数值计算笔记

Fiddie

2022 年 5 月 10 日

目　　录

CHAPTER 1

误差分析

§1.1 基本概念

误差来源: (1)**数据误差(观测误差)**: 所用数据的初始值是近似的. (2)**截断误差**: 求和、求极限的时候仅取有限项来近似; (3)**离散误差**: 用一个近似的离散公式来求一个连续问题的解(比如算积分).

定义 1.1.1: 绝对误差

设某个准确值为x, 近似值为$\bar{x}$, 则x与$\bar{x}$的差$e_{\bar{x}} = x - \bar{x}$为近似值$\bar{x}$的**绝对误差**. 如果有一个正数$\varepsilon$使得

$$|e_{\bar{x}}| = |x - \bar{x}| \leq \varepsilon,$$

则称ε为近似值$\bar{x}$的一个**绝对误差界**.

定义 1.1.2: 相对误差

定义$r_{\bar{x}} = \dfrac{x - \bar{x}}{\bar{x}}$为**相对误差**. 如果有一个正数$\delta$使得

$$|r_{\bar{x}}| = \frac{|x - \bar{x}|}{|\bar{x}|} \leq \delta,$$

则称δ为近似值$\bar{x}$的一个**相对误差界**.

在十进制下, 把一个实数写成无限小数的形式: $a = \pm a_0 a_1 \cdots a_m . a_{m+1} \cdots a_n a_{n+1} \cdots$. 其中$a_i \in \{0, 1, \cdots, 9\}$且$m \neq 0$时$a_0 \neq 0$. 如果$a$近似$\bar{a}$取$n - m$位小数时, 有

$$|a - \bar{a}| \leq \frac{1}{2} \times 10^{-(n-m)}. \tag{1.1}$$

按照这个规律得到的误差是**舍入误差**. 对于近似数$\bar{a} = \pm a_0 a_1 \cdots a_m . a_{m+1} \cdots a_n$, 若$\bar{a}$的绝对误差满足式(1.1), 且$a_s$是$\bar{a}$的第一位非零数字, 则从$a_s$起到最右边的数字$a_n$为止都是$\bar{a}$的**有效数字**, $\bar{a}$具有$n + 1 - s$位有效数.

定理 1.1.1

若形如$\bar{a}=\pm a_0a_1\cdots a_m.a_{m+1}\cdots a_n$的近似数有$n+1-s$位有效数字, 则相对误差有估计式

$$\left|\frac{a-\bar{a}}{\bar{a}}\right|\leq\frac{1}{2a_s}\times10^{-(n-s)}.$$

其中$a_s\neq0$是$\bar{a}$的第一位有效数字.

证明: 用式(1.1)以及$|\bar{a}|\geq a_s\times10^{-s}$即可. □

注: 定理表明, 一个近似数的有效数字越多, 相对误差越小, 精确度越高.

算术运算中误差的传播: 设x,y的近似值是$\bar{x},\bar{y}$, 计算函数$z=f(x,y)$产生的误差可表示为: (微分中值定理)

$$e_{\bar{z}}=z-\bar{z}=f(x,y)-f(\bar{x},\bar{y})\approx\left(\frac{\partial f}{\partial x}\right)_{(\bar{x},\bar{y})}e_{\bar{x}}+\left(\frac{\partial f}{\partial y}\right)_{(\bar{x},\bar{y})}e_{\bar{y}}.$$

对于加减乘除运算, 有下面的结论:

(1) $z=x\pm y$: $e_{\bar{z}}=e_{\bar{x}\pm\bar{y}}=e_{\bar{x}}\pm e_{\bar{y}}$; (2) $z=xy$: $e_{\bar{z}}=e_{\bar{x}\bar{y}}=\bar{y}e_{\bar{x}}+\bar{x}e_{\bar{y}}$; (3) $z=\dfrac{x}{y}$: $e_{\bar{z}}=e_{\bar{x}/\bar{y}}=\dfrac{\bar{y}e_{\bar{x}}-\bar{x}e_{\bar{y}}}{\bar{y}^2}$.

另外作$r_{\bar{z}}=\dfrac{e_{\bar{z}}}{\bar{z}}$可求得相对误差(略) 根据求得的相对误差, 尽量不用较小的数做分母; 更重要的是尽量避免两个较小的数相减(会产生相减相消).

§1.2 机器误差

计算机中数x只是有限位小数, p进制数x可表示为$\boxed{x=\pm p^J\sum_{k=1}^{t}d_kp^{-k}}$. 其中, $d_k\in\{0,1,\cdots,p-1\}$. 又记$a=\sum_{k=1}^{t}d_kp^{-k}$, 则$x=\pm a\times p^J$, 这里$\boxed{a=0.d_1d_2\cdots d_t}$.

定义 1.2.1

把上面的p称为**基数**, a称为x的**尾数**, t为**字长**, J为x的**阶**. 规定阶J的范围为$-L\leq J\leq U$, 其中L和U为正整数或零.

若J是固定不变的, 称上面为**定点表示**; 若J可变, 称为**浮点表示**. 如果浮点表示中尾数的第一位$d_1\neq0$, 这种数叫**规格化浮点数**. 从而k进制中尾数满足关系式为$\dfrac{1}{k}\leq a<1$. 所有规格化浮点数组成的集合F叫**规格化浮点数系**, 它是个离散的有限集合.

定理 1.2.1

规格化浮点数系F中共有$2(p-1)p^{t-1}(L+U+1)+1$个浮点数.

证明: 非零数$x=\pm p^J\sum_{k=1}^{t}d_kp^{-k}$中, 正负有2种; d_1有$p-1$种取法, $d_2,\cdots,d_t$有p种取法, J有$L+U+1$种取法. 再加上0, 一共有$2(p-1)p^{t-1}(L+U+1)+1$个浮点数. □

下面看浮点数的运算: 计算机采取断位与舍入是不同的. 作加减运算时, 首先要进行对阶(把小数点对齐, 使其阶相等), 对阶方法是把阶小的数的尾数右移, 每移一位其阶就加1直到两数的阶相等. 最后将对阶后的两数相加或相减. 对乘法运算时不需要对阶.

做三个以上数的加法运算时, 需要考虑相加的两个同号数的阶数尽量相近.

定理 1.2.2: 浮点运算的误差

$fl(x+y)=(x+y)(1+\varepsilon_1), fl(x-y)=(x-y)(1+\varepsilon_2), fl(xy)=(xy)(1+\varepsilon_3), fl(\frac{x}{y})=\frac{x}{y}(1+\varepsilon_4)$, 其中$|\varepsilon_i|=eps$, 采用舍入的方法时$eps=\frac{1}{2}p^{1-t}$, p为进制.(断位就没有$\frac{1}{2}$)

习题

1. 计算球的体积$V=\frac{4}{3}\pi r^3$时, 为使V的相对误差不超过0.3%, 半径r的误差应满足_____.
2. 如果用$\pi^*=\frac{2198810}{699903}$作为$\pi$的近似值, 分析$\pi^*$有几位有效数字?
3. 给出在字长为十进制二位计算机上用浮点运算分别从左到右与从右到左计算

$$1+0.4+0.3+0.2+0.04+0.03+0.02+0.01$$

的结果, 并对你的结果作出解释.

CHAPTER 2

范数理论

我们知道, 在实数范围内可以用绝对值来刻画两个数之间的“距离”. 如果要刻画两个向量或者两个矩阵之间的距离, 需要引入“范数”的概念. 当然, $\mathbb{R}^n$与赋予$\mathbb{R}^n$上的范数的度量空间也有邻域、开集、收敛性、连续性的概念.

在用计算机求解线性方程组时, 会对算法中的矩阵进行各种分析, 比如, 如果系数矩阵“差不多接近奇异”, 那么一个线性方程组的解可能很“不稳定”, 即对系数矩阵或者右端向量作一点扰动以后, 得到的解与真解差别很大(计算机无法精确存储大部分的数据). 此时需要引入矩阵范数来对矩阵进行度量.

求线性方程组几乎贯穿了整个数值计算, 比如求特征值的迭代法、最小二乘法、微分方程数值解等都需要用到求线性方程组的思想, 当然也涉及到范数理论, 所以这部分可谓是最重要的.

§2.1 向量范数

下面的定义的$\mathbb{R}^n$可改为$\mathbb{C}^n$.

定义 2.1.1: 向量范数

设$x \in \mathbb{R}^n$ 若在$\mathbb{R}^n$上定义了个实值函数$\|x\|$, 满足下面三个条件:
(1)非负性: $\|x\| > 0, \forall x \in \mathbb{R}^n, x \neq 0$;
(2)齐次性: $\|\lambda x\| = |\lambda| \|x\|, \forall x \in \mathbb{R}^n, \lambda \in \mathbb{R}$.
(3)三角不等式: $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{R}^n$,
则称$\|x\|$为x的一种**范数(norm)**, 并说$\mathbb{R}^n$是赋以范数$\|x\|$的**赋范线性空间**.

记$f_p(x) = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$, 根据民科夫斯基(Minkowski)不等式可知$f_p(x)$是$\mathbb{R}^n$的一种范数.

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{1/p},$$

定义 2.1.2: l_p范数

把上述$f_p(x)$称为向量x的l_p范数, 记作

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

常用的几种范数是:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$
$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} = (x^T x)^{1/2},$$
$$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p = \max_{1\le i\le n} |x_i|.$$

在$\mathbb{R}^n$中的内积记作$(x,y) = y^T x$. 通常把$\|x\|_2 = \sqrt{(x,x)}$称为**Euclid范数(Euclid长度)**. 同样在$\mathbb{C}^n$中的内积记作$(x,y) = y^H x$, 用Euclid范数可以推出Cauchy-Schwarz不等式:

$$|(x,y)| \le \|x\|_2 \|y\|_2, \forall x, y \in \mathbb{R}^n$$

当且仅当x, y线性相关时等号成立.

定理 2.1.1

$\mathbb{R}^n$中的向量范数都是等价的, 即任两种范数$\|x\|_\alpha, \|x\|_\beta$, 总存在两个与$x$无关的正常数$C_1, C_2 \in \mathbb{R}$使得

$$C_1\|x\|_\beta \le \|x\|_\alpha \le C_2\|x\|_\beta, \forall x \in \mathbb{R}^n.$$

证明: 只需证任一种范数$\|x\|_\alpha$与$\|x\|_2$等价. 由Cauchy-Schwarz不等式,

$$\|x\|_\alpha \le \sum_{i=1}^{n} |x_i| \|e_i\|_\alpha \le M\|x\|_2,$$

则$|\|x\|_\alpha - \|y\|_\alpha \le \|x-y\|_\alpha \le M\|x-y\|_2$. 所以范数$\|x\|_\alpha$关于$l_2$范数是$x$的连续函数.

由于$\mathbb{R}^n$中单位球面$S = \{x | \|x\|_2 = 1, x \in \mathbb{R}^n\}$是有界闭集, 则$\|x\|_\alpha$可在$S$上达到最大值$C_2$与最小值$C_1$, 对任意的$x \in \mathbb{R}^n$, 令$y = \dfrac{x}{\|x\|_2}$, 则$\|x\|_\alpha = \|x\|_2\|y\|_\alpha$. 因为$\|y\|_2 = 1$, 所以$C_1 \le \|y\|_\alpha \le C_2$, 所以$C_1\|x\|_2 \le \|x\|_\alpha \le C_2\|x\|_2$. □

推论 2.1.2

$\mathbb{R}^n$中的向量范数$\|x\|_\alpha$关于任意一个范数$\|x\|_\beta$是x的一致连续函数, 即对$\forall \varepsilon > 0$, 存在$\delta > 0$, 当$\|y-x\|_\beta < \delta$时, 恒有$|\|y\|_\alpha - \|x\|_\alpha| < \varepsilon$.

定理 2.1.3

$\mathbb{R}^n$空间中的l_1, l_2, l_∞范数满足如下的关系式:

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

习题:

1. 证明$\lim\limits_{p\to\infty} \|x\|_p = \max\limits_{1\le i\le n} |x_i|$.
2. 设$\mathbf{Q} \in \mathbb{R}^{n\times n}$是正交矩阵, $x \in \mathbb{R}^n$, 则
$$\|\mathbf{Q}x\|_2 = \|x\|_2.$$
即正交变换是保持二范数的. (正交变换可以看作旋转与反射, 这两种操作不会改变向量的(欧氏)长度.)

3. 设A是给定的n阶实对称正定矩阵, $x \in \mathbb{R}^n$, 则$\|x\|_A = (x^T Ax)^{1/2}$是$\mathbb{R}^n$一种向量范数.
4. 设$v \in \mathbb{R}^n$, 证明或证伪
$$\|v\|_1 \|v\|_\infty \leq \frac{1+\sqrt{n}}{2}\|v\|_2^2.$$
5. **(2016Team, 3)**设$A = (a_{ij})_{i,j=1}^{m,n}$是$m \times n$矩阵, 秩为$r \leq n-1$, 且矩阵中的元素均为整数, 且模不超过$H$, 即
$$|a_{ij}| \leq H, 1 \leq i \leq m, 1 \leq j \leq n.$$
证明: 存在一个非零整向量$\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{Z}^n$, 使得$A\boldsymbol{x} = 0$且
$$\|\boldsymbol{x}\|_\infty \leq (2nH)^{n-1}.$$
(提示：Cramer法则, 考虑A的极大线性无关组.)

§2.2 矩阵范数

2.2.1 矩阵范数的定义、相容性、从属性

定义 2.2.1: 矩阵范数

设$\mathbb{R}^{n\times n}$表示全体$n\times n$阶实矩阵构成的线性空间,$A\in\mathbb{R}^{n\times n}$, 若在$\mathbb{R}^{n\times n}$中定义了一个实值函数, 记作$\|A\|$, 满足如下条件:

(1)$\|A\|>0,\forall A\in\mathbb{R}^{n\times n},A\neq O$,

(2)$\|\lambda A\|=|\lambda|\|A\|,\forall A\in\mathbb{R}^{n\times n},\lambda\in\mathbb{R}$,

(3)$\|A+B\|\leq\|A\|+\|B\|,\forall A,B\in\mathbb{R}^{n\times n}$,

(4)$\|AB\|\leq\|A\|\cdot\|B\|,\forall A,B\in\mathbb{R}^{n\times n}$.

则称$\|A\|$为矩阵A的一种**范数**.

定义 2.2.2: 相容

若在$\mathbb{R}^{m\times n}$中定义了一种矩阵范数$\|\cdot\|_\beta$, 在$\mathbb{R}^n$中定义了一种向量范数$\|\cdot\|_\alpha$, 若不等式

$$\|\mathbf{A}x\|_\alpha\leq\|\mathbf{A}\|_\beta\|x\|_\alpha$$

对任意$\mathbf{A}\in\mathbb{R}^{m\times n},x\in\mathbb{R}^n$恒成立, 则说上述矩阵范数和向量范数**相容**.

定理 2.2.1

设$\|\cdot\|_\beta$是$\mathbb{R}^{n\times n}$中任意一种矩阵范数, 则在$\mathbb{R}^n$中存在向量范数$\|\cdot\|_\alpha$, 使得$\|\cdot\|_\beta$和$\|\cdot\|_\alpha$相容.

证明: 设$x=[x_1,\cdots,x_n]^T\in\mathbb{R}^n$, 取

$$\|x\|_\alpha=\left\|\begin{pmatrix}x_1&0&\cdots&0\\x_2&0&\cdots&0\\\vdots&\vdots&&\vdots\\x_n&0&\cdots&0\end{pmatrix}\right\|$$

就是这样的一种向量范数. 3个条件不难验证, 相容性也不难验证. □

2.2.2 几种常用的矩阵范数

定理 2.2.2: 矩阵的1,2,∞范数

把前面定理$\|A\|=\max\limits_{\|x\|_\alpha=1}\|Ax\|_\alpha$中的向量范数取特定的$l_p$范数, 可以得到具体的矩阵范数, 相应记为$\|A\|_p$. 记$A=(a_{ij})_{m\times n}$, 则

$$\|A\|_1=\max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{ij}|,\qquad\text{(列和的max)}$$

$$\|A\|_2=\sqrt{\lambda_1},\qquad\lambda_1\text{是}A^TA\text{的最大特征值}$$

$$\|A\|_\infty=\max_{1\leq i\leq m}\sum_{j=1}^{n}|a_{ij}|,\qquad\text{(行和的max)}.$$

$\|A\|_2$又称为**谱范数**.

注: 张强老师记法: “1”是竖着的, 所以是列和的max; “∞”是横着的, 所以是行和的max.

证明: (1)对于1范数, 记$A=(a_1,\cdots,a_n), a_j=(a_{1j},\cdots,a_{mj})^T, j=1,\cdots,n$. 则$\max\limits_{1\le j\le n}\sum\limits_{i=1}^{m}|a_{ij}|=\max\limits_{1\le j\le n}\|a_j\|_1$. 对任意的$x\in\mathbb{R}^n$, 有

$$\begin{aligned}\|Ax\|_1&=\|x_1a_1+x_2a_2+\cdots+x_na_n\|_1\\&\le|x_1|\|a_1\|_1+\cdots+|x_n|\|a_n\|_1\\&\le(|x_1|+\cdots+|x_n|)\max_{1\le j\le n}\|a_j\|_1\\&=\|x\|_1\max_{1\le j\le n}\|a_j\|_1,\end{aligned}$$

当$\|x\|_1=1$时, 有$\|Ax\|_1\le\max\limits_{1\le j\le n}\|a_j\|_1\max\limits_{1\le j\le n}\sum\limits_{i=1}^{m}|a_{ij}|$.

另一方面, 下面我们要让上式的等号成立, 设k是列和最大时的列数, 即$\max\limits_{1\le j\le n}\|a_j\|_1=\|a_k\|_1$. 取$x=e_k$, 则$\|e_k\|_1=1$, 且$\|Ae_k\|_1=\|a_k\|_1=\max\limits_{1\le j\le n}\sum\limits_{j=1}^{m}|a_{ij}|$, 故$\|A\|_1=\max\limits_{\|x\|_1=1}\|Ax\|_1=\max\limits_{1\le j\le n}\sum\limits_{j=1}^{m}|a_{ij}|$. □

(2)对于2范数, 由于矩阵A^TA是实对称正定或半正定的, 特征值皆为实数且非负, 因此A^TA的最大特征值λ_1存在. 由于

$$\|Ax\|_2=\sqrt{(Ax,Ax)}=\sqrt{(Ax)^T(Ax)}=\sqrt{x^TA^TAx},$$

根据实二次型的极性, $\max\limits_{\|x\|_2=1}x^TA^TAx=\lambda_1$, 便得到$\|A\|_2=\max\limits_{\|x\|_2=1}\|Ax\|_2=\sqrt{\lambda_1}$. □

(3)对于∞范数, 类似(1), 容易证明

$$\max_{\|x\|_\infty=1}\|Ax\|_\infty\le\max_{1\le i\le m}\sum_{j=1}^{n}|a_{ij}|,$$

再证等号成立, 要构造一个可以使得它取最大值的项. 记k为行和最大时的行数注意$a_{kj}=|a_{kj}|\mathrm{sgn}(a_{kj})$, 取

$$x_0=(x_1^{(0)},x_2^{(0)},\cdots,x_n^{(0)})^T,\text{其中}x_j^{(0)}=\begin{cases}1, & a_{kj}\ge0,\\-1, & a_{kj}<0.\end{cases},$$

则$\|x_0\|_\infty=1$. 以下步骤省略(证等号成立). □

注: $\|A\|_1=\|A^T\|_\infty$.

