数值分析

2018 年 6 月 23 日

目录

基本信息

评分 考试 60%; 作业和实验和课堂测验 40%.

1 引论

1.1 基本概念

数值分析的对象 对于某个实际问题, 研究将其数学模型通过数值计算方法, 编写程序求解.

1.2 误差分析

模型和数据的错误或偏差不在数值分析的研究范围内.

- 截断 (方法) 误差. 如使用 $|x|$ 代替 $\sin x$.
- 舍入 (浮点) 误差

1.2.1 误差定义

记 x 为准确, x^* 为近似值. 有以下三种方法描述近似数和误差.

绝对误差 则定义 $e^* = x^* - x$ 为绝对误差, 其可正可负; 称 e^* 的上界为绝对误差限, 记为 ϵ^*. 绝对误差的值通常是不可得到的, 但是绝对误差限通常可以通过度量仪器的参数得到.

相对误差 定义相对误差 $e_r^* = \frac{e^*}{x}$; 但在相对误差较小时, 常取 $e_r^* = \frac{e^*}{x^*}$ 为相对误差. 同样地定义相对误差限 $\epsilon_r^* = |\frac{\epsilon^*}{x^*}|$.

有效数字 准确数通过四舍五入原则得到的近似数, 其前几位都为有效数字. 误差一定不会超过有效数字末位单位的一半, 如 $|3.14 - \pi| \leq \frac{1}{2} \times 0.01$. 定义近似数 x^* 的规范化表示为 $x^* = m \times 10^l$, 其中 $1 \leq m < 10$, $m = \sum_{-n<k\leq 0} a_k 10^{-k}$. 则 $|e^*| \leq \frac{1}{2} \times 10^{l-n+1}$.

有效数字和误差的关系 有效数字确定了相对误差限的上界, 相对误差限确定了有效数字的下界.

1.2.2 运算过程的误差分析

对于 $A = f(\vec{x})$, $A^* = f(x^*)$, 对于 Taylor 展开取线性项

$$A^* - A \approx \sum \frac{\partial f^*}{\partial x_i} x_i^*$$

同理可以得到相对误差的计算

1.3 定性分析

1.3.1 病态问题和条件数

对于函数的计算, 微小的输入误差导致很大的输出误差, 则称其为病态问题. 形式化地, 定义条件数 C_p 为输出相对误差和输入相对误差的比, 如果 C_p 很大, 相当于病态问题.

1.3.2 数值稳定性

1.4 算法设计技术

多项式求值 秦九韶算法. 减少乘除法.

2 插值法

2.1 基本概念

插值问题 给定点 $x_0, x_1 \dots x_n$, 以及 $y_0, y_1 \dots y_n$, 并且 $y_i = f(x_i)$, f 是某一个未知的函数. 求一条曲线 $y = p(x)$ 其严格通过 $(x_0, y_0), (x_1, y_1) \dots$.

其中称 p 是 f 的插值函数, $\langle x_i \rangle$ 称为插值节点.

注意插值问题和拟合问题的区别.

插值方法 通常有多项式插值, 三角函数, 有理函数, 样条函数等等. 事实上任何函数簇, 只要在被插值节点的值线性无关, 都可以用于插值.

2.2 高次多项式插值

求一个多项式 p 作为 f 的插值函数.

朴素方法 设 $p(x) = \sum_{0 \le i < n} a_i x^i$ 后解线性方程组.

2.2.1 Lagrange 插值

形如 $p(x) = L_n(x) = \sum_{i=0}^{n} y_i l_i(x)$ 的插值多项式称为 Lagrange 插值多项式, 其中 $l_i(x_j) = \delta_{ij}$. 易得 $l_i(x) = \prod_{j \ne i} \frac{x - x_j}{x_i - x_j}$.

记 $\omega_{n+1}(x) = \prod_i (x - x_i)$, 则有 $l_i(x) = \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x)}$.

存在唯一性 当 x_i 互异时, 由 Vandermonde 矩阵的可逆性, $x^0, \dots x^{n-1}$ 的线性组合组成的插值多项式必定存在唯一.

由代数基本定理, 此多项式一定和我们 $L_n(x)$ 相等.

误差估计 余项 $R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)$. 其中 f 在插值区间上 n 次连续可微, $n+1$ 次可微, ξ 与 x 有关.

2.2.2 Newton 插值

均差 均差 $f[x_0, x_1, \dots x_n]$ 定义如下

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$
$$f[x_0 \dots x_n] = \frac{f[x_1, x_2, \dots x_n] - f[x_0, x_1, \dots, x_{n-1}]}{x_n - x_0}$$

均差有如下性质

- $f[x_0, x_1, \dots x_n] = \sum_{0 \le i \le n} \frac{f(x_i)}{\prod_{j \ne i}(x_i - x_j)}$
  因此诸 x_i 的顺序对 $f[x_0, x_1, \dots x_n]$ 的值没有影响.
- $f[x_0, \dots x_n] = \frac{f^{(n)}(\xi)}{n!}$
  因此若 $\deg f < n$, 则 $f[x_0, \dots x_n] = 0$.

Newton 插值 通过均差的定义可以得到 Newton 插值基本公式

$$\begin{aligned} f(x) =& f(x_0) + f[x_0, x_1](x - x_0) + \dots + f[x_0, \dots x_n] \prod_{0 \le i < n} (x - x_i) \\ &+ f[x, x_0, \dots x_n] \prod_i (x - x_i) \end{aligned}$$

其中第一行 $P_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + \dots + f[x_0, \dots x_n] \prod_{0 \le i < n}(x - x_i)$ 为插值多项式, 第二行 $R_n(x) = f[x, x_0, \dots x_n] \prod_i (x - x_i)$ 为余项.

Newton 后项插值 对于特殊的 x_i 等间距的情况 $x_i = x_0 + ih$, 记 $f_k = f(x_0 + kh)$. 定义如下算子

$$\begin{aligned} \mathbf{I} f_k &= f_k \\ \mathbf{E} f_k &= f_{k+1} \\ \Delta f_k &= (\mathbf{E} - \mathbf{I}) f_k = f_{k+1} - f_k \end{aligned}$$

归纳易得 $f[x_0, x_1, \dots x_n] = \frac{1}{n!} h^{-n} \Delta^n f_k$.

代入原始 Newton 插值有

$$P_n(x_0 + th) = \sum_{0 \le i \le n} \Delta^i f_0 \frac{t^{\underline{i}}}{i!}$$

或者, 不严密地, 使用算子

$$\begin{aligned} f \circ \mathbf{E}^t &= \mathbf{E}^t \circ f \\ &= (\mathbf{I} + \Delta)^t \circ f \\ &= \sum_{n \geq 0} \binom{t}{n} \Delta^n \circ f \end{aligned}$$

2.2.3 Hermite 插值

概念 在插值节点 (x_i, y_i) 上要求 $p(x_i) = y_i$ 以外, 还要求导数值相等即 $p'(x_k) = f'(x_k)$.

不给出一般的 Hermite 插值的讨论, 但是显然可以对特定的问题求 Hermite 插值.

2.2.4 高次插值的一般性质

以上的高次插值可以看出有如下的优点

- 易于构造
- 使用方便
- 光滑性好 (C^n 连续)

但是缺点也如

- 不收敛, 如对于 Runge 的经典例子 $\frac{1}{1+x^2}$.
- 引入不需要的驻点, 凹凸性不好
- 数值稳定性不好, 计算系数误差变大

2.3 分段低次插值

2.3.1 分段线性插值

将 $(x_0, y_0), (x_1, y_1), \ldots (x_n, y_n)$ 用折线依次连接, 每个区间 $[x_i, x_{i+1}]$ 都是一条线段 $(x_i, y_i) \to (x_{i+1}, y_{i+1})$.

有一致收敛性.

