dots, \quad p_{i,j}(h_j) = T_{n_j}(f),$$

由 Neville 算法 (2.16), 有递推关系: $p_{i,i}(h) = T_{n_i}(f)$,

$$p_{i,j}(h) = \frac{(h^2 - h_i^2)p_{i+1,j}(h) + (h_j^2 - h^2)p_{i,j-1}(h)}{h_j^2 - h_i^2},$$

定理 4.3.2: Richardson 外推方法

将外推值 $p_{0,i}(0)$ 作为 $I(f)$ 的一个近似, 记 $T_{i,j} := p_{i,j}(0)$, 有递推关系

$$T_{i,j} = T_{i+1,j} + \frac{T_{i+1,j} - T_{i,j-1}}{n_j^2/n_i^2 - 1}. \quad (4.25)$$

4.3.3 Romberg 求积方法**定理 4.3.3: Romberg 求积方法**

对区间 $[a, b]$ 进行 $n_k = 2^k$ 等分 ($k = 0, 1, \dots$), 利用复合梯形公式计算 T_1^k , 再递推

$$T_{i+1}^k = \frac{4^i T_i^{k+1} - T_i^k}{4^i - 1}. \quad (4.26)$$

直到 T_{i+1}^0 满足精度要求.

特别地, T_1^0 为梯形公式, T_2^k 为复合 Simpson 公式.

4.4 Gauss 求积公式

考虑一般的带权积分 $\rho \neq 1$.

定理 4.4.1

给定插值节点 x_0, \dots, x_n , ℓ_0, \dots, ℓ_n 是插值基函数, 则当且仅当

$$A_k^{(n)} = I(\ell_k) = \int_a^b \ell_k(x) \rho(x) dx$$

时 $I_n(\cdot)$ 的代数精确度至少为 n .

证明. 充分性: $\forall f \in \mathcal{P}_n$, 有

$$f(x) = \sum_{k=0}^n f(x_k) \ell_k(x),$$

于是

$$I(f) = \sum_{k=0}^n f(x_k) I(\ell_k) = \sum_{k=0}^n A_k^{(n)} f(x_k) = I_n(f);$$

必要性: 若 $\forall f \in \mathcal{P}_n$ 均有 $I(f) = I_n(f)$, 则

$$I(f) - I_n(f) = \sum_{k=0}^n [I(\ell_k) - A_k^{(n)}] f(x_k) = 0,$$

特别地, 取 $f(x) = 1, x, \dots, x^n$, 得到一个 Vandermonde 矩阵, 从而 $A_k^{(n)} = I(\ell_k)$. □

定理 4.4.2

$n+1$ 个求积节点的求积公式 $I_n(\cdot)$ 代数精确度不超过 $2n+1$.

证明. 给定节点 x_0, \dots, x_n 定义

$$\omega_{n+1}(x) := \prod_{k=0}^n (x - x_k),$$

则 $I(\omega_{n+1}^2) > 0 = I_n(\omega_{n+1}^2)$, 说明代数精确度 $< 2n+2$. □

如何达到最大可能代数精确度 由 (2.7), 余项

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \omega_{n+1}(x).$$

故积分误差为

$$E_n(f) = \frac{1}{(n+1)!} I(f^{(n+1)}(\xi(x)) \omega_{n+1}(x)).$$

若 I_n 具有 $2n+1$ 次代数精确度, 即 $\forall f \in \mathcal{P}_{2n+1}$ 均有 $E_n(f) = 0$, 由于 $f^{(n+1)} \in \mathcal{P}_n$, 故

$$\omega_{n+1} \perp \mathcal{P}_n. \quad (4.27)$$

即 ω_{n+1} 为 $[a, b]$ 上权 ρ 的 $n+1$ 次正交多项式, 这可以通过选取节点得到.

定理 4.4.3: Gauss 求积公式

取 $[a, b]$ 上权 ρ 的 $n+1$ 次正交多项式的根 x_0, \dots, x_n 作为求积节点, 则

$$I_n(f) = \sum_{k=0}^n I(\ell_k) f(x_k),$$

具有 $2n+1$ 阶代数精确度, 称为 Gauss 公式, 节点称为 Gauss 点.

定理 4.4.4

Gauss 求积公式是一致稳定的.

证明. 略

□

第五章 线性方程组的直接解法

求解线性代数方程组就相当于求解矩阵式 $Ax = b$:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}. \quad (5.1)$$

因此有必要了解一些对矩阵的操作.

5.1 矩阵操作

定义 5.1.1: 稀疏矩阵

如果一个矩阵绝大多数元素是 0, 则称其是稀疏的 (sparse).

定义 5.1.2: 秩一矩阵

若一个矩阵可以表示成 $A = uv^\dagger$, 则其秩为一.

定理 5.1.1: 奇异值分解

根据奇异值分解, 可以将矩阵表示成秩一矩阵的线性组合:

$$A = U\Sigma V^\dagger = \sum_i \sigma_i u_i v_i^\dagger, \quad (5.2)$$

其中 U, V 为幺正矩阵.

定理 5.1.2: 矩阵的 Hierarchical 表示

考虑分块矩阵

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

A 一般不是稀疏的, 但非对角元 A_{12}, A_{22} 是稀疏的, 则 Ax 可以分块地写成:

$$Ax = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 \\ A_{21}x_1 + A_{22}x_2 \end{bmatrix},$$

对 $A_{11}x, A_{22}x_2$ 递归处理.

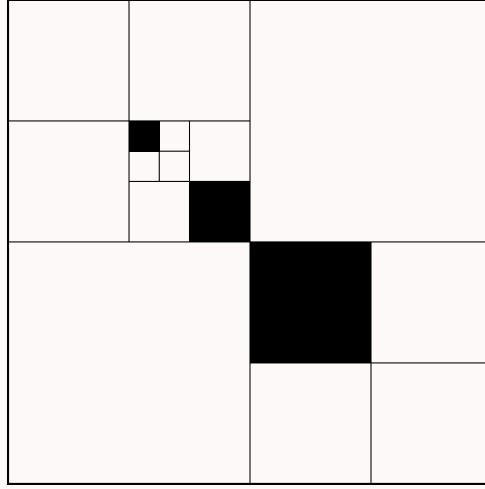


图 5.1: Hierarchical 算法示意图，白块表示稀疏部分

定理 5.1.3: 矩阵乘法的 Strassen 算法

对矩阵乘法 $C = AB$ 分块得：

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

定义

$$M_1 = (A_{11} + A_{22})(B_{11} + B_{22}), \quad (5.3a)$$

$$M_2 = (A_{21} + A_{22})B_{11}, \quad (5.3b)$$

$$M_3 = A_{11}(B_{12} - B_{22}), \quad (5.3c)$$

$$M_4 = A_{22}(B_{21} - B_{11}), \quad (5.3d)$$

$$M_5 = (A_{11} + A_{12})B_{22}, \quad (5.3e)$$

$$M_6 = (A_{21} - A_{11})(B_{11} + B_{12}), \quad (5.3f)$$

$$M_7 = (A_{12} - A_{22})(B_{21} + B_{22}), \quad (5.3g)$$

则

$$C_{11} = M_1 + M_4 - M_5 + M_7, \quad (5.4a)$$

$$C_{12} = M_3 + M_5, \quad (5.4b)$$

$$C_{21} = M_2 + M_4, \quad (5.4c)$$

$$C_{22} = M_1 - M_2 + M_3 + M_6. \quad (5.4d)$$

Strassen 算法将分块矩阵的乘法从直接法的 8 次降低到了 7 次，由此分而治之，矩阵乘法的时间复杂度便从 $\mathcal{O}(n^3)$ 降低到了 $\mathcal{O}(n^{\log_2 7}) = \mathcal{O}(n^{2.807})$.

定义 5.1.3: 离散 Fourier 变换

式 (2.27) 中定义了一个线性变换, 称为离散 Fourier 变换 (discrete Fourier transform, DFT)

$$X_n = \sum_{m=0}^{N-1} x_m \omega^{mn},$$

其中 $\omega := e^{-i2\pi/N}$ 是 N 次单位根, 对应的变换矩阵为

$$F = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & (\omega^2)^2 & \cdots & (\omega^2)^{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & (\omega^{N-1})^2 & \cdots & (\omega^{N-1})^{N-1} \end{bmatrix} \quad (5.5)$$

事实上 F 上只有 n 个不同的元素. 其逆变换为

$$F^{-1} = \frac{1}{N} \bar{F}. \quad (5.6)$$

定理 5.1.4: Cooley-Tukey 快速 Fourier 变换

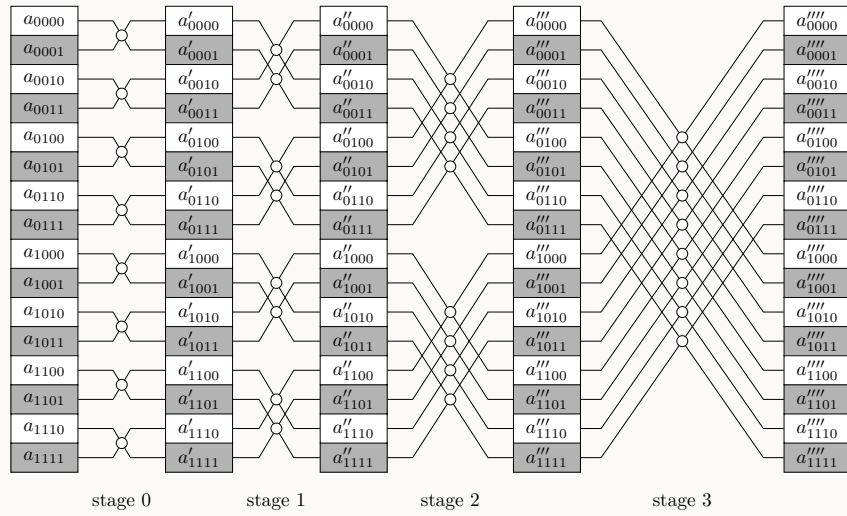
以基 2 (radix-2) 的情形为例, 即 $N = 2^M$. 将 X_n 的求和分成偶数项 E_n 和奇数项 O_n

$$\begin{aligned} X_n &= \sum_{k=0}^{N/2-1} x_{2k} \omega^{2kn} + \sum_{k=0}^{N/2-1} x_{2k+1} \omega^{(2k+1)n} \\ &= \sum_{k=0}^{N/2-1} x_{2k} \omega^{2kn} + \omega^n \sum_{k=0}^{N/2-1} x_{2k+1} \omega^{2kn} =: E_n + \omega^n O_n, \end{aligned}$$

由于 $\omega^N = 1$, 注意到

$$\begin{aligned} X_{n+N/2} &= \sum_{k=0}^{N/2-1} x_{2k} \omega^{2k(n+N/2)} + \sum_{k=0}^{N/2-1} x_{2k+1} \omega^{(2k+1)(n+N/2)} \\ &= \sum_{k=0}^{N/2-1} x_{2k} \omega^{2kn} - \omega^n \sum_{k=0}^{N/2-1} x_{2k+1} \omega^{2kn} = E_n - \omega^n O_n, \end{aligned}$$

由此便将 N 个 X_n 求和 (N^2) 转化成了 $N/2$ 个 E_n, O_n 求和 ($N^2/2$). 采用分而治之的算法思想, 可以将 DFT 的时间复杂度从矩阵向量乘法的 $\mathcal{O}(N^2)$ 优化到 $\mathcal{O}(N \log N)$, 这称为快速 Fourier 变换 (fast Fourier transform, FFT).

图 5.2: FFT 算法示意图 ($N = 2^3 = 8$), 作者: Alexandros Tsagkaropoulos

对于非基 2 的情形, ω^k 的周期不是基 2 的, 做处理:

$$X_n = \sum_{m=0}^{N-1} x_m \omega^{-mn} = \sum_{m=0}^N x_m \omega^{[(m-n)^2 - m^2 - n^2]/2}.$$

定义 $\nu_k := \omega^{k^2/2}$, 记 $Y_n := \nu_n X_n$, $z_m := \nu_m^{-1} x_m$, 则有卷积形式:

$$Y_n = \sum_{m=0}^{N-1} z_m \nu_{m-n}.$$

可以用 0 将 z_m, ν_m 延拓, 使其周期是一个比 N 大的基 2 数 N' . 再在两端做 DFT:

$$\begin{aligned} \sum_{n=0}^{N'-1} Y_n \omega_{N'}^{nk} &= \sum_{n=0}^{N'-1} \sum_{m=0}^{N'-1} z_m \nu_{m-n} \omega_{N'}^{nk} \\ &= \sum_{m=0}^{N'-1} z_m \omega_{N'}^{mk} \sum_{n=0}^{N'-1} \nu_{m-n} \omega_{N'}^{(n-m)k} = \sum_{m=0}^{N'-1} z_m \omega_{N'}^{mk} \sum_{m=0}^{N'-1} \nu_m \omega_{N'}^{mk} \end{aligned}$$

因此通过对 z_m, ν_m 做两次 DFT、一次向量分量积、一次逆 DFT 便可得到 Y_n .