定义 2.2.3: Frobenius范数

把矩阵A的Frobenius范数定义为

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

即所有元素的平方和再开方. 容易验证它是个矩阵范数.

注: 容易证明这是矩阵范数. 此外由于n阶单位矩阵$\mathbb{I}$满足$\|\mathbb{I}\|_F=\sqrt{n}$, 故它不从属于任何范数.

定理 2.2.3

Frobenius范数与l_2范数相容.

证明: 注意到

$$\|\mathbf{A}x\|_2^2=\sum_{i=1}^{m}\left|\sum_{j=1}^{n}a_{ij}x_j\right|^2\le\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|a_{ij}|^2\right)\left(\sum_{j=1}^{n}|x_j|^2\right)=\|\mathbf{A}\|_F^2\|x\|_2^2,\forall x\in\mathbb{R}^n.$$

因此$\|\mathbf{A}x\|_2 \le \|\mathbf{A}\|_F\|x\|_2$. □

引理 2.2.4

$\mathrm{tr}(A^TA)$是A所有元素的平方和, 即$\mathrm{tr}(A^TA) = \|A\|_F$.

证明: 高等代数简单习题. □

定理 2.2.5

设A是n阶实对称方阵, 特征值为$\lambda_1, \cdots, \lambda_n$, 则$\|A\|_F^2 = \lambda_1^2 + \cdots + \lambda_n^2$.

证明: 由于A是对称的, 则$\lambda_i(A^TA) = \lambda_i(A^2) = \lambda_i^2(A)$, 所以

$$\lambda_1^2 + \cdots + \lambda_n^2 = \sum_{i=1}^{n} \lambda_i(A^TA) = \mathrm{tr}(A^TA) = \|A\|_F^2$$

注: Frobenius范数与谱范数满足

$$\|A\|_2 \le \|A\|_F \le \sqrt{\min\{m,n\}}\|A\|_2.$$

左边利用相容性可证, 右边是因为(注意前面的引理)

$$\|A\|_2^2 = \max_{1\le i\le n} \lambda_1 \ge \frac{1}{n}(\lambda_1 + \cdots + \lambda_n) = \frac{1}{n}\mathrm{tr}(A^TA) = \frac{1}{n}\|A\|_F^2.$$
$$\|A\|_2^2 = \max_{1\le i\le m} \mu_1 \ge \frac{1}{m}(\mu_1 + \cdots + \mu_n) = \frac{1}{m}\mathrm{tr}(AA^T) = \frac{1}{m}\|A\|_F^2.$$

其中$\{\lambda_i\}$是A^TA的特征值, $\{\mu_i\}$是AA^T的特征值.

定理 2.2.6

设$A \in \mathbb{C}^{m\times n}$, 在酉变换下, 谱范数$\|\cdot\|_2$和F-范数$\|\cdot\|_F$保持不变, 即设$\mathbf{P} \in \mathbb{C}^{m\times m}, \mathbf{Q} \in \mathbb{C}^{n\times n}, \mathbf{P}^H\mathbf{P} = \mathbf{I}, \mathbf{Q}^H\mathbf{Q} = \mathbf{I}$, 则

$$\|\mathbf{A}\|_2 = \|\mathbf{AQ}\|_2 = \|\mathbf{PA}\|_2,$$
$$\|\mathbf{A}\|_F = \|\mathbf{AQ}\|_F = \|\mathbf{PA}\|_F.$$

证明: (1)由于$\mathbf{Q}^H\mathbf{Q} = \mathbf{I}$, 则

$$\|\mathbf{AQ}\|_2 \le \|\mathbf{A}\|_2\|\mathbf{Q}\|_2 = \|\mathbf{A}\|_2,$$

以及

$$\|\mathbf{A}\|_2 = \|\mathbf{AQQ}^H\|_2 \le \|\mathbf{AQ}\|_2\|\mathbf{Q}^H\|_2 = \|\mathbf{AQ}\|_2,$$

所以$\|\mathbf{A}\|_2 = \|\mathbf{AQ}\|_2$, 同理$\|\mathbf{A}\|_2 = \|\mathbf{PA}\|$.

(2)设$x \in \mathbb{C}^n$, 则

$$\|\mathbf{Q}x\|_2^2 = x^H\mathbf{Q}^H\mathbf{Q}x = x^Hx = \|x\|_2^2.$$

这样在酉变换下向量的Euclid范数保持不变. (可以看作向量作了旋转以后, 到原点的距离不变). 记

$$\mathbf{A} = [a_1, a_2, \cdots, a_n], \mathbf{QA} = [\mathbf{Q}a_1, \cdots, \mathbf{Q}a_n],$$

则

$$\|\mathbf{A}\|_2^2 = \sum_{j=1}^{n} \|a_j\|_2^2 = \sum_{j=1}^{n} \|\mathbf{Q}a_j\|_2^2 = \|\mathbf{QA}\|_F^2.$$

因此

$$\|\mathbf{A}\| = \|\mathbf{QA}\|_F.$$

另一个同理. □

2.2.3 矩阵范数的等价性

类似向量范数, 矩阵范数也有等价性.

引理 2.2.7

设$A \in \mathbb{R}^{n\times n}$. 则$\|A\|_M = n \max\limits_{1\le i,j\le n} |a_{ij}|$是矩阵范数.

定理 2.2.8

设$\|\cdot\|_\alpha, \|\cdot\|_\beta$是两个矩阵范数, 则存在正常数$c_1, c_2 \in \mathbb{R}$使得$c_1\|A\|_\beta \le \|A\|_\alpha \le c_2\|A\|_\beta, \forall A \in \mathbb{R}^{n\times n}$.

证明: 只需证明任意一种矩阵范数$\|\cdot\|_\alpha$, 存在正常数$d_1, d_2 \in \mathbb{R}$, 使得

$$d_1\|A\|_M \le \|A\|_\alpha \le d_2\|A\|_M.$$

其中$\|\cdot\|_M$的定义如前一引理. 证明过程类似于向量范数的等价性. □

习题:

1. 设$A \in \mathbb{R}^{n\times n}$. 则$\|A\|_M = n \max\limits_{1\le i,j\le n} |a_{ij}|$是矩阵范数.
2. 设$A \in \mathbb{C}^{n\times n}$, 则$\|A\|_2^2 \le \|A\|_1\|A\|_\infty$.
3. 设$A \in \mathbb{R}^{m\times n}, \operatorname{rank} A = n$, 则$\|A(A^TA)^{-1}A^T\|_2 = 1$.
4. 设$A \in \mathbb{R}^{m\times n}$, 证明:
   (1) $\max\limits_{i,j}|a_{ij}| \le \|A\|_2 \le \sqrt{mn}\max\limits_{i,j}|a_{ij}|$.
   (2) $\frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{m}\|A\|_\infty$.
   (3) $\frac{1}{\sqrt{m}}\|A\|_1 \le \|A\|_2 \le \sqrt{n}\|A\|_1$.
5. 设$0 \ne s \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$, 则$\left\|\mathbf{A}\left(\mathbb{I} - \frac{ss^T}{s^Ts}\right)\right\|_F^2 = \|\mathbf{A}\|_F^2 - \frac{\|\mathbf{A}s\|_2^2}{s^Ts}$.
6. 设$u \in \mathbb{R}^m, v \in \mathbb{R}^n$. 若$E = uv^T$, 则$\|E\|_F = \|E\|_2 = \|u\|_2\|v\|_2$, 且$\|E\|_\infty \le \|u\|_\infty\|v\|_1$.
7. 设$A \in \mathbb{R}^{m\times n}$. B是A的子矩阵, 即选择A的其中μ行与ν列构成的$\mu\times\nu$矩阵, 其中$\mu \le m, \nu \le n$. $1 \le p \le \infty$, 则$\|B\|_p \le \|A\|_p$. (提示: 把A删掉一行/一列会怎样?)

§ 2.3 谱半径

2.3.1 谱半径的定义与基本性质

定义 2.3.1: 谱半径

设$\boldsymbol{A} \in \mathbb{C}^{n\times n}$, 把$\boldsymbol{A}$的所有特征值的最大模称为$\boldsymbol{A}$的**谱半径(spectral radius)**, 记作

$$\rho(\boldsymbol{A}) = \max_{1\le i\le n}|\lambda_i|$$

其中λ_i是$\boldsymbol{A}$的特征值.

注: 谱半径的引入可以把特征值转化为求范数. 回顾矩阵2范数的定义, 可知$\rho(\boldsymbol{A}^T\boldsymbol{A}) = \|\boldsymbol{A}\|_2^2$.

定理 2.3.1

对于$\mathbb{C}^{n\times n}$中任何矩阵范数$\|\cdot\|$, 恒有$\rho(\boldsymbol{A}) \le \|\boldsymbol{A}\|$.

证明: 由前面的定理, 存在向量范数$\|\cdot\|_\alpha$使得$\|\boldsymbol{A}x\|_\alpha \leq \|\boldsymbol{A}\|\|x\|_\alpha$, 对任意$\boldsymbol{A} \in \mathbb{C}^{n\times n}, x \in \mathbb{C}^n$成立. 设$\lambda$是$\boldsymbol{A}$的特征值, x_λ是对应的特征向量, 则

$$\|\boldsymbol{A}x_\lambda\|_\alpha = |\lambda|\|x_\lambda\|_\alpha.$$

因此有

$$|\lambda| \leq \|\boldsymbol{A}\|, \forall \lambda \text{是} \boldsymbol{A} \text{的特征值.}$$

所以$\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|$. □

定理 2.3.2

设$A \in \mathbb{C}^{n\times n}$, 对于任意给定正数$\varepsilon > 0$, 在$\mathbb{C}^{n\times n}$中至少存在一种矩阵范数$\|\cdot\|_\beta$使得

$$\|A\|_\beta \leq \rho(A) + \varepsilon.$$

证明: $\boldsymbol{A} \in \mathbb{C}^{n\times n}$相似于Jordan标准型$\boldsymbol{J}$, 即存在可逆矩阵$\boldsymbol{P}$使得$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{J}$. 令$\boldsymbol{D} = \mathrm{diag}(1, \varepsilon, \cdots, \varepsilon^{n-1})$, 并设$\tilde{\boldsymbol{J}} = \boldsymbol{D}^{-1}\boldsymbol{J}\boldsymbol{D}$, 则$\tilde{\boldsymbol{J}}$相当于把非对角元素1换成$\varepsilon$. 所以

$$\tilde{\boldsymbol{J}} = \boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q}, \boldsymbol{Q} = \boldsymbol{P}\boldsymbol{D}.$$

由于

$$\|\boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q}\|_1 = \|\tilde{\boldsymbol{J}}\|_1 \leq \rho(\boldsymbol{A}) + \varepsilon,$$

且$\|\boldsymbol{A}\|_\beta = \|\boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q}\|_1$构成矩阵范数, 则欲证不等式成立. □

2.3.2 向量和矩阵序列的极限

一列向量中每个分量取极限得到的新向量就是这一列向量的极限.

定理 2.3.3

$\mathbb{C}^n$空间中向量序列$\{x_k\}$收敛于x的充分必要条件是对任意一种范数$\|\cdot\|$, $\lim\limits_{k\to\infty}\|x_k - x\| = 0$.

证明: 对l_∞范数验证即可. □

定理 2.3.4

$\mathbb{C}^n$空间中矩阵序列$A_1, \cdots, A_k, \cdots$收敛于矩阵A的充分必要条件是对任意一种矩阵范数$\|\cdot\|$, $\lim\limits_{k\to\infty}\|A_k - A\| = 0$.

习题:

1. 证明: 对$\mathbb{C}^{n\times n}$中任意范数$\|\cdot\|$, 都有$\|I\| \geq 1$.
2. 对$\mathbb{C}^{n\times n}$中任意范数$\|\cdot\|$, 若A是非奇异矩阵, 证明$\|A^{-1}\| \geq \dfrac{1}{\|A\|}$.
3. 设$A \in \mathbb{C}^{n\times n}$, 则$\lim\limits_{k\to\infty} A^k = O$的充分必要条件是$\rho(A) < 1$. (提示: 反证法, 结合定理2.3.4和定理2.3.2.)

§ 2.4 条件数与摄动理论初步

定义 2.4.1: 条件数

对于A, 定义它所用范数的**条件数**为$\mathrm{cond}(A) = \|A\| \cdot \|A^{-1}\|$, 当$A$为谱范数时, 称$\mathrm{cond}(A)_2 \triangle K(A) = \|A\|_2\|A^{-1}\|_2$为**谱条件数**.

定理 2.4.1

设A是n阶非奇异实对称矩阵, 则$\|A^{-1}\|_2 = \dfrac{1}{|\lambda_n|}$, 其中$\lambda_n$是$A$的按绝对值最小的特征值.

证明: 由条件, A^TA是正定矩阵, 它的最大特征值为λ_1^2, 最小特征值为λ_n^2. 则$(A^TA)^{-1}$的最大特征值为λ_n^{-2}, 最小特征值为λ_1^{-2}. 所以

$$\|A^{-1}\|_2 = \rho((A^{-1})^TA^{-1}) = \rho((A^TA)^{-1}) = \frac{1}{|\lambda_n|}.$$

□

推论 2.4.2

设A是n阶非奇异实对称矩阵, λ_1, λ_n分别是A的按绝对值最大,最小的特征值. 则$K(A) = \dfrac{|\lambda_1|}{|\lambda_n|}$.

注: 一般来说条件数的求解需要求逆矩阵, 但是谱条件数用特征值就可以表示出来, 所以谱条件数比较常用.

在讲线性方程组的扰动问题之前, 先介绍一个很常用的Banach引理:

引理 2.4.3: Banach

设$B \in \mathbb{C}^{n\times n}$, $\rho(B) < 1$, 则矩阵$I \pm B$非奇异且对任意满足$\|I\| = 1$的范数$\|\cdot\|$, 若有$\|B\| = 1$, 则

$$\frac{1}{1+\|B\|} \le \|(I \pm B)^{-1}\| \le \frac{1}{1-\|B\|}.$$

证明: 由于$\rho(B) < 1$, 则$I \pm B$的特征值位于区域$\{\lambda : |\lambda - 1| < 1\}$中, 所以$I \pm B$特征值非零, 从而是非奇异的.

由于$I = (I+B)^{-1}(I+B)$, 则

$$\begin{aligned} 1 = \|I\| &\le \|(I+B)^{-1}\|\|I+B\| \\ &\le \|(I+B)^{-1}\|(\|I\| + \|B\|) \\ &= \|(I+B)^{-1}\|(1+\|B\|), \end{aligned}$$

这就是左边的不等式.

又由于$I = (I+B)^{-1}(I+B) = (I+B)^{-1} + (I+B)^{-1}B$, 则

$$\|(I+B)^{-1}\| \le \|I\| + \|(I+B)^{-1}\|B\|,$$

所以$(1-\|B\|)\|(I+B)^{-1}\| \le 1$, 这就是右边不等式. □

定理 2.4.4

设矩阵A非奇异, 线性方程组$Ax = b$的右端b有扰动δ_b, 解变为$x + \delta_x$, 即

$$Ax = b\text{且}A(x+\delta_x) = b + \delta_b$$

把$\dfrac{\|\delta_x\|}{\|x\|}$称为对$y$作扰动的相对误差. 则有相对误差估计式

$$\frac{\|\delta_x\|}{\|x\|} \le \mathrm{cond}(A)\frac{\|\delta_b\|}{\|b\|}.$$

证明: 由条件, $A\delta_x = \delta_b$, 所以

$$\|\delta_x\| = \|A^{-1}\delta_b\| \le \|A^{-1}\|\|\delta_b\|,$$

由于$\|b\| \leq \|A\|\|x\|$, 两式相乘可得

$$\frac{\|\delta_x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta_b\|}{\|b\|} = \mathrm{cond}(A)\frac{\|\delta_b\|}{\|b\|}.$$

□

定理 2.4.5

设线性方程组$Ax = b$的系数矩阵A有扰动δ_A(叫摄动矩阵), 其中$\|A^{-1}\delta_A\| < 1$, $\|A^{-1}\|\|\delta_A\| < 1$, 解变为$x + \delta_x$, 则有相对误差估计式$\dfrac{\|\delta_x\|}{\|x\|} \leq \dfrac{cond(A)\dfrac{\|\delta_A\|}{\|A\|}}{1 - \mathrm{cond}(A)\dfrac{\|\delta_A\|}{\|A\|}}$.

证明: 注意$A + \delta A = A(I + A^{-1}\delta A)$, 所以可以化得$\delta x = -(I + A^{-1}\delta A)^{-1}A^{-1}\delta Ax$, 根据Banach引理, 有

$$\|\delta x\| \leq \|(I + A^{-1}\delta_A)^{-1}\| \cdot \|A^{-1}\delta_A\| \cdot \|x\| \leq \frac{\|A^{-1}\delta_A\|}{1 - \|A^{-1}\delta_A\|}\|x\|,$$

由于$\|A^{-1}\|\|\delta_A\| < 1$, 则

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\delta_A\|}{1 - \|A^{-1}\|\|\delta_A\|},$$

作简单化简就证完. □

当矩阵A的条件数很大时, 通常称A是**坏条件的(ill-conditioned)**. 有些条件并不坏, 但是初始数据误差和计算过程产生的误差的传播积累对计算的影响较大, 那么就是数值不稳定的.

条件数的定义依赖于范数. 我们可以证明, 任意两个条件数$\mathrm{cond}_\alpha(\cdot)$与$\mathrm{cond}_\beta(\cdot)$是等价的, 即存在$c_1, c_2$使得

$$c_1\mathrm{cond}_\alpha(A) \leq \mathrm{cond}_\beta(A) \leq c_2\mathrm{cond}_\alpha(A), A \in \mathbb{R}^{n\times n}.$$

例如, 对于n阶方阵A, 有

$$\begin{aligned}
&\frac{1}{n}\kappa(A) \leq \mathrm{cond}_1(A) \leq n\kappa(A),\\
&\frac{1}{n}\mathrm{cond}_\infty(A) \leq \kappa(A) \leq n\mathrm{cond}_\infty(A),\\
&\frac{1}{n^2}\mathrm{cond}_1(A) \leq \mathrm{cond}_\infty(A) \leq n^2\mathrm{cond}_1(A),
\end{aligned}$$

所以如果矩阵在$\alpha-$范数下是坏条件, 那么在$\beta-$范数下也是坏条件的.

注: 对任意l^p范数都有$\mathrm{cond}_p(A) \geq 1$. 小条件数的矩阵就是好的, 例如Euclid范数下的正交矩阵Q有完美的条件数, 因为$\kappa(Q) = \|Q\|_2\|Q^{-1}\|_2 = 1$.

例 2.4.1 (行列式与“靠近奇异的程度”无关) 我们知道, 如果矩阵$\boldsymbol{A}$满足$\det \boldsymbol{A} = 0$, 那么$\boldsymbol{A}$是奇异的. 所以很容易联想到如果$\det \boldsymbol{A} \approx 0$是否会有$\boldsymbol{A}$“接近奇异”?