2.3.2 分段三次 Hermite 插值

要求不仅给出 $y_i = f(x_i)$, 还要给出 $y_i' = f'(x_i)$. 在此前提下, 要求插值函数 I_{n+1} 满足 $I(x_i) = y_i$, $I'(x_i) = y_i'$, 且在每个区间 $[x_i, x_{i+1}]$ 上 I 都是三次函数 $a_i x^3 + b_i x^2 + c_i x + d_i$.

分区间表示 如式 (5.3), 即对于每个 $[x_i, x_{i+1}]$ 都给出一个表达式.

整体表示 $I(x) = \sum_{i=0}^{n} f_i \alpha_i(x) + f_i' \beta_i(x)$ 和对 $n+1$ 个节点整体插值的结果形式一样, 但是 α, β 不同整体插值的 α, β 是高次多项式, 而分段 Hermite 是逐段三次多项式. 则要求有

$$\begin{aligned} \alpha_i(x_j) &= \delta_{ij} \\ \beta_i(x_j) &= 0 \\ \alpha_i'(x_j) &= 0 \\ \beta_i'(x_j) &= \delta_{ij} \end{aligned}$$

余项 通过 Hermite 插值的分析, 容易得到 $\max |R(x)| \leq \frac{h^4}{384} \max |f^{(4)}(x)|$, 其中 $h = \max(x_{k+1} - x_k)$.

问题 实际中很难给出 y_k', 通常只知道 y_k.

2.3.3 三次样条插值

给定 (x_i, y_i), 要求插值函数满足

- 在每段区间中是次数不大于 3 的多项式
- 在插值区间上 C^2 连续 (即在 x_i 满足 C^2 连续即可)

3 函数逼近

3.1 函数线性空间

3.1.1 范数

线性空间 S 的元素到 $\mathbb{R}$ 的映射 $\|\cdot\|$, 满足如下性质

$$\begin{aligned}\|x\| \geq 0, \quad \|x\| = 0 \Leftrightarrow x = \mathbf{0}\\ \|ax\| = |a| \cdot \|x\|,\ a \in \mathbb{R}\\ \|x+y\| \leq \|x\| + \|y\|\end{aligned}$$

则称为 S 上的范数. S 称为赋范数空间.

连续函数的范数 常见地, 对 $f \in C[a,b]$ 有定义

$$\|f\|_n = \left(\int_a^b |f^n(x)|\,\mathrm{d}\,x\right)^{1/n}$$

特别的, $\|f\|_\infty = \max|f(x)|$.

3.1.2 内积

考虑 $\mathbb{R}$ 或者 $\mathbb{C}$ 上的线性空间 S, 内积 $(\cdot,\cdot)$ 将 $S\times S$ 映射到 F 满足

$$\begin{aligned}(u,v) &= \overline{(v,u)}\\ (au+bv,w) &= a(u,w)+b(v,w)\\ \forall u\ &: (u,u)\in\mathbb{R}, \quad (u,u)\geq 0, \quad (u,u)=0 \Leftrightarrow u=\mathbf{0}\end{aligned}$$

Cauchy-Schwartz 定理 由构造判别式法易证 Cauchy-Schwartz 定理

$$|(u,v)|^2 \leq (u,u)(v,v)$$

权函数 $[a,b]$ 上的函数 $\rho(x)$ 满足

$$\begin{aligned}\rho(x) &\geq 0\\ \int_a^b \rho(x)x^k\,\mathrm{d}\,x &\in \mathbb{R}\\ \forall g(x)\in C[a,b],\ g(x)\geq 0\ &:\ \int_a^b g(x)\rho(x)\,\mathrm{d}\,x = 0 \Leftrightarrow g(x)\equiv 0\end{aligned}$$

3.1.3 最佳逼近

对于 $f \in C[a,b]$, 考虑用 $[a,b]$ 上的 (不超过) n 次的多项式 $P(x)$ 逼近.

若 $\|f - P^*\|_n = \min_{P \in H_n} \|f - P\|_n$, 其中 H_n 表示次数界为 n 的多项式集合, 则称 P^* 是 f 的最佳逼近多项式.

当 $n = \infty$ 时称为最优一致逼近多项式, $n = 2$ 称最优平方逼近多项式.

3.1.4 正交函数族

若函数族 φ_k 满足

$$(\varphi_i, \varphi_j) = \int_a^b \rho(x)\varphi_i(x)\varphi_j(x)\,\mathrm{d}\,x = A_i\delta_{ij}$$

则称 φ_k 是 $[a,b]$ 上的关于 ρ 的正交函数族.

常见的如 $\langle 1, \sin x, \cos x, \sin 2x, \cos 2x \ldots \rangle$.

正交多项式 若正交函数族 $\varphi_k, k \in \mathbb{N}$ 满足 $\deg \varphi_k = k$ 则称 φ_n 为 n 次正交多项式. φ_n 可以容易地构造出. 在某个 $[a,b]$, 给定 φ_0 则可以惟一确定 φ_n.

可以证明正交多项式 φ_n 在 $[a,b]$ 上有且仅有 n 个零点.

3.2 最佳平方逼近

一般函数族最佳平方逼近 考虑的是 $[a,b]$ 上用一般的函数族 $\langle \varphi_0, \ldots \varphi_n \rangle$ 逼近 f 能达到的最佳平方逼近 $\min \|f - \sum a_i\varphi_i\|_2$.

对这个函数求偏微分可得到法方程

$$\forall i \ : \ \sum_j a_j(\varphi_i, \varphi_j) = (f, \varphi_i)$$

注意之后需要证明这样的 $\sum a_i\varphi_i$ 确实是最佳平方逼近, 因为驻点不是充要条件.

记 S^* 为最佳平方逼近, 则法方程的重要推论是 $(f, S^*) = (S^*, S^*)$.

多项式平方逼近 当 $\varphi_i = x^i$, $0 \le i \le n$, $\rho \equiv 1$ 时的情况. 同上, 需要解方程 $\mathbf{Ha} = \mathbf{d}$, 其中 $H_{ij} = \frac{1}{i+j-1}$ 称为 Hilbert 矩阵, $d_i = (f, \varphi_i)$.

但是容易看出, $\lim_{n\to\infty} \det \mathbf{H} = 0$, 因此求解 $\mathbf{Ha} = \mathbf{d}$ 是不稳定的.

正交函数族平方逼近 考虑 φ_i 正交的情况. 法方程变形为

$$a_i = \frac{(f, \varphi_i)}{(\varphi_i, \varphi_i)}$$

这种情况下, 误差界分析推出 Bessel 公式 $\sum_i \left(a_i^* \|\varphi_i\|_2\right)^2 \|f\|_2^2$

3.3 曲线拟合的最小二乘法

给定数据点 $\langle (x_i, y_i) \rangle$, 从函数族 $\langle \varphi_0, \varphi_1 \ldots \rangle$ 中取出一个函数 S^*, 最小化 $\sum_i (S^*(x_i) - y_i)^2$.

和最佳平方逼近的联系 定义离散点上的内积, 给定 $\langle x_i \rangle$, 则定义其离散点上内积为 $(f, g) = \sum_i \rho_i f(x_i) g(x_i)$, 其中 ρ 是权序列. 同样可以定义二次范数 $\|f\|_2 = \sqrt{(f, f)}$.

则最小二乘法变为与最佳平方逼近同样的形式 $\min \|f - S\|_2$, 其中 $f(x_i) = y_i$. 求解也和最佳平方逼近是一样的.

基函数的选择 选择 $\langle \varphi_i \rangle$, 可以根据数据规律手动选取, 或者使用一些通用的正交 / 非正交函数族. 对于多项式的情况, 一般选择正交多项式而非 $\langle x^k \rangle$, 因后者的行列式病态.

4 数值积分

给定函数 f 和积分区间 $[a, b]$, 求 $I[f] = \int_a^b f(x) \,\mathrm{d}\, x$.