定理 5.1.5: 周期 Toeplitz 变换

n 阶矩阵 A 若满足 $a_{ij} = c_{i-j}$ 且序列 c_k 周期为 n , 则 A 称为周期 Toeplitz 矩阵, 且

$$A = F^{-1} \Lambda F, \quad (5.7)$$

其中 $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{n-1})$ 且

$$\lambda_i = \sum_{j=0}^{n-1} c_j \omega^{-ij}. \quad (5.8)$$

注. n 阶矩阵 A 若满足 $a_{ij} = c_{i-j}$, 则该矩阵可以扩展成 $2n$ 阶的周期 Toeplitz 矩阵.

$$\begin{bmatrix} A & * \\ * & * \end{bmatrix}$$

5.2 Gauss 消元法

5.2.1 Gauss 消元法

如何求解 n 元线性方程组? Cramer 法则? 时间复杂度 $\mathcal{O}(n \cdot (n+1)!)$ 这是不可接受的.

定理 5.2.1: Gauss 消元法

线性方程组形如

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + \cdots + a_{2n}^{(1)}x_n = b_2^{(1)} \\ \vdots \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}$$

如果 $a_{11}^{(1)} \neq 0$, 可将第一行的 $-a_{i1}^{(1)}/a_{11}^{(1)}$ 倍加到第 i 行 ($i = 2, 3, \dots, n$), 得到一个等价方程组

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \vdots \\ a_{n2}^{(2)}x_2 + \cdots + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases}$$

如果 $a_{22}^{(2)} \neq 0$, 便可以此类推……最终得到一个等价的上三角线性方程组:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \vdots \\ a_{nn}^{(n)}x_n = b_n^{(n)} \end{cases}$$

便不难从下至上地解得: $x_n = b_n^{(n)}/a_{nn}^{(n)}$,

$$x_i = \frac{1}{a_{ii}^{(i)}} \left(b_i^{(i)} - \sum_{j=i+1}^n a_{ij}^{(i)} x_j \right), \quad i = n-1, \dots, 2, 1. \quad (5.9)$$

Gauss 消元法的复杂度为 $\mathcal{O}(n^3)$.

注. 从上面的算法过程中可见, 一旦第 k 步 $a_{kk}^{(k)} = 0$, (顺序) Gauss 消元法就不能继续进行下去. 但是当第 $k+1, \dots, n$ 行中存在 $a_{ik}^k \neq 0$, 就可以交换 k, i 行, 从而使算法继续.

此外, 即使 $a_{kk}^{(k)} \neq 0$ 但 $|a_{kk}^{(k)}| \ll 1$, 也会出现大数除小数导致精度下降的问题.

定理 5.2.2: 列主元方法

在第 k 步消去之前, 找到绝对值最大的主元 (pivot):

$$i_k = \arg \max_{k \leq i \leq n} |a_{ik}^{(k)}|, \quad (5.10)$$

然后交换第 k, i_k 行.

5.2.2 LU 分解

下面我们从矩阵角度考察 Gauss 消元法.

定义 5.2.1: 初等矩阵

给定 (实) 向量 u, v 和标量 σ , 形如

$$E = I - \sigma uv^\top \quad (5.11)$$

的称为 (实) 初等矩阵 (elementary matrix).

推论. 初等矩阵的逆也是同类型的初等矩阵:

$$(I - \sigma uv^\top)^{-1} = I - \frac{\sigma}{\sigma v^\top u - 1} uv^\top. \quad (5.12)$$

例 5.2.1

- 初等排列矩阵:

$$P_{ij} = I - (e_i - e_j)(e_i - e_j)^\top = I - e_{ii} - e_{jj} + e_{ij} + e_{ji}.$$

左乘初等排列矩阵即互换第 i, j 行, 右乘即互换第 i, j 列;

- 倍加矩阵:

$$I + \alpha e_i e_j^\top = I + \alpha e_{ij},$$

左乘即将第 i 行的 α 倍加到第 j 行上.

定义 5.2.2: 初等单位下三角矩阵

形如

$$L_i = I + \ell_i e_i^\top = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \ell_{i+1,i} & 1 & \\ & & \vdots & & \ddots \\ & & \ell_{ni} & & & 1 \end{bmatrix} \quad (5.13)$$

称为第 i 列的初等单位下三角矩阵. 其中向量 ℓ_i 的前 i 个分量为 0.

推论. 初等单位下三角矩阵的逆也是初等单位下三角矩阵

$$(I + \ell_i e_i^\top)^{-1} = I - \ell_i e_i^\top. \quad (5.14)$$

推论. 对角元均为 1 的下三角矩阵称为单位下三角矩阵, 可以写成

$$L = L_1 L_2 \cdots L_{n-1} = \begin{bmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ \ell_{n1} & \ell_{n2} & \cdots & 1 \end{bmatrix}.$$

定理 5.2.3: 矩阵的 LU 分解

根据 Gauss 消元法的过程, 记

$$A^{(k)} := \begin{bmatrix} a_{11}^{(1)} & \cdots & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\ & \ddots & \vdots & \ddots & \vdots \\ & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & \vdots & \ddots & \vdots \\ & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix}, \quad b^{(k)} := \begin{bmatrix} b_1^{(1)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{bmatrix}, \quad \ell_k := -\frac{1}{a_{kk}^{(k)}} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ a_{k+1,k}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{bmatrix}.$$

相应的初等单位下三角矩阵为 $L_k = I + \ell_k e_k^\top$. 增广矩阵有递推关系:

$$[A^{(k+1)} \quad b^{(k+1)}] = L_k [A^{(k)} \quad b^{(k)}], \quad (5.15)$$

即

$$A^{(n)} = L_{n-1} \cdots L_1 A^{(1)}, \iff A^{(1)} = L_1^{-1} \cdots L_{n-1}^{-1} A^{(n)},$$

则 $L = L_1^{-1} \cdots L_{n-1}^{-1}$ 为单位下三角矩阵, $U = A^{(n)}$ 为上三角矩阵, 这样就将 $A^{(1)}$ 分解成了下上三角矩阵的乘积.

注. 根据 Gauss 消元法的过程可见, LU 分解的前提是 $a_{11}^{(1)}, \dots, a_{nn}^{(n)}$ 均不为 0.

定理 5.2.4: 三角分解定理

给定矩阵 A , 若其顺序主子式

$$\Delta_i := \begin{vmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ii} \end{vmatrix}, \quad i = 1, \dots, n \quad (5.16)$$

均不为 0, 则存在唯一的单位下三角矩阵 L 和上三角矩阵 U 使得 $A = LU$.

证明. 通过数学归纳法, 可证明: $a_{11}^{(1)}, \dots, a_{ii}^{(i)} \neq 0 \iff \Delta_1, \dots, \Delta_i \neq 0$, 此时

$$\Delta_i = a_{11}^{(1)} \cdots a_{ii}^{(i)}, \quad (5.17)$$

若存在 L_1, U_1 和 L_2, U_2 使得

$$A = L_1 U_1 = L_2 U_2,$$

两边左乘 L_1^{-1} , 右乘 U_2^{-1} 得到:

$$U_1 U_2^{-1} = L_1^{-1} L_2,$$

此式左边为上三角矩阵, 右边为单位下三角矩阵, 故只能是单位矩阵 I , 即 $U_1 = U_2, L_1 = L_2$. \square

例 5.2.2: LU 分解的例子

$$\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & 3 & 1 & \\ 3 & 4 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{bmatrix}.$$

注. L 有 $n(n-1)/2$ 个变量, U 有 $n(n+1)/2$ 个变量, 可以直接由 A 确定 L, U .

定理 5.2.5: 直接 LU 分解法 (Doolittle 分解法)

写成分块矩阵的形式:

$$L^{(k)} = \begin{bmatrix} 1 & \\ \ell_{n-k+1} & L^{(k-1)} \end{bmatrix}, \quad U^{(k)} = \begin{bmatrix} u_{u-k+1, u-k+1} & u_{u-k+1}^\top \\ & U^{(k-1)} \end{bmatrix},$$

则

$$A^{(k)} = L^{(k)} U^{(k)} = \begin{bmatrix} u_{u-k+1, u-k+1} & u_{u-k+1}^\top \\ u_{u-k+1, u-k+1} \ell_{n-k+1} & A^{(k-1)} + \ell_{n-k+1} u_{u-k+1}^\top \end{bmatrix}.$$

由此可确定 u, ℓ , 同时 $A^{(k)}$ 的阶数减 1.

注. 三角分解定理 5.2.4 给出了 LU 分解的条件. 而对于一般的可逆矩阵, 也可以通过换行实现 LU 分解.

定理 5.2.6: 一般三角分解定理

若 A 可逆, 则存在排列矩阵 P 、单位下三角矩阵 L 和上三角矩阵 U 使得

$$PA = LU. \quad (5.18)$$

证明. 考虑列主元方法的 Gauss 消元法, 第 k 步交换 k, i_k 行, 则

$$A^{(k+1)} = L_k I_{ki_k} A^{(k)},$$

即

$$A^{(n)} = L_{n-1} I_{n-1, i_{n-1}} \cdots L_1 I_{1, i_1} A^{(1)}, \iff A^{(1)} = I_{1, i_1} L_1^{-1} \cdots I_{n-1, i_{n-1}} L_{n-1}^{-1} A^{(n)},$$

定义

$$P_k = I_{n-1, i_{n-1}} \cdots I_{ki_k},$$

则 $P_k^\top P_{k+1} = I_{k i_k}$, 进而

$$P_1 A^{(1)} = P_2 L_1^{-1} P_2^\top P_3 L_2^{-1} \cdots P_{n-1} L_{n-1}^{-1} A^{(n)},$$

易得 $P_{k+1} e_k = e_k$, 故

$$L'_k := P_{k+1} L_k^{-1} P_{k+1}^\top = P_{k+1} (I - \ell_k e_k^\top) P_{k+1}^\top = I - P_{k+1} \ell_k e_k^\top,$$

仍然是第 k 列的初等单位上三角矩阵, 令 $P = P_1$, $L = L'_1 \cdots L'_{n-2} L_{n-1}$, $U = A^{(n)}$ 即得. \square

5.2.3 Cholesky 分解

下面再看对称矩阵的三角分解.

定理 5.2.7: Cholesky 分解

若 A 实对称正定, 则存在唯一的对角元素为正的下三角矩阵 L 使得

$$A = LL^\top. \quad (5.19)$$

证明. 采用加边 Cholesky 分解法: 写成分块矩阵的形式

$$L_i = \begin{bmatrix} L_{i-1} & \\ \ell_{i-1}^\top & \ell_{ii} \end{bmatrix}, \quad A_i = \begin{bmatrix} A_{i-1} & a_{i-1} \\ a_{i-1}^\top & a_{ii} \end{bmatrix}$$

满足 $A_i = L_i L_i^\top$, 可得

$$\ell_{i-1} = L_{i-1}^{-1} a_{i-1}, \quad (5.20a)$$

$$\ell_{ii} = \sqrt{a_{ii} - \ell_{i-1}^\top \ell_{i-1}}. \quad (5.20b)$$

从 $\ell_{11} = \sqrt{a_{11}}$ 出发便可迭代得到整个 L . \square

注. 这种算法特别适合稀疏矩阵.

5.2.4 Thomas 方法

考虑线性方程组

$$\begin{cases} b_1 x_1 + c_1 x_2 = d_1, \\ a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \quad i = 2, \dots, n-1 \\ a_n x_{n-1} + b_n x_n = d_n \end{cases}$$

系数矩阵为三对角矩阵:

$$\begin{bmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & a_n & b_n & \end{bmatrix}.$$

定理 5.2.8: Thomas 方法

容易验证有如下三角分解形式:

$$A = LU = \begin{bmatrix} 1 & & & \\ \ell_2 & 1 & & \\ & \ddots & \ddots & \\ & & \ell_n & 1 \end{bmatrix} \begin{bmatrix} u_1 & c_1 & & \\ & u_2 & \ddots & \\ & & \ddots & c_{n-1} \\ & & & u_n \end{bmatrix},$$

可直接乘开得到 $u_1 = b_1$

$$\ell_i = \frac{a_i}{u_{i-1}}, \quad u_i = b_i - \ell_i c_{i-1}. \quad (5.21)$$

5.3 稳定性分析

在用直接法求解 $Ax = b$ 的过程中, 由于舍入误差的存在, 必然会导致结果产生误差. 因而有必要对可能产生的误差作一估计.

例 5.3.1: 数据的微小变化导致解的巨大变化

方程组

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix} \Rightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

对向量 b 数据做微小的修改

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \end{bmatrix} = \begin{bmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{bmatrix} \Rightarrow \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \end{bmatrix} = \begin{bmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{bmatrix}.$$

对系数矩阵 A 数据做微小的修改

$$\begin{bmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{bmatrix} \begin{bmatrix} x''_1 \\ x''_2 \\ x''_3 \\ x''_4 \end{bmatrix} = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix} \Rightarrow \begin{bmatrix} x''_1 \\ x''_2 \\ x''_3 \\ x''_4 \end{bmatrix} = \begin{bmatrix} -81 \\ 137 \\ -34 \\ 22 \end{bmatrix}.$$

可见数据的微小变化会导致解的巨大变化, 这是因为系数矩阵的条件数 $\text{cond}(A) = 32825/11$ 很大.