考虑矩阵

$$\boldsymbol{B}_n = \begin{pmatrix} 1 & -1 & \cdots & -1 \\ 0 & 1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

它的行列式为1, 但是$\mathrm{cond}_\infty(\boldsymbol{B}_n) = n \cdot 2^{n-1}$. 所以它是坏条件的.

另一方面, 好条件的矩阵的行列式可能很小, 比如

$$\boldsymbol{D}_n = \mathrm{diag}(10^{-1}, \cdots, 10^{-1}) \in \mathbb{R}^{n\times n}$$

满足$\mathrm{cond}_p(\boldsymbol{D}_n)=1$, 但是$\det(\boldsymbol{D}_n)=10^{-n}$. □

例 2.4.2 方程组$\mathbb{A}x=b$, 其中$\mathbb{A}$是$n\times n$阶对称且非奇异矩阵, 设$\mathbb{A}$有误差$\delta_{\mathbb{A}}$, 则原方程组变成

$$(\mathbb{A}+\delta_{\mathbb{A}})(x+\delta_x)=b$$

其中δx是解的误差向量. 证明:

$$\frac{\|\delta_x\|_2}{\|x+\delta_x\|_2}\leq\left|\frac{\lambda_1}{\lambda_n}\right|\frac{\|\delta_{\mathbb{A}}\|_2}{\|\mathbb{A}\|_2},$$

其中λ_1,λ_n分别是$\mathbb{A}$的按模最大和最小特征值.

证明: 由条件可知

$$\mathbb{A}\delta_x+\delta_{\mathbb{A}}(x+\delta_x)=0,$$

因此

$$\delta_x=-\mathbb{A}^{-1}\delta_{\mathbb{A}}(x+\delta_x),$$

由$\mathbb{A}$对称非奇异, 则$\|A\|_2=|\lambda_1|,\|A^{-1}\|_2=|\lambda_n|^{-1}$. 从而

$$\|\delta_x\|_2\leq\|\mathbb{A}^{-1}\|_2\|\delta_{\mathbb{A}}\|_2\|x+\delta_x\|_2=\frac{1}{|\lambda_n|}\|\delta_{\mathbb{A}}\|_2\|x+\delta_x\|_2=\frac{|\lambda_1|}{|\lambda_n|}\|\delta_{\mathbb{A}}\|_2\|x+\delta_x\|_2\frac{1}{\|A\|_2}.$$

稍作整理即得欲证结论. □

条件数反映了线性方程组$Ax=b$的解对于A和b的扰动的稳定程度.

例 2.4.3 (Wilson) 设线性方程组$Ax=b$的系数矩阵A和常数向量b分别为

$$A=\begin{pmatrix}10&7&8&7\\7&5&6&5\\8&6&10&9\\7&5&9&10\end{pmatrix},b=\begin{pmatrix}32\\23\\33\\31\end{pmatrix}.$$

方程组$Ax=b$的解为$(x_1,x_2,x_3,x_4)^T=(1,1,1,1)^T$. 但是如果把$b$扰动为

$$b+\delta_b=(32.1,22.9,33.1,30.9)^T,$$

则解会变为

$$x+\delta x=(9.2,-12.6,4.5,-1.1)^T.$$

此例表明对常数项b的微小**相对扰动**

$$\frac{\|\delta_b\|}{\|b\|}\approx\frac{1}{200},$$

线性方程组解的相对改变量为$\dfrac{\|\delta x\|}{\|x\|}\approx 2000$. 这充分表明了线性方程组的解关于常数项的扰动非常敏感.

类似地扰动系数阵A也成立, 例如让b不变, A变为

$$A+\delta A=\begin{pmatrix}10&7&8.1&7.2\\7.08&5.04&6&5\\8&5.98&9.89&9\\6.99&4.99&9&9.98\end{pmatrix}$$

则解变为

$$x+\delta x=(-81,137,-34,22)^T.$$

与原来的$x=(1,1,1,1)^T$相比相差更大了. 事实上, 由于A是正定对称阵, $\lambda_1(A)=30.2887, \lambda_4(A)=0.01015$, 它的条件数是

$$K(A)=\frac{\lambda_1(A)}{\lambda_4(A)}\approx 2984.$$

它很大! 难怪A,b的微小扰动会引起解的巨大变化. 例子充分表明, 在求解线性方程组时, 系数阵的条件数很重要. (在计算机存储数据的时候不可能精确存储一个有理数, 会有误差)

2.4.1 Sherman-Morrison-Woodbury公式

恒等式

$$\boldsymbol{B}^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{B}^{-1}(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{A}^{-1}$$

可以显示一个矩阵发生变化之后, 其逆的变化情况.

Sherman-Morrison-Woodbury公式可以给出矩阵$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{V}^T)$的逆的表示方式, 其中$\boldsymbol{A}\in\mathbb{R}^{n\times n}, \boldsymbol{U},\boldsymbol{V}\in\mathbb{R}^{n\times k}$, 而且$\boldsymbol{A}$非奇异, $(\boldsymbol{I}+\boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U})$也是非奇异:

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{V}^T)^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I}+\boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}.$$

对一个矩阵作秩-k修正也会让逆得到一个秩-k修正.

当$k=1$时的情况非常好用. 如果$\boldsymbol{A}\in\mathbb{R}^{n\times n}$, $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n, \alpha=1+\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}\neq 0$, 则

$$(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^T)^{-1}=\boldsymbol{A}^{-1}-\frac{1}{\alpha}\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}.$$

这个公式叫**Sherman-Morrison公式**.

习题:

1. 若$B\in\mathbb{C}^{n\times n}$, 则级数$\sum_{k=0}^{\infty}B^k$收敛的充分必要条件是$\rho(B)<1$, 且收敛时极限为$(I-B)^{-1}$.
2. 设A,E是同阶矩阵, 设A非奇异, $r\equiv\|A^{-1}E\|<1$, 则$A+E$非奇异, 且

$$\|(A+E)^{-1}-A^{-1}\|\leq\|A^{-1}\|\frac{r}{1-r}.$$

   (提示: 用Banach引理)
3. **(2014Team, 5)** (1)设$x_0=0$, 请把

$$x_k=2x_{k-1}+b_k, k=1,2,\cdots,n.$$

   写成矩阵形式$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$. 对于$b_1=-\frac{1}{3}, b_k=(-1)^k, k=2,3,\cdots,n$, 验证$x_k=\frac{(-1)^k}{3}, k=1,2,\cdots,n$是精确解.

   (2)求A^{-1}, 并计算A在l^1范数下的条件数.
4. **(Golub, 2.6.1题)**若$\|\boldsymbol{I}\|\geq 1$, 则$\mathrm{cond}(\boldsymbol{A})\geq 1$.
5. **(Golub, 2.6.2题)**对给定的范数, 证明

$$\mathrm{cond}(\boldsymbol{A}\boldsymbol{B})\leq\mathrm{cond}(\boldsymbol{A})\mathrm{cond}(\boldsymbol{B}).$$

   而且对非零的α都有

$$\mathrm{cond}(\alpha\boldsymbol{A})=\mathrm{cond}(\boldsymbol{A}).$$

6. **(Golub, 2.6.3题)**分别建立矩阵$\boldsymbol{X}\in\mathbb{R}^{m\times n}(m\geq n)$ 的谱条件数与下面两个矩阵谱条件数的不等式关

系:

$$\boldsymbol{B}=\begin{pmatrix}\boldsymbol{I}_m & \boldsymbol{X}\\ 0 & \boldsymbol{I}_n\end{pmatrix},\text{和}\boldsymbol{C}=\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{I}_n\end{pmatrix}.$$

7. **(Golub, 2.1.6题)**若$\boldsymbol{A}\in\mathbb{R}^{n\times n}$对称非奇异, 记

$$\boldsymbol{B}=\boldsymbol{A}+\alpha(\boldsymbol{u}\boldsymbol{u}^T+\boldsymbol{v}\boldsymbol{v}^T)+\beta(\boldsymbol{u}\boldsymbol{v}^T+\boldsymbol{v}\boldsymbol{u}^T),$$

其中$\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n,\alpha,\beta\in\mathbb{R}$. 设$\boldsymbol{B}$非奇异, 用Sherman-Morrison-Woodburg公式导出$\boldsymbol{B}^{-1}$的表达式.
8. **(Golub, 2.1.7题)**给出对称版本的Sherman-Morrison-Woodbury公式来刻画$\boldsymbol{A}+\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$的逆, 其中$\boldsymbol{A}\in\mathbb{R}^{n\times n}$和$\boldsymbol{S}\in\mathbb{R}^{k\times k}$都是对称矩阵, 而$\boldsymbol{U}\in\mathbb{R}^{n\times k}$.
9. **(Golub, 2.6.5题)**设$\boldsymbol{A}\in\mathbb{R}^{n\times n}$非奇异, $\boldsymbol{b}\in\mathbb{R}^n$, $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$, $\boldsymbol{C}=\boldsymbol{A}^{-1}$. 用Sherman-Morrison公式证明:

$$\frac{\partial x_k}{\partial a_{ij}}=-x_jc_{ki}.$$

§2.5 部分习题解答

例 2.5.1 设$\mathbf{Q}\in\mathbb{R}^{n\times n}$是正交矩阵, $x\in\mathbb{R}^n$, 则

$$\|\mathbf{Q}x\|_2=\|x\|_2.$$

即正交变换是保持二范数的.

证明: $\|\mathbf{Q}x\|_2^2=(\mathbf{Q}x)^T(\mathbf{Q}x)=(x^T\mathbf{Q}^T)(\mathbf{Q}x)=x^T(\mathbf{Q}^T\mathbf{Q})x=x^Tx=\|x\|_2^2$. □

例 2.5.2 设A是给定的n阶实对称正定矩阵, $x\in\mathbb{R}^n$, 则$\|x\|_A=(x^TAx)^{1/2}$是$\mathbb{R}^n$一种向量范数.

证明: 验证范数的三个条件即可. 对于第三个条件, 由于A实对称正定, 则A合同于单位阵, 从而$A=LL^T$, 其中L可逆. 从而$\|x\|_A=(x^TLL^Tx)^{1/2}=\|L^Tx\|_2$, 从而

$$\|x+y\|_A=\|L^T(x+y)\|_2\le\|L^Tx\|_2+\|L^Ty\|_2=\|x\|_A+\|y\|_A.$$

这样就证明了$\|\cdot\|_A$是范数. □

例 2.5.3 设$v\in\mathbb{R}^n$, 证明或证伪

$$\|v\|_1\|v\|_\infty\le\frac{1+\sqrt{n}}{2}\|v\|_2^2.$$

答: 这个命题是错误的. 考虑$v=(1,1,\cdots,1)^T$, 则$\|v\|_1=n,\|v\|_\infty=1,\|v\|_2=\sqrt{n}$. 此时不等式为

$$n\le\frac{\sqrt{n}+n}{2},$$

当$n>1$时这是不成立的. □

例 2.5.4 设$A\in\mathbb{C}^{n\times n}$, 则$\|A\|_2^2\le\|A\|_1\|A\|_\infty$.

证明: 注意$\|A\|_2^2=\rho(A^TA)\le\|A^TA\|_1\le\|A^T\|_1\|A\|_1=\|A\|_\infty\|A\|_1$. □

例 2.5.5 设$A\in\mathbb{R}^{m\times n},\operatorname{rank}A=n$, 则$\|A(A^TA)^{-1}A^T\|_2=1$.

证明: 记$B = A(A^TA)^{-1}A^T$, 则$B^T = B$, 则$\|B\|_2 = \rho(B)$. 容易验证$B^2 = B$, 则B的特征值只能为0或1. 任取$0 \neq x \in \mathbb{R}^n$, 由$\operatorname{rank} A = n$可知$Ax \neq 0$, 由于

$$B(Ax) = A(A^TA)^{-1}A^TAx = A[(A^TA)^{-1}A^TA]x = Ax = 1 \cdot Ax,$$

则1是B的一个特征值, $\rho(B) = 1$, 即$\|B\|_2 = 1$. □

例 2.5.6 设$A \in \mathbb{R}^{m\times n}$, 证明:

(1) $\max\limits_{i,j}|a_{ij}| \leq \|A\|_2 \leq \sqrt{mn}\max\limits_{i,j}|a_{ij}|$.

(2) $\dfrac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty$.

(3) $\dfrac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1$.

例 2.5.7 (Golub, 2.3.7题) 设$0 \neq s \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$, 则

$$\left\|\mathbf{A}\left(\mathbb{I} - \frac{ss^T}{s^Ts}\right)\right\|_F^2 = \|\mathbf{A}\|_F^2 - \frac{\|\mathbf{A}s\|_2^2}{s^Ts}.$$

证明: 注意到$\operatorname{tr}(ABC) = \operatorname{tr}(BCA)$以及$\|A\|_F = \operatorname{tr}(A^TA)$, 则

$$\begin{aligned}
\text{左边} &= \operatorname{tr}\left(\left(\mathbb{I} - \frac{ss^T}{s^Ts}\right)^T\mathbf{A}^T\mathbf{A}\left(\mathbb{I} - \frac{ss^T}{s^Ts}\right)\right)\\
&= \operatorname{tr}\left(\mathbf{A}^T\mathbf{A}\left(\mathbb{I} - \frac{ss^T}{s^Ts}\right)\left(\mathbb{I} - \frac{ss^T}{s^Ts}\right)^T\right)\\
&= \operatorname{tr}\left[\mathbf{A}^T\mathbf{A}\left(\mathbb{I} + \frac{ss^Tss^T}{(s^Ts)^2} - 2\frac{ss^T}{s^Ts}\right)\right]\\
&= \operatorname{tr}\left[\mathbf{A}^T\mathbf{A}\left(\mathbb{I} + \frac{ss^T}{(s^Ts)^2} - 2\frac{ss^T}{s^Ts}\right)\right]\\
&= \operatorname{tr}\left[\mathbf{A}^T\mathbf{A}\left(\mathbb{I} - \frac{ss^T}{s^Ts}\right)\right]\\
&= \operatorname{tr}(\mathbf{A}^T\mathbf{A}) - \operatorname{tr}\left(\frac{\mathbf{A}^T\mathbf{A}ss^T}{s^Ts}\right)\\
&= \operatorname{tr}(\mathbf{A}^T\mathbf{A}) - \operatorname{tr}\left(\frac{s^T\mathbf{A}^T\mathbf{A}s}{s^Ts}\right)\\
&= \|\mathbf{A}\|_F^2 - \frac{\|\mathbf{A}s\|_2^2}{s^Ts}.
\end{aligned}$$

□

例 2.5.8 (Golub, 2.3.8题) 设$u \in \mathbb{R}^m, v \in \mathbb{R}^n$. 若$E = uv^T$, 则$\|E\|_F = \|E\|_2 = \|u\|_2\|v\|_2$, 且$\|E\|_\infty \leq \|u\|_\infty\|v\|_1$.

证明: (1)$\operatorname{rank}(E) = 1$, 记$E$的非零特征值为$\lambda$, 则其余特征值都是0($n-1$重), 所以根据前面的定理,

$$\|E\|_F^2 = \lambda^2 = \|E\|_2^2.$$

(2)$\|E\|_F^2 = \operatorname{tr}((uv^T)^T(uv^T)) = \operatorname{tr}(vu^Tuv^T) = (u^Tu)\operatorname{tr}(vv^T) = (u^Tu)(v^Tv) = \|u\|_2^2\|v\|_2^2$.

(3)$\|E\|_\infty = \max\limits_{\|x\|=1}\dfrac{\|uv^Tx\|_\infty}{\|x\|_\infty} \leq \max\limits_{\|x\|=1}\dfrac{\|u\|_\infty\|v^T\|_\infty\|x\|_\infty}{\|x\|_\infty} = \|u\|_\infty\|v^T\|_\infty = \|u\|_\infty\|v\|_1$. □

例 2.5.9 设$A \in \mathbb{R}^{m\times n}$. B是A的子矩阵, 即选择A的其中μ行与ν列构成的$\mu\times\nu$矩阵, 其中$\mu \leq m, \nu \leq n$. $1 \leq p \leq \infty$, 则$\|B\|_p \leq \|A\|_p$.

证明: 如果把$A \triangleq A_{m,n}$的某一列删掉, 会得到一个$m \times (n-1)$阶子矩阵$A_{m,n-1}$, 我们证明$\|A_{m,n-1}\|_p \leq \|A_{m,n}\|_p$. 设$A_{m,n-1}$是由$A$删掉第$j$列得到的, 满足

$$A_{m,n-1} = A_{m,n}C_{n,n-1},$$

其中

$$C = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \\ & & & 1 & & & \\ & & & & \ddots & & \\ & & & & & 1 & \end{pmatrix},$$

即第j行全为0. 则

$$\|C\|_p = \max_{\|x\|_p=1} \|Cx\|_p = \max_{\|x\|_p=1} \left(\sum_{i=1}^{j-1} |x_i|^p + \sum_{i=j+1}^{n} |x_i|^p \right)^{1/p} \leq \max_{\|x\|_p=1} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} = \|x\|_p = 1.$$

于是

$$\|A_{m,n-1}\|_p = \|A_{m,n}C_{n,n-1}\|_p \leq \|A_{m,n}\|_p \|C_{n,n-1}\|_p \leq \|A_{m,n}\|_p.$$

同理可证

$$\|A_{m-1,n}\|_p \leq \|A_{m,n}\|_p.$$

利用归纳法可以证明$\|B\|_p \leq \|A\|_p$. □

例 2.5.10 (2016Team, 3) 设$A = (a_{ij})_{i,j=1}^{m,n}$是$m \times n$矩阵, 秩为$r \leq n-1$, 且矩阵中的元素均为整数, 且模不超过$H$, 即

$$|a_{ij}| \leq H, 1 \leq i \leq m, 1 \leq j \leq n.$$

证明: 存在一个非零整向量$\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{Z}^n$, 使得$A\boldsymbol{x} = 0$且

$$\|\boldsymbol{x}\|_\infty \leq (2nH)^{n-1}.$$

例 2.5.11 (2014Team, 5) (1)设$x_0 = 0$, 请把

$$x_k = 2x_{k-1} + b_k, k = 1, 2, \cdots, n.$$

写成矩阵形式$A\vec{x} = \vec{b}$. 对于$b_1 = -\frac{1}{3}, b_k = (-1)^k, k = 2, 3, \cdots, n$, 验证$x_k = \frac{(-1)^k}{3}, k = 1, 2, \cdots, n$是精确解. (2)求$A^{-1}$, 并计算$A$在$l^1$范数下的条件数.

证明: (1)

$$A = \begin{pmatrix} 1 & & & \\ -2 & 1 & & \\ & \ddots & \ddots & \\ & & -2 & 1 \end{pmatrix}.$$

验证精确解: 略.

(2)易知

$$A^{-1}=\begin{pmatrix}1&&&\\2&1&&\\&\ddots&\ddots&\\&&2&1\end{pmatrix}.$$

$\|A\|_1=\|A^{-1}\|_1=3$, 则条件数是$\mathrm{cond}_1(A)=\|A\|_1\|A^{-1}\|_1=9$.

引理 2.5.1

设$A=(a_{ij})$是n阶Hermite矩阵, 特征值为$\lambda_1\geq\cdots\geq\lambda_n$, 则$\lambda_n\leq a_{ii}\leq\lambda_i, i=1,2,\cdots,n$.