4.1 插值积分

代数精度 若数值积分方法 I 满足在 $[a, b]$ 上有 $\forall k \in [0, m] \cap \mathbb{Z} \ : \ I[x^k] = \int_a^b x^k \,\mathrm{d}\, x$, 但 $I[x^{m+1}] \neq \int_a^b x^{m+1} \,\mathrm{d}\, x$, 则称数值积分方法 I 在 $[a, b]$ 上有代数精度 m.

4.1.1 插值求积方法

$I_n[f] = \int_a^b L_n(x) \,\mathrm{d}\, x$, 其中 $L_n(x)$ 是 f 在 $[a, b]$ 上 $n{+}1$ 个点 $\langle x_0, x_1 \ldots x_n \rangle$ 的 Lagrange 插值. $\langle x_k \rangle$ 按照某种与 f 无关的方法确定.

插值求积方法的余项 考虑 Lagrange 插值的余项是 $R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)$, 故有插值求积的余项是 $\int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)\,\mathrm{d}\,x$.

由此易证插值求积 $I_n[f]$ 的代数精度至少为 n. 但是可以构造一种选取 $\langle x_k\rangle$ 的方法使得其代数精度大于 n.

余项表达式 如果 $I[f]$ 有 m 阶代数精度, 则其余项 $R[f] = \int_a^b f(x)\,\mathrm{d}\,x - I[f]$ 可写成

$$R[f] = Kf^{(m+1)}(\eta)$$

其中 K 是和 f 无关的数. 通常带入 x^{m+1} 就可求出 K.

收敛 不严格的, 考虑差值方法 $I[f]$ 的两个参数 n 和 $h = \max\{x_{i+1} - x_i\}$. 若 $\lim_{h\to 0} R[f] = 0$, 则称积分公式 I 收敛.

4.1.2 函数值的加权表示

$I_n[f] = \sum_{k=0}^n A_k f(x_k)$, 其中 A_k, x_k 是以一种与 f 无关的方法选择的.

若 $I_n[f] = \sum_{k=0}^n A_k f(x_k)$ 在上代数精度至少为 n, 则 $I_n[f] = \int_a^b L_n(x)$. 即这样的 I_n 是插值求积方法.

证明可以考虑 $I_n[l_k]$, 有 $A_k = \int_a^b l_k(x)\,\mathrm{d}\,x$, 再带入 $I_n[f] = \sum_{k=0}^n A_k f(x_k)$ 即证.

4.1.3 Newton-Cotes 积分公式

基本概念 Newton-Cotes 公式是等距选取插值节点 $x_k = a + kh,\ h = \frac{b-a}{n}$ 时的 Lagrange 插值积分. 等距情况下, Lagrange 插值积分简化如下

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}\,x &\approx \int_a^b L_n(x)\,\mathrm{d}\,x \\ &= \int_a^b \sum_{k=0}^{n} f(x_k) l_k(x)\,\mathrm{d}\,x \\ &= \sum_{k=0}^{n} f(x_k) \int_a^b \frac{\prod_{j\neq k} x - x_j}{\prod_{j\neq k} x_k - x_j}\,\mathrm{d}\,x \\ &= \sum_{k=0}^{n} f(x_k) \int_0^n \frac{\prod_{j\neq k} t - j}{\prod_{j\neq k} k - j} h\,\mathrm{d}\,t \\ &= \sum_{k=0}^{n} \frac{(-1)^{n-k} h f(x_k)}{k!(n-k)!} \int_0^n \prod_{0\leq j\leq n, j\neq k} (t-j)\,\mathrm{d}\,t\end{aligned}$$

精度 N-C 公式是插值公式, 代数精度至少是 n. 对于 n 是偶数的情况, 还容易证明 N-C 公式至少有 $n+1$ 阶代数精度.

Simposon 公式 通常, N-C 公式只使用 $n = 1, 2, 4$ 的情形. $n = 1$ 时就是梯形公式, 而 $n = 2$ 时的公式

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

称为 Simposon 公式.

Simposon 公式代数精度为 3 阶, 可有前面公式计算余项

$$R[f] = -\frac{b-a}{180}\left(\frac{b-a}{2}\right)^4 f^{(4)}(\eta)$$

4.1.4 复合公式

基本思想为, 将 $[a, b]$ 分割为若干小区间 $[x_i, x_{i+1}]$. 在每个小区间上, 应用插值公式.

主要是复合梯形公式和复合 Simposon 公式.

4.2 Romberg 公式

主要讨论梯形公式的 Romberg 公式.

4.2.1 梯形公式的递推

使用递推可以有效减少通过倍增方法细化分割时需要的计算代价.

不妨设将 $[a,b]$ 等分为 $[x_i,x_{i+1}]$, $0\le i\le n$, $n+1$ 个节点, 共 $n+1$ 个小区间, 使用复合梯形积分公式, 得到积分值为 T_n.

考虑细化划分, 增加 $x_{i+1/2}=\frac{x_i+x_{i+1}}{2}$, $2n+1$ 个节点, 变成 $2n$ 个小区间. 每个小区间的积分从 $I_{n,i}=h\frac{f_i+f_{i+1}}{2}$ 变为 $I_{2n,i}=h\frac{f_i+2f_{i+1/2}+f_{i+1}}{4}=\frac{1}{2}I_{n,i}+\frac{1}{2}hf_{i+1/2}$. 整体积分

$$T_{2n}=\frac{1}{2}T_n+\frac{h}{2}\sum_{i=0}^{n-1}x_{i+1/2}$$

使用上式较重新计算需要更少的计算量.

4.2.2 Romberg 外推计算

复合梯形公式的余项 可以证明复合梯形公式的余项为 $R_n=\sum_{i>0}a_ih_n^{2i}=o(h_n^2)$.

Romberg 外推公式 容易得到 $h_{2n}=\frac{1}{2}h_n$. 因此 $R_{2n}=\frac{1}{4}a_ih_n^2+\sum_{i>1}a_i\left(\frac{h_n}{2}\right)^{2i}$, 因此令

$$T_{2n}^{(1)}=\frac{4T_{2n}-T_n}{3}$$

容易得到 $R_{2n}^{(1)}=\sum_{i>1}a_ih_{2n}^{2i}=o(h_n^4)=o(h_{2n})^4$.

同样, 令

$$T_{2n}^{(2)}=\frac{4^2T_{2n}^{(1)}-T_n^{(1)}}{4^2-1}$$

得到 $R_{2n}^{(2)}=o(h^6)$.

Richardson 外推法 即上面的公式一般形式, 上标表示做了多少次外推, 下标表示做了多少次分割

$$T_n^{(m)}=\frac{4^mT_{2n}^{(m-1)}-T_n^{(m-1)}}{4^m-1},\quad n=2^k,k\ge 0,m\ge 0$$

初始值 $T_n^{(0)}$ 是, 将 $[a,b]$ 等分为 $[x_i,x_{i+1}]$, $0\le i\le n$ 共 n 个小区间, 使用复合梯形公式的求积结果.

利用递推加速有 (其中 $x_{2k+1} = a + (2k+1)\frac{b-a}{2n}$)

$$T_{2n}^{(0)} = \frac{1}{2}T_n^{(0)} + \frac{h_n}{2}\sum_{k=0}^{n-1} f(x_{2k+1})$$

4.3 Gauss 公式

基本思想 Gauss 公式形如

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

其中 A_k, x_k 是实现选取和 f 无关的参数 (但是和 a, b 有关). Gauss 公式通过适当地选取 A_k, x_k 来提高 I 的代数精度, 可达 $2n+1$ 阶.

朴素构造 从 $x^0, x^1, x^2 \ldots$ 开始尝试, 每次解一个关于 A_k, x_k 的高次方程组

$$\sum_{k=0}^{s} A_k x_k^s = \int_a^b x^s \,\mathrm{d}\, x, \quad 0 \le s < \ldots$$

直到无解. 问题是, 高次方程求解是困难的.