定义 5.3.1: 条件数

给定诱导的矩阵范数 $\|\cdot\|$, 可逆矩阵 A 的条件数 (condition number) 为

$$\text{cond}(A) \equiv \|A\| \|A^{-1}\|. \quad (5.22)$$

推论. 条件数的性质:

- $\text{cond}(A) \geq 1$;
- $\text{cond}(A^{-1}) = \text{cond}(A)$;
- $\text{cond}(cA) = \text{cond}(A)$;
- 若 U 为正交矩阵, 则 $\text{cond}_2(U) = 1$, 且

$$\text{cond}_2(A) = \text{cond}_2(AU) = \text{cond}_2(UA); \quad (5.23)$$

- 若 λ_1, λ_n 是 A 模最大与最小的特征值, 则

$$\text{cond}(A) \geq \frac{|\lambda_1|}{|\lambda_n|}, \quad (5.24)$$

若 A 对称, 则 $\text{cond}_2(A) = |\lambda_1| / |\lambda_n|$;

- 由范数的等价性, 可知条件数的等价性:

$$\frac{1}{n} \text{cond}_2(A) \leq \text{cond}_1(A) \leq n \text{cond}_2(A), \quad (5.25a)$$

$$\frac{1}{n} \text{cond}_\infty(A) \leq \text{cond}_2(A) \leq n \text{cond}_\infty(A), \quad (5.25b)$$

$$\frac{1}{n^2} \text{cond}_1(A) \leq \text{cond}_\infty(A) \leq n^2 \text{cond}_1(A). \quad (5.25c)$$

定理 5.3.1: 解的扰动定理

给定可逆矩阵 A 和微小扰动 ΔA , 满足

$$\frac{\|\Delta A\|}{\|A\|} < \frac{1}{\text{cond}(A)},$$

则 $(A + \Delta A)$ 也可逆, 考察线性方程组 $Ax = b$ 及其扰动方程组

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

则有

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \|A^{-1}\| \|\Delta A\|} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right). \quad (5.26)$$

证明. 由扰动定理定理 1.4.6 知 $(A + \Delta A)$ 可逆且

$$\|(A + \Delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|}. \quad (5.27)$$

由

$$\begin{aligned} \Delta x &= (A + \Delta A)^{-1}(b + \Delta b) - x \\ &= (A + \Delta A)^{-1}(b + \Delta b - (A + \Delta A)x) \\ &= (A + \Delta A)^{-1}(\Delta b - \Delta Ax), \end{aligned}$$

两边取范数

$$\begin{aligned}\|\Delta x\| &\leq \|(A + \Delta A)^{-1}\| (\|\Delta b\| + \|\Delta A\| \|x\|) \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|} \left(\frac{\|\Delta A\|}{\|A\|} \|A\| \|x\| + \frac{\|\Delta b\|}{\|b\|} \|A\| \|x\| \right).\end{aligned}\quad \square$$

注. 因此条件数可以看成扰动方程组相对误差的放大倍数.

定理 5.3.2: 矩阵相对奇异性的度量

若 A 可逆, 定义所有使得 $(A + \Delta A)$ 不可逆的 ΔA 构成集合 S , 则

$$\min_{\Delta A \in S} \frac{\|\Delta A\|_2}{\|A\|_2} = \frac{1}{\text{cond}_2(A)}, \quad (5.28)$$

证明. 当 $\|A^{-1}\| \|\Delta A\| < 1$ 时, $(A + \Delta A)$ 可逆, 故

$$\min_{\Delta A \in S} \|\Delta A\|_2 \geq \frac{1}{\|A^{-1}\|_2}.$$

由 $\|\cdot\|_2$ 的定义, $\exists x$ 且 $\|x\|_2 = 1$ 使得 $\|A^{-1}x\|_2 = \|A^{-1}\|_2$, 令 $y = A^{-1}x/\|A^{-1}\|_2$, 并取

$$\Delta A = -\frac{xy^\top}{\|A^{-1}\|_2},$$

则 $\|y\|_2 = 1$ 且

$$(A + \Delta A)y = \frac{x}{\|A^{-1}\|_2} - \frac{xy^\top y}{\|A^{-1}\|_2} = 0,$$

故 $(A + \Delta A)$ 不可逆, 又

$$\|\Delta A\|_2 = \max_{\|z\|_2=1} \|\Delta Az\|_2 = \frac{\|x\|_2}{\|A^{-1}\|_2} \max_{\|z\|_2=1} |y^\top z| = \frac{1}{\|A^{-1}\|_2}.\quad \square$$

注. 因此可逆矩阵到最接近的奇异矩阵的相对距离在 2 - 范数意义下就是 2 - 条件数的倒数. 当条件数很大时, 矩阵与奇异矩阵的相对距离很小, 称为病态 (ill conditioned).

定理 5.3.3: 近似解的相对误差

若 x, x' 分别是方程组 $Ax = b$ 的精确解和近似解, x' 的剩余 $r = b - Ax'$, 则

$$\frac{1}{\text{cond}(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|x' - x\|}{\|x\|} \leq \text{cond}(A) \frac{\|r\|}{\|b\|}. \quad (5.29)$$

证明. 由 $A(x' - x) = -r$ 和 $\|A\| \|x\| \geq \|b\|$ 可得

$$\|r\| \leq \|A\| \|x' - x\| = \frac{\text{cond}(A)}{\|A^{-1}\|} \|x' - x\|,$$

两边除 $\|b\|$, 由 $x = A^{-1}b$ 可得

$$\frac{\|r\|}{\|b\|} \leq \text{cond}(A) \frac{\|x' - x\|}{\|A^{-1}\| \|b\|} \leq \text{cond}(A) \frac{\|x' - x\|}{\|x\|};$$

另一方面, 由 $x' - x = -A^{-1}r$ 可得

$$\|x' - x\| = \|A^{-1}r\| \leq \|A^{-1}\| \|r\| = \text{cond}(A) \frac{\|r\|}{\|A\|}.$$

两边同除 $\|x\|$, 由 $Ax = b$ 可得

$$\frac{\|x' - x\|}{\|x\|} \leq \text{cond}(A) \frac{\|r\|}{\|A\| \|x\|} \leq \text{cond}(A) \frac{\|r\|}{\|b\|}.$$

综上, 两边不等式均得证. \square

注. 这说明当方程组病态时, 即使剩余 $\|r\|$ 比较小, 解的相对误差仍可能很大.

例 5.3.2: Hilbert 矩阵

Hilbert 矩阵

$$H_n := \begin{bmatrix} 1 & 1/2 & \cdots & 1/(n+1) \\ 1/2 & 1/3 & \cdots & 1/(n+2) \\ \vdots & \vdots & \ddots & \vdots \\ 1/(n+1) & 1/(n+2) & \cdots & 1/(2n+1) \end{bmatrix}$$

的条件数增长很快:

$$\text{cond}_2(H_n) = \mathcal{O}\left(\frac{(1+\sqrt{2})^{4n}}{\sqrt{n}}\right).$$

定理 5.3.4: 条件数与数值精度

用直接法解方程组 $Ax = b$, A, b 的元素有效位数为 s 而 $\text{cond}(A)$ 的数量级为 t , 则求得 x 分量有效位数约为 $s - t$.

病态方程组的解法 除采用更高精度的运算外, 另一个更有效的方法是对原方程进行预处理:

$$Ax = b, \iff PAQ(Q^{-1}x) = Pb,$$

从而降低系数矩阵的条件数: $\text{cond}(PAQ) \ll \text{cond}(A)$. 一般 P, Q 可选择为三角矩阵或对角矩阵.

例 5.3.3: 预处理例子

方程组

$$\begin{bmatrix} 10 & 10^5 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10^5 \\ 2 \end{bmatrix} \implies \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{9999} \begin{bmatrix} 10000 \\ 9998 \end{bmatrix}.$$

系数矩阵的条件数 $\text{cond}_2(A) = 100010$ 很大, 左乘 $D = \text{diag}(10^{-5}, 1)$ 平衡:

$$\begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

系数矩阵的条件数 $\text{cond}_2(DA) = 940/359$ 得到了有效降低.

第六章 线性方程组的迭代解法

定义 6.0.1: 线性方程组的迭代解法

构造一个向量序列，使其尽快收敛到线性方程组 $Ax = b$ 的解。

6.1 迭代法基本概念

6.1.1 向量序列和矩阵序列的极限

定义 6.1.1: 向量序列的极限

给定赋范空间 $(\mathbb{F}^n, \|\cdot\|)$ 以及一组向量序列 $x^{(1)}, x^{(2)}, \dots$ ，若 $\exists x \in \mathbb{F}^n$ 使得

$$\lim_{i \rightarrow \infty} \|x^{(i)} - x\| = 0, \iff \lim_{i \rightarrow \infty} x^{(i)} = x. \quad (6.1)$$

则称 $x^{(i)}$ 收敛于 x 。

注. 由于有限维线性空间中所有范数都是等价的，向量序列的收敛性不依赖于范数的选择。

定理 6.1.1: 向量序列的收敛性判定

向量序列收敛等价于其各分量序列收敛：

$$\lim_{i \rightarrow \infty} x^{(i)} = x, \iff \lim_{i \rightarrow \infty} x_j^{(i)} = x_j, \quad \forall j = 1, \dots, n. \quad (6.2)$$

证明. 选取无穷范数：

$$\lim_{i \rightarrow \infty} \|x^{(i)} - x\|_{\infty} = 0, \iff \lim_{i \rightarrow \infty} \max_j |x_j^{(i)} - x_j| = 0, \iff \lim_{i \rightarrow \infty} |x_j^{(i)} - x_j| = 0, \quad \forall j. \quad \square$$

注. 类似地，我们可以定义矩阵序列的收敛性。

定理 6.1.2: 收敛到零矩阵

$$\lim_{k \rightarrow \infty} A^{(k)} = O \iff \forall x \in \mathbb{F}^n \text{ 均有 } \lim_{k \rightarrow \infty} A^{(k)}x = 0.$$

证明. 必要性：对于任一矩阵的从属范数，

$$\|A^{(k)}x\| \leq \|A^{(k)}\| \|x\| \rightarrow 0;$$

充分性：其矩阵各分量

$$|a_{ij}^{(k)}| = |e_i^T A^{(k)} e_j| \leq \|A^{(k)} e_j\|_\infty \rightarrow 0.$$

□

定理 6.1.3: 矩阵幂序列的收敛性

给定 $B \in \mathbb{F}^{n \times n}$ ，下面三个命题等价：

1. $\lim_{k \rightarrow \infty} B^k = O$;
2. $\rho(B) < 1$;
3. 存在一个矩阵范数 $\|\cdot\|$ 使得 $\|B\| < 1$.

证明. (1) \Rightarrow (2): 采取反证法, 若 λ 为 B 的特征值且 $|\lambda| \geq 1$ 其对应特征向量为 x , 则

$$\|B^k x\| = |\lambda|^k \|x\| \not\rightarrow 0;$$

(2) \Rightarrow (3): 由 $\rho(B) < 1$, 给定 $\epsilon = \frac{1 - \rho(B)}{2} > 0$, 由定理 1.4.4, 存在一个矩阵范数使得 $\|B\| \leq \rho(B) + \epsilon < 1$;

(3) \Rightarrow (1): 由 $\|B^k\| \leq \|B\|^k \rightarrow 0$ 即得.

□

定理 6.1.4: 矩阵幂的矩阵范数与谱半径

给定 $B \in \mathbb{F}^{n \times n}$ 和任一矩阵范数 $\|\cdot\|$, 均有

$$\lim_{k \rightarrow \infty} \|B^k\|^{1/k} = \rho(B). \quad (6.3)$$

证明. 采用夹逼定理. 一方面, 由定理 1.4.3,

$$\rho(B) = \rho(B^k)^{1/k} \leq \|B^k\|^{1/k};$$

另一方面, $\forall \epsilon > 0$, 记

$$B_\epsilon = \frac{B}{\rho(B) + \epsilon},$$

则 $\rho(B_\epsilon) < 1$, 由定理 6.1.3, $\lim_{k \rightarrow \infty} B_\epsilon^k = O$, 即 $\exists N = N(\epsilon) > 0$ 使得 $\forall k > N$,

$$\|B_\epsilon^k\| = \frac{\|B^k\|}{(\rho(B) + \epsilon)^k} < 1, \iff \|B^k\|^{1/k} < \rho(B) + \epsilon, \iff \lim_{k \rightarrow \infty} \|B^k\|^{1/k} \leq \rho(B).$$

综上, 极限相等.

□

6.1.2 迭代公式的构造

定理 6.1.5: 线性定常迭代

若线性方程 $Ax = b$ 与方程 $x = Bx + f$ 同解, 则可以构造一个定常 (stationary) 的迭代公式:

$$x^{(k+1)} = Bx^{(k)} + f, \quad (6.4)$$

其中 B 称为迭代矩阵. 若 $x^{(k)}$ 收敛, 则收敛值 x^* 就是方程的解.

定理 6.1.6: 迭代法的收敛性分析

迭代法 $x^{(k+1)} = Bx^{(k)} + f$ 收敛与以下命题等价:

1. $\lim_{k \rightarrow \infty} B^k = O$;
2. $\rho(B) < 1$;
3. 存在一种矩阵范数 $\|\cdot\|$ 使得 $\|B\| < 1$.