引理 2.5.2

设A是n阶Hermite矩阵, 特征值为$\lambda_1\geq\cdots\geq\lambda_n$. 则对任意$1\leq j_1\leq\cdots\leq j_k\leq n$, 都有

$$\sum_{i=1}^{k}a_{j_ij_i}\leq\sum_{i=1}^{k}\lambda_i, k=1,\cdots,n-1.$$

当$k=n$时, 这就是熟知的等式$\mathrm{tr}(A)=\sum_{i=1}^{n}\lambda_i$.

证明: 由于A是Hermite矩阵, 则存在酉矩阵Q, 使得$A=Q\Lambda Q^H$. 其中$\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$. 取$x_i=(0,\cdots,0,1,0,\cdots,0)$, 其中第$j_i$个分量为1, 其余分量为0. 则

$$\begin{aligned}\sum_{i=1}^{n}a_{j_ij_i}&=\sum_{i=1}^{n}x_i^HAx_i=\sum_{i=1}^{n}x_i^HQ\Lambda Q^Hx_i\\&=\mathrm{tr}(P^H\Lambda P),\text{其中}P\text{的第}i\text{列为}Q^Hx_i,\text{各个列向量正交}\\&=\mathrm{tr}(\Lambda PP^H)=\sum_{i=1}^{n}\lambda_ib_{ii}.\end{aligned}$$

其中$B=PP^H$满足$\mathrm{rank}(B)=k$. 当$k=n$时, 这就是熟知的等式; 当$k\leq n-1$时,

$$\sum_{i=1}^{n}b_{ii}=\mathrm{tr}B=\mathrm{tr}(PP^H)=\mathrm{tr}(P^HP)=k.$$

注意B是幂等矩阵, 特征值只能为0或1, 由前面的结论可知$0\leq b_{ii}\leq1, i=1,\cdots,n$. 所以

$$\begin{aligned}\sum_{i=1}^{n}\lambda_ib_{ii}&\leq\sum_{i=1}^{k}\lambda_ib_{ii}+\left(\sum_{i=k+1}^{n}b_{ii}\right)\lambda_{k+1}\\&=\sum_{i=1}^{k}\lambda_ib_{ii}+\left(\mathrm{tr}B-\sum_{i=1}^{k}b_{ii}\right)\lambda_{k+1}\\&=\sum_{i=1}^{k}\lambda_ib_{ii}+\left(k-\sum_{i=1}^{k}b_{ii}\right)\lambda_{k+1}\\&=\sum_{i=1}^{k}(\lambda_i-\lambda_{k+1})b_{ii}+k\lambda_{k+1}\\&\leq\sum_{i=1}^{k}(\lambda_i-\lambda_{k+1})+k\lambda_{k+1}\\&=\sum_{i=1}^{k}\lambda_i.\end{aligned}$$

定理 2.5.3: Neumann不等式

设A, B是n阶Hermite阵, 特征值分别为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n 与 \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n.$$

则

$$\mathrm{tr}(AB) \leq \sum_{i=1}^{n} \lambda_i \mu_i.$$

证明: 设$\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$. 由于$A$是Hermite矩阵, 则存在酉矩阵$Q$, 使得$A = Q\Lambda Q^T$. 于是

$$\mathrm{tr}(CD) = \mathrm{tr}Q\Lambda Q^T B = \mathrm{tr}\Lambda Q^T B Q = \sum_{i=1}^{n} \lambda_i c_{ii},$$

其中$C = (c_{ij}) = Q^T B Q$. 正交变换不改变特征值, 则B的特征值也是C的特征值. 所以

$$\begin{aligned}
\mathrm{tr}(CD) &= \sum_{i=1}^{n-1} \left[(\lambda_i - \lambda_{i+1}) \sum_{j=1}^{i} c_{jj} \right] + \lambda_n \sum_{j=1}^{n} c_{jj} (\text{Abel求和}) \\
&\leq \sum_{i=1}^{n-1} \left[(\lambda_i - \lambda_{i+1}) \sum_{j=1}^{i} \mu_j \right] + \lambda_n \sum_{j=1}^{n} \mu_j \\
&= \sum_{i=1}^{n} \lambda_i \mu_i.
\end{aligned}$$

(等号成立当且仅当c_{jj}恰为C的特征值μ_j, 等价于$B = \sum_{i=1}^{n} \mu_i \varphi_i \varphi_i^T$, 其中$\varphi_i$为$\Phi$的列分量) □

例 2.5.12 (2012Individual, 4) 设$C, D \in \mathbb{C}^{n\times n}$是*Hermite*矩阵, 特征值为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n 与 \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n.$$

证明:

$$\sum_{i=1}^{n} (\lambda_i - \mu_i)^2 \leq \|C - D\|_F^2.$$

证明: 由于

$$\|C - D\|_F^2 = \mathrm{tr}((C-D)^H (C-D)) = \mathrm{tr}(C^H C + D^H D - 2C^H D) = \sum_{i=1}^{n} (\lambda_i^2 + \mu_i^2) - 2\mathrm{tr}(C^H D),$$

只需证明

$$\mathrm{tr}(CD) \leq \sum_{i=1}^{n} \lambda_i \mu_i.$$

这就是Neumann不等式. □

CHAPTER 3

解线性方程组的直接方法

只要会算法和简单的分析方法就可以了. 这部分简单回顾一下.

§3.1 Gauss消去法

最基本的Gauss消去法就是化为上三角阵方程组来求解, 详见高代课本.

Gauss列主元消去法: 在第k步中, 选取$a_{kk}^{(k-1)},\cdots,a_{nk}^{(k-1)}$中绝对值最大的作为主元, 比如$a_{rk}^{(k-1)}$绝对值最大, 交换第$k$与第$r$行元素, 就是Gauss列主元消去法.

Gauss按比例列主元消去法: 消元之前先计算每行最大的元素, 第i行最大的记为s_i, 在第k步, 求$\frac{|a_{ik}|}{s_i}$最小的行, 记为第r行: $\frac{|a_{rk}|}{s_r}$, 然后交换第k与第r行（相应地s_k与s_r也要交换）.

Gauss-Jordan消去法: 每一步消元都把非主元化为0, 如果要选列主元那就交换两行.

能用Gauss消元法的条件就是各个主元都不为0. 有如下定理:

定理 3.1.1

在Gauss消去法中，主元$a_{11},a_{22}^{(1)},\cdots,a_{kk}^{(k-1)}$全不为0的充分必要条件是：矩阵$A$的顺序主子矩阵

$$A_1=(a_{11}),A_2=\begin{pmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{pmatrix},\cdots,A_k=\begin{pmatrix}a_{11}&a_{12}&\cdots&a_{1k}\\a_{21}&a_{22}&\cdots&a_{2k}\\\vdots&\vdots&&\vdots\\a_{k1}&a_{k2}&\cdots&a_{kk}\end{pmatrix}$$

都是非奇异, 其中$k\le n$.

注: 这里没有选过主元, $a_{11},a_{22}^{(1)},\cdots,a_{kk}^{(k-1)}$都是消元后的结果.

§3.2 直接三角分解法

定义 3.2.1: LU分解

若方阵A可分解为一个下三角阵L与一个上三角阵U的乘积, 即$A=LU$, 称这种分解为方阵A的LU分解. 特别地, 若L为单位下三角阵时为**Doolittle分解**, U为上三角阵时为**Crout分解**.

定义 3.2.2: LDR分解

为了解决唯一性问题, 把A分解为$A = LDR$, L, R分别为下、上三角阵, D是对角阵, 这种分解叫**LDR分解**.

定理 3.2.1

n阶矩阵A有唯一的LDR分解的充分必要条件是A的顺序主子矩阵$A_1, A_2, \cdots, A_{n-1}$都非奇异.

如果对线性方程组作分解$Ax = LUx = b$, 可以先解$Ly = b$, 再解$Ux = y$, 这就是**直接三角分解法**.

Crout方法: 先作Crout分解$A = LU$, 用分量形式表示, 得到关系式, 然后先计算第一列元素, 再计算第一行元素; 然后算第2, 3,$\cdots$列与行的元素. 这样就进行了分解. 然后再反解方程组.

列选主元Crout方法: 每一步都先把列主元求出来再交换两行(包括右端向量b).

LL^T(Cholesky)分解: 对于**实对称正定矩阵**A, 存在非奇异下三角阵L, 使得$A = LL^T$, 且当L的主对角元都正的时候, 这种分解唯一. (证明方法: 记$A = LDR, A^T = (LDR)^T = LDR = A$, 由分解的唯一性立证.)

LDL^T分解: 对于**实对称正定矩阵**A, 有非奇异单位下三角阵L与非奇异对角阵D使得$A = LDL^T$. 作LDL^T分解时, 可以把右端DL^T合在一起, 记作G, 可减小运算量. 该方法与Cholesky方法相比不用开方.

追赶法: 对于三对角线性方程组$Ax = b$, 可以分解为如下形式:

$$A = \begin{pmatrix} d_1 & c_1 & & & & \\ a_2 & d_2 & c_2 & & & \\ & a_3 & d_3 & c_3 & & \\ & & \cdots & \cdots & \cdots & \\ & & & a_{n-1} & d_{n-1} & c_{n-1} \\ & & & & a_n & d_n \end{pmatrix} = \begin{pmatrix} p_1 & & & & \\ a_2 & p_2 & & & \\ & a_3 & p_3 & & \\ & & \cdots & \cdots & \\ & & & a_n & p_n \end{pmatrix} \begin{pmatrix} 1 & q_1 & & & \\ & 1 & q_2 & & \\ & & 1 & q_3 & \\ & & & \cdots & \\ & & & & q_{n-1} \\ & & & & 1 \end{pmatrix}.$$

定理 3.2.2

如果三对角矩阵满足优对角条件: $|d_1| > |c_1| > 0, |d_k| \geq |a_k| + |c_k|$且$a_k c_k \neq 0$, $|d_n| > |a_n| > 0$, 则$p_1, \cdots, p_n$非零.

证明: 用归纳法证$|q_k| < 1$, 再证$|p_k| > 0$. □

CHAPTER 4

解非线性方程的数值方法

§4.1 基本概念

定义 4.1.1

如果从任何可取的初始值出发都能保证收敛, 称为**大范围收敛**, 如果必须选取初始值充分接近于所要求的根, 称为**局部收敛**.

定义 4.1.2: 收敛阶数

设一个迭代法收敛, $\lim\limits_{k\to\infty} x_k = p$, p是方程$f(x) = 0$的一个根. 令$e_k = x_k - p$, 若存在实数λ和非零常数C使得

$$\limsup_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^\lambda} \le C,$$

则称该迭代法为λ阶收敛. 若λ为整数, 则上式可以去绝对值.

注: $\lambda = 1$称为线性收敛; 迭代法必须收敛才有收敛阶数的概念.

定义 4.1.3: 迭代终止准则

当k足够大时, x_k充分接近于x, 所以要给出**迭代终止准则**, 到达某些条件后停止迭代过程.

常见的迭代过程有: ①$|f(x_n)| < TOL$(误差容限)、②$|x_n - x_{n-1}| < TOL$(绝对误差较小)、③$\dfrac{|x_n - x_{n-1}|}{|x_n|} < TOL$(相对误差较小). 不同方法有各自缺点, 要看情况选择.

通常局部收敛方法比大范围收敛方法收敛的快. 合理的算法是先用大范围方法(如二分法)求接近于根的近似值, 到一定程度再改用局部收敛方法(如Newton法).

§4.2 二分法

基本思想: 设$f(x)$在区间$[a,b]$连续, $f(a)f(b) < 0$, 根据介值定理, 方程$f(x) = 0$在区间$[a,b]$至少有一个根. 记$[a,b] = [a_1,b_1]$, p_1是中点, 若$|f(p_1)| < \delta$(足够小的量), 则p_1是根的近似值; 否则, 若$f(p_1)f(b_1) < 0$, 则区间$[p_1,b_1]$至少有一个根, 取$a_2 = p_1, b_2 = b_1$. 若$f(p_1)f(b_1) > 0$, 取$a_2 = a_1, b_2 = p_1$, ……这样就能每次把区间缩小一半.

注意事项:

- 计算中点时, 采用$c \leftarrow a + (b-a)/2$而不是$c \leftarrow (b+a)/2$. 比如只有二位精度的数中如$[0.67, 0.69]$, $\dfrac{a+b}{2} = \dfrac{1.36}{2} = \dfrac{0.14 \times 10}{2} = 0.70$. 数值计算的通用策略：把一个小的修正项加到先前的近似值来计算一个量是最佳的.
- 采用$\operatorname{sgn}(w) \neq \operatorname{sgn}(u)$, 而不是$wu < 0$来确定函数在区间上是否变号, 因为后者需要不必要的乘法且可能引起溢出.

定理 4.2.1

设函数$f(x)$在区间$[a,b]$连续且端点异号, 用$[a_k, b_k]$表示各区间, 则a_n与b_n极限存在且相等, 且极限是f的一个零点. 若$\lim\limits_{n\to\infty} p_n = p$且$p_n = \dfrac{a_n + b_n}{2}$. 则$|p - p_n| \leq \dfrac{1}{2}(b_n - a_n) = 2^{-n}(b-a)(*)$.

注: 定理说明二分法是大范围收敛的. (*)是一个先验的绝对误差界. 若令ε表示给定误差容限, 要$\dfrac{1}{2^n}(b-a) \leq \varepsilon$即$n > \dfrac{\lg \dfrac{b-a}{\varepsilon}}{\lg 2}$, 取$n$是大于$\lg \dfrac{b-a}{\varepsilon} / \lg 2$的最小整数, 这个提供了一个迭代终止准则, 在第$n$步停止迭代即可.

注: 该法近似解的误差下降速度不快, 但方法简单且可靠. 可以用来求根的初始值.

§ 4.3 不动点迭代与加速迭代收敛

4.3.1 不动点迭代

定义 4.3.1

对于方程$f(x) = 0$的解, 把它转化为等价方程$x = g(x)$, 给定初始值x_0, 令$x_k = g(x_{k-1})$得到序列$\{x_k\}$, 该类迭代法叫**不动点迭代法**或**Picard迭代**, $g(x)$叫迭代函数.

定理 4.3.1

假设$g(x)$为定义在$[a,b]$上的一个实函数, 满足:

(1) $g(x) \in [a,b], \forall x \in [a,b]$,

(2) $g(x)$满足Lipschitz条件, 其Lipschitz常数< 1.

则对任意初始值$x_0 \in [a,b]$, 由Picard迭代产生的序列, 都收敛于g的唯一不动点g, 且有误差估计式

$$|e_k| \leq \frac{L^k}{1-L}|x_1 - x_0|,$$

其中$e_k = x_k - p$.

推论 4.3.2

若将条件(2)改为$g(x)$的导数$|g'(x)| \leq L < 1, \forall x \in [a,b]$, 则定理结论依然成立.

定理 4.3.3

在定理4.3.1的条件下, 若$g(x)$在区间$[a,b]$上为$m(\geq 2)$次连续可微, 且在p处有$g^{(j)}(p) = 0, j = 1, \cdots, m-1, g^{(m)}(p) = 0$, 则Picard迭代为$m$阶收敛.

证明: 用Taylor展开证明. □

定理 4.3.4

设p是方程$f(x) = 0$的根,$g(x)$在p的某邻域内m次连续可微, 且$g^{(j)}(p) = 0, j = 1, \cdots, m-1, g^{(m)}(p) = 0$, 则存在$r > 0$, 当$x_0 \in [p-r, p+r]$时Picard迭代收敛, 且阶数为$m$.

证明: 用连续函数性质证明, 存在r使得$|g'(x)| \leq L < 1, x \in [p-r, p+r]$, 再用前面定理. □

注: 这里没说用定理4.3.1的条件, 所以要在p附近的区间找满足定理4.3.1的Lipschitz常数(利用连续可微性以及微分中值定理)从而可以应用定理4.3.1.

4.3.2 迭代法加速技术

下面设$\{x_n\}$线性收敛于p, 则$\dfrac{x_{n+1}-p}{x_n-p}\simeq\dfrac{x_{n+2}-p}{x_{n+1}-p}$. 化简得$p\simeq x_n-\dfrac{(x_{n+1}-x_n)^2}{x_{n+2}-2x_{n+1}+x_n}, n=0,1,2,\cdots$.

定义 4.3.2: Aitken加速迭代

把式子$\tilde{x}_{n+1}=x_n-\dfrac{(x_{n+1}-x_n)^2}{x_{n+2}-2x_{n+1}+x_n}$叫**Aitken加速方法**.

定理 4.3.5

设序列$\{x_n\}$线性收敛于p, 且对所有足够大的n有$(x_n-p)(x_{n+1}-p)\neq 0$, 则由Aitken加速方法产生的序列$\{\tilde{x}_n\}$比$\{x_n\}$更快收敛于p, 即$\lim\limits_{n\to\infty}\dfrac{\tilde{x}_{n+1}-p}{x_n-p}=0$.

下面把Aitken加速技巧用于不动点迭代得到的线性收敛序列. 迭代公式为

定义 4.3.3: Steffensen迭代

$x_{k+1}=x_k-\dfrac{(g(x_k)-x_k)^2}{g(g(x_k))-2g(x_k)+x_k}$叫**Steffensen迭代法**.

定理 4.3.6

设方程$x=g(x)$有解p, 若存在一正数r, 使得对所有$x\in[p-r,p+r]$, $g(x)$连续三次可微, 则Steffensen迭代法对任一初始值$x_0\in[p-r,p+r]$是二阶收敛的.

§4.4 Newton法

解非线性方程$f(x)=0$最著名且有效的方法之一. 若初始值充分接近于根, Newton法的收敛速度很快. 将$f(x)$在初值x_0处Taylor展开, 取线性部分作为$f(x)$的近似, 有

$$f(x_0)+f'(x_0)(x-x_0)\approx 0.$$

若$f'(x_0)\neq 0$, 则有$x=x_0-\dfrac{f(x_0)}{f'(x_0)}$. 用迭代序列

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

就是Newton方法.

令$g(x)=x-\dfrac{f(x)}{f'(x)}$, 则方程$f(x)=0$与$g(x)=x$是等价的.

定义 4.4.1: Newton

把$x_{k+1}=x_k-\dfrac{f(x_k)}{f'(x_k)}$称为**Newton迭代公式**.

定理 4.4.1

设函数$f(x)$有$m(>2)$阶连续导数, p是方程$f(x)=0$的单根, 则当x_0充分接近于p时, Newton法至少二阶收敛.

证明: 根据Newton法, $e_{n+1}=x_{n+1}-p=\cdots=\dfrac{e_nf'(x_n)-f(x_n)}{f'(x_n)}$, 用Taylor公式得

$$0=f(p)=f(x_n-e_n)=f(x_n)-e_nf'(x_n)+\frac{1}{2}e_n^2f''(\xi_n).$$

则$e_{n+1}=\dfrac{f''(\xi_n)}{2f'(x_n)}e_n^2\approx\dfrac{1}{2}\dfrac{f''(p)}{f'(p)}e_n^2=Ce_n^2$, 由$k$阶收敛定义, 牛顿法二次收敛. □

注: Newton方法的收敛性依赖于x_0的选取.

注: 当p为n重根时, Newton法仅为一次收敛.