优化求解 由前所述, Gauss 公式一定是插值的. 如上公式中, 如果我们求出诸插值节点 x_k, 则只需要解 A_k 的线性方程组即可.

关于 x_k, 有如下定理

$I[f]$ 有 $2n+1$ 阶代数精度 $\Leftrightarrow$

$$n+1 \text{ 次多项式 } \prod_{k=0}^{n}(x - x_k) \text{ 和任何次数界为 } n \text{ 的多项式正交}$$

证明需要考虑插值方法的余项 $R[f] = \int_a^b \frac{f^{(n+1)}(\psi)}{(n+1)!}\omega_{n+1}(x)\,\mathrm{d}\, x$.

5 解线性方程的直接法

5.1 Gauss 消元法

基本概念 算法略. 我们的语言中, 第 k 次消元完成后, 矩阵应当如下形态

$$\begin{bmatrix} a_{11}^{(k)} & \cdots & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} \\ & \ddots & \vdots & \vdots & \vdots \\ & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & \vdots & \ddots & \vdots \\ & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix}$$

收敛条件 诸顺序主子式 $D_i \neq 0$.

列主元的 Gauss 消元 第 k 步选取使得 $|a_{jk}^{(k-1)}|$ 最大的 $j \geq k$, 与第 k 行交换之后再继续.

分解地理解 原始 Gauss (行) 消元即, 对于顺序主子式非 0 的 $\mathbf{A}$, 有唯一的分解

$$\mathbf{A} = \mathbf{LU}$$

其中 $\mathbf{L}$ 是单位下三角阵, $\mathbf{U}$ 是上三角阵.

选列主元的 Gauss 消元就是

$$\mathbf{PA} = \mathbf{LU}$$

其中 $\mathbf{P}$ 是排列阵.

5.2 矩阵分解

5.2.1 Doolittle 分解

就是 LU 分解. Doolittle 分解是唯一的.

$$\mathbf{A} = \mathbf{LU}$$

直接可得到

$$u_{ri} = a_{ri} - \sum_{k=1}^{r-1} l_{rk}u_{ki}, \quad i \geq r$$
$$l_{ir} = \frac{a_{ir} - \sum_{k=1}^{r-1} l_{ik}u_{kr}}{u_{rr}}, \quad i > r$$

Doolittle 分解中我们也可以选主元.

5.2.2 通过 Gauss 消元计算 LU 分解

LU 分解 假设 Gauss 消元过后, 有

$$\mathbf{L}_n\mathbf{L}_{n-1}\dots\mathbf{L}_1\mathbf{A} = \mathbf{U}$$

那么 $\mathbf{A}$ 的 LU 分解是

$$\mathbf{L} = \mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\dots\mathbf{L}_n^{-1}$$
$$\mathbf{U} = \mathbf{U}$$

事实上, 对于消元过程中每一次 "第 i 行乘以 s_{ij} 后加到第 j 行", 都有一个 $l_{ij} = -s_{ij}$.

PLU 分解 亦称为选主元的 LU 分解. 假设选列主元的 Gauss 消元后, 有

$$\mathbf{L}_n\mathbf{P}_n\dots\mathbf{L}_2\mathbf{P}_2\mathbf{L}_1\mathbf{P}_1\mathbf{A} = \mathbf{U}$$

其中 $\mathbf{P}$ 是排列矩阵或者单位阵. 那么从 $\prod_{k=1}^{j}\mathbf{L}_k$ 右下子阵是单位子阵一点出发, 容易证明上式即

$$\mathbf{L}_n\dots\mathbf{L}_2\mathbf{L}_1\mathbf{P}_n\dots\mathbf{P}_2\mathbf{P}_1\mathbf{A} = \mathbf{U}$$

因此有

$$\mathbf{P} = \mathbf{P}_1^{-1}\mathbf{P}_2^{-1}\dots\mathbf{P}_n^{-1}$$
$$\mathbf{L} = \mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\dots\mathbf{L}_n^{-1}$$
$$\mathbf{U} = \mathbf{U}$$

5.2.3 Cholesky 分解

对于对称正定阵 $\mathbf{A}$, 存在下三角矩阵 $\mathbf{L}$,

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

若要求 $l_{ii} > 0$, 则 $\mathbf{L}$ 唯一.

直接可以得到

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}}, \quad i > j$$

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}$$

5.2.4 追赶法

考虑非奇异对角占优[1]三对角阵

$$\mathbf{A} = \begin{bmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & a_n & b_n \end{bmatrix}$$

其中 $|b_i| > |a_i| + |c_i|$.

[1] 注意是 $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$

可分解为 $\mathbf{A}=\mathbf{LU}$, 其中

$$\mathbf{L}=\begin{bmatrix}\alpha_1 & & & & & \\ r_2 & \alpha_2 & & & & \\ & r_3 & \alpha_3 & & & \\ & & \ddots & \ddots & & \\ & & & r_{n-1} & \alpha_{n-1} & \\ & & & & r_n & \alpha_n\end{bmatrix}$$

$$\mathbf{U}=\begin{bmatrix}1 & \beta_1 & & & & \\ & 1 & \beta_2 & & & \\ & & 1 & \beta_3 & & \\ & & & \ddots & \ddots & \\ & & & & 1 & \beta_{n-1} \\ & & & & & 1\end{bmatrix}$$

同样直接容易求得

$$\begin{aligned} r_i &= a_i \\ \alpha_i &= b_i - a_i\beta_{i-1} \\ \beta_i &= \frac{c_i}{\alpha_i} \end{aligned}$$

5.3 误差分析

5.3.1 向量内积

向量内积 $(\mathbf{u},\mathbf{v})$ 需要满足

线性性 $(\alpha\mathbf{u}+\beta\mathbf{v},\mathbf{w})=\alpha(\mathbf{u},\mathbf{w})+\beta(\mathbf{v},\mathbf{w})$.

正定性 $(\mathbf{u},\mathbf{u})\geq 0$, 当且仅当 $\mathbf{u}=\mathbf{0}$ 等号成立.

对称性 $(\mathbf{u},\mathbf{v})=(\mathbf{v},\mathbf{u})$, 对于复数情况是对称共轭.

5.3.2 向量范数

向量范数 $\|\cdot\|:\mathbb{R}^n\to\mathbb{R}$ 需要满足

线性性 $\|\alpha\mathbf{v}\|=|\alpha|\|\mathbf{v}\|$

正定性 $\|\mathbf{v}\| \geq 0$, 当且仅当 $\mathbf{v} = \mathbf{0}$ 等号成立

三角不等式 $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$

范数的几何意义可以理解为长度.

常用的 p 范数定义为 $\|\mathbf{v}\|_p = \sqrt[p]{\sum_i v_i}$, 特别地有 $\|\mathbf{v}\|_\infty = \max_i v_i$.

范数的连续性

$$\lim_{\mathbf{u}\to\mathbf{v}} \|\mathbf{u}\| - \|\mathbf{v}\| = 0$$

证明直接考虑证明 $\|\mathbf{u} - \mathbf{v}\| \to 0$.

范数的等价性 任意两种向量范数 $\|\mathbf{v}\|_{\nu_1}$ 和 $\|\mathbf{v}\|_{\nu_2}$, 存在 $0 < c_1 \leq c_2$ 使得

$$c_1\|\mathbf{v}\|_{\nu_1} \leq \|\mathbf{v}\|_{\nu_2} \leq c_2\|\mathbf{v}\|_{\nu_1}$$

证明考虑用 $\nu_1 = \infty$ 做跳板.