证明. 记误差向量 $e^{(k)} := x^{(k)} - x^*$, 则 $e^{(k+1)} = Be^{(k)}$. 迭代法收敛意味着 $\forall e^{(0)} \in \mathbb{F}^n$ 均有

$$\lim_{k \rightarrow \infty} e^{(k)} = \lim_{k \rightarrow \infty} B^k e^{(0)} = 0,$$

由定理 6.1.2, $\lim_{k \rightarrow \infty} B^k = O$, 由定理 6.1.3 知三个命题等价. □

定理 6.1.7: 迭代法的误差分析

若迭代法收敛, 第 k 次迭代的误差满足

$$\|e^{(k)}\| \leq \frac{\|B\|}{1 - \|B\|} \|x^{(k)} - x^{(k-1)}\| \leq \frac{\|B\|^k}{1 - \|B\|} \|x^{(1)} - x^{(0)}\|. \quad (6.5)$$

证明. 由 $e^{(k)} = Be^{(k-1)} = B(e^{(k-1)} - e^{(k)}) + Be^{(k)}$, 故

$$\|e^{(k)}\| \leq \|B(e^{(k-1)} - e^{(k)})\| + \|Be^{(k)}\| \leq \|B\| (\|e^{(k-1)} - e^{(k)}\| + \|e^{(k)}\|),$$

移项即得第一个不等号; 又 $e^{(k)} - e^{(k-1)} = B(e^{(k-1)} - e^{(k-2)})$ 可得

$$\|e^{(k)} - e^{(k-1)}\| \leq \|B\| \|e^{(k-1)} - e^{(k-2)}\|,$$

反复迭代即得第二个不等号. □

定理 6.1.8: 迭代法的收敛速度

若给定 $\epsilon > 0$, 希望

$$\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \leq \epsilon,$$

则要求迭代次数

$$k \geq \frac{-\ln \epsilon}{-\ln \|B^k\|^{1/k}}. \quad (6.6)$$

证明. 按矩阵从属范数的定义

$$\|B^k\| = \max_{e^{(0)} \neq 0} \frac{\|B^k e^{(0)}\|}{\|e^{(0)}\|} = \max_{e^{(0)} \neq 0} \frac{\|e^{(k)}\|}{\|e^{(0)}\|}.$$

因此只需 $\|B^k\| \leq \epsilon$ 即可. □

注. 式 (6.6) 并不好处理, 可以用近似:

$$k \approx \frac{-\ln \epsilon}{-\ln \rho(B)}.$$

定义 6.1.2: 平均收敛率

迭代法的平均收敛率为

$$R_k(B) := -\ln \|B^k\|^{1/k}. \quad (6.7)$$

其渐进收敛率 $R(B) := R_\infty(B) = -\ln \rho(B)$.

6.2 (单步) 定常线性迭代**6.2.1 Richardson 迭代法****例 6.2.1: Richardson 迭代法**

最简单的情形 $x = (I - A)x + b$, 称为 Richardson 迭代:

$$x^{(k+1)} = (I - A)x^{(k)} + b, \quad (6.8)$$

收敛条件为 $\rho(I - A) < 1$. 定义残差 $r^{(k)} := b - Ax^{(k)}$, 则有递推关系:

$$r^{(k+1)} = (I - A)r^{(k)},$$

$$x^{(k+1)} = x^{(k)} + r^{(k)},$$

继而

$$x^{(k)} = x^{(0)} + r^{(0)} + \cdots + r^{(k-1)} = x_0 + \sum_{i=0}^{k-1} (I - A)^i r^{(0)}.$$

定义 6.2.1: Krylov 子空间

给定非零向量 v 和非零方阵 A , 定义

$$\mathcal{K}_n(v) := \text{span}(v, Av, \dots, A^{n-1}v),$$

称为 Krylov 子空间.

推论. 显然

$$\mathcal{K}_1(v) \subset \mathcal{K}_2(v) \subset \cdots \subset \mathcal{K}_n(v) \subset \cdots$$

最大维度 $\sup_{n \rightarrow \infty} \dim \mathcal{K}_n(v)$ 称为 v 关于 A 的度.

定义 6.2.2: 预处理矩阵

若 M 可逆且 M^{-1} 是 A^{-1} 的一个较好的近似, 对 $M^{-1}Ax = M^{-1}b$ 使用 Richardson 迭代:

$$x^{(k+1)} = (I - M^{-1}A)x^{(k)} + M^{-1}b = x^{(k)} + M^{-1}r^{(k)}, \quad (6.9)$$

矩阵 M 称为预处理矩阵 (preconditioning matrix).

注. 预处理矩阵的选择依据: 方便计算、好近似.

6.2.2 Jacobi 迭代法和 Gauss-Seidel 迭代法

将 A 拆解为 $A = D + L + U$ 的形式, 其中 D 是对角部分, L 是严格下三角部分, U 是严格上三角部分.¹

例 6.2.2: Jacobi 迭代法

预处理矩阵 $M = D$, 迭代矩阵 $B_J = -D^{-1}(L + U)$, 迭代的分量形式为:

$$x_{i+1}^{(k)} = \frac{1}{a_{kk}} \left(b_k - \sum_{j \neq k} a_{kj} x_i^{(j)} \right). \quad (6.10)$$

例 6.2.3: Gauss-Seidel 迭代法

预处理矩阵 $M = D + L$, 迭代矩阵 $B_G = -(D + L)^{-1}U$, 迭代的分量形式为:

$$x_{i+1}^{(k)} = \frac{1}{a_{kk}} \left(b_k - \sum_{j=1}^{k-1} a_{kj} x_{i+1}^{(j)} - \sum_{j=k+1}^n a_{kj} x_i^{(j)} \right). \quad (6.11)$$

注. Jacobi 法和 G-S 法收敛 $\iff \rho(B_J) < 1$ 和 $\rho(B_G) < 1$. 另外, 有时容易验证矩阵 A 满足收敛的充分条件.

定理 6.2.1: Jacobi 法和 G-S 法收敛的充分条件: 对角占优

若 A 严格对角占优或不可约弱对角占优, 则 Jacobi 法和 G-S 法收敛.

证明. 对于 Jacobi 法, 若 B_J 有特征值 λ 且 $|\lambda| \geq 1$, 则

$$\det(\lambda I - B_J) = 0, \implies \det(D + \lambda^{-1}(L + U)) = 0,$$

若 A 严格对角占优 (或不可约弱对角占优), 则 $D + \lambda^{-1}(L + U)$ 也为严格对角占优 (或不可约弱对角占优), 矛盾! 故 $\rho(B_J) < 1$, Jacobi 法收敛.

对于 G-S 法, 设 B_G 有特征值 λ 且 $|\lambda| \geq 1$, 则

$$\det(\lambda I - B_G) = 0, \implies \det(D + L + \lambda^{-1}U) = 0,$$

若 A 严格对角占优 (或不可约弱对角占优), 则 $D + L + \lambda^{-1}U$ 也为严格对角占优 (或不可约弱对角占优), 矛盾! 故 $\rho(B_G) < 1$, G-S 法收敛. \square

定理 6.2.2: 对称矩阵的 Jacobi 法和 G-S 法收敛的充要条件

若 A 对称且对角元均为正, 则 Jacobi 法和 G-S 法收敛 $\iff D \pm (L + U)$ 均正定.

¹本节对 L, U 的定义与教材上的定义相差一个负号.

6.2.3 超松弛迭代法

例 6.2.4: 逐次超松弛迭代

预处理矩阵 $M = D/\omega + L$, 迭代矩阵 $L_\omega = (D + \omega L)^{-1}[(1 - \omega)D - \omega U]$,

$$x_{i+1}^{(k)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{kk}} \left(b_k - \sum_{j=1}^{k-1} a_{kj}x_{i+1}^{(j)} - \sum_{j=k+1}^n a_{kj}x_i^{(j)} \right). \quad (6.12)$$

称为逐次超松弛迭代法 (successive overrelaxation, SOR), 其中 ω 称为松弛因子.

定理 6.2.3: $\rho(L_\omega)$ 与 ω 的关系

若 A 可逆且对角元非 0, 则

$$\rho(L_\omega) \geq |1 - \omega|. \quad (6.13)$$

证明. 设 L_ω 的特征值是 $\lambda_1, \dots, \lambda_n$, 由于 $L_\omega = (D + \omega L)^{-1}[(1 - \omega)D - \omega U]$ 可以写成下三角矩阵和上三角矩阵的乘积:

$$\lambda_1 \cdots \lambda_n = \det(L_\omega) = \det(D^{-1}) \det((1 - \omega)D) = (1 - \omega)^n,$$

故

$$\rho(L_\omega) = \max_i |\lambda_i| \geq |\lambda_1 \cdots \lambda_n|^{1/n} = |1 - \omega|. \quad \square$$

推论. SOR 收敛的必要条件是 $0 < \omega < 2$.

定理 6.2.4: 对称正定矩阵的 SOR 收敛的充要条件

若 A 对称正定, 则 $0 < \omega < 2 \iff$ SOR 收敛.

证明. 只需证明充分性. 设 λ 和 x 是 L_ω 的任意一个特征值和对应的特征向量, 则

$$[(1 - \omega)D - \omega U]x = \lambda(D + \omega L)x,$$

上式两边与 x 内积得到

$$\lambda = \frac{(1 - \omega) \langle Dx, x \rangle - \omega \langle Ux, x \rangle}{\langle Dx, x \rangle + \omega \langle Lx, x \rangle},$$

由 A 正定, D 也正定, 记 $\langle Dx, x \rangle =: p > 0$, $\langle Lx, x \rangle = a + ib$, 由 A 对称, $L^\top = U$,

$$\langle Ux, x \rangle = \langle L^\top x, x \rangle = \langle x, Lx \rangle = \overline{\langle Lx, x \rangle} = a - ib,$$

可将 λ 用 ω, p, a, b 表示出来, 继而

$$|\lambda|^2 - 1 = \frac{[(1 - \omega)p - \omega a]^2 + \omega^2 b^2}{(p + \omega a)^2 + \omega^2 b^2} - 1 = -\frac{p\omega(2 - \omega)(p + 2a)}{(p + \omega a)^2 + \omega^2 b^2}$$

由 A 正定,

$$\langle Ax, x \rangle = \langle Dx, x \rangle + \langle Lx, x \rangle + \langle Ux, x \rangle = p + 2a > 0,$$

又 $2 - \omega > 0$, 故 $|\lambda|^2 < 1$, 从而 $\rho(L_\omega) < 1$, SOR 收敛. \square

最佳松弛因子 希望选取最优的 ω 使得迭代矩阵谱半径最小:

$$\omega^* = \arg \min_{\omega} \rho(L_{\omega}). \quad (6.14)$$

定理 6.2.5: 最佳松弛因子

若 A 为对称正定的 (分块) 三对角矩阵, 则 $\rho(B_G) = \rho(B_J)^2 < 1$ 且 SOR 的最佳松弛因子为

$$\omega^* = \frac{2}{1 + \sqrt{1 - \rho(B_G)}}. \quad (6.15)$$

且 $\rho(L_{\omega^*}) = \omega^* - 1$.

注. 对于对称正定三对角矩阵, G-S 法收敛速度是 Jacobi 法的两倍, SOR 的收敛速度最快.

例 6.2.5: Poisson 方程的五点差分格式

讨论二元函数 $u = u(x, y)$ 满足

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y),$$

边界条件 $u|_{\partial\Omega} = 0$. 采用差分法, 在均匀网格点 (x_i, y_i) 附近作 Taylor 展开:

$$u_{i\pm 1,j} = u_{ij} \pm h \left. \frac{\partial u}{\partial x} \right|_{ij} + \frac{h^2}{2} \left. \frac{\partial^2 u}{\partial x^2} \right|_{ij} \pm \frac{h^3}{6} \left. \frac{\partial^3 u}{\partial x^3} \right|_{ij} + \mathcal{O}(h^4).$$

可得

$$\left. \frac{\partial^2 u}{\partial x^2} \right|_{ij} = \frac{u_{i+1,j} + u_{i-1,j} - 2u_{ij}}{h^2} + \mathcal{O}(h^2)$$

同理可得 $\partial^2 u / \partial^2 y$. 代入原 Poisson 方程得到

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij} = h^2 f_{ij}. \quad (6.16)$$

若网格是 $m \times n$ 的, 则可以写成线性方程组 $Ax = b$ 的形式, 其中

$$x = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}, \quad u_i = \begin{bmatrix} u_{1i} \\ \vdots \\ u_{mi} \end{bmatrix}; \quad b = -h^2 \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}, \quad f_i = \begin{bmatrix} f_{1i} \\ \vdots \\ f_{mi} \end{bmatrix};$$

$$A = \begin{bmatrix} A_1 & -I & & \\ -I & A_2 & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & A_n \end{bmatrix}, \quad A_i = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}.$$

其中 A 是 mn 阶方阵, A_i 是 m 阶方阵. 故 A 是分块三对角正定矩阵, B_J 的 mn 个特征值为

$$\lambda_{ij} = \frac{\cos(i\pi h) + \cos(j\pi h)}{2}, \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

谱半径 $\rho(B_J) = \cos(\pi h)$, 则 $\rho(B_G) = \cos^2(\pi h)$

$$\omega^* = \frac{2}{1 + \sin(\pi h)},$$

因此, Jacobi 法和 SOR 的收敛速度分别为

$$R(B_J) = -\ln(\cos(\pi h)) = \frac{\pi^2}{2}h^2 + \frac{\pi^4}{12}h^4 + \mathcal{O}(h^6),$$

$$R(L_{\omega^*}) = -2\ln(\cos(\pi h)) + 2\ln(1 + \sin(\pi h)) = 2\pi h + \frac{\pi^3}{3}h^3 + \mathcal{O}(h^5).$$

可见 Jacobi 法和 SOR 的收敛速度差了一个数量级.