一个有用的套路: 设$f(x)$有足够阶连续导时候, p是$f(x)=0$的$q(\geq 2)$重根, 则可以将$f(x)$表示为$f(x)=(x-p)^qh(x)$. 其中$h(p)\neq 0$. (见书)

4.4.1 Newton法的改进

方法一: 构造$F(x)$代替$f(x)$.

当p为n重根时, 对f,f'进行Taylor展开, 可得

$$\frac{f(x)}{f'(x)}=\frac{1}{q}(x-p)\frac{f^{(q)}(\xi_1)}{f^{(q)}(\xi_2)}.$$

令$F(x)=\dfrac{f(x)}{f'(x)}$, 则p是$F(x)$的单根, 从而有新的迭代公式$x_k=x_{k-1}-\dfrac{F(x_{k-1})}{F'(x_{k-1})}$.

缺点: 增加二阶导数的计算且迭代过程中计算更复杂, 但重根的出现将产生严重的舍入误差问题. 为了避免算二阶导数, 可以利用Steffensen方法加速Newton法.

方法二: Steffensen加速, 略.

方法三: 加一个常数λ, $x_{k+1}=x_k-\lambda\dfrac{f(x_k)}{f'(x_k)}$变为二阶收敛(习题).

证明: p为n重零点时, 构造$f(x)=(x-p)^ng(x)$, 用$g(x)$代替$f(x)$, 可以反求λ. 把λ求出来后, 再验证是否为二阶收敛.

定理 4.4.2

设函数$f(x)$在有限区间$[a,b]$存在二阶导数, 且满足$(1)f(a)f(b)<0$, $(2)f'(x)\neq 0, x\in[a,b]$, $(3)f''(x)$在$[a,b]$上不变号, $(4)\left|\dfrac{f(a)}{f'(a)}\right|<b-a$且$\left|\dfrac{f(b)}{f'(b)}\right|<b-a$. 则Newton法对任意的初始值$x_0\in[a,b]$都收敛于方程$f(x)=0$的唯一解$p$, 收敛阶数为2.

注: 前两个条件保证方程$f(x)=0$在$[a,b]$只有一个根, 条件(3)保证图形是凸的或者凹的, 条件(4)保证$x_0\in[a,b]$时Newton序列$\{x_k\}$在(a,b)中.

§4.5 割线法

考虑经过点$(x_{k-1},f(x_{k-1}))$的割线C_k代替曲线, 将割线C_k与x轴的交点的横坐标作为方程$f(x)=0$的近似解. 割线C_k的方程为

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

令$y=0$可得C_k与x轴交点的横坐标是$x_k-\dfrac{f(x_k)(x_k-x_{k-1})}{f(x_k)-f(x_{k-1})}$.

定义 4.5.1: 割线法

把$x_{k+1}=x_k-\dfrac{f(x_k)(x_k-x_{k-1})}{f(x_k)-f(x_{k-1})}$叫割线法.

割线法可以看作Newton法中用$f[x_{k-1},x_k]=\dfrac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$来代替$f'(x_k)$. 迭代速度比Newton法稍微慢, 但不需要算导数.

定理 4.5.1

令I表示区间$(p-r,p+r)$, p是方程$f(x)=0$的根, $r>0$. 设函数$f(x)$在I中有足够阶连续导数, 满足$(1)f'(x)\neq 0,x\in I$, $(2)\left|\dfrac{f''(\xi)}{2f'(\eta)}\right|\leq M,\forall\xi,\eta\in I$, $(3)d=Mr<1$. 则对任意的初始值$x_0,x_1\in I$, 由割线法产生的序列$\{x_n\}$都收敛于p, 且

$$\lim_{k\to\infty}\frac{|e_{k+1}|}{|e_k|^q}=K^{\frac{1}{q}},$$

其中$K=\left|\dfrac{f''(p)}{2f'(p)}\right|,q=\dfrac{1}{2}(1+\sqrt{5})$.

证明: 首先证明$|e_k|<r$, 迭代过程可以进行下去, 要用Newton插值公式

$$f(x)=f(x_k)+(x-x_k)f[x_{k-1},x_k]+\frac{1}{2}(x-x_{k-1})(x-x_k)f''(\xi_k).$$

并用微分中值定理, 可化为$e_{k+1}=M_ke_ke_{k-1}$. 紧接着作换元$d_k=M|e_k|$, 并证明$d_k\leq d^k<1$, 从而$\{x_k\}$收敛于p. 最后证明收敛次数. □

注: 割线法只需要算一次函数值, 且不需要计算导数. (牛顿法要算两次), 如果割线法算两次函数值, 收敛阶数达到了$q^2=\dfrac{3+\sqrt{5}}{2}$. 这比牛顿法的二次收敛好得多(在工作量相等的情况下)

推论 4.5.2

设p是方程$f(x)=0$的一个根, $f'(p)\neq 0$, 且$f''(x)$在p附近连续, 则存在$r>0$, 对任意初始值$x_0,x_1\in[p-r,p+r]$, 由割线法产生的序列$\{x_k\}$都收敛于p.

注: 已经证明没有一种迭代法只算一次函数值就可以达到二阶收敛, 然而割线法的一种变形

$$x_{k+1}=x_k-\frac{(f(x_k))^2}{f(x_k+f(x_k))-f(x_k)},k=1,2,\cdots$$

是二阶收敛的, 它要求计算两个函数值而不需要计算导数值. 而$\dfrac{f(x_k)}{f(x_k+f(x_k))-f(x_k)}$可以看作是$f'(x_k)$的近似.

若令$f(x)=g(x)-x$, 就变成了

$$x_{k+1}=x_k-\frac{(g(x_k)-x_k)^2}{g(g(x_k))-2g(x_k)+x_k},k=1,2,\cdots$$

这个就是Steffensen迭代法.

§4.6 多项式求根

4.6.1 降次法

给定多项式$p(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$, 计算$p(x)$和$p'(x)$的值. 最有效的方法是**Horner算**

法(秦九韶算法): 记$q(x)=b_nx^{n-1}+b_{n-1}x^{n-2}+\cdots+b_2x+b_1$, 且$p(x)=(x-x_0)q(x)+b_0$, 则比较系数可得

$$\begin{aligned}b_n&=a_n,\\ b_{n-j}&=a_{n-j}+b_{n-j+1}x_0.\end{aligned}$$

而$p(x_0)=b_0$即可算出来. 另外, 对两边求导数有$p'(x)=(x-x_0)q'(x)+q(x)$, 则$p'(x_0)=q(x_0)$. 仿照上面用递推公式

$$\begin{aligned}c_n&=b_n,\\ c_{n-j}&=b_{n-j}+c_{n-j+1}x_0.\end{aligned}$$

可以算得$p'(x_0)=c_1$. 把他们算出来后, 用Newton法$x_k=x_{k-1}-\dfrac{p(x_{k-1})}{p'(x_{k-1})}$可求根.

注: 用降次法由于误差的存在, 不能保证所得结果与标准结果的根一样, 可能会有坏情况.

4.6.2 Muller法

给定三个初值$(x_0,f(x_0)),(x_1,f(x_1)),(x_2,f(x_2))$, 取经过这3个点的抛物线, 它与$x$轴的交点是$x_3$来作为根的一个近似值, 仿照该步骤再作抛物线求$x_4$,$\cdots$

设二次多项式$q(x)=a(x-x_2)^2+b(x-x_2)+c$经过这3个点, 则

$$\begin{aligned}f(x_0)&=a(x_0-x_2)^2+b(x_0-x_2)+c,\\ f(x_1)&=a(x_1-x_2)^2+b(x_1-x_2)+c,\\ f(x_2)&=c.\end{aligned}$$

可以解得

$$\begin{aligned}c&=f(x_2),\\ a&=\frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0},\\ b&=f[x_2,x_1]+(x_2-x_1)a.\end{aligned}$$

根据求根公式, 用$x_3-x_2=\dfrac{-2c}{b\pm\sqrt{b^2-4ac}}$可求根. 为了让分母绝对值较大, 可改写为

$$x_3=x_2-\frac{2c}{b+\operatorname{sgn}(b)\sqrt{b^2-4ac}}.$$

以此类推取$x_4,x_5\cdots$.

注: 一般Muller法比割线法收敛得快. 在一定条件下, Muller法的收敛阶数是1.84, 比割线法1.618好.

CHAPTER 5

函数插值理论

§5.1 多项式插值

如果给定了$n+1$个点$(x_i,y_i)_{i=0}^n$, 需要找一个次数尽可能低的多项式p, 使得

$$p(x_i)=y_i, 0\le i\le n$$

这样的多项式叫做这组数据点的**插值多项式**.

定理 5.1.1: 多项式插值定理

若$x_0,x_1,\cdots,x_n$是不同的实数, 则对任意数值$y_0,y_1,\cdots,y_n$, 存在唯一的次数至多是n次的多项式p_n, 使得

$$p_n(x_i)=y_i, 0\le i\le n.$$

证明: (1)**(唯一性)**. 假设有两个这样的多项式p_n,q_n, 则$(p_n-q_n)(x_i)=0,(0\le i\le n)$. 由于$\deg(p_n-q_n)\le n$, 若它不是零多项式, 则最多有$n$个零点. 由于$x_i$互不相同, 则$p_n-q_n$有$n+1$个零点, 所以它一定是零多项式, 则$p_n\equiv q_n$.

(2)**(存在性)**. 用数学归纳法证明. 当$n=0$时, 可以选择一个常值函数p_0使得$p_0(x_0)=y_0$.

假设已经得到一个次数不超过$k-1$次的多项式p_{k-1}, 使得$p_{k-1}(x_i)=y_i, 0\le i\le k-1$, 尝试构造如下形式的$p_k$:

$$p_k(x)=p_{k-1}(x)+c(x-x_0)(x-x_1)\cdots(x-x_{k-1}),$$

这是次数至多为k次的多项式.

由于$p_k(x_i)=p_{k-1}(x_i)=y_i(0\le i\le k-1)$, 所以$p_k$插值了$p_{k-1}$插值的那些数据, 下面来根据条件$p_k(x_k)=y_k$确定$c$, 由这个条件可知

$$p_{k-1}(x_k)+c(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})=y_k,$$

用上式可以求出c. □

5.1.1 Newton插值多项式

上述证明递归过程中的多项式$p_0,p_1,\cdots,p_n$的每个p_k可以由p_{k-1}添加一个简单项得到, 因此在递归过程最

后, p_n是$p_0,\cdots,p_{n-1}$的线性组合. 每个p_k都形如

$$p_k(x)=c_0+c_1(x-x_0)+c_2(x-x_0)(x-x_1)+\cdots+c_k(x-x_0)\cdots(x-x_{k-1}).$$

或者改写为

$$p_k(x)=\sum_{i=0}^{k}c_i\prod_{j=0}^{i-1}(x-x_j).$$

(当$m<0$时, 约定$\prod_{j=0}^{m}(x-x_j)=1$). 最初的几项是

$$\begin{aligned}p_0(x)&=c_0,\\p_1(x)&=c_0+c_1(x-x_0)\\p_2(x)&=c_0+c_1(x-x_0)+c_2(x-x_0)(x-x_1).\end{aligned}$$

以此类推, 这些多项式叫**Newton插值多项式**.

Newton多项式的计算, 可以用Horner算法（秦九韶算法）来计算. 对于式子

$$u=\sum_{i=0}^{k}c_i\prod_{j=0}^{i-1}d_j=c_0+c_1d_0+c_2d_0d_1+\cdots+c_kd_0d_1\cdots d_{k-1},$$

可以把上式改写为

$$u=(\cdots(((c_k)d_{k-1}+c_{k-1})d_{k-2}+c_{k-2})d_{k-3}+\cdots+c_1)d_0+c_0.$$

计算u时, 从最里面的括号开始, 用$u_k,u_{k-1},\cdots,u_0$表示括号中的量, 则计算方法为

$$\begin{aligned}u_k&\leftarrow c_k\\u_{k-1}&\leftarrow u_kd_{k-1}+c_{k-1}\\u_{k-2}&\leftarrow u_{k-1}d_{k-2}+c_{k-2}\\&\vdots\\u_0&\leftarrow u_1d_0+c_0\end{aligned}$$

用Horner算法可以得到c_k的公式是

$$c_k=\frac{y_k-p_{k-1}(x_k)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})}.$$

不难发现这与前面定理的证明结果是一致的.

5.1.2 Lagrange插值

前面的多项式表示定理说明了插值多项式是唯一的, 但是可能换种表示方法.

Lagrange插值法是把点$\{(x_k,y_k)\}_{k=0}^n$的插值多项式$p(x)$表示为

$$p(x)=\sum_{k=0}^{n}y_kl_k(x),$$

其中$\{l_k\}_{k=0}^n$是一列多项式, 只依赖于$\{x_k\}_{k=0}^n$, 与$\{y_k\}_{k=0}^n$无关.

定义**Kronecker** δ**函数**为

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

如果定义$l_k(x_i) := \delta_{ki}$, 则有

$$y_i = \sum_{k=0}^{n} y_k l_k(x_i) = p(x_i),$$

满足Lagrange插值的条件$p(x_j) = y_j, (\forall j)$.

根据这个定义, $l_i(x)$是n次多项式且在点$\{x_k\}_{k=0}^n \setminus \{x_i\}$处的值是0, 那么$l_i(x)$必定具有形式

$$l_i(x) = c\prod_{j\neq i}(x - x_j). \qquad (*)$$

由于$l_i(x)$点x_i处的值是1, 则

$$1 = c\prod_{j\neq i}(x_i - x_j).$$

把上式代回$(*)$, 可以得到$l_i(x)$的表达式:

$$l_i(x) = \prod_{j\neq i}\frac{x - x_j}{x_i - x_j}.$$

这些多项式叫做**基函数**, 且**Lagrange插值多项式**为

$$p(x) = \sum_{k=0}^{n} y_k l_k(x).$$

注: 在一般情况下, 记

$$\boldsymbol{y} = (y_0, y_1, \cdots, y_n)^T, \boldsymbol{p} = (p_0, p_1, \cdots, p_n)^T, \mathbb{L} = (l_i(x_j))_{ij}.$$

把上面$n+1$条标量式

$$y_i = f(x_i) = \sum_{k=0}^{n} y_k l_k(x_i)$$

改为向量形式可得

$$\boldsymbol{y} = \mathbb{L}^T\boldsymbol{y}.$$

在Lagrange插值多项式中$l_k(x_i) = \delta_{ki}$, 此时$\mathbb{L} = \mathbb{I}$(单位阵). 从而

$$\boldsymbol{p} = \boldsymbol{y} = \mathbb{L}^T\boldsymbol{y}.$$

因此可以得到任意次数至多为n次的多项式都满足

$$p(x) = \sum_{i=0}^{n} p(x_i) l_i(x).$$

5.1.3 误差估计

下面设

$$w(x) = \prod_{i=0}^{n}(x - x_i).$$

定理 5.1.2: 插值多项式误差定理

设f是$C^{n+1}[a,b]$中的函数, 多项式p是函数f在区间$[a,b]$的$n+1$个不同点$x_0,\cdots,x_n$上次数不超过n的插值多项式. 对$[a,b]$中的每个x, 都有(a,b)中的一点ξ_x与之对应, 使得

$$f(x)-p(x)=\frac{1}{(n+1)!}f^{(n+1)}(\xi_x)w(x).$$

证明: 当$x=x_i$时, 上式等号两端都为0, 此时定理结论显然成立. 下面**固定**x(异于结点), 令

$$w(x)=\prod_{i=0}^{n}(x-x_i),\phi(t)=f(t)-p(t)-\lambda\omega(t),$$

其中λ与x有关, 使得$\phi(x)=0$, 那么

$$\lambda=\frac{f(x)-p(x)}{w(x)}.$$

由于$\phi\in C^{n+1}[a,b]$, 且在$n+2$个点$x,x_0,x_1,\cdots,x_n$处取0, 由Rolle定理, ϕ'在区间(a,b)中至少有$n+1$个不同的零点. 又由Rolle定理, ϕ''在区间(a,b)中至少有n个不同的零点, 不断重复用Rolle定理, 可以断定$\phi^{(n+1)}$在区间(a,b)中至少有一个零点, 记为ξ_x. 由于

$$\phi^{(n+1)}=f^{(n+1)}-p^{(n+1)}-\lambda w^{(n+1)}=f^{(n+1)}-(n+1)!\lambda,$$

则

$$0=\phi^{(n+1)}(\xi_x)=f^{(n+1)}(\xi_x)-(n+1)!\lambda=f^{(n+1)}(\xi_x)-(n+1)!\frac{f(x)-p(x)}{w(x)}.$$

这是欲证结论. □

§5.2 Chebyshev插值多项式

插值多项式误差定理里面有一项可以被**优化**, Chebyshev在研究蒸汽机车联动装置的运动时引出了Chebyshev多项式, 成为应用数学的重要组成部分.

5.2.1 基本定义与性质

Chebyshev多项式**(第一类)递归定义**如下:

$$\begin{aligned}&T_0(x)=1,T_1(x)=x,\\&T_{n+1}(x)=2xT_n(x)-T_{n-1}(x),n\geq 1.\end{aligned}$$

不难计算出

$$\begin{aligned}T_2(x)&=2x^2-1,\\T_3(x)&=4x^3-3x,\\T_4(x)&=8x^4-8x^2+1,\\T_5(x)&=16x^5-20x^3+5x,\\T_6(x)&=32x^6-48x^4+18x^2-1.\end{aligned}$$

这些多项式有许多性质.

定理 5.2.1

设$x \in [-1,1]$, 则Chebyshev多项式可以写成

$$T_n(x) = \cos(n \arccos x), n \geq 0.$$

证明: 由于$\cos(A+B) = \cos A \cos B - \sin A \sin B$, 则

$$\cos(n+1)\theta = \cos\theta \cos n\theta - \sin\theta \sin n\theta,$$
$$\cos(n-1)\theta = \cos\theta \cos n\theta + \sin\theta \sin n\theta,$$

两式相加然后移项可得

$$\cos(n+1)\theta = 2\cos\theta \cos n\theta - \cos(n-1)\theta.$$

令$\theta = \arccos x (x = \cos\theta)$, 则下面定义的函数

$$f_n(x) = \cos(n \arccos x)$$

满足

$$f_0(x) = 1, f_1(x) = x,$$
$$f_{n+1}(x) = 2x f_n(x) - f_{n-1}(x), n \geq 1.$$

因此对所有的n, 都有$f_n = T_n$. □

注: 根据这个定理可以推出Chebyshev多项式的其他性质: 如

$$|T_n(x)| \leq 1 \qquad (-1 \leq x \leq 1)$$
$$T_n\left(\cos\frac{j\pi}{n}\right) = (-1)^j, \quad (0 \leq j \leq n)$$
$$T_n\left(\cos\frac{2j-1}{2n}\pi\right) = 0, \quad (1 \leq j \leq n)$$

把Chebyshev多项式变成**首一多项式**也很常用, 根据定义可知$T_n(x)$最高次项是$2^{n-1}x^n$, 因此当$n > 0$时, $2^{1-n}T_n$是首一多项式.