5.3.3 矩阵范数

矩阵范数 $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ 需要满足

正定性 $\|\mathbf{A}\| \geq 0$, 当且仅当 $\mathbf{A} = \mathbf{0}$ 等号成立

线性性 $\|\alpha\mathbf{A}\| = |\alpha|\|\mathbf{A}\|$

三角不等式 $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$

乘法 $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$

从属和相容 对于向量范数 $\|\cdot\|_\nu$ 定义其从属的矩阵范数是

$$\|\mathbf{A}\|_\nu = \max_{\mathbf{x}} \frac{\|\mathbf{Ax}\|_\nu}{\|x\|_\nu}$$

显然从属范数满足相容性条件

$$\|\mathbf{Ax}\|_\nu \leq \|\mathbf{A}\|_\nu\|\mathbf{x}\|_\nu$$

常见矩阵范数

- $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}}$ 称为 Frobenius 范数
- $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$
- $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$
- $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$

矩阵范数性质

- $\|\mathbf{A}\| \geq \rho(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$
  这里注意, 不能想当然的认为 $\rho(\mathbf{AB}) \leq \rho(\mathbf{A})\rho(\mathbf{B})$.
- $\forall \epsilon > 0 \ : \ \exists \nu \ : \ \|\mathbf{A}\|_\nu \leq \rho(\mathbf{A}) + \epsilon$
  证明略复杂. 注意这里的 $\rho(\mathbf{A}) = \max_i |\lambda_i|$.
- 若 $\|\mathbf{B}\| < 1$ 且 $\mathbf{I} \pm \mathbf{B}$ 非奇异, 则 $\|(\mathbf{I} \pm \mathbf{B})^{-1}\| \leq \frac{1}{1-\|\mathbf{B}\|}$

5.3.4 病态矩阵

对于方程组求解问题 $\mathbf{Ax} = \mathbf{b}$, 若微小的系数变动造成最后解的巨大波动, 则称矩阵 $\mathbf{A}$ 是病态的.

摄动条件 对于 $\delta\mathbf{b}$ 微小的扰动, 求解 $\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$, 有

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

对于 $\delta\mathbf{A}$ 微小的扰动, 需要假设 $\|\mathbf{A}^{-1}\delta\mathbf{A}\| < 1$, 求解 $(\mathbf{A} + \delta\mathbf{A})\mathbf{x} = \mathbf{b}$, 有

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1 - \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}$$

矩阵的条件数 定义 $\mathrm{cond}(\mathbf{A})_\nu = \|\mathbf{A}^{-1}\|_\nu\|\mathbf{A}\|_\nu$ 为 $\mathbf{A}$ 的条件数. 一般取 $\nu = 1, 2, \infty$. 条件数反映矩阵的病态程度, 条件数越大, 矩阵越病态.

条件数性质

- $\mathrm{cond}(\mathbf{A}) \geq 1$
- $\mathrm{cond}(\mathbf{A})_2 = \left|\frac{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}{\lambda_{\min}(\mathbf{A}\mathbf{A}^T)}\right| = \left|\frac{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}{\lambda_{\min}(\mathbf{A}^T\mathbf{A})}\right|$
  证明第二个等号可以使用经典的神奇等式 $|\mathbf{I}-\mathbf{AB}| = |\mathbf{I}-\mathbf{BA}|$, 或者直接证明 $\mathbf{AB}$ 特征值和 $\mathbf{BA}$ 相同.
- $\mathrm{cond}(\mathbf{AB}) \leq \mathrm{cond}(\mathbf{A})\,\mathrm{cond}(\mathbf{B})$

6 解线性方程的迭代法

6.1 通用迭代法

对于方程 $\mathbf{Ax} = \mathbf{b}$, 通过将其改写为

$$\mathbf{x} = \mathbf{Bx} + \mathbf{f}$$

之后按照 $\mathbf{x}^{(n+1)} = \mathbf{Bx}^{(n)} + \mathbf{f}$ 迭代.

若 $\lim_{n\to\infty}\mathbf{x}^{(n)} = \mathbf{x}^*$ 存在, 则称迭代法收敛. 显然 $\mathbf{x}^*$ 必为原方程的解.

实际中通常将 $\mathbf{A}$ 改写为 $\mathbf{M}+\mathbf{N}$, $\mathbf{M}$ 容易求逆, 之后

$$\mathbf{x} = -\mathbf{M}^{-1}\mathbf{Nx} + \mathbf{M}^{-1}\mathbf{b}$$

6.1.1 收敛条件

考虑初始误差 $\mathbf{e}^{(0)} = \mathbf{x}^{(0)} - \mathbf{x}^*$. 显然 $\mathbf{e}^{(n)} = \mathbf{B}^n\mathbf{e}^{(0)}$. 故有, 迭代法收敛, 当且仅当

$$\lim_{n\to\infty}\mathbf{x}^{(n)} = \mathbf{x}^* \Leftrightarrow \lim_{n\to\infty}\mathbf{B}^n = \mathbf{0}$$

等价的收敛条件 上述收敛条件等价于

1. $\rho(\mathbf{B}) < 1$
2. $\exists$从属范数ν : $\|\mathbf{B}\|_\nu < 1$

证明需要大量使用如下性质

$$\forall\epsilon \,:\, \exists\nu \,:\, \|\mathbf{B}\|_\nu \in [\rho(\mathbf{B}), \rho(\mathbf{B})+\epsilon]$$

收敛速度 简单的考虑, 设对于 ν 有 $\|\mathbf{B}\|_\nu = q < 1$, 则容易证明

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

平均收敛速度 显然 $\|\mathbf{e}^{(k)}\| \le \|\mathbf{e}^{(0)}\| \cdot \|\mathbf{B}^k\|$, 如欲让

$$\frac{\|\mathbf{e}^{(k)}\|}{\|\mathbf{e}^{(0)}\|} < \sigma$$

则 $\|\mathbf{B}^k\| < \sigma$. 即

$$k \ge \frac{-\ln \sigma}{-\ln\left(\|\mathbf{B}^k\|^{1/k}\right)}$$

称 $-\ln\left(\|\mathbf{B}^k\|^{1/ke}\right)$ 为平均收敛速度, 记为 $R_k(\mathbf{B})$.

称 $-\ln \rho(\mathbf{B})$ 为渐进收敛速度, 显然 $\lim_{k\to\infty} R_k(\mathbf{B}) = R(\mathbf{B})$.

6.2 具体迭代方法

6.2.1 Jacobi 方法

改写 $\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}$, 其中 $\mathbf{D}$ 是对角矩阵, $\mathbf{L}$ 是对角线为 0 的下三角阵, $\mathbf{U}$ 是对角线为 0 的上三角阵. 令

$$\mathbf{M} = \mathbf{D}$$

则有

$$\mathbf{x} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}$$

即

$$x_i^{(k+1)} = \frac{b_i - \sum_{j\ne i} a_{ji} x_i^{(k)}}{a_{ii}}$$

收敛条件 若 $\mathbf{A}$ 为严格对角占优阵, 则 Jacobi 法收敛.

若 $\mathbf{A}$ 对称, 且 $\mathbf{A}$, $2\mathbf{D} - \mathbf{A}$ 正定, 则 Jacobi 法收敛. 证明需要使用如下性质: $\mathbf{A}$对称正定 $\Rightarrow a_{ii} > 0$.

6.2.2 Gauss-Seidel 方法

同上, 令

$$\mathbf{M} = \mathbf{D} + \mathbf{L}$$

但是迭代的时候, 计算 $x_i^{(k)}$ 要使用 $x_{<i}^{(k)}$.

$$\mathbf{D}\mathbf{x}^{(k+1)} = -\mathbf{L}\mathbf{x}^{(k+1)} - \mathbf{U}\mathbf{x}^{(k)} + \mathbf{b}$$

即

$$x_i^{(k+1)} = \frac{b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)}}{a_{ii}}$$

收敛条件 若 $\mathbf{A}$ 为严格对角占优阵, 则 G-S 法收敛.

若 $\mathbf{A}$ 对称正定, 则 G-S 方法收敛.