6.3 非线性迭代方法

定义 6.3.1: 算子

Hilbert 空间 V 中的可逆对称算子 $A: V \rightarrow V$ 满足: $\forall x, y \in V$,

- $\langle Ax, y \rangle = \langle x, Ay \rangle$;
- $\langle Ax, x \rangle \geq 0$ 且 $\langle Ax, x \rangle = 0 \iff x = 0$.

可定义范数 $\|x\|_A := \sqrt{\langle Ax, x \rangle}$.

例 6.3.1

\mathbb{R}^n 中的内积和 n 阶对称正定矩阵 M 一一对应:

$$\langle x, y \rangle_M = \langle Mx, y \rangle.$$

定理 6.3.1: Ritz 变分问题

考虑算子方程

$$Ax = b, \tag{6.17}$$

考虑 $\varphi: V \rightarrow \mathbb{R}$

$$\varphi(x) := \frac{1}{2} \|x\|_A^2 - \langle b, x \rangle. \tag{6.18}$$

若算子 A 对称正定, 则

$$Ax^* = b \iff x^* = \arg \min_{x \in V} \varphi(x). \tag{6.19}$$

证明. 由 $b = Ax^*$, $\varphi(x^*) = -\|x^*\|_A^2/2$. $\forall x \in V$

$$\varphi(x) - \varphi(x^*) = \frac{1}{2} (\|x\|_A^2 - 2\langle Ax^*, x \rangle + \|x^*\|_A^2) = \frac{1}{2} \|x - x^*\|_A^2.$$

故 $\varphi(x) \geq \varphi(x^*)$ 且 $\varphi(x) = \varphi(x^*) \iff x = x^*$. □

因此, 迭代法的选择策略为: 给定 $x^{(k)}$, 确定 $x^{(k+1)}$ 使得

$$\varphi(x^{(k+1)}) < \varphi(x^{(k)}).$$

6.3.1 子空间方法

定理 6.3.2: 子空间搜索

设 $x + \Phi$ 是一个仿射 (affine) 平面, 则

$$y^* = \arg \min_{y \in \Phi} \varphi(x + y) \iff b - A(x + y^*) \perp \Phi.$$

证明. 由于

$$\varphi(x + y) - \varphi(x^*) = \frac{1}{2} \|x + y - x^*\|_A^2$$

故对于 y^* 应有

$$\forall y \in \Phi, \langle x^* - (x + y^*), y \rangle_A = 0 \iff x^* - (x + y^*) \perp_A \Phi. \quad \square$$

定理 6.3.3: 一维极小搜索方法

给定 $x^{(k)}$ 和搜索方向 $p^{(k)} \neq 0$, 考虑一维极小值问题

$$\alpha_k = \arg \min_{\alpha \in \mathbb{R}} \varphi(x^{(k)} + \alpha p^{(k)}) = \frac{\langle r^{(k)}, p^{(k)} \rangle}{\|p^{(k)}\|_A^2}. \quad (6.20)$$

并使

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}, \quad (6.21)$$

若 $\alpha_k \neq 0$, 则 $\varphi(x^{(k+1)}) < \varphi(x^{(k)})$, 且 $r^{(k+1)} \perp p^{(k)}$.

证明. 残差 $r^{(k)} = b - Ax^{(k)}$, 则

$$\frac{d}{d\alpha} \varphi(x^{(k)} + \alpha p^{(k)}) = -\langle r^{(k)}, p^{(k)} \rangle + \alpha \|p^{(k)}\|_A^2 = 0. \quad \square$$

6.3.2 最速下降法

定理 6.3.4: 最速下降法

搜索方向选择 φ 在 $x^{(k)}$ 处下降最快的方向: $p^{(k)} = r^{(k)}$, 称为最速下降法 (steepest descent method):

$$r^{(k)} = b - Ax^{(k)}, \quad (6.22a)$$

$$\alpha_k = \frac{\|r^{(k)}\|^2}{\|r^{(k)}\|_A^2}, \quad (6.22b)$$

$$x^{(k+1)} = x^{(k)} + \alpha_k r^{(k)}. \quad (6.22c)$$

则 $\varphi(x^{(k)})$ 单调下降且有界 $\varphi(x^*)$, 极限 $\lim_{k \rightarrow \infty} x^{(k)} = x^*$.

推论. 相邻两次搜索方向是正交的: $\langle r^{(k+1)}, r^{(k)} \rangle = 0$, 且

$$r^{(k+1)} = (I - \alpha_k A)r^{(k)}.$$

定理 6.3.5: 最速下降法的误差收敛性

记误差 $e^{(k)} := x^{(k)} - x^*$, 则误差的 A -范数满足

$$\|e^{(k)}\|_A \leq \left(\frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1} \right)^k \|e^{(0)}\|_A. \quad (6.23)$$

引理. $\forall x \neq 0$,

$$\frac{\|x\|_A^2 \|x\|_{A^{-1}}^2}{\|x\|^4} \leq \frac{(\text{cond}_2(A) + 1)^2}{4 \text{cond}_2(A)}. \quad (6.24)$$

证明. 设 $\lambda_{\max}, \lambda_{\min}$ 分别是对称正定算子 A 的最大最小特征值. 则 $\forall \lambda_i$ 为 A 的特征值, 满足

$$\frac{1}{\lambda_i} \leq \frac{1}{\lambda_{\min}} + \frac{1}{\lambda_{\max}} - \frac{\lambda_i}{\lambda_{\min} \lambda_{\max}}.$$

则 $\forall x$ 且 $\|x\| = 1$,

$$\begin{aligned} \frac{\|x\|_A^2 \|x\|_{A^{-1}}^2}{\|x\|^4} &= \frac{\langle Ax, x \rangle \langle A^{-1}x, x \rangle}{\langle x, x \rangle^2} = \sum_i \lambda_i x_i^2 \sum_i \frac{1}{\lambda_i} x_i^2 \\ &\leq \sum_i \lambda_i x_i^2 \left(\frac{1}{\lambda_{\min}} + \frac{1}{\lambda_{\max}} - \frac{1}{\lambda_{\min} \lambda_{\max}} \sum_i \lambda_i x_i^2 \right) \leq \frac{(\lambda_{\max} + \lambda_{\min})^2}{4 \lambda_{\max} \lambda_{\min}}. \end{aligned}$$

由 A 对称, $\text{cond}_2(A) = |\lambda_{\max}| / |\lambda_{\min}|$, 由 A 正定, $\lambda_{\min} > 0$. □

证明. 由 $Ae^{(k)} = -r^{(k)}$, 有 $\langle e^{(k)}, r^{(k)} \rangle_A = -\|r^{(k)}\|^2$, $\|e^{(k)}\|_A = \|r^{(k)}\|_{A^{-1}}$, 故

$$\begin{aligned} \|e^{(k+1)}\|_A^2 &= \|e^{(k)} + \alpha_k r^{(k)}\|_A^2 = \|e^{(k)}\|_A^2 + \alpha_k^2 \|r^{(k)}\|_A^2 + 2\alpha_k \langle e^{(k)}, r^{(k)} \rangle_A \\ &= \|e^{(k)}\|_A^2 + \alpha_k^2 \|r^{(k)}\|_A^2 - 2\alpha_k \|r^{(k)}\|^2 = \|e^{(k)}\|_A^2 - \frac{\|r^{(k)}\|^4}{\|r^{(k)}\|_A^2} \\ &= \|e^{(k)}\|_A^2 \left(1 - \frac{\|r^{(k)}\|^4}{\|r^{(k)}\|_A^2 \|r^{(k)}\|_{A^{-1}}^2} \right) \leq \|e^{(k)}\|_A^2 \left(1 - \frac{4 \text{cond}_2(A)}{(\text{cond}_2(A) + 1)^2} \right) \\ &= \|e^{(k)}\|_A^2 \left(\frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1} \right)^2. \end{aligned}$$

迭代即可得到. □

注. 虽然最速下降法总是收敛的, 但当 A 病态时, $\text{cond}_2(A) \gg 1$, 收敛速度非常慢; 当 $\|r^{(k)}\|$ 很小时, 计算不稳定.

6.3.3 共轭梯度法

仍然考虑对称正定矩阵 A 和一维极小搜索, 但不再沿着正交的 $r^{(0)}, r^{(1)}, \dots$, 寻找另一组 $p^{(0)}, p^{(1)}, \dots$ 对于第 k 次搜索:

$$\alpha_k = \arg \min_{\alpha} \varphi(x^{(k)} + \alpha p^{(k)}), \implies x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)},$$

由于

$$x^{(k+1)} = x^{(0)} + \alpha_0 p^{(0)} + \dots + \alpha_k p^{(k)}.$$

我们希望 $p^{(k)}$ 的选取能使得确定 α_k 后, $x^{(k+1)}$ 是 $x^{(0)} + \text{span}(p^{(0)}, \dots, p^{(k)})$ 中最优的, 即一维极小搜索的结果与 $(k+1)$ 维搜索结果相同:

$$(\alpha_0, \dots, \alpha_k) = \arg \min_{(\alpha_0, \dots, \alpha_k)} \varphi(x^{(0)} + \alpha_0 p^{(0)} + \dots + \alpha_k p^{(k)}).$$

定义 6.3.2: A - 正交向量组

若向量组 p_1, \dots, p_n, \dots 满足 $\forall i \neq j$ 均有

$$\langle p_i, p_j \rangle_A = 0, \iff p_i \perp_A p_j. \quad (6.25)$$

则称为 A - 正交向量组, 也称 A - 共轭向量组.

推论. 给定一个线性无关的向量组, 可按照 Gram-Schmidt 正交化得到对应的 A - 正交向量组.

定理 6.3.6: 多维子空间搜索问题

记 $y \in \text{span}(p^{(0)}, \dots, p^{(k-1)})$, $x = x^{(0)} + y + \alpha p^{(k)}$,

$$\varphi(x) = \varphi(x^{(0)} + y) + \alpha \langle y, p^{(k)} \rangle_A - \alpha \langle r^{(0)}, p^{(k)} \rangle + \frac{\alpha^2}{2} \|p^{(k)}\|_A^2.$$

上式出现了交叉项 $\langle y, p^{(k)} \rangle_A$, 为了简化问题, 自然可考虑使交叉项 $\langle y, p^{(k)} \rangle_A = 0$, 这要求 $p^{(k)}$ 与 $\text{span}(p^{(0)}, \dots, p^{(k-1)})$ 是 A - 正交的.

由此便分解成了两个极小问题:

$$\min_x \varphi(x) = \min_y \varphi(x^{(0)} + y) + \min_{\alpha} \left[-\alpha \langle r^{(0)}, p^{(k)} \rangle + \frac{\alpha^2}{2} \|p^{(k)}\|_A^2 \right]$$

前者满足 $x^{(0)} + y^* = x^{(k)}$, 后者得到

$$\alpha_k = \frac{\langle r^{(0)}, p^{(k)} \rangle}{\|p^{(k)}\|_A^2} = \frac{\langle r^{(k)}, p^{(k)} \rangle}{\|p^{(k)}\|_A^2}. \quad (6.26)$$

由于 $\langle r^{(0)}, p^{(k)} \rangle = \langle r^{(k)}, p^{(k)} \rangle$, 故形式其实与 (6.20) 相同. 并且有

$$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)}. \quad (6.27)$$

推论. 由子空间搜索定理 6.3.2, $r^{(k)} \perp \text{span}(p^{(0)}, \dots, p^{(k-1)})$.

注. 与 $p^{(0)}, \dots, p^{(k-1)}$ A - 正交的 $p^{(k)}$ 取法并不唯一.