定理 5.2.2: 首一多项式定理

若p是n次首一多项式, 则

$$\|p\|_\infty \triangleq \max_{-1 \leq x \leq 1} |p(x)| \geq 2^{1-n}.$$

证明: (反证)设

$$|p(x)| < 2^{1-n}, |x| \leq 1.$$

令$q = 2^{1-n}T_n, x_j = \cos\frac{j\pi}{n}$. 由于$q$首一, 则

$$(-1)^j p(x_j) \leq |p(x_j)| < 2^{1-n} = (-1)^j q(x_j),$$

则

$$(-1)^j [q(x_j) - p(x_j)] > 0, 0 \leq i \leq n.$$

所以在区间$[-1,1]$上, 多项式$q-p$的符号在正负之间变动$n+1$次, 从而由介值定理可知在区间$(-1,1)$内$q-p$至少有n个根, 但这是不可能的, 因为$q-p$次数至多是$n-1$(注意p,q都是首一多项式, $q-p$不会出现x^n). □

5.2.2 用Chebyshev多项式选取结点

下面记$C[-1,1]$空间的无穷范数为

$$\|f\|_\infty \triangleq \max_{x\in[-1,1]} |f(x)|.$$

根据插值多项式误差定理, 有误差估计

$$\|f(x)-p(x)\|_\infty \le \frac{1}{(n+1)!}\|f^{(n+1)}(x)\|_\infty \max_{|x|\le 1}\left|\prod_{i=0}^{n}(x-x_i)\right|.$$

给定了f, 如果能找到“最佳逼近”, 即选取插值点$x_0,\cdots,x_n$得到的插值多项式$p(x)$可以让$\|w\|$尽可能小, 那么就是我们想要的多项式.

根据首一多项式定理, 对任意结点集$\{x_k\}_{k=0}^n$, 我们有

$$\max_{|x|\le 1}\left|\prod_{i=0}^{n}(x-x_i)\right| \ge 2^{1-n},$$

等号成立的条件是

$$\prod_{i=0}^{n}(x-x_i) = 2^{1-n}T_{n+1}$$

那么结点就是多项式T_{n+1}的根! 它们就是

$$x_j = \cos\frac{2j+1}{2n+2}\pi, 0\le j\le n.$$

于是可以得到

定理 5.2.3

若结点x_i是多项式T_{n+1}的根, $p(x)$是$\{(x_i,y_i)\}$的插值多项式, 则对$|x|\le 1$, 有

$$|f(x)-p(x)| \le \frac{1}{2^n(n+1)!}\|f^{(n+1)}\|_\infty.$$

注: 关于函数逼近, 在后续章节会进一步介绍.

§5.3 均差

【待更】

§5.4 Hermite插值

Hermite插值的基本想法是: 求插值多项式$p(x)$使得插值结点处的函数与(高阶)导数都保持一致.

例 5.4.1 下面求函数f和它的导数f'在两个不同结点x_0,x_1上的一个次数最低的插值多项式.

解: 欲求的插值多项式应该满足4个条件:

$$p(x_i)=f(x_i),\qquad p'(x_i)=f'(x_i),\qquad (i=0,1).$$

由于有4个条件, 那么自然要在次数不超过3的全体多项式空间$P^3[a,b]$中求出满足条件的插值多项式, 而$P^3[a,b]$中的多项式有4个系数. 但我们不是把$p(x)$写成$1,x,x^2,x^3$的组合形式, 而是写成

$$p(x)=a+b(x-x_0)+c(x-x_0)^2+d(x-x_0)^2(x-x_1),$$

这是因为这样写可以简化工作, 减小运算量. 这样

$$p'(x) = b + 2c(x - x_0) + 2d(x - x_0)(x - x_1) + d(x - x_0)^2.$$

令$h = x_1 - x_0$, 代入四个条件可知

$$\begin{aligned} f(x_0) &= a, \\ f'(x_0) &= b, \\ f(x_1) &= a + bh + ch^2, \\ f'(x_1) &= b + 2ch + dh^2. \end{aligned}$$

根据这个线性方程组, 可以解出a, b, c, d. 因此无论怎么取值$f(x_i), f'(x_i)$, 这个问题总是有解的.

一般情况下, 如果我们要用一个多项式去插值一个函数与某些导数的值, 由于关于多项式系数的线性方程组可能是奇异的, 所以可能会遇到一些困难. 例如:

例 5.4.2 *求多项式p使得$p(0) = 0, p(1) = 1, p'\left(\frac{1}{2}\right) = 2$.*

解: 有三个条件, 那就考虑用二次多项式来插值:

$$p(x) = a_0 + a_1 x + a_2 x^2.$$

代入题中条件可知

$$\begin{aligned} 0 &= p(0) = a_0, \\ 1 &= p(1) = a_1 + a_2, \\ 2 &= p'\left(\frac{1}{2}\right) = a_1 + a_2. \end{aligned}$$

显然这个问题没有二次多项式解, 因为它的系数矩阵是奇异的. 但是如果我们用三次多项式插值:

$$p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

代入题中条件可知

$$\begin{aligned} 0 &= p(0) = a_0, \\ 1 &= p(1) = a_1 + a_2 + a_3, \\ 2 &= p'\left(\frac{1}{2}\right) = a_1 + a_2 + \frac{3}{4}a_3. \end{aligned}$$

则这个方程组的解是$a_0 = 0, a_3 = -4, a_1 + a_2 = 5$. □

注: 这类问题又称Birkhoff插值问题, 它是Hermite插值的推广, 在Hermite插值过程中放弃在某些点处的某些阶导数取值的要求.

我们下面主要研究具有唯一解的插值问题, 这类问题叫**Hermite插值**, 在Hermite插值问题中, 假设结点x_i处的j阶导数$p^{(j)}(x_i)$给定, 且它的低阶导数$p^{(j-1)}(x_i), \cdots, p'(x_i), p(x_i)$也是给定的.

记k_i表示在结点x_i上给定的插值条件个数, $x_0, \cdots, x_n$是插值结点, 在x_i上给定插值条件为

$$p^{(j)}(x_i) = c_{ij}, 0 \le j \le k_i - 1, 0 \le i \le n \tag{5.1}$$

把多项式$p(x)$上的插值条件总数记为$m + 1$, 则

$$m + 1 = k_0 + k_1 + \cdots + k_n. \tag{5.2}$$

定理 5.4.1: Hermite插值多项式的唯一性

在次数不超过m的多项式组成的空间P^m中, 存在唯一的多项式p满足$(*)$式条件.

证明: P^m中的多项式有$m+1$个系数. 由(5.1), 关于多项式p的条件一共是$m+1$个, 所以我们需要求解一个具有$m+1$个未知量和$m+1$个方程的方程组, 并且希望保证它的系数矩阵非奇异. 要证明一个方阵非奇异, 只需要证明齐次线性方程组$Au=0$只有零解, 即可证明存在唯一性.

在下面讨论的插值问题中, 齐次问题就是求$p\in P^m$, 使得

$$p^{(j)}(x_i)=0,\qquad 0\le j\le k_i-1, 0\le i\le n.$$

其中$x_i(0\le i\le n)$是这个多项式的k_i重零点, 因此它一定是下列多项式q的倍数:

$$q(x)=\prod_{i=0}^{n}(x-x_i)^{k_i},$$

但是q的次数是

$$m+1=\sum_{i=0}^{n}k_i,$$

而p的次数至多是m, 所以$p=q=0$. □

例 5.4.3 *当只有一个结点时, Hermite插值是Taylor展开式:*

$$p(x)=c_{00}+c_{01}(x-x_0)+\frac{c_{02}}{2!}(x-x_0)^2+\cdots+\frac{c_{0k}}{k!}(x-x_0)^k.$$

下面, 我们只考虑特殊情形(又叫**Lagrange型**): 假设$f(x)$在插值结点$x_0,\cdots,x_n$处的函数值是

$$f(x_0),f(x_1),\cdots,f(x_n),$$

一阶导数值为

$$f'(x_0),f'(x_1),\cdots,f'(x_n).$$

欲求插值多项式$p(x)\in P^{2n+1}$使得

$$p(x_i)=f(x_i),p(x_i)=f'(x_i),i=0,1,\cdots,n. \tag{5.3}$$

(根据前面定理可知$p(x)$最多是$2n+1$次的)

我们希望Hermite插值拥有能够像Lagrange插值一样的插值公式. 类似Lagrange插值, 定义

$$\boxed{p(x)=\sum_{i=0}^{n}f(x_i)A_i(x)+\sum_{i=0}^{n}f'(x_i)B_i(x)} \tag{5.4}$$

其中A_i,B_i是具有某些特殊性质的多项式.

回顾Lagrange插值多项式, 下面的函数满足(5.4)式, 也满足(5.3)式(留作作业):

$$\begin{cases}A_i(x_j)=\delta_{ij},\\ A_i'(x_j)=0,\end{cases},\begin{cases}B_i(x_j)=0,\\ B_i'(x_j)=\delta_{ij},\end{cases} \tag{5.5}$$

根据Hermite插值多项式的唯一性, 这就是我们要求的Hermite插值多项式.

接下来想办法把A_i, B_i都求出来, 求法完全类似于Lagrange插值多项式. 定义

$$l_i(x) = \prod_{j\neq i} \frac{x-x_j}{x_i-x_j}, 0 \leq i \leq n.$$
$$w(x) = \prod_{i=1}^{n}(x-x_i).$$

则

$$l_i'(x_i) = \sum_{j\neq i} \frac{1}{x_i-x_j}.$$

由上面对A, B性质的刻画可知$x_0, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$都是$B_i(x)$的二重根, 且$x_i$是$B_i(x)$的单根, 则

$$B_i(x) = b_i(x-x_0)^2(x-x_1)^2\cdots(x-x_{i-1})^2(x-x_i)(x-x_{i+1})^2\cdots(x-x_n)^2,$$

代入$B_i'(x_i) = 1$可知b_i的值, 最终得到

$$\boxed{B_i(x) = (x-x_i)l_i^2(x)}$$

下面确定$A_i(x)$. $x_0, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$都是$A_i(x)$的二重根, 可以令

$$A_i(x) = a_i(x+c_i)(x-x_0)^2(x-x_1)^2\cdots(x-x_{i-1})^2(x-x_{i+1})^2\cdots(x-x_n)^2,$$

代入$A_i(x_i) = 1$与$A_i'(x_i) = 0$, 可得a_i, c_i的值. 最终得到

$$\boxed{A_i(x) = [1-2(x-x_i)l_i'(x_i)]l_i^2(x)}$$

定理 5.4.2: 误差估计

设$x_0, x_1, \cdots, x_n$是区间$[a,b]$中不同的结点, $f \in C^{2n+1}[a,b]$. 若次数至多为$2n+1$次的多项式p满足

$$p(x_i) = f(x_i), p'(x_i) = f'(x_i), 0 \leq i \leq n,$$

则$\forall x \in [a,b], \exists \xi \in (a,b)$, 使得

$$f(x) - p(x) = \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}w^2(x).$$

证明: 证明也完全类似于Lagrange插值多项式的证明. 固定$x \in [a,b]$(不是结点), 定义ϕ为

$$\phi = f - p - \lambda w^2,$$

其中选取λ使得$\phi(x) = 0$. 则ϕ在$[a,b]$中有$n+2$个零点: $x, x_0, x_1, \cdots, x_n$, 反复用多次Rolle定理, 可知$\phi^{(2n+2)}$在区间(a,b)内有一个零点ξ_x, 从而

$$0 = \phi^{(2n+2)}(\xi_x) = f^{(2n+2)}(\xi_x) - p^{(2n+2)}(\xi_x) - \lambda[w^2(x)]^{(2n+2)}|_{x=\xi_x},$$

由于p是至多$2n+1$次多项式, 则$p^{(2n+2)} = 0$; 由于w^2的首项是t^{2n+2}, 则$[w^2(x)]^{(2n+2)} \equiv (2n+2)!$. 再代入$\lambda = \dfrac{f(x)-p(x)}{w(x)}$, 可以得到

$$0 = f^{(2n+2)}(\xi_x) - \frac{f(x)-p(x)}{w(x)}(2n+2)!,$$

稍微整理可得欲证结论. □

作业:

1. 验证(5.5)满足(5.4), 也满足(5.3).
2. 证明$l_i'(x_i) = \sum_{j\neq i} \frac{1}{x_i - x_j}$.
3. 求出a_i, b_i, c_i的表达式.
4. 设$f \in C^n[a,b]$, f在点a有m重根, 在点b有k重根, $m \geq 1, k \geq 1$且$m + k - 1 = n$. 证明: $f^{(n)}$在区间(a,b)中至少有一个零点.

§5.5 样条插值

主要介绍三次样条插值. 个人认为林成森书上这部分写得比较乱, 在这里仅回顾一下它的基本思想.

5.5.1 基本概念

样条函数是由一些具有某些光滑性条件的子区间上的分段多项式构成, 给定$n+1$个点$t_0, t_1, \cdots, t_n$, 满足$t_0 < t_1 < \cdots, t_n$, 这些点称为**结点**. 如果指定一个整数$k \geq 0$, 具有结点$t_0, \cdots, t_n$的一个k次样条函数S满足如下条件:

- 在每个子区间$[t_{i-1}, t_i)$上, S是次数不超过k的多项式.
- 在$[t_0, t_n]$上, S有$k-1$阶连续导数.

零次样条函数就是分段常值函数

$$S(x) = \begin{cases} c_0, & x \in [t_0, t_1) \\ c_1, & x \in [t_1, t_2) \\ \vdots & \\ c_{n-1}, & x \in [t_{n-1}, t_n) \end{cases}$$

一次样条函数是连续的分段一次函数:

$$S(x) = \begin{cases} a_0x + b_0, & x \in [t_0, t_1) \\ a_1x + b_1, & x \in [t_1, t_2) \\ \vdots & \\ a_{n-1}x + b_{n-1}, & x \in [t_{n-1}, t_n) \end{cases}, \text{满足} S_i(t_{i+1}) = S_{i+1}(t_{i+1}).$$

5.5.2 三次样条

考虑

x	t_0	t_1	$\cdots$	t_n
y	y_0	y_1	$\cdots$	y_n

对这个表值插值可以构造一个三次样条函数S. 在每个区间$[t_i, t_{i+1}]$上, S都是不同的三次多项式, 把在$[t_i, t_{i+1}]$上表示的多项式记为S_i, 那么

$$S(x) = \begin{cases} S_0(x), & x \in [t_0, t_1], \\ S_1(x), & x \in [t_1, t_2], \\ \vdots & \\ S_{n-1}(x), & x \in [t_{n-1}, t_n], \end{cases}$$

满足边界条件

$$S_{i-1}(t_i) = y_i = S_i(t_i), 1 \le i \le n-1.$$

则S是连续的. 进一步假设S', S''也连续.

> S, S', S''的连续性能否确定三次样条? 因为有n个三次多项式, 每个多项式有4个系数, 所以这个分段三次多项式有$4n$个系数. 而在每个子区间$[t_i, t_{i+1}]$有两个插值条件$S(t_i) = y_i, S(t_{i+1}) = y_{i+1}$, 这样根据$S$的连续性得到了$2n$个条件.
>
> 在每个内部结点上, S'的连续性可以确定$S'_{i-1}(t_i) = S'_i(t_i)$, 而$S''$的连续性也可以给出$S''_{i-1}(t_i) = S''_i(t_i)$, 一共有$2n-2$个条件. 这样就总共可以确定$4n$个系数的$4n-2$个条件. 剩余的**自由度**是2, 需要合理利用这些自由度.

下面推导$[t_i, t_{i+1}]$上$S_i(x)$的表达式.

Step 1. 首先定义$z_i = S''(t_i)$, 由于S_i是$[t_i, t_{i+1}]$上的三次多项式, 所以S''_i是满足$S''_i(t_i) = z_i$与$S''_i(t_{i+1}) = z_{i+1}$的线性函数, 则S''_i是z_i, z_{i+1}之间的直线:

$$S''_i(x) = \frac{z_i}{h_i}(t_{i+1} - x) + \frac{z_{i+1}}{h_i}(x - t_i).$$

其中$h_i = t_{i+1} - t_i$.

Step 2. 把这个函数积分两次可以得到

$$S_i(x) = \frac{z_i}{6h_i}(t_{i+1} - x)^3 + \frac{z_{i+1}}{6h_i}(x - t_i)^2 + C(x - t_i) + D(t_{i+1} - x).$$

其中, C, D是常数.

Step 3. 再把插值条件$S_i(t_i) = y_i$与$S_i(t_{i+1}) = y_{i+1}$代入上述表达式, 可以确定C, D. 得到

$$S_i(x) = \frac{z_i}{6h_i}(t_{i+1} - x)^3 + \frac{z_{i+1}}{6h_i}(x - t_i)^2 + \left(\frac{y_{i-1}}{h_i} - \frac{z_{i+1}h_i}{6}\right)(x - t_i) + \left(\frac{y_i}{h_i} - \frac{z_i h_i}{6}\right)(t_{i+1} - x).$$

Step 4. 下面确定$z_1, z_2, \cdots, z_{n-1}$. 在内结点$t_i$上, 有$S'_{i-1}(t_i) = S'_i(t_i)$, 对上式求导可得$S'_i(x)$, 从而

$$S'_i(t_i) = -\frac{h_i}{3}z_i - \frac{h_i}{6}z_{i+1} - \frac{y_i}{h_i} + \frac{y_{i+1}}{h_i},$$

同理可得

$$S'_{i-1}(t_i) = -\frac{h_{i-1}}{6}z_{i-1} + \frac{h_{i-1}}{3}z_i - \frac{y_{i-1}}{h_{i-1}} + \frac{y_i}{h_{i-1}},$$

联立两个等式, 得到

$$h_{i-1}z_{i-1} + 2(h_i + h_{i-1})z_i + h_i z_{i+1} = \frac{6}{h_i}(y_{i+1} - y_i) - \frac{6}{h_{i-1}}(y_i - y_{i-1}), i = 1, 2, \cdots, n-1.$$

这样就得到了**含有$n+1$个未知量的$n-1$阶线性方程组**, 可以任意选择z_0, z_n来求解这个方程组, 一个很好的选择是$z_0 = z_n = 0$, 这样得到的样条函数叫**自然三次样条**.

对于$1 \le i \le n-1, z_0 = z_n = 0$, 线性方程组是对称、三对角、对角占优的, 它可以写成

$$\begin{pmatrix} u_1 & h_1 & & & & \\ h_1 & u_2 & h_2 & & & \\ & h_2 & u_3 & h_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & h_{n-3} & u_{n-2} & h_{n-2} \\ & & & & h_{n-2} & u_{n-1} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_{n-2} \\ z_{n-1} \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{n-2} \\ v_{n-1} \end{pmatrix}$$

其中

$$\begin{aligned} h_i &= t_{i+1} - t_i \\ u_i &= 2(h_i + h_{i+1}) \\ b_i &= \frac{6}{h_i}(y_{i+1} - y_i) \\ v_i &= b_i - b_{i-1}. \end{aligned}$$

利用追赶法, 就可以求出所有z_i了. □

5.5.3 三次样条的性质

自然三次样条是“最光滑”的插值函数.