6.2.3 超松弛迭代法

同上, 引入超参数 ω. 设每步 G-S 方法求出了 $\bar{\mathbf{x}}^{(k+1)}$, 则令

$$\mathbf{x}^{(k+1)} = (1-\omega)\mathbf{x}^{(k)} + \omega\bar{\mathbf{x}}^{(k+1)}$$

相当于

$$\mathbf{M} = \omega^{-1}\mathbf{D} + \mathbf{L}$$

显然 $\omega = 1$ 时 SOR 就是 G-S 方法, 通常将 $\omega < 1$ 的情况成为欠松弛.

收敛条件 若 SOR 收敛, 则 $0 < \omega < 2$. 证明: 直接考虑 $\rho(\mathbf{B}) < 1$ 即可, 利用 $|\mathbf{A}| = \prod_i \lambda_i$.

若 $\mathbf{A}$ 为对称正定阵, 且 $0 < \omega < 2$, 则 SOR 收敛.

若 $\mathbf{A}$ 为严格对角占优阵, 且 $0 < \omega \leq 1$ 则 SOR 收敛.

7 非线性方程和方程组的数值解

7.1 基本概念

方程求根 给定 $f : \mathbb{R} \to \mathbb{R}$, f 连续, 求一个 x : $f(x) = 0$. 称该 x 为 f 的根, 通常记为 x^*.

多重根 令 m 是最大的自然数使得 $\lim_{x\to x^*}\frac{f(x)}{(x-x^*)^{m-1}}=0$, 称 x^* 是 f 的 m 重根.

二分法 二分法是通用的求根方法, 简单而且保证收敛, 但是速度太慢.

7.2 不动点迭代法

7.2.1 基本概念

不动点法 将方程求根问题 $f(x)=0$ 改为不动点寻找问题 $x=\varphi(x)$, 通过迭代 $x_{k+1}=\varphi(x_k)$ 求根. 如果 $\lim_{n\to\infty}x_n=x^*$ 则称不动点法收敛.

全局收敛定理 若 $\varphi\in C[a,b]$, 且 $\forall x\in[a,b]$: $\varphi(x)\in[a,b]$, 并且 $\exists L<1$: $|\varphi(x_1)-\varphi(x_2)|\le L|x1-x2|$, 则 $[a,b]$ 中有唯一 x^* : $\varphi(x^*)=x^*$, 并且不动点法收敛.

局部收敛 如果 $\forall x_0\in B(x^*,\delta)$, 不动点法都收敛, 则称不动点法在 x^* 局部收敛.

局部收敛定理 若在 x^* 附近, φ 连续且 $|\varphi'(x)|<1$, 则不动点法在 x^* 局部收敛.

7.2.2 收敛阶

若 $x_{k+1}=\varphi(x_k)$ 收敛到 x^*, 并且

$$\exists p\ge 1\ :\ \frac{x_{k+1}-x^*}{(x_k-x^*)^p}=C\neq 0$$

则称 φ 在 x^* 是 p 阶收敛的.

收敛定理 若 $\varphi^{(p)}$ 在 x^* 领域连续, 且存在正整数 p,

$$\varphi(x^*)=\varphi'(x^*)=\ldots=\varphi^{(p-1)}(x^*)=0$$
$$\varphi^{(p)}(x^*)\neq 0$$

则 φ 在 x^* 是 p 阶收敛的.

7.3 Newton 法

7.3.1 基本 Newton 法

$x_{k+1} = \varphi(x_k)$, 迭代方程

$$\varphi(x) = x - \frac{f(x)}{f'(x)}$$

基本 Newton 法局部二阶收敛, 但是如果 x^* 是原函数的重根那么只有一阶收敛.

基本 Newton 法严重依赖初值的选取, 初值选好了收敛非常快, 初值选取不好不收敛.

7.3.2 简化 Newton 法

简化 Newton 法不用每次求导数 $f'(x_k)$

$$\varphi(x) = x - \frac{f(x)}{f'(x_0)}$$

简化 Newton 法局部线性收敛.

7.3.3 Newton 下山法

主要解决的 Newton 法对于初值非常依赖的问题. Newton 法求出 $\bar{x}_{k+1} = \varphi(x_k)$, 之后比较 $|f(\bar{x}_{k+1})|$ 和 $|f(x_k)|$. 若 $|f(\bar{x}_{k+1})| < |f(x_k)|$, 直接令 $x_{k+1} = \bar{x}_{k+1}$, 否则选取 $\lambda \in (0, 1]$, 令 $x_{k+1} = \lambda\bar{x}_{k+1} + (1-\lambda)x_k$, 要求 $|f(x_{k+1})| < |f(x_k)|$.

7.3.4 弦截法

不用求导数. 需要两个初值 x_1, x_0, 计算 x_{k+1} 使用 x_k, x_{k-1} 如下

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

弦截法局部 $\frac{\sqrt{5}+1}{2}$ 阶收敛.

7.4 方程求根的敏感性

病态方程 如果代数方程 $p(x)=\sum_{i=0}^{n}a_ix^i$ 对于很小的系数扰动, 根的变化很大, 则称 $p(x)$ 是病态的.

设系数摄动是 $\epsilon q(x)$, 则通常称

$$\frac{q(x^*)}{p'(x^*)}$$

为方程求根的条件数, 越大表示方程越病态.

8 矩阵特征值计算

8.1 特征值基本概念

Rayleigh 商 对于 $\mathbf{A}$, 定义其 Rayleigh 商为 $R(\mathbf{x})=\frac{(\mathbf{Ax},\mathbf{x})}{(\mathbf{x},\mathbf{x})}$. 设 $\mathbf{A}$ 的特征值是 $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_n$, 则有

$$\lambda_n\leq R(\mathbf{x})\leq\lambda_1$$
$$\lambda_n=\min R(\mathbf{x})$$
$$\lambda_1=\max R(\mathbf{x})$$

Gershgorin 圆盘 对于 $\mathbf{A}$, 令 $r_i=\sum_{j\neq i}a_{ij}$, $D_i=\{z\in\mathbb{C}\mid|z-a_{ii}|<r_i\}$ 成为 Gershgorin 圆盘. 设 $\mathbf{A}$ 的特征值是 $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_n$, 则有,

$$\forall\lambda_i\ :\ \lambda_i\in\cup_iD_i$$

即特征值一定在 Gershgorin 圆盘中.

另外, 考虑 $\cup_iD_i$ 中某连通区域 Ω, 如果其由 k 个圆盘构成, 则 Ω 中有且仅有 k 个特征值.

可以通过相似矩阵特征值不变的特性来精细化 Gershgorin 圆盘, 如 $\mathbf{D}^{-1}\mathbf{AD}$ 和 $\mathbf{A}$ 的特征值相同, Gershgorin 区域不同.

8.2 幂法求主特征值

8.2.1 主特征值

设 $\mathbf{A}$ 有特征值 $|\lambda_1|\geq|\lambda_2|\geq\ldots\geq|\lambda_n|$, 称 λ_1 为 $\mathbf{A}$ 的主特征值 (允许主特征值有实重根). 幂法可求出主特征值和主特征向量.

8.2.2 幂法

任取 $\mathbf{v}_0$, 使得 $\mathbf{v}_0 = \sum_i \alpha_i \mathbf{x}_i$, 其中诸 $\mathbf{x}$ 是特征向量. 要求 $\alpha_1 \neq 0$.
易证 (特别注意最后的 $x_k \neq 0$)

$$\forall k \in [1, n] \ : \ \lim_{m\to\infty} \frac{v_{m+1\,k}}{v_{mk}} = \lambda_1, \qquad x_k \neq 0$$

其中

$$\mathbf{v}_i = \mathbf{A}^i \mathbf{v}_0$$

简单地, 即如下. 其中 $\mathbf{a} \cdot \mathbf{b}$ 为向量点乘.

$$\lambda_1 = \| \lim_{k\to\infty} \mathbf{v}_{k+1} \cdot \mathbf{v}_k^{-1} \|_\infty$$

8.2.3 收敛速度

假设 $|\lambda_1| = |\lambda_2| \ldots > |\lambda_k|$, 则幂法的收敛速度由 $\frac{|\lambda_k|}{|\lambda_1|}$ 决定. 比值越小, 收敛越快.