共轭梯度法 当 $r^{(k)} \neq 0$ 时, $p^{(0)}, \dots, p^{(k-1)}, r^{(k)}$ 线性无关, 可通过 Gram-Schmidt 正交化得到 $p^{(k)}$:

$$p^{(k)} = r^{(k)} - \sum_{i=0}^{k-1} \frac{\langle r^{(k)}, p^{(i)} \rangle_A}{\|p^{(i)}\|_A^2} p^{(i)}, \quad (*)$$

定理 6.3.7

$$r^{(k)} \perp_A \text{span}(p^{(0)}, \dots, p^{(k-2)}).$$

证明. $r^{(k)} \perp \text{span}(p^{(0)}, \dots, p^{(k-1)}) = \text{span}(r^{(0)}, \dots, r^{(k-1)})$, 由式 (6.27)

$$Ap^{(i)} = \frac{r^{(i+1)} - r^{(i)}}{\alpha_i} \perp r^{(k)}, \quad i = 0, \dots, k-2. \quad \square$$

推论. $p^{(k)}$ 为 $r^{(k)}$ 和 $p^{(k-1)}$ 的线性组合, 式 (*) 变为

$$p^{(k)} = r^{(k)} - \frac{\langle r^{(k)}, p^{(k-1)} \rangle_A}{\|p^{(k-1)}\|_A^2} p^{(k-1)} =: r^{(k)} + \beta_{k-1} p^{(k-1)}.$$

可以简化系数 β_k 的计算:

$$\beta_k = -\frac{\langle r^{(k+1)}, p^{(k)} \rangle_A}{\|p^{(k)}\|_A^2} = \frac{\|r^{(k+1)}\|^2}{\alpha_k \|p^{(k)}\|_A^2} = \frac{\|r^{(k+1)}\|^2}{\|r^{(k)}\|^2}.$$

定理 6.3.8: 共轭梯度法

共轭梯度法 (conjugate gradient method): $p^{(0)} = r^{(0)} = b - Ax^{(0)}$,

$$\alpha_k = \frac{\|r^{(k)}\|^2}{\|p^{(k)}\|_A^2}, \quad (6.28a)$$

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}, \quad (6.28b)$$

$$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)}; \quad (6.28c)$$

$$\beta_k = \frac{\|r^{(k+1)}\|^2}{\|r^{(k)}\|^2}, \quad (6.28d)$$

$$p^{(k+1)} = r^{(k+1)} + \beta_k p^{(k)}. \quad (6.28e)$$

若 $r^{(k)} = 0$, 则 $x^{(k)} = x^*$, 计算终止.

推论. 由于 $r^{(0)}, \dots, r^{(k)}$ 两两正交, 理论上最多 $n = \dim(V)$ 步可以得到精确解.

定理 6.3.9: 梯度下降法的误差收敛性

误差 $e^{(k)} := x^{(k)} - x^*$ 的 A -范数满足

$$\|e^{(k)}\|_A \leq 2 \left(\frac{\sqrt{\text{cond}_2(A)} - 1}{\sqrt{\text{cond}_2(A)} + 1} \right)^k \|e^{(0)}\|_A. \quad (6.29)$$

其中 $\lambda_{\max}, \lambda_{\min}$ 分别是对称正定算子 A 的最大最小特征值.

证明. 记 $\alpha = \lambda_{\min}/\lambda_{\max}$, 取

$$p_k^*(x) = T_k \left(\frac{2x - (1 + \alpha)}{1 - \alpha} \right) / T_k \left(-\frac{1 + \alpha}{1 - \alpha} \right),$$

其中 T_k 是 Chebyshev 多项式, 由于 $|t| \geq 1$ 时,

$$T_k(t) = \frac{1}{2} \left[\left(t + \sqrt{t^2 - 1} \right)^k + \left(t - \sqrt{t^2 - 1} \right)^k \right] \approx \frac{1}{2} \left(|t| + \sqrt{t^2 - 1} \right)^k.$$

则 $\forall x \in (\alpha, 1)$,

$$\|p_k^*\|_\infty = 1 / T_k \left(-\frac{1+\alpha}{1-\alpha} \right) \leq 2 \left(\frac{1+\alpha}{1-\alpha} - \sqrt{\left(\frac{1+\alpha}{1-\alpha} \right)^2 - 1} \right)^k = 2 \left(\frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}} \right)^k. \quad \square$$

6.3.4 共轭梯度法的预处理

从误差估计 (6.29) 可以看出, 共轭梯度法的收敛速度强烈依赖于 A 的条件数, 当 A 病态时收敛得很慢. 为了改善收敛速度, 可以先进行预处理 (preconditioning), 降低矩阵的条件数.

定理 6.3.10: 预处理的共轭梯度算法

给定预处理矩阵 M 且对称正定,

$$Ax = b \iff M^{-1}Ax = M^{-1}b,$$

对预处理后的方程应用共轭梯度算法: $p^{(0)} = M^{-1}r^{(0)}$,

$$\alpha_k = \frac{\|r^{(k)}\|_{M^{-1}}^2}{\|p^{(k)}\|_A^2}, \quad (6.30a)$$

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}, \quad (6.30b)$$

$$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)}; \quad (6.30c)$$

$$\beta_k = \frac{\|r^{(k+1)}\|_{M^{-1}}^2}{\|r^{(k)}\|_{M^{-1}}^2}, \quad (6.30d)$$

$$p^{(k+1)} = M^{-1}r^{(k+1)} + \beta_k p^{(k)}. \quad (6.30e)$$

当 $r^{(k)} = 0$ 时计算终止.

推论. 其误差仍满足式 (6.29), 但条件数改善为 $\text{cond}_2(M^{-1}A)$.

注. 预处理矩阵的选择依据: 方便计算、好近似.

- 代数方法: 分解 $A = M + N$, 其中 $M = LL^\top$ 对称正定、 L 尽可能稀疏, 而 N 尽可能小, 称为 A 的不完全 Cholesky 分解.

比如取为对角部分 $M = D$, 在 A 的对角元素相差较大时收敛速度会大大提高.

- 分析方法: 谱等价, $\forall x$

$$c_1 \|x\|_A^2 \leq \|x\|_M^2 \leq c_2 \|x\|_A^2,$$

则 $\rho(M^{-1}A) \in [c_2^{-1}, c_1^{-1}]$.

第七章 非线性方程和方程组的数值解法

定义 7.0.1: 非线性方程 (组)

给定 (非线性) 光滑 (smooth) 映射 $f: \Omega \rightarrow \mathbb{F}^n$, 其中 $\Omega \subset \mathbb{F}^n$, 寻找 $x \in \Omega$ 满足方程 $f(x) = 0$, 称 x 是方程的根 (root).

例 7.0.1: 非线性方程的例子

- 代数方程: $p \in \mathcal{P}_n \subset \mathcal{C}^\infty(\mathbb{C})$

$$p(x) = a_0 + a_1x + \cdots + a_nx^n = 0,$$

- 超越方程: 如 $3x^2 - e^x = 0$.

7.1 非线性方程的不动点迭代法

迭代法是求解非线性代数方程的重要方法. 其技术路线是将 $f(x) = 0$ 的根转化为求解 $\varphi(x, \dots, x)$ 的不动点 (fixed point). 再利用不动点迭代:

$$x_{i+1} = \varphi(x_i, x_{i-1}, \dots, x_{i-m}), \quad (7.1)$$

特别地, $x_{i+1} = \varphi(x_i)$ 称为单步方法.

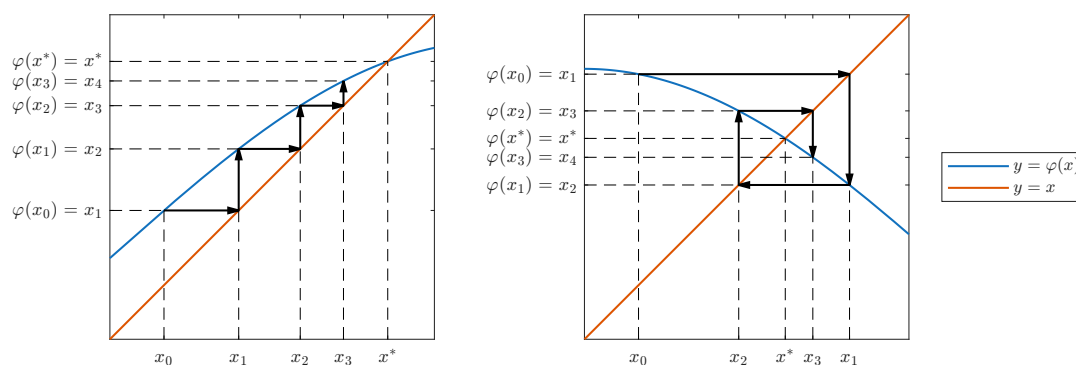


图 7.1: 不动点迭代法示意图

7.1.1 迭代法的收敛性

定理 7.1.1: 压缩映像原理

若连续函数 $\varphi \in \mathcal{C}[a, b]$ 满足:

- 像集 $\varphi([a, b]) \subset [a, b]$;
- Lipschitz 条件: $\exists L < 1$, 使得 $\forall x_1, x_2 \in [a, b]$ 均有

$$|\varphi(x_2) - \varphi(x_1)| \leq L |x_2 - x_1|.$$

则 $\forall x_0 \in [a, b]$, 迭代序列 $x_{i+1} = \varphi(x_i)$ 收敛到唯一的不动点 x^* . 且

$$|x_k - x^*| \leq \frac{1}{1-L} \min\{|x_{k+1} - x_k|, L|x_k - x_{k-1}|, \dots, L^k|x_1 - x_0|\}. \quad (7.2)$$

证明. 不动点的存在性可由 $\psi(x) = \varphi(x) - x$ 的介值定理可得, 唯一性由反证法得到.

对于不等式, 由迭代方程 $x_{i+1} = \varphi(x_i)$, 可得

$$|x_i - x_{i+1}| = |\varphi(x_{i-1}) - \varphi(x_i)| \leq L |x_{i-1} - x_i|,$$

对 $x_i - x_{i+p}$ 进行裂项可得,

$$|x_i - x_{i+p}| \leq |x_i - x_{i+1}| + \dots + |x_{i+p-1} - x_{i+p}| \leq (1 + L + \dots + L^{p-1}) |x_i - x_{i+1}|.$$

令 $p \rightarrow \infty$, 则 $x_{i+p} = x^*$, 即得. \square

推论. 对于连续函数 $\varphi \in \mathcal{C}^1[a, b]$, 将 Lipschitz 条件换为 $\|\varphi'\|_\infty \leq L < 1$, 同样可得 φ 在 $[a, b]$ 上存在唯一不动点.

注. 定理 7.1.1 给出了 φ 在区间 $[a, b]$ 上的全局收敛性. 但很多情况下全局收敛性并不容易检验. 由此引出 x^* 邻域上的局部收敛性概念.

定义 7.1.1: 局部收敛性

若 x^* 是 φ 的不动点, 且存在 x^* 的邻域 U , 使得 $\forall x^{(0)} \in U$, 迭代序列 $x^{(k+1)} = \varphi(x^{(k)}) \in U$ 且 $x^{(k)} \rightarrow x^*$, 则称 x^* 是局部收敛的 (local convergence), U 为其局部收敛域.

推论. 若 x^* 为 φ 的不动点, φ 在 x^* 某个邻域 U 上连续且 $|\varphi'(x^*)| < 1$, 则 x^* 局部收敛.

定义 7.1.2: 收敛阶

给定收敛序列 $x_n \rightarrow x$, 若

$$0 < \liminf_{n \rightarrow \infty} \frac{|x_{n+1} - x|}{|x_n - x|^p} \leq \limsup_{n \rightarrow \infty} \frac{|x_{n+1} - x|}{|x_n - x|^p} < +\infty, \quad (7.3)$$

则称序列 x_n 是 p 阶收敛的, 特别地, $p = 1$ 称为线性收敛, $p = 2$ 称为平方收敛;
若下极限可能为 0, 则称序列 x_n 是至少 p 阶收敛的; 若上极限为 0, 则称序列 x_n 是超 p 阶收敛的.

例 7.1.1

- 调和序列 $\{1/n\}$ 和等比序列 $\{a^n\}$ ($|a| < 1$) 线性收敛;
- 序列 $\{e^{-n^2}\}$ 超线性收敛, 但 $\forall \epsilon > 0$ 都不是超 $(1 + \epsilon)$ 阶收敛;
- 序列 $\{e^{-e^n}\}$ 是 e 阶收敛的.

定义 7.1.3: 收敛效率

若序列 $\{x_n\}$ 通过某种算法收敛于 x , 收敛阶为 p , 计算每个 x_n 的计算量为 θ (不依赖于 n), 定义

$$EI := p^{1/\theta},$$

表示收敛效率, 即单位计算量下算法的收敛率.

定理 7.1.2: 单步迭代法的收敛阶

若 φ 在其不动点 x^* 上的一个邻域 U 上足够光滑, 则 $x_{i+1} = \varphi(x_i)$

$$\text{在 } x^* \text{ 处 } p \text{ 阶局部收敛} \iff \varphi'(x^*) = \cdots = \varphi^{(p-1)}(x^*) = 0, \varphi^{(p)}(x^*) \neq 0$$

证明. 在 x^* 上作 Taylor 展开, 对于 x_k , $\exists \xi \in U$ 介于 x_k, x^* 之间使得

$$\varphi(x_k) = \varphi(x^*) + \varphi'(x^*)(x_k - x^*) + \cdots + \frac{\varphi^{(p-1)}(x^*)}{(p-1)!}(x_k - x^*)^{p-1} + \frac{\varphi^{(p)}(\xi)}{p!}(x_k - x^*)^p,$$

而 $\varphi'(x^*) = \cdots = \varphi^{(p-1)}(x^*) = 0$, 故

$$\frac{x_{k+1} - x^*}{(x_k - x^*)^p} = \frac{\varphi(x_k) - \varphi(x^*)}{(x_k - x^*)^p} = \frac{\varphi^{(p)}(\xi)}{p!} \rightarrow \frac{\varphi^{(p)}(x^*)}{p!}.$$

□

7.1.2 Newton 法

定理 7.1.3: Newton 迭代法

若 x_k 是 $f(x) = 0$ 根 x^* 的一个近似, 由 Taylor 展开,

$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \mathcal{O}(x^* - x_k)^2,$$

若 $f'(x_k) \neq 0$, 由 $f(x^*) = 0$, 可得

$$x^* = x_k - \frac{f(x_k)}{f'(x_k)} + \mathcal{O}(x^* - x_k)^2,$$

忽略二阶项, 剩余项作为 x^* 新的近似 x_{k+1} , 可得 Newton 迭代法:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \quad (7.4)$$

推论. 对应迭代函数

$$\varphi(x) = x - \frac{f(x)}{f'(x)}, \implies \varphi'(x) = \frac{f(x)f''(x)}{f'(x)^2},$$

故 $\varphi'(x^*) = 0$, Newton 迭代法是超线性收敛的.