定理 5.5.1: 最优性定理

设f''在$[a,b]$内连续且$a = t_0 < t_1 < \cdots < t_n = b$, 若$S$是$f$在结点$t_i$上的自然三次样条插值, $0 \le i \le n$, 则

$$\int_a^b [S''(x)]^2 \mathrm{d}x \le \int_a^b [f''(x)]^2 \mathrm{d}x.$$

证明: 设$g = f - S$, 则$g(t_i) = 0$, 且

$$\int_a^b (f'')^2 \mathrm{d}x = \int_a^b (S'')^2 \mathrm{d}x + \int_a^b (g'')^2 \mathrm{d}x + 2\int_a^b S''g'' \mathrm{d}x.$$

只需证明

$$\int_a^b S''g'' \mathrm{d}x \ge 0.$$

事实上, 由分部积分、条件$S''(t_0) = S''(t_n) = 0$以及$[t_{i-1}, t_i]$上S'''是常数c_i, 可知

$$\begin{aligned} \int_a^b S''g'' &= \sum_{i=1}^n \int_{t_{i-1}}^{t_i} S''g'' \mathrm{d}x \\ &= \sum_{i=1}^n \left\{ (S''g')(t_i) - (S''g')(t_{i-1}) - \int_{t_{i-1}}^{t_i} S'''g' \mathrm{d}x \right\} \\ &= -\sum_{i=1}^n c_i \int_{t_{i-1}}^{t_i} g' \mathrm{d}x \\ &= -\sum_{i=1}^n c_i [g(t_i) - g(t_{i-1})] = 0. \end{aligned}$$

定理证毕. □

§5.6 (*)附录: Bernstein多项式

本节带(*)表示是课程之外的补充内容.

5.6.1 Bernstein多项式的定义

Serge Bernstein在1912年引进了下面的Bernstein多项式:

$$B_k^n(x) = \binom{n}{k} x^k (1-x)^{n-k}$$

把B_k^n叫做区间$[0,1]$上次数为n的Bernstein多项式.

Bernstein多项式满足下面的性质: (其中P_n表示不超过n次的所有多项式构成的函数空间)

- 非负性: $B_k^n(x) \geq 0, x \in [0,1]$;
- 规范性: $\sum_{k=0}^{n} B_k^n(x) = 1,\ x \in \mathbb{R}$;
- 积分相等: $\int_0^1 B_k^n(x)\mathrm{d}x = \frac{1}{n+1}, k = 0, 1, \cdots, n.$
- 导数: $(B_k^n(x))' = n(B_{k-1}^{n-1}(x) - B_k^{n-1}(x))$.
- 对称关系式: $B_k^n(x) = B_{n-k}^n(1-x),\ k = 0, 1, \cdots, n;$
- 递推关系式: $B_0^n(x) = (1-x)B_0^{n-1}(x),\ B_n^n(x) = tB_{n-1}^{n-1}(x),\ \forall x \in \mathbb{R}, n \in \mathbb{N}.$
- 递推关系式: $B_k^n(x) = xB_{k-1}^{n-1}(x) + (1-x)B_k^{n-1}(x),\ x \in \mathbb{R},\ n \in \mathbb{N},\ k = 1, 2, \cdots, n-1.$
- $x = 0$是B_k^n的k次零点, $x = 1$是B_k^n的$n-k$次零点.
- B_k^n在$x = \frac{k}{n}$处取最大值.
- 多项式$B_0^n, \cdots, B_n^n$是P_n的一组基.

5.6.2 逼近性质

利用Bernstein多项式作为基, 我们可以用来逼近连续函数f:

$$(B_n f)(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) B_k^n(x),$$

上面的$B_n : C[0,1] \to C[0,1]$可以看做是**线性算子**, 即有如下表示:

$$B_n(af + bg) = aB_n f + bB_n g, \qquad a, b \in \mathbb{R}; f, g \in C[0,1].$$

除此之外, B_n是**正算子**, 即如果$f \geq 0$, 则$B_n f \geq 0$.

下面的定理解释了Bernstein多项式的逼近性质, 基于此我们可以很快证明Weierstrass多项式逼近定理.

定理 5.6.1: Bohman-Korovskin

设$L_n (n \geq 1)$是定义在$C[a,b]$上的正线性算子序列, 若对于三个函数$f(x) = 1, x, x^2$, $\|L_n f - f\|_\infty \to 0$, 则对所有$f \in C[a,b]$此结论也成立.

证明: 在$C[a,b]$中, 我们可以取绝对值. 若L是正线性算子, 则L具有单调性, 即

$$f \geq g \Rightarrow f - g \geq 0 \Rightarrow L(f-g) \geq 0 \Rightarrow Lf - Lg \geq 0 \Rightarrow Lf \geq Lg.$$

由于$|f| \geq f$, $|f| \geq -f$, 所以$L(|f|) \geq Lf$, $L(|f|) \geq -Lf$, 从而$L(|f|) \geq |Lf|$.

对于$k = 0, 1, 2$, 我们设$h_k(x) = x^k$, 并且定义函数$\alpha_n, \beta_n, \gamma_n$为

$$\alpha_n = L_n h_0 - h_0, \quad \beta_n = L_n h_1 - h_1, \quad \gamma_n = L_n h_2 - h_2.$$

由定理假设我们可以断定

$$\|\alpha_n\|_\infty \to 0, \quad \|\beta_n\|_\infty \to 0, \quad \|\gamma_n\|_\infty \to 0.$$

对任意$f \in C[a,b]$与$\varepsilon > 0$, 下证存在整数N使得当$n \geq N$时, $\|L_n f - f\|_\infty < 3\varepsilon$. 由于$f$在一个紧区间上连续, 所以$f$一致连续. 所以存在$\delta > 0$, 使得对任意$x, y \in [a,b]$, 当$|x-y| < \delta$时, $|f(x) - f(y)| < \varepsilon$.

取$c = \dfrac{2\|f\|_\infty}{\delta^2}$, 有

$$|x-y| \geq \delta \Rightarrow |f(x) - f(y)| \leq 2\|f\|_\infty \leq 2\|f\|_\infty \frac{(x-y)^2}{\delta^2} = c(x-y)^2.$$

所以对任意$x, y \in [a,b]$, 有

$$|f(x) - f(y)| \leq \varepsilon + c(x-y)^2,$$

这个不等式可以写成

$$|f - f(y)h_0| \leq \varepsilon h_0 + c[h_2 - 2yh_1 + y^2 h_0], \qquad \forall x \in [a,b].$$

根据h_k的定义以及简单代换x可以得到上述结果. 由定理证明一开始的叙述, 有

$$|L_n f - f(y)L_n h_0| \leq \varepsilon L_n h_0 + \varepsilon[L_n h_2 - 2yL_n h_1 + y^2 L_n h_0],$$

这是函数之间的不等式, 我们可以对y作换元:

$$\begin{aligned}
|(L_n f)(y) - f(y)(L_n h_0)(y)| &\leq \varepsilon(L_n h_0)(y) + c[(L_n h_2)(y) - 2y(L_n h_1)(y) + y^2(L_n h_0)(y)] \\
&= \varepsilon[1 + \alpha_n(y)] + c[y^2 + \gamma_n(y) - 2y(y + \beta_n(y)) + y^2(1 + \alpha_n(y))] \\
&= \varepsilon + \varepsilon\alpha_n + c\gamma_n(y) - 2cy\beta_n(y) + cy^2\alpha_n(y) \\
&\leq \varepsilon + \varepsilon\|\alpha_n\|_\infty + c\|\gamma_n\|_\infty + 2c\|h_1\|_\infty\|\beta_n\|_\infty + c\|h_2\|_\infty\|\alpha_n\|_\infty.
\end{aligned}$$

取N使得当$n \geq N$时上述不等式最后的右端小于2ε. 则对$n \geq N$, 有

$$\|L_n f - f \cdot L_n h_0\|_\infty \leq 2\varepsilon.$$

最后, 我们有

$$\begin{aligned}
\|L_n f - f\|_\infty &\leq \|L_n f - f \cdot L_n h_0\|_\infty + \|f \cdot L_n h_0 - f \cdot h_0\|_\infty \\
&\leq 2\varepsilon + \|f\|_\infty\|\alpha_n\|_\infty.
\end{aligned}$$

如果需要的话, 可以增加N使得当$n \geq N$时, $\|f\|_\infty\|\alpha_n\|_\infty < \varepsilon$, 那么上述不等式中的最后项不超过$3\varepsilon$. □

习题:

1. 证明: $\displaystyle\sum_{k=0}^{n} \frac{k}{n} B_k^n(x) = x$, $x \in \mathbb{R}$.
2. 证明: $\displaystyle\sum_{k=0}^{n} \frac{k^2}{n^2} B_k^n(x) = \frac{n-1}{n}x^2 + \frac{x}{n}$, $x \in \mathbb{R}$.

CHAPTER 6

数值积分

给定函数f在$n+1$个点$x_0, x_1, \cdots, x_n$上的值, 是否可以利用这些值近似计算$f'(c)$或者$\int_a^b f(x)dx$?

如果仅知道$f \in C[a,b]$与函数值$f(x_0), \cdots, f(x_n)$, 满足这样的函数有很多, 仅知道$f(x_i)$几乎是无用的.

如果知道f是次数最多为n次的多项式, 那么根据前面的插值理论, $n+1$个点的函数值可以完全确定f, 此时可以求出f并且可以求出$f'(c)$与$\int_a^b f(x)dx$的精确值.

在大多数情况下, 已掌握的信息无法完全确定f, 它的导数或积分的数值可能误差很大, 除非给出某些相应误差的界.

§ 6.1 数值微分与Richardson外推

6.1.1 数值微分

给定一个函数$f(x)$, 如何数值求它的导数? 最简单的思路: 根据$f'(x)$的定义:

$$f'(x) \approx \frac{1}{h}[f(x+h) - f(x)]. \qquad (*)$$

对于线性函数$f(x) = ax + b$, $f'(x)$的值是**精确**的. 在其他情况下该公式可能精确, 但也只是极为偶然的情况. 所以我们把它的误差求出来. 求误差一般都用Taylor展开**(在学习偏微分方程数值解这门课的时候会频繁用到Taylor展开)**,

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\xi).$$

其中$\xi \in (x, x+h)$, 且f充分光滑. 整理上式得到

$$f'(x) - \frac{1}{h}[f(x+h) - f(x)] = -\frac{h}{2}f''(\xi). \qquad (**)$$

对上述的一大类函数, 余项与基本的数值公式出现在同一条式子, 这条式子比$(*)$更有用. 可以发现, 当$h \to 0$时, 余项中的h会使得误差也会趋于0. 为了精确计算$f'(x)$, 步长h要取得很小.

> **名人名言:** 不会就Taylor展开. ——dyb老师

把$(**)$的项$-\frac{h}{2}f''(\xi)$称为**截断误差**. 如果截断误差形如Ch^k, 则把**截断误差的阶**记为$O(h^k)$. 这是Taylor公式截掉高阶项后得到的误差. 截断误差与舍入误差有着同样重要的作用.

公式(**)的误差是$O(h)$, 能否把它提高到$O(h^2)$? 可以, 一个较好的公式是

$$f'(x) \approx \frac{1}{2h}[f(x+h) - f(x-h)],$$

根据Taylor公式,

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{3!}f'''(\xi_1),$$
$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{3!}f'''(\xi_2),$$

两式相减可得

$$f'(x) = \frac{1}{2h}[f(x+h) - f(x-h)] - \frac{h^2}{12}[f'''(\xi_1) + f'''(\xi_2)].$$

截断误差的阶变成了$O(h^2)$, 所以这个是更受欢迎的结果.

如果增加一个小小的假定: 函数$f'' \in C[x-h, x+h]$, 根据介值定理, 存在$\xi \in [x-h, x+h]$使得$f''(\xi) = \frac{1}{2}[f'''(\xi_1) + f'''(\xi_2)]$. 这样,

$$f'(x) = \frac{1}{2h}[f(x+h) - f(x-h)] - \frac{h^2}{6}f'''(\xi).$$

同样可以用差商来估计$f''(x)$. 由Taylor展开,

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{3!}f'''(x) + \frac{h^4}{4!}f^{(4)}(\xi_1),$$
$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{3!}f'''(x) + \frac{h^4}{4!}f^{(4)}(\xi_2),$$

那么

$$f''(x) = \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)] - \frac{h^2}{12}f^{(4)}(\xi),$$

其中$\xi \in (x-h, x+h)$. 此时也常用在二阶微分方程的数值解中.

作业:

1. 证明二阶求导近似公式

$$f''(x) = \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)] - \frac{h^2}{12}f^{(4)}(\xi).$$

2. 证明求导近似公式

$$f'(x) \approx \frac{1}{12h}[-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)]$$

的误差项是$O(h^4)$的.

3. 证明二阶求导近似公式

$$f''(x) \approx \frac{1}{12h^2}[-f(x+2h) + 16f(x+h) - 30f(x) + 16f(x-h) - f(x-2h)]$$

的误差项是$O(h^4)$的.

6.1.2 Richardson外推

Richardson外推可以改进数值格式的精度. 由Taylor展开,

$$f(x+h)=\sum_{k=0}^{\infty}\frac{1}{k!}h^k f^{(k)}(x),$$
$$f(x-h)=\sum_{k=0}^{\infty}\frac{1}{k!}h^k(-1)^k f^{(k)}(x),$$

第一条式子减去第二条式子即可消去所有k是偶数的项:

$$f(x+h)-f(x-h)=2hf'(x)+\frac{2}{3!}h^3f'''(x)+\frac{2}{5!}h^5f^{(5)}(x)+\cdots$$

整理得

$$f'(x)=\frac{1}{2h}[f(x+h)-f(x-h)]-\left[\frac{1}{3!}h^2f^{(3)}(x)+\frac{1}{5!}h^4f^{(5)}(x)+\cdots\right]$$

这个等式形如

$$L=\varphi(h)+a_2h^2+a_4h^4+a_6h^6+\cdots. \qquad (a)$$

下面设计的数值过程用于估计L. 直观上(形式上), $\lim\limits_{h\to0}L$是精确值. 对于$h>0$, 可以计算$\varphi(h)$的值, 但我们不能计算$\varphi(0)$, 因而只能计算L的近似值, 方法就是让h趋于0.

对于每个$h>0$, 误差为$a_2h^2+a_4h^4+\cdots$. 若$a_2\neq0$, 则当h充分小时, a_2h^2大于其他项, 所以要设法消去a_2h^2.

如果用$\frac{h}{2}$代替$(*)$中的h, 可得

$$L=\varphi\left(\frac{h}{2}\right)+a_2\frac{h^2}{4}+a_4\frac{h^4}{16}+a_6\frac{h^6}{64}+\cdots \qquad (b)$$

让(a)式减去4倍(b)式可以消去a_2h^2, 得到

$$3L=4\varphi\left(\frac{h}{2}\right)-\varphi(h)-3a_4\frac{h^4}{4}-15a_6\frac{h^6}{16}-\cdots.$$

因此

$$L=\frac{4}{3}\varphi\left(\frac{h}{2}\right)-\frac{1}{3}\varphi(h)-a_4\frac{h^4}{4}-5a_6\frac{h^6}{16}-\cdots. \qquad (c)$$

这条式子提供了用$\varphi(h)$与$\varphi(h/2)$来计算L的方法, 它的误差是$O(h^4)$的.

当然, 上述过程可以不断进行下去. 令

$$\psi(h)=\frac{4}{3}\varphi\left(\frac{h}{2}\right)-\frac{1}{3}\varphi(h),$$

并把(c)式改写为

$$L=\psi(h)+b_4h^4+b_6h^6+\cdots, \qquad (d)$$

用$\frac{h}{2}$代替h得

$$L=\psi\left(\frac{h}{2}\right)+b_4\frac{h^4}{16}+b_6\frac{h^6}{64}+\cdots, \qquad (e)$$

让(d)式减去16倍的(e)可得

$$L=\frac{16}{15}\psi\left(\frac{h}{2}\right)-\frac{1}{15}\psi(h)-\frac{b_5h^6}{20}-\cdots$$

如果再令$\theta(h)=\frac{16}{15}\psi\left(\frac{h}{2}\right)-\frac{1}{15}\psi(h)$, 用同样的方法可得

$$L=\frac{64}{63}\theta\left(\frac{h}{2}\right)-\frac{1}{63}\theta(h)-3c_8\frac{h^8}{252}-\cdots.$$

事实上可以执行任意多步得到不断增加精确度的公式, 这个算法就是**Richardson外推法**.

作业:

1. 验证
$$L=\frac{64}{63}\theta\left(\frac{h}{2}\right)-\frac{1}{63}\theta(h)-3c_8\frac{h^8}{252}-\cdots.$$
2. 设一个用$\varphi(h)$来逼近L的数值过程为
$$L=\varphi(h)+\sum_{j=1}^{\infty}a_jh^j,$$
这个格式的误差是$O(h)$的. 用Richardson外推法构造一个误差达到$O(h^3)$的公式.

§6.2 插值积分

数值计算积分

$$\int_a^b f(x)dx$$

的一个有效策略是用函数g来逼近f, 其中g的积分容易计算. 这样

$$\int_a^b f(x)dx\approx\int_a^b g(x)dx.$$

选取g为不超过一定次数的多项式是个不错的选择.

例 6.2.1 计算$\int_0^1 e^{x^2}dx$的近似值时, 根据*Taylor*公式, $e^{x^2}\approx 1+x^2+\frac{x^4}{2}+\frac{x^6}{6}(x\approx 0)$, 所以

$$\int_0^1 e^{x^2}\mathrm{d}x\approx\int_0^1\left(1+x^2+\frac{x^4}{2}+\frac{x^6}{6}\right)\mathrm{d}x=\frac{51}{35}=1.45714\cdots$$

而真实值为$1.46265\cdots$.

回顾前一章的插值, 定义

$$l_i(x)=\prod_{j\neq i}\frac{x-x_j}{x_i-x_j},0\le i\le n,$$

则$f(x)$在结点$x_0,x_1,\cdots,x_n$的Lagrange插值多项式为

$$p(x)=\sum_{i=0}^{n}f(x_i)l_i(x).$$

用$p(x)$来近似$f(x)$, 可以简单写出

$$\int_a^b f(x)dx\approx\int_a^b p(x)dx=\sum_{i=0}^{n}f(x_i)\int_a^b l_i(x)dx.$$

由于$l_i(x)$是多项式, 它的积分很容易就可以被求出来. 记$A_i = \int_a^b l_i(x)dx$, 则积分公式可以改写为

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i).$$

如果结点$x_0, \cdots, x_n$是**等距**的, 那么把上面的公式称作**Newton-Cotes型积分公式**.

6.2.1 梯形公式

取$n = 1$, 结点取$x_0 = a, x_1 = b$, 则

$$l_0(x) = \frac{b-x}{b-a}, l_1(x) = \frac{x-a}{b-a},$$

从而

$$A_0 = \int_a^b l_0(x)dx = \frac{1}{2}(b-a) = \int_a^b l_1(x)dx = A_1.$$

那么求积公式为

$$\boxed{\int_a^b f(x)dx \approx \frac{b-a}{2}[f(a)+f(b)]}.$$

这条公式叫做**梯形公式**.

下面确定梯形公式的误差. 由Lagrange插值多项式的误差公式,

$$f(x) - p_1(x) = \frac{1}{2}f''(\xi_x)(x-a)(x-b),$$

两边积分可得

$$\int_a^b f(x)dx - \int_a^b p_1(x)dx = \frac{1}{2}\int_a^b f''(\xi_x)(x-a)(x-b)dx = -\frac{1}{12}(b-a)^3 f''(\xi), \text{其中}\xi \in (a,b).$$

因此用插值多项式积分来代替$f(x)$的积分的误差为$-\frac{1}{12}(b-a)^3 f''(\xi)$.