8.2.4 变动

使用归一化的向量 令

$$\mathbf{v}_i = \mathbf{A}\mathbf{u}_{i-1}$$
$$\mathbf{u}_i = \|\mathbf{v}_i\|_\infty^{-1} \mathbf{v}_i$$

那么 $\lambda_1 = \lim_{m\to\infty} \|\mathbf{v}_m\|_\infty$.

原点平移法 如果 $\frac{|\lambda_k|}{|\lambda_1|}$ 很接近 1, 那么收敛速度很慢. 这时考虑求 $\mathbf{A} - p\mathbf{I}$ 的特征值 $\lambda_1 - p, \lambda_2 - p \ldots$. 通过选择 p, 使得次大特征值和主特征值的比尽量小.

如果已知 $\lambda_1 > \lambda_2 \geq \ldots > \lambda_n$, 则最后的主特征值只可能是 $\lambda_1 - p$ 或者 $\lambda_n - p$, 只需要在 $|\lambda_n - p| < |\lambda_1 - p|$ 的前提下, 令 $p^* = \frac{\lambda_2 + \lambda_n}{2}$ 即可.

反幂法 幂法用来求模最大的特征值 λ_1, 反幂法用来求模最小的特征值 λ_n. 显然, 考虑 $\mathbf{A}$, 针对 $\mathbf{A}^{-1}$ 做幂法就能得到 λ_n^{-1}.

8.3 QR 分解

8.3.1 Household 变换

定义 若 $\mathbf{w}^T\mathbf{w} = 1$, 则称

$$\mathbf{H} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^T$$

为 Household 变换矩阵, 亦称单位反射阵.

性质

- $\mathbf{H}$ 是对称阵

$$\mathbf{H}^T = \mathbf{H}$$

- $\mathbf{H}$ 是正交阵

$$\mathbf{H}^T\mathbf{H} = \mathbf{I}$$

  注意对称阵的乘积不一定对称, 所以多个单位反射阵的积只是正交, 并不一定对称.

- 反射性: 乘 $\mathbf{H}$ 就相当于作了平面 $\mathbf{w}^T\mathbf{x} = 0$ 的镜像.

$$\forall \mathbf{v} = \mathbf{x} + \mathbf{y},\ \mathbf{w}^T\mathbf{x} = 0,\ \mathbf{y} \parallel \mathbf{w}\ :\ \mathbf{H}\mathbf{v} = \mathbf{x} - \mathbf{y}$$

- 反射性: 可以在等模的前提下做任何 $\mathbf{x}$ 到 $\mathbf{y}$ 的变换

$$\forall \mathbf{x}\ :\ \forall \mathbf{y}, \|x\|_2 = \|y\|_2\ :\ \exists \text{单位反射阵}\, \mathbf{H}\ :\ \mathbf{H}\mathbf{x} = \mathbf{y}$$

  容易从几何意义得到, $\mathbf{w} = \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2}$.

约化定理 对于 $\mathbf{x} \neq \mathbf{0}$, 存在 $\mathbf{H}$ 使得

$$\mathbf{H}\mathbf{x} = \sigma\mathbf{e}_1$$

其中

$$\sigma = \begin{cases} \operatorname{sgn}(x_1)\|\mathbf{x}\|_2 & x_1 \neq 0 \\ \|\mathbf{x}\|_2 & x_1 = 0 \end{cases}$$

$$\mathbf{u} = \mathbf{x} + \sigma\mathbf{e}_1$$

$$\beta = \frac{\|\mathbf{u}\|_2^2}{2}$$

$$\mathbf{H} = \mathbf{I} - \beta^{-1}\mathbf{u}\mathbf{u}^T$$

事实上, 约化定理就是反射性第二条的简单应用. 基本目的是将 $\mathbf{x}$ 变成 $\alpha\mathbf{e}_1$. 这种情况下只能 $\alpha = \pm\|\mathbf{x}\|_2$, 又为了 $\mathbf{x} - \alpha\mathbf{e}_1 \neq \mathbf{0}$, 所以 $\alpha = -\mathrm{sgn}(x_1)\|\mathbf{x}\|_2$. 之后自然地就得到了 $\mathbf{H}$.

8.3.2 QR 分解

对于任何非奇异的矩阵 $\mathbf{A}$, 一定存在正交阵 $\mathbf{Q}$, 满足 $\mathbf{QA} = \mathbf{R}$, 其中 $\mathbf{R}$ 是上三角阵. 如果要求 $\mathbf{R}$ 对角元素为正, 那么 QR 分解唯一.

分解过程 一般地, 考虑

$$\mathbf{H}_i\mathbf{A} = \begin{bmatrix} \mathbf{U}_i & \mathbf{B}_i \\ \mathbf{0} & \mathbf{C}_i \end{bmatrix}$$

其中 $\mathbf{H}_i \in \mathbb{R}^{n\times n}$, $\mathbf{U}_i \in \mathbb{R}^{i\times i}$ 为上三角阵, $\mathbf{B}_i \in \mathbb{R}^{i\times(n-i)}$, $\mathbf{C}_i \in \mathbb{R}^{(n-i)\times(n-i)}$ 且 $\mathbf{C}$ 可逆. 由约化定理,

$$\exists\tilde{\mathbf{H}}_{i+1} \in \mathbb{R}^{(n-i)\times(n-i)} \ : \ \tilde{\mathbf{H}}_{i+1}\mathbf{C}_i = \begin{bmatrix} u_{i+1} & \tilde{\mathbf{B}}_{i+1} \\ \mathbf{0} & \mathbf{C}_{i+1} \end{bmatrix}$$

那么令

$$\mathbf{H}_{i+1} = \begin{bmatrix} \mathbf{I}_i & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{H}}_{i+1} \end{bmatrix}\mathbf{H}_i$$

则有

$$\mathbf{H}_{i+1}\mathbf{A} = \begin{bmatrix} \mathbf{U}_{i+1} & \mathbf{B}_{i+1} \\ \mathbf{0} & \mathbf{C}_{i+1} \end{bmatrix}$$

其中 $\mathbf{H}_{i+1} \in \mathbb{R}^{n\times n}$, $\mathbf{U}_{i+1} \in \mathbb{R}^{i+1\times i+1}$ 为上三角阵, $\mathbf{B}_{i+1} \in \mathbb{R}^{i+1\times(n-i-1)}$, $\mathbf{C}_{i+1} \in \mathbb{R}^{(n-i-1)\times(n-i-1)}$ 且 $\mathbf{C}$ 可逆.

如上, 从 $\mathbf{H}_0 = \mathbf{I}_n$ 开始, 最后就可以得到 $\mathbf{H}_n\mathbf{A} = \mathbf{U}_n$.

QR 分解定理 非奇异矩阵 $\mathbf{A} \in \mathbb{R}^{n\times n}$, 存在上三角阵 $\mathbf{R}$ 和正交阵 $\mathbf{Q}$, 使得 $\mathbf{A} = \mathbf{QR}$, 称为 $\mathbf{A}$ 的 QR 分解. 如果要求 $\mathbf{R}$ 对角线元素为正, 则 $\mathbf{A}$ 的 QR 分解唯一.

证明考虑正定矩阵 $\mathbf{A}^T\mathbf{A}$ 的 Cholesky 分解是唯一的.

9 常微分方程数值解

9.1 基本概念

常微分方程描述 对于

$$y' = f(x, y)$$

其中 $x \in [a, b]$, $y \in \mathbb{R}$, 给定初值

$$y(x_0) = y_0$$

称为一个常微分方程 (ode).

解的存在唯一性 若 f 在 $[a, b] \times \mathbb{R}$ 连续, 且满足 Lipschitz 条件

$$|f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2|$$

则如上描述的常微分方程对于任意 $x_0 \in [a, b], y_0 \in \mathbb{R}$ 有唯一解.