注. Newton 法的几何解释: f 在 x_k 处的切线

$$y = f(x_k) + f'(x_k)(x - x_k),$$

与 x 轴的交点作为 x^* 新的近似 x_{k+1} .

定理 7.1.4: Newton 迭代法的局部收敛性

若根 x^* 处导数 $f'(x^*) \neq 0$ 且 f'' 在 x^* 邻域 U 上连续, 则 Newton 法至少二阶收敛.

证明. 由 Taylor 展开, $\exists \xi \in U$ 介于 x^*, x_k 之间使得

$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi)}{2}(x^* - x_k)^2,$$

而 $f(x^*) = 0$, 故

$$\frac{x_{k+1} - x^*}{(x_k - x^*)^2} = \frac{f''(\xi)}{2f'(x_k)} \rightarrow \frac{f''(x^*)}{2f'(x^*)}.$$

故 Newton 法至少二阶收敛; 当 $f''(x^*) \neq 0$ 时, Newton 法就是二阶收敛的. \square

注. Newton 法至少二阶收敛的前提之一是 $f'(x^*) \neq 0$, 这相当于说 x^* 是 f 的单重根.

定义 7.1.4: r - 重根

若

$$f(x^*) = f'(x^*) = \cdots = f^{(r-1)}(x^*) = 0, \quad f^{(r)}(x^*) \neq 0,$$

则称 x^* 是 f 的一个 r - 重根 (multiple root).

推论. 若 x^* 是 f 的 r - 重根, 则 $\exists \xi, \eta$ 介于 x_k, x^* 之间使得,

$$\varphi(x_k) = x_k - \frac{f(x^*) + f'(x^*)(x_k - x^*) + \cdots + f^{(r)}(\xi)(x - x^*)^r/r!}{f'(x^*) + f''(x^*)(x_k - x^*) + \cdots + f^{(r)}(\xi)(x - x^*)^r/(r-1)!},$$

故

$$\frac{x_{k+1} - x^*}{x_k - x^*} = 1 - \frac{1}{r} \frac{f^{(r)}(\xi)}{f^{(r)}(\eta)} \rightarrow 1 - \frac{1}{r} \geq \frac{1}{2}.$$

Newton 法变为线性收敛的! 故需要对 Newton 法进行改进.

定理 7.1.5: Newton 法的改进

若 x^* 为 f 的 r - 重根, 则依前文推论将 Newton 法 (7.4) 改进为二阶收敛的:

$$x_{k+1} = x_k - r \frac{f(x_k)}{f'(x_k)}, \quad (7.5)$$

对

$$u(x) := \frac{f(x)}{f'(x)}, \quad (7.6)$$

使用 Newton 法 (7.4), 易于验证, x^* 是 u 的单重根.

7.1.3 割线法

下面再介绍一种多步迭代方法.

定理 7.1.6: 割线法

以 f 在 x_k, x_{k-1} 的割线斜率 (差商)

$$f[x_k, x_{k-1}] = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

代替 Newton 法 (7.4) 中的导数 $f'(x_k)$, 称为割线法 (secant method):

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k), \quad (7.7)$$

定理 7.1.7: 割线法的收敛性

若在根 x^* 的邻域 $U(x^*, \delta) = [x^* - \delta, x^* + \delta]$ 上 $f'(x) \neq 0$ 且 $f \in \mathcal{C}^2(U)$, 若

$$M = \frac{\max_{x \in U} |f''(x)|}{2 \min_{x \in U} |f'(x)|} < +\infty,$$

则给定 $x_0, x_1 \in U(x^*, \min(\delta, 1/M))$, 割线法 $(1 + \phi)$ 阶收敛到 x^* .

证明. 记 $e_k = x_k - x^*$, 则

$$e_{k+1} = e_k - \frac{f(x_k) - f(x^*)}{f[x_k, x_{k-1}]} = e_k \frac{f[x_k, x_{k-1}] - f[x_k, x^*]}{f[x_k, x_{k-1}]},$$

故 $\exists \xi, \eta \in U$ 使得

$$e_{k+1} = e_k e_{k-1} \frac{f[x_k, x_{k-1}, x^*]}{f[x_k, x_{k-1}]} = e_k e_{k-1} \frac{f''(\xi)}{2f'(\eta)},$$

故

$$|e_{k+1}| \leq |e_k| |e_{k-1}| M \leq |e_k| M \min(\delta, 1/M)$$

由于 e_k 至少线性收敛, 故

$$|e_{k+1}| \approx |e_k| |e_{k-1}| M^*, \quad M^* = \frac{|f''(x^*)|}{2|f'(x^*)|},$$

其中 $a_n \approx b_n$ 表示 $\lim_{n \rightarrow \infty} a_n - b_n = 0$. 对上式取对数得到

$$\ln(M^* |e_{k+1}|) \approx \ln(M^* |e_k|) + \ln(M^* |e_{k-1}|),$$

当 $k \rightarrow \infty$ 时, 上式变为等号, 符合 Fibonacci 数列递推式, 故阶数为 $1 + \phi = \frac{1 + \sqrt{5}}{2}$. \square

注. Newton 法与割线法比较:

- 割线法不用求导;
- 当 x_k 接近 x^* 时舍入误差对割线法影响较大;
- 割线法的效率 $EI_1 = 1 + \phi$; Newton 法的效率 $EI_2 = 2^{1/\theta}$, 其中 θ 是计算 $f'(x)$ 相对差分的计算量. 当

$$\theta > \frac{1}{\log_2(1 + \phi)} > 1.44$$

时, 割线法效率更高.

7.1.4 Aitken 加速方法

由定理 7.1.2, 若 $|\varphi'(x^*)| \in (0, 1)$, 则迭代收敛速度为线性的. 由于

$$x_{k+2} - x^* = \varphi'(\xi)(x_{k+1} - x^*) \approx \varphi'(x^*)(x_{k+1} - x^*),$$

下一项 $x_{k+1} - x^* \approx \varphi'(x^*)(x_k - x^*)$, 故可联立消掉 $\varphi'(x^*)$, 得到:

$$x^* \approx x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k} \equiv x_k - \frac{\Delta x_k^2}{\Delta^2 x_k}. \quad (7.8)$$

其中差分 $\Delta x_k = x_{k+1} - x_k$, 二阶差分 $\Delta^2 x_k = \Delta x_{k+1} - \Delta x_k$. 和微分记号一样, 约定 $\Delta x_k^2 \equiv (\Delta x_k)^2$, 而 $\Delta(x_k^2)$ 需要加括号.

定理 7.1.8: Aitken 加速方法

若 x_k 至少线性收敛到 x^* , 则可定义

$$x'_k := x_k - \frac{\Delta x_k^2}{\Delta^2 x_k}, \quad (7.9)$$

并且 x'_k 收敛比 x_k 快:

$$\lim_{k \rightarrow \infty} \frac{x'_k - x^*}{x_k - x^*} = 0.$$

证明. 记 $e_k := x_k - x^*$, 由 x_k 至少线性收敛, $\exists \lambda$ 且 $|\lambda| < 1$ 使得

$$e_{k+1} = (\lambda + \delta_k)e_k, \quad \delta_k \rightarrow 0.$$

则

$$\Delta x_k = e_{k+1} - e_k = e_k[(\lambda - 1) + \delta_k],$$

$$\Delta^2 x_k = e_{k+2} - 2e_{k+1} + e_k = e_k[(\lambda + \delta_{k+1})(\lambda + \delta_k) - 2(\lambda + \delta_k) + 1] = e_k[(\lambda - 1)^2 + \mu_k],$$

其中 $\mu_k = \lambda(\delta_{k+1} + \delta_k) - 2\delta_k + \delta_k\delta_{k+1} \rightarrow 0$. 当 k 充分大时, $\Delta^2 x_k \neq 0$, 可定义 x'_k , 且

$$\lim_{k \rightarrow \infty} \frac{x'_k - x^*}{x_k - x^*} = \lim_{k \rightarrow \infty} 1 - \frac{[(\lambda - 1) + \delta_k]^2}{(\lambda - 1)^2 + \mu_k} = 0. \quad \square$$

注. Aitken 加速方法只需给定序列 x_k , 其公式 (7.9) 与序列 x_k 的产生方法 φ 无关.

7.1.5 Steffensen 迭代法

定理 7.1.9: Steffensen 迭代法

对 Aitken 加速公式 (7.9) 利用 $x_{k+1} = \varphi(x_k)$, $x_{k+2} = \varphi(\varphi(x_k))$ 作为新的迭代公式:

$$\hat{x}_{k+1} = \hat{x}_k - \frac{(\varphi(\hat{x}_k) - \hat{x}_k)^2}{\varphi(\varphi(\hat{x}_k)) - 2\varphi(\hat{x}_k) + \hat{x}_k}.$$

这实际上根据 φ 定义了新的迭代函数

$$\psi(x) = x - \frac{(\varphi(x) - x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x} = \frac{x\varphi(\varphi(x)) - \varphi(x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x}. \quad (7.10)$$

若 x^* 是 ψ 的不动点, 则 x^* 也是 φ 的不动点;

若 x^* 是 φ 的不动点, $\varphi \in C^1(U)$ 且 $\varphi'(x^*) \neq 1$, 则 x^* 是 ψ 的不动点.

证明. 由 ψ 表达式可得, $\psi(x^*) = x^* \implies \varphi(x^*) = x^*$; 而 $\varphi(x^*) = x^*$ 时, $\psi(x^*)$ 出现 $0/0$ 型, 由 L'Hôpital 法则

$$\psi(x^*) = x^* - \lim_{x \rightarrow x^*} \frac{(\varphi(x) - x)^2}{\varphi(\varphi(x)) - 2\varphi(x) + x} = x^* - \frac{2(\varphi(x^*) - x^*)(\varphi'(x^*) - 1)}{\varphi'(\varphi(x^*))\varphi'(x^*) - 2\varphi'(x^*) + 1} = x^*. \quad \square$$

注. 证明中分母出现了 $\varphi'(x^*) - 1$, 故要求 $\varphi'(x_k) \neq 1$, 即 x^* 不是 $\varphi(x) - x = 0$ 的重根. 但事实上 $\varphi'(x^*) = 1$ 也可以证明 $\psi(x^*) = x^*$.

定理 7.1.10: Steffensen 迭代法的收敛阶

若 $x_{k+1} = \varphi(x_k)$ 是 p 阶收敛的 ($p > 1$), 则 $x_{k+1} = \psi(x_k)$ 是 $(2p - 1)$ 阶收敛的.

若 $p = 1$ 且 $\varphi'(x^*) \neq 1$, 则 $x_{k+1} = \psi(x_k)$ 是二阶收敛的.

证明. 由 φ 是 p 阶收敛的: $\varphi(x) - x^* = \Theta(x - x^*)^p$, $\varphi(\varphi(x)) - x^* = \Theta(x - x^*)^{2p}$,

$$\begin{aligned} \psi(x) - x^* &= \frac{(x - x^*)(\varphi(\varphi(x)) - x^*) - (\varphi(x) - x^*)^2}{\varphi(\varphi(x)) - x^* - 2(\varphi(x) - x^*) + x - x^*} \\ &= \frac{\Theta(x - x^*)^{2p+1} - \Theta(x - x^*)^{2p}}{\Theta(x - x^*)^{2p} - \Theta(x - x^*)^p + x - x^*} = \Theta(x - x^*)^{2p-1}. \quad \square \end{aligned}$$

注. 若 $\varphi \in C^2(U)$ 且 $\varphi'(x^*) \neq 1$, 则 Steffensen 法不但可以提高收敛速度 (至少二阶收敛), 在 $|\varphi'(x^*)| > 1$ 时还可以把不收敛的方法改进为二阶收敛的方法. 但当 $p > 1$ 时, Steffensen 法一般好处不大, 故其多用于改进线性收敛的情形.

例 7.1.2

特别地, $\varphi(x) = x + f(x)$, 可得

$$\psi(x) = x - \frac{f(x)^2}{f(x + f(x)) - f(x)}. \quad (7.11)$$

若 $f(x^*) = 0$, 可得 $\psi(x^*) = x^*$, $\psi'(x^*) = 0$ 故 $x_{k+1} = \psi(x_k)$ 至少二阶收敛.