对初值的敏感性 对于给定初值有唯一解的 ode, 记 $y(s, x)$ 是给定初值 $y(x_0) = s$ 时的解, 则

$$|y(s_1, x) - y(s_2, x)| \leq e^{L|x-x_0|}|s_1 - s_2|$$

其中 L 为 f 的 Lipschitz 常数.

微分方程的数值解法 给定 $x_0 < x_1 < \ldots < x_{n-1} < x_n$, 以及初值 y_0, 要求 y_i. 通常认为 $|x_{i+1} - x_i| = h$ 是常量. 下面使用 $y_i = \varphi(\mathbf{x}, \mathbf{y}, h)$ 来表示一个数值解法.

局部截断误差 对于 $y_i = \varphi(\mathbf{x}, \mathbf{y}, h)$, 定义其截断误差为

$$T_i = y(x_i) - \varphi(\mathbf{x}, \langle y(x_0), y(x_1) \ldots y(x_n)\rangle, h)$$

精度 若 $y_i = \varphi(\mathbf{x}, \mathbf{y}, h)$ 满足 $T_n = o(h^{p+1})$, 则称它有 p 阶精度.

9.2 Euler 方法

9.2.1 显式 Euler 法

显式单步法. 基本方程

$$y_{i+1} = y_i + hf(x_i, y_i)$$

局部截断误差

$$T_i = h^2 \frac{y''(x_i)}{2} + o(h^3)$$

精度为 1 阶.

9.2.2 隐式 Euler 法

隐式法. 基本方程

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$$

精度为 1 阶.

通常需要由显式 Euler 法给出一个 $\mathbf{y}$ 的初值, 然后再由隐式 Euler 法迭代.

如果 $|hL| < 1$, L 是 Lipschitz 常数, 那么迭代一定收敛到隐式法的迭代方程, 虽然不一定收敛到原 ode 的解.

9.2.3 梯形法

隐式法. 基本方程

$$y_{i+1} = y_i + h\frac{f(x_i, y_i) + f(x_{i+1}, y_{i+1})}{2}$$

局部阶段误差

$$T_i = -\frac{h^3}{12} y^{(3)}(x_i) + o(h^4)$$

精度为 2 阶.

显式 Euler 法提供初值时, 称为 “改进的 Euler 法”.

9.3 Runge-Kuta 方法

9.3.1 基本叙述

Runge-Kuta 方法是单步显式法. 对于单步显式法 $y_{i+1} = y_i + h\varphi(x_i, y_i, h)$, 局部截断误差是

$$T_n = \int_{x_n}^{x_{n+1}} f(x, y(x))\,\mathrm{d}\,x - h\varphi(x_n, y_n, h)$$

因此主要使得 $h\varphi(x_n, y_n, h)$ 接近积分即可. 这里就可以使用数值积分的方法.

9.3.2 低阶情况

考虑二阶 Runge-Kuta 方法

$$\begin{aligned} y_{i+1} &= y_i + h(c_1K_1 + c_2K_2) \\ K_1 &= f(x_n, y_n) \\ K_2 &= f(x_n + \lambda_2 h, y_n + \mu_{21}hK_1) \end{aligned}$$

分解得到, $c_1, c_2, \lambda_2, \mu_{21}$ 满足

$$\begin{aligned} c_1 + c_2 &= 1 \\ c_2\lambda_2 &= \frac{1}{2} \\ c_2\mu_{21} &= \frac{1}{2} \end{aligned}$$

精度为 2 阶. 如 $c_1 = c_2 = \frac{1}{2}$ 就是梯形法, $c_1 = 0, c_2 = 1$ 就是 “中点公式”.

9.4 收敛和稳定

9.4.1 收敛性

概念 对于某种数值方法, 如果 $\lim_{h\to\infty} y_n = y(x_n)$, 那么称其收敛.

单步法的收敛性 对于单步法 $y_{n+1} = y_n + h\varphi(x_n, y_n, h)$, 如果其具有 p 阶精度, 且 φ 关于 y 有 Lipschitz 条件, 那么只要初值 y_0 是准确的, 整体截断误差 $T_n = y_n - y(x_n) = o(h^p)$.

通过单步法收敛定理, 可以直接通过 φ 是否满足 Lipschitz 条件判断单步法是否收敛.

相容性 如果单步法满足 $\varphi(x,y,0)=f(x,y)$, 那么称单步法和原 ode 相容. 当且仅当单步法有 $p\geq 1$ 阶精度, 单步法和原 ode 相容.

9.4.2 稳定性

若

$$|y_n-y(x_n)|<\epsilon\Rightarrow\forall m>n\ :\ |y_m-y(x_m)|<\epsilon$$

则称这种数值方法稳定.

基本概念 稳定性不仅受原 f 的影响, 而且还受 h 的影响. 为了标准化研究, 只考虑如下的 ode

$$f(x,y)=\lambda y$$

解析解为 $y=e^{\lambda x}$, 要求 $\Re\lambda<0$.

使得迭代法收敛的 $|h\lambda|$ 构成的 $\mathbb{C}$ 区域称为绝对收敛域, 实轴上部分成为绝对收敛区间.

以显式 Euler 法为例子, 上述 ode 的迭代式为 $y_{n+1}=(1+h\lambda)y_n$. 因此 $\epsilon_{n+1}=(1+h\lambda)\epsilon_n$, 只要 $|1+h\lambda|<1$ 显式 Euler 法就针对初值的波动是稳定的.

9.5 线性多步法

9.5.1 基本概念

考虑显式的多步法, k 步的方程为

$$y_{n+k}=\sum_{i=0}^{k-1}\alpha_i y_{n+i}+h\sum_{i=0}^{k-1}\beta_i f(x_{n+i},y_{n+i})$$

9.5.2 精度推导

考虑 T_{n+k} 的局部截断误差, 有

$$\begin{aligned}
T_{n+k} =& y(x_{n+k}) - \sum_{i=0}^{k-1} \alpha_i y_{n+i} - h \sum_{i=0}^{k-1} \beta_i f(x_{n+i}, y_{n+i}) \\
=& y(x_n + kh) - \sum_{i=0}^{k-1} \alpha_i y(x_n + ih) - h \sum_{i=0}^{k-1} \beta_i y'(x_n + ih) \\
=& \sum_{j\geq 0} \frac{(kh)^j}{j!} y^{(j)}(x_n) \\
& - \sum_{i=0}^{k-1} \alpha_i \sum_{j\geq 0} \frac{(ih)^j}{j!} y^{(j)}(x_n) \\
& - h \sum_{i=0}^{k-1} \beta_i \sum_{j\geq 0} \frac{(ih)^j}{j!} y^{(j+1)}(x_n) \\
=& \sum_{j\geq 0} y^{(j)}(x_n) \frac{h^j}{j!} c_j
\end{aligned}$$

其中

$$c_j = k^j - \sum_{i=0}^{k-1} \alpha_i i^j - j \sum_{i=0}^{k-1} \beta_i i^{j-1}$$

如果 $c_0 = c_1 = \ldots = c_p = 0$, 则显然 α, β 确定的线性多步法有 p 阶精度.

9.5.3 Adams 公式

Adam 公式形如

$$y_{n+k} = y_{n+k-1} + h \sum_{i=0}^{k} \beta_i f(x_{n+i})$$

注意求和上界是 k. 如果 $\beta_k = 0$, 则 Adams 公式为显式的, 否则为隐式的.

只需求出 β_k. 如上 $\alpha_{k-1} = 1$, 方程为

$$\sum_{i=0}^{k} j i^{j-1} \beta_i = k^j - k^{j-1}, \quad 1 \leq j \leq k+1$$

约定 $0^0 = 1$. 令 $\beta_k = 0$ 并丢弃最后一个方程, 即得显式情况的方程.

显式公式的精度是 k, 隐式公式是 $k+1$.