7.2 非线性方程组的不动点迭代法

定义 7.2.1: 向量值函数

向量值函数是一个映射 $F: D \rightarrow \mathbb{R}^n$, 其中 $D \subset \mathbb{R}^n$, 其分量是多元标量函数 $f_i: D \rightarrow \mathbb{R}$:

$$F(x) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{bmatrix} \quad (7.12)$$

定义 7.2.2: 向量值函数的连续性

给定点 $x_0 \in D$, 若 $\forall \epsilon > 0, \exists \delta > 0$ 使得领域内 $\forall x \in B(x_0, \delta) \subset D$ 均有

$$\|F(x) - F(x_0)\| < \epsilon,$$

则称 F 在 x_0 处连续.

定义 7.2.3: 向量值函数的导数

给定 $x \in D$ 和充分小 h 使得 $x + h \in D$, 若存在矩阵 $A(x) \in \mathbb{R}^{m \times n}$ 满足

$$\lim_{h \rightarrow 0} \frac{\|F(x+h) - F(x) - A(x)h\|}{\|h\|} = 0, \quad (7.13)$$

则称 F 在 x 处可微, $A(x) =: F'(x)$ 称为 F 的导数, 事实上就是 Jacobi 矩阵:

$$F'(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad (7.14)$$

定理 7.2.1: 压缩映像原理 (向量形式)

连续函数 $\varphi: D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ 满足:

- 闭性: D 是闭集;
- 映内性: $\varphi(D) \subset D$;
- 压缩性: $\exists L < 1$, 使得 $\forall x_1, x_2 \in D$,

$$\|\varphi(x_2) - \varphi(x_1)\| \leq L \|x_2 - x_1\|,$$

则 $\forall x_0 \in D$, 迭代序列 $x_{k+1} = \varphi(x_k)$ 收敛到 φ 唯一的不动点 x^* .

注. 压缩性依赖于范数的选择. 如 $\varphi(x) = Ax$, 其中

$$\varphi'(x) = A = \begin{bmatrix} 0.1 & 0.9 \\ 0 & 0.1 \end{bmatrix}.$$

在 2-范数下 $\|A\|_2 < 1$ 可压缩, 但在无穷范数下 $\|A\|_\infty = 1$ 不可压缩.

定义 7.2.4: 局部收敛性 (向量形式)

给定 $\varphi(x^*)$ 的不动点, 若存在 x^* 的一个邻域 $S \subset D$ 使得 $\forall x^{(0)} \in S, x^{(k+1)} = \varphi(x^{(k)}) \in S$ 且 $x^{(k)} \rightarrow x^*$, 则称 x^* 是局部收敛的.

定理 7.2.2

若 $\exists L \in (0, 1)$ 和某个范数 $\|\cdot\|$ 使得 $\forall x \in S(x^*, \delta)$ 均有

$$\|\varphi(x) - \varphi(x^*)\| \leq L \|x - x^*\|,$$

则 x^* 是局部收敛的.

推论. 一个充分条件: 若存在一个算子范数使得 $\|\varphi'(x^*)\| < 1$, 则 x^* 是局部收敛的.

定理 7.2.3: Newton 法 (向量形式)

给定 $x^{(0)}$, 计算

$$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{-1} F(x^{(k)}).$$

注. Newton 法的一个显著缺点是需要反复计算 F' . 可借鉴割线法改进.

定理 7.2.4: 修正的 Newton 法

将 Jacobi 矩阵 $F'(x^{(k)})$ 替换为矩阵 $A^{(k)}$:

$$x^{(k+1)} = x^{(k)} - A^{(k)-1} F(x^{(k)}),$$

矩阵 $A^{(k)}$ 满足拟 Newton 方程:

$$A^{(k)}(x^{(k)} - x^{(k-1)}) = F(x^{(k)}) - F(x^{(k-1)}), \quad (7.15)$$

显然 $A^{(k)}$ 的取法并不唯一, 可以迭代地写成

$$A^{(k+1)} = A^{(k)} + \Delta A^{(k)},$$

且 $\text{rank}(A^{(k)})$ 一般为 1 或 2, 即 $A^{(k+1)}$ 是 $A^{(k)}$ 的一个低秩修正.

定理 7.2.5: Broyden 秩 1 方法

令 $p^{(k)} = x^{(k+1)} - x^{(k)}$, $q^{(k)} = F(x^{(k+1)}) - F(x^{(k)})$, 设 $\Delta A^{(k)} = u^{(k)} v^{(k)\top}$ 是秩 1 矩阵, 则

$$(A^{(k)} + u^{(k)} v^{(k)\top}) p^{(k)} = q^{(k)}, \implies u^{(k)} = \frac{q^{(k)} - A^{(k)} p^{(k)}}{v^{(k)\top} p^{(k)}},$$

可令 $v^{(k)} = p^{(k)}$, 则

$$x^{(k+1)} = x^{(k)} - A^{(k)-1} F(x^{(k)}), \quad (7.16a)$$

$$\Delta A^{(k)} = \frac{(q^{(k)} - A^{(k)} p^{(k)}) p^{(k)\top}}{\|p^{(k)}\|^2}, \quad (7.16b)$$

$$A^{(k+1)} = A^{(k)} + \Delta A^{(k)}. \quad (7.16c)$$

第八章 矩阵特征值问题的数值方法

第九章 常微分方程初值问题的数值解法

9.1 常微分方程初值问题

定义 9.1.1: 一阶非线性常微分方程初值问题

考虑函数 $y : [x_0, b] \rightarrow \mathbb{R}^d$ 满足

$$y'(x) = f(x, y(x)), \quad (9.1a)$$

$$y(x_0) = y_0. \quad (9.1b)$$

特别地, 若 $f(x, y) = f(y)$ 与 x 无关, 则称为自治问题 (autonomous problem).

例 9.1.1: Hamilton 方程

给定相空间上的 Hamilton 量 $H : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, 广义坐标 $q \in \mathbb{R}^d$ 和广义动量 $p \in \mathbb{R}^d$ 随时间的演化满足 Hamilton 方程 (或正则方程, canonical equations):

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad (9.2a)$$

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}. \quad (9.2b)$$

比如调和谐振子 $H = kx^2/2 + p^2/2m$, 令角频率 $\omega = \sqrt{k/m}$, 则 Hamilton 方程解得

$$\begin{cases} \dot{x} = \frac{p}{m}, \\ \dot{p} = -kx. \end{cases} \implies \begin{cases} x = x_0 \cos(\omega t) + \frac{p_0}{m} \frac{\sin(\omega t)}{\omega}, \\ p = p_0 \cos(\omega t) - kx_0 \frac{\sin(\omega t)}{\omega}. \end{cases}$$

定义 9.1.2: 半线性高阶常微分方程

考虑函数 $y : [x_0, b] \rightarrow \mathbb{R}^d$ 满足

$$y^{(n)} = f(x, y, y', \dots, y^{(n-1)}), \quad (9.3a)$$

$$y(x_0) = y_0, \quad y'(x_0) = y_1, \quad \dots, \quad y^{(n-1)}(x_0) = y_{n-1}. \quad (9.3b)$$

可转化为一阶常微分方程:

$$Y := \begin{bmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{bmatrix}, \quad Y' = \begin{bmatrix} O & I & & \\ & \ddots & \ddots & \\ & & \ddots & I \\ & & & O \end{bmatrix} Y + \begin{bmatrix} O \\ \vdots \\ O \\ -f(x, Y) \end{bmatrix} \quad (9.4)$$

定理 9.1.1: Lipschitz 条件

若 $f(x, y)$ 在 $[a, b] \times \mathbb{R}^d$ 上连续, 且存在不依赖 x 的常数 $L > 0$ 使得 $\forall y_1, y_2 \in \mathbb{R}^d$,

$$\|f(x, y_1) - f(x, y_2)\| \leq L \|y_1 - y_2\|, \quad (9.5)$$

则 $\forall y_0$, 初值问题存在唯一解.

注. 我们总是假设 $f(x, y)$ 满足 Lipschitz 条件, 以保证初值问题存在唯一解.

定义 9.1.3: 适定性

若初值问题 (9.1) 存在唯一解 y , 考虑其扰动

$$y'(x) = f(x, y(x)) + \Delta f(x), \quad (9.6a)$$

$$y(x_0) = y_0 + \Delta y_0, \quad (9.6b)$$

若 $\forall \epsilon > 0, \exists \delta$ 使得当 $\|\Delta f\| < \delta, \|\Delta y_0\| < \delta$ 时, 扰动问题 (9.6) 存在唯一解 \tilde{y} 且满足

$$\|\tilde{y} - y\| < \epsilon, \quad (9.7)$$

则称初值问题 (9.1) 是适定的 (well-posed).

注. 适定性说明初值问题 (9.1) 的解存在唯一且连续依赖于 f 和初值 y_0 .

9.2 Euler 方法

定理 9.2.1: Euler 方法

考虑区间 $[x_0, b]$ 的一个划分:

$$a = x_0 < x_1 < \cdots < x_N = b$$

其中 $h_n := x_n - x_{n-1}$ 称为步长. 对于初值问题 (9.1), 若 $h := \max_n h_n$ 充分小, 则可用差商近似代替导数:

$$\frac{y(x_{n+1}) - y(x_n)}{h_{n+1}} \approx y'(x_n) = f(x_n, y(x_n)).$$

即

$$y_{n+1} = y_n + h_{n+1}f(x_n, y_n). \quad (9.8)$$

这称为 Euler 方法，是一种一阶显式单步法。

注. 必须关注的三个问题:

1. 收敛性: 当 $h \rightarrow 0$ 时, $\sup_n \|y_n - y(x_n)\| \rightarrow 0$;
2. 收敛速度;
3. 稳定性: 舍入误差的影响能否控制.

定理 9.2.2: 初值问题的等价积分形式

考虑划分区间上的初值问题 (9.1) 的等价积分形式

$$y(x) = y_n + \int_{x_n}^x f(t, y(t)) dt.$$

即可通过数值积分近似计算 $y(x_{n+1})$ 作为 y_{n+1} .

推论. 用左矩形法则近似积分, 即

$$y(x_{n+1}) = y(x_n) + h_{n+1}f(x_n, y(x_n)) + R_{n+1},$$

舍去积分余项 R_{n+1} , 即得 Euler 公式 (9.8).

例 9.2.1: 隐式 Euler 方法

用右矩形法则近似积分, 即

$$y(x_{n+1}) = y(x_n) + h_{n+1}f(x_{n+1}, y(x_{n+1})) + R_{n+1},$$

舍去积分余项 R_{n+1} , 即得隐式 Euler 公式

$$y_{n+1} = y_n + h_{n+1}f(x_{n+1}, y_{n+1}). \quad (9.9)$$

例 9.2.2: 梯形公式

用梯形法则近似积分, 即得梯形公式

$$y_{n+1} = y_n + \frac{h_{n+1}}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})]. \quad (9.10)$$

定义 9.2.1: 单步法的一般形式

结合 (9.8), (9.9) 和 (9.10), 初值问题 (9.1) 的单步法一般形式为

$$y_{n+1} = y_n + h_{n+1}\varphi(x_n, x_{n+1}, y_n, y_{n+1}).$$

若 φ 显含 y_{n+1} , 则称为隐式方法; 否则称为显式方法.

定义 9.2.2: 局部截断误差

单步法在 x_{n+1} 处的局部截断误差定义为积分余项:

$$R_{n+1} = y(x_{n+1}) - y(x_n) - h_{n+1}\varphi(x_n, x_{n+1}, y(x_n), y(x_{n+1})).$$

定义 9.2.3: 相容性

若 $\lim_{h \rightarrow 0} R = 0$, 则称单步法是相容的.

若存在不依赖 x 的常数 $M > 0$ 和整数 $p \geq 1$, 使得

$$\|y(x+h) - y(x) - h\varphi(x, x+h, y(x), y(x+h))\| \leq Mh^{p+1}, \quad (9.11)$$

则称单步法至少 p 阶相容的. 若局部截断误差可以展开成

$$R = \psi(x, y)h^{p+1} + \mathcal{O}(h^{p+2}),$$

称 $\psi(x, y)h^{p+1}$ 为局部截断误差的主项.

例 9.2.3: Euler 方法的局部截断误差

对于 Euler 方法 (9.8), 局部截断误差为

$$\begin{aligned} R_{n+1} &= y(x_{n+1}) - y(x_n) - h_{n+1}f(x_n, y(x_n)) \\ &= y(x_n) + h_{n+1}y'(x_n) + \frac{h_{n+1}^2}{2}y''(x_n) + \mathcal{O}(h_{n+1}^3) - y(x_n) - h_{n+1}y'(x_n) \\ &= \frac{h_{n+1}^2}{2}y''(x_n) + \mathcal{O}(h_{n+1}^3). \end{aligned}$$

因此 Euler 方法是一阶相容的, 局部截断误差的主项为 $h_{n+1}^2 y''(x_n)/2$.

例 9.2.4: 梯形方法的局部截断误差

对于梯形方法 (9.10), 局部截断误差为

$$\begin{aligned} R_{n+1} &= y(x_{n+1}) - y(x_n) - \frac{h_{n+1}}{2}[f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))] \\ &= -\frac{h_{n+1}^3}{12}y'''(x_n) + \mathcal{O}(h_{n+1}^4). \end{aligned}$$

因此梯形方法是二阶相容的, 局部截断误差的主项为 $-h_{n+1}^3 y'''(x_n)/12$.

定理 9.2.3: 单步法相容

单步法相容 $\iff f(x, y) = \varphi(x, x, y, y)$.