如果a, b相距比较远, 梯形公式的误差将会很大(因为误差项有$(b-a)^3$), 此时可以考虑把$[a,b]$划分:

$$a = x_0 < x_1 < \cdots < x_n = b,$$

在每个子区间上用梯形公式(注意结点未必等距), 这样可以得到**复合梯形公式**:

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)dx \\ &\approx \frac{1}{2}\sum_{i=1}^{n}(x_i - x_{i-1})[f(x_{i-1}) + f(x_i)].\end{aligned}$$

注: 后面准备提到的积分公式都可以用在划分之后的子区间上, 这样可以得到许多**复合法则**.

如果把划分结点取为**等距**, 即$h = \frac{b-a}{n}, x_i = a + ih$, 那么复合梯形公式变为(推导留作作业)

$$\int_a^b f(x)dx \approx \frac{h}{2}\left[f(a) + 2\sum_{i=1}^{n-1} f(a+ih) + f(b)\right]$$

此时误差项为$-\frac{1}{12}(b-a)h^2 f''(\xi)$.

作业:

1. 推导等距划分结点的复合梯形公式与误差项.
2. 在Newton-Cotes方法中, 如果取$n=2,[a,b]=[0,1]$, 可得

$$\int_0^1 f(x)dx \approx \frac{1}{6}f(0)+\frac{2}{3}f\left(\frac{1}{2}\right)+\frac{1}{6}f(1).$$

(把$l_i(x)$求出来, 然后积分即可)

3. 推导基于结点$0,\frac{1}{3},\frac{2}{3},1$的$\int_0^1 f(x)dx$的Newton-Cotes公式.

6.2.2 待定系数法与Simpson公式

计算$l_i(x)$的过程无疑是痛苦的.

在Newton-Cotes积分公式中, 可以立即看出这个积分公式对次数不超过n的多项式精确成立. 此时称积分公式具有n次的**代数精度**. 这样, 用**待定系数法**就可以确认Newton-Cotes积分公式的系数.

例 6.2.2 *寻找一个Newton-Cotes积分公式*

$$\int_0^1 f(x)dx \approx A_0 f(0)+A_1 f\left(\frac{1}{2}\right)+A_2 f(1)$$

使得它对所有次数不超过2的多项式精确成立.

解: 代入$f(x)=1,x,x^2$可得

$$\begin{aligned}1&=\int_0^1 dx = A_0+A_1+A_2\\ \frac{1}{2}&=\int_0^1 xdx=\frac{1}{2}A_1+A_2\\ \frac{1}{3}&=\int_0^1 x^2dx=\frac{1}{4}A_1+A_2.\end{aligned}$$

这个方程组的解是$A_0=\frac{1}{6},A_1=\frac{2}{3},A_2=\frac{1}{6}$. 而且对所有二次多项式$f(x)=c_0+c_1x+c_2x^2$这个积分公式都是精确成立的. □

注: 利用待定系数法, 如果给定了$n+1$个插值点(可以不等距), 且插值多项式存在, 那么最多可以找到一个多项式使得它对次数不超过n的多项式都成立.

如果把这个例题的$[0,1]$换成$[a,b]$, 实际上就得到了**Simpson公式**:

$$\boxed{\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right].}$$

对于所有次数不超过2的多项式, Simpson公式都精确成立. 出乎意料的是, 它对次数为3的多项式也精确成立!

对于Simpson公式, 利用积分中值定理, 存在$\eta\in(a,b)$, 使得

$$\text{RHS}=\frac{1}{24}f^{(3)}(\eta)\int_a^b(x-a)\left(x-\frac{a+b}{2}\right)(x-b)\mathrm{d}x=0.$$

这个分析方式不能像梯形公式那样行得通. 不过可以用其他方法证明Simpson公式的误差项是$-\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi)$.

当然, 与梯形公式一样, Simpson公式也有它的复合版本. 设n是**偶数**, 且结点是等距的. 令

$$x_i=a+ih, h=\frac{b-a}{n}, 0\le i\le n,$$

则

$$\int_a^b f(x)dx = \sum_{i=1}^{n/2} \int_{2i-2}^{2i} f(x)dx,$$

在每个子区间用Simpson公式可得

$$\int_a^b f(x) \approx \frac{h}{3} \sum_{i=1}^{n/2} [f(x_{2i-2}) + 4(x_{2i-1}) + f(x_{2i})].$$

为了避免项的重复, 这个公式的右端项可以改写为

$$\frac{h}{3}\left[f(x_i) + 2\sum_{i=2}^{n/2} f(x_{2i-2}) + 4\sum_{i=1}^{n/2} f(x_{2i-1}) + f(x_n)\right].$$

复合Simpson公式的误差项是$-\frac{1}{180}(b-a)h^4 f^{(4)}(\xi)$.

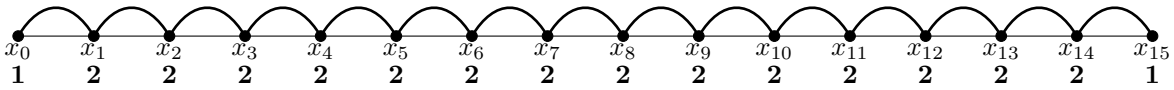


图1: 复合梯形公式

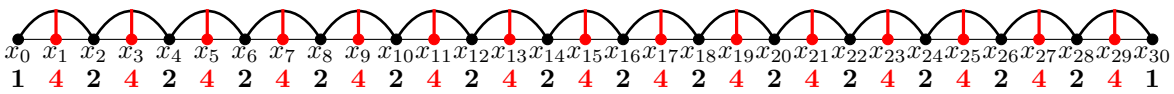


图2: 复合Simpson公式

例 6.2.3 设$f \in C^1[a,b]$并且2阶导数存在. 考虑积分$\int_a^b f(x)\mathrm{d}x$的等距结点$x_0, \cdots, x_n$上的复合梯形公式

$$I_n(f) = h\left[\frac{1}{2}f_0 + f_1 + \cdots + f_{n-1} + \frac{f_n}{2}\right].$$

如果$f'(a)$与$f'(b)$已知, 给出一个能改善精度的求积公式.

证明: 由于

$$E_n(f) = \int_a^b f(x)\mathrm{d}x - I_n(f) = \sum_{j=1}^{n} -\frac{1}{12}h^3 f''(\xi_j), \xi_j \in [x_{j-1}, x_j], \qquad (*)$$

并且我们有误差估计式

$$\int_{x_{i-1}}^{x_i} f''(x)\mathrm{d}x - (x_i - x_{i-1})f''(\xi_i) = O((x_i - x_{i-1}))^2,$$

(这个式子用Riemann积分的定义容易证明.) 化简得

$$f'(x_i) - f'(x_{i-1}) - hf''(\xi_i) = O(h^2),$$

所以

$$f'(b) - f'(a) = \sum_{i=1}^{n}[f'(x_i) - f'(x_{i-1})] = h\sum_{i=1}^{n} f''(\xi_i) + O(h).$$

所以$(*)$可以化为

$$E_n(f) = -\frac{1}{12}h^2[f'(b) - f'(a) + O(h)] = -\frac{1}{12}h^2[f'(b) - f'(a)] + O(h^3).$$

我们给出改善的求积公式

$$T_n(f) = h\left[\frac{1}{2}f_0 + f_1 + \cdots + f_{n-1} + \frac{f_n}{2}\right] - \frac{1}{12}h^2[f'(b) - f'(a)] = I_n(f) - \frac{1}{12}h^2[f'(b) - f'(a)],$$

那么

$$\tilde{E}_n(f) := \int_a^b f(x)\mathrm{d}x - T_n(f) = O(h^3),$$

可以提升精度. □

注: 如果$f \in C^2[a,b]$, 那么用Euler-Maclaurin公式可以证明我们的求积公式$T_n(f)$的精度达到$O(h^4)$.

6.2.3 带权的插值积分

如果把Lagrange插值公式

$$f(x) = \sum_{i=0}^{n} f(x_i)l_i(x)$$

两边乘上**权函数**$W(x)$, 再积分, 可得

$$\int_a^b f(x)W(x)dx \approx \sum_{i=0}^{n} A_i f(x_i),$$

其中

$$A_i = \int_a^b l_i(x)W(x)dx.$$

例 6.2.4 *求公式*

$$\int_{-\pi}^{\pi} f(x)\cos x dx \approx A_0 f\left(-\frac{3\pi}{4}\right) + A_1 f\left(-\frac{\pi}{4}\right) + A_2 f\left(\frac{\pi}{4}\right) + A_3 f\left(\frac{3\pi}{4}\right).$$

当f是一个三次多项式时它是精确成立的.

解: 代入$f(x) = 1, x, x^2, x^3$就可以确定系数. 由对称性可知$A_0 = A_3, A_1 = A_2$, 且结果简化为

$$\begin{aligned} 0 &= \int_{-\pi}^{\pi} = 2A_0 + 2A_1, \\ -4\pi &= \int_{-\pi}^{\pi} x^2\cos x dx = 2A_0\left(\frac{3\pi}{4}\right)^2 + 2A_1\left(\frac{\pi}{4}\right)^2 \end{aligned}$$

解得$A_1 = A_2 = -A_0 = -A_3 = \dfrac{4}{\pi}$, 则求积公式为

$$\int_{-\pi}^{\pi} f(x)\cos x dx \approx \frac{4}{\pi}\left[-f\left(-\frac{3\pi}{4}\right) + f\left(-\frac{\pi}{4}\right) + f\left(\frac{\pi}{4}\right) - f\left(\frac{3\pi}{4}\right)\right].$$

作业:

1. 证明Simpson公式的误差项是$-\dfrac{1}{90}\left(\dfrac{b-a}{2}\right)^5 f^{(4)}(\xi)$.
2. 复合Simpson公式的误差项是$-\dfrac{1}{180}(b-a)h^4 f^{(4)}(\xi)$.
3. 下面公式对次数不超过4的多项式精确成立:

$$\int_0^1 f(x)dx \approx \frac{1}{90}\left[7f(0) + 32f\left(\frac{1}{4}\right) + 12f\left(\frac{1}{2}\right) + 32f\left(\frac{3}{4}\right) + 7f(1)\right].$$

4. 求公式
$$\int_0^1 f(x)dx \approx A_0 f(0) + A_1 f(1),$$
使得它对所有形如$f(x) = ae^x + b\cos\frac{\pi x}{2}$的函数精确成立.
5. 求公式
$$\int_0^{2\pi} f(x)dx \approx A_1 f(0) + A_2 f(\pi)$$
使得它对任何具有形式$f(x) = a + b\cos x$的函数精确成立. 证明所得到的公式对任何形如
$$f(x) = \sum_{k=0}^{n}[a_k \cos(2k+1)x + b_k \sin kx]$$
的函数也是精确成立的.

§6.3 内积空间与正交多项式

内积在最佳逼近、Gauss积分都会用到. 这里仅介绍**实内积空间**.

定义 6.3.1: 内积公理

设V是线性空间, 满足下面几个条件的$(\cdot,\cdot)$叫做内积, 此时V也是内积空间.
- 对称性: $(f,g) = (g,f), \forall f,g \in V$.
- 线性性: $(f, \alpha g + \beta h) = \alpha(f,g) + \beta(f,h), \forall f,g,h \in V, \alpha,\beta \in \mathbb{R}$.
- 正定性: 若$f \neq 0$, 则$(f,f) > 0$. 且$(f,f) = 0$当且仅当$f = 0$.
- 如果V是赋范线性空间, 赋予范数$\|\cdot\|$, 那么
$$\|\cdot\| = \sqrt{(f,f)}.$$

例 6.3.1 $\mathbb{R}^n$上可以定义内积
$$(x,y) = \sum_{i=1}^{n} x_i y_i.$$

例 6.3.2 $[a,b]$上带权W的连续函数空间为$C_W[a,b]$, 定义内积
$$(f,g) = \int_a^b f(x)g(x)W(x)dx.$$
其中W是连续的正函数.

设G是内积空间. 如果$(f,g) = 0$, 那么记$f \perp g$; 如果任意$g \in G$都有$f \in g$, 那么记$f \perp G$.
在内积空间中常常需要找标准正交系, 如果$\{f_n\}$满足
$$(f_i, f_j) = 0, i \neq j$$
则称$\{f_n\}$是**正交的**. 如果对所有i,j都有
$$(f_i, f_j) = \delta_{ij},$$
则称$\{f_n\}$是**标准正交的**.

关于内积空间的性质, 这里不介绍太多(详细可以参考泛函分析教材). 这里介绍内积仅为了推导Legendre多项式.

定理 6.3.1: 正交多项式定理

如下归纳定义的多项式序列是正交的:

$$p_n(x) = (x - a_n)p_{n-1}(x) - b_n p_{n-2}(x), n \geq 2,$$

其中,

$$p_0(x) = 1, p_1(x) = x - a_1,$$

并且

$$a_n = \frac{(xp_{n-1}, p_{n-1})}{(p_{n-1}, p_{n-1})}, b_n = \frac{(xp_{n-1}, p_{n-2})}{(p_{n-2}, p_{n-2})},$$

证明: 根据递推式可知p_n是首一多项式, 从而不恒为0. 从而a_n, b_n良好定义. 下面对n归纳来证明

$$(p_n, p_i) = 0, \forall 1 \leq i \leq n-1.$$

当$n = 1$时, 由a_1的定义可知

$$(p_1, p_0) = ((x - a_1)p_0, p_0) = (xp_0, p_0) - a_1(p_0 - p_0) = 0.$$

假设对$n-1$的情况结论正确, $n \geq 2$, 那么

$$(p_n, p_{n-1}) = (xp_{n-1}, p_{n-1}) - a_n(p_{n-1}, p_{n-1}) - b_n(p_{n-2}, p_{n-1}) = 0,$$
$$(p_n, p_{n-2}) = (xp_{n-1}, p_{n-2}) - a_n(p_{n-1}, p_{n-2}) - b_n(p_{n-2}, p_{n-2}) = 0.$$

对任一$i = 0, 1, \cdots, n-3$, 有

$$(p_n, p_i) = (xp_{n-1}, p_i) - a_n(p_{n-1}, p_i) - b_n(p_{n-2}, p_i) = (p_{n-1}, xp_i) = (p_{n-1}, p_{i+1} + a_{i+1}p_i + b_{i+1}p_{i-1}) = 0.$$

(最后一步中, 如果$i = 0$, 那么$xp_0 = p_1 + a_1 p_0$). □

注: 如果p_n是其中一个n次正交多项式, 则$p_n \perp P^{n-1}[a, b]$.

注: 这也是丘赛【**2011Individual, 1(1)**】的题目.

例 6.3.3 (Legendre多项式) 由前一定理, 设$p_0(x) = 1, p_1(x) = x, W(x) \equiv 1, [a, b] = [-1, 1]$, 可以推导出(首一的) *Legendre*多项式.

解: 最初几步计算如下:

$$p_0(x) = 1, a_1 = \frac{(xp_0, p_0)}{(p_0, p_0)} = 0,$$
$$p_1(x) = x, a_2 = \frac{(xp_1, p_1)}{(p_1, p_1)} = 0, b_2 = \frac{(xp_1, p_0)}{(p_0, p_0)} = \frac{1}{3},$$
$$p_2(x) = x^2 - \frac{1}{3}.$$

接下来的三个Legendre多项式是

$$p_3(x) = x^3 - \frac{3}{5}x,$$
$$p_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35},$$
$$p_5(x) = x^5 - \frac{10}{9}x^3 + \frac{5}{21}x.$$

注: 这里的Legendre多项式与林成森教材的差了个常数倍, 但是林成森教材的讲授思路是首先定义

$$p_0(x)=1, p_n(x)=\frac{1}{2^n n!}\frac{d^n(x^2-1)^n}{dx^n}, n=1,2,\cdots,$$

此时的递推式为

$$p_{n+1}(x)=\frac{2n+1}{n+1}xp_n(x)-\frac{n}{n+1}p_{n-1}(x).$$

由此可以得到

$$\begin{aligned}p_0(x)&=1,\\ p_1(x)&=x,\\ p_2(x)&=\frac{1}{2}(3x^2-1),\\ p_3(x)&=\frac{1}{2}(5x^3-3x),\\ p_4(x)&=\frac{1}{8}(35x^4-30x^2+3).\end{aligned}$$

反正在学习的过程中, 哪种方式对自己来说更好理解, 就采用哪个.

例 6.3.4 (Chebyshev多项式) *由前一定理, 设$p_0(x)=1, p_1(x)=x, W(x)=\frac{1}{\sqrt{1-x^2}}, [a,b]=[-1,1]$, 那么可以得到Chebyshev多项式.*

证明: 此时作变量代换$x=\cos\theta$以后, 内积可以如下变化:

$$(f,g)=\int_{-1}^{1}f(x)g(x)\frac{dx}{\sqrt{1-x^2}}=\int_0^{\pi}f(\cos\theta)g(\cos\theta)d\theta.$$

代入$T_n(x)=\cos(n\arccos x)$, 立即得到

$$(T_n,T_m)=\int_0^{\pi}\cos n\theta\cos m\theta\mathrm{d}\theta=\delta_{mn}.$$

当$m\neq n$时, $(T_n,T_m)=0$. □

6.3.1 数值积分的区间变换

可以把某一个区间的数值积分公式推导出任一其他区间上的其他公式. 如果对于某些次数的多项式, 第一个公式精确成立, 那么第二个公式也会精确成立.

给定数值积分公式

$$\int_c^d f(t)dt\approx\sum_{i=0}^{n}A_if(t_i).$$

现在不关心这个公式来自哪里, 但是要假设对于次数$\leq m$的多项式精确成立.

如果需要另外某个区间如$[a,b]$的公式, 那最直接的想法就是把$[c,d]$作变量代换变到$[a,b]$去. 映射$\lambda:[c,d]\to[a,b]$可以定义为

$$\lambda(t)=\frac{b-a}{d-c}t+\frac{ad-bc}{d-c}.$$

此时积分变为

$$\int_a^b f(x)dx=\frac{b-a}{d-c}\int_c^d f(\lambda(t))dt\approx\frac{b-a}{d-c}\sum_{i=0}^{n}A_if(\lambda(t_i)).$$

若f是多项式, 则$f(\lambda(t))$也是多项式, 且次数相同. 这样对于m次的多项式, 新的求积公式也精确成立.

注: 映射$\lambda(t)$的记忆比较困难, 其实可以如下操作: 先把$[c,d]$变为中心对称区间$\left[-\frac{d-c}{2},\frac{d-c}{2}\right]$, 再乘$\frac{b-a}{d-c}$倍, 缩放为$\left[-\frac{b-a}{2},\frac{b-a}{2}\right]$, 最后再平移到$[a,b]$. 当$[c,d]=[-1,1]$或$[0,1]$时这样的操作就非常方便.

作业:

1. Legendre多项式的系数满足
$$a_n=0, b_n=\frac{(n-1)^2}{(2n-1)(2n-3)}.$$
2. 如果内积为$(f,g)=\int_{-a}^{a}f(x)g(x)W(x)dx$, 其中$W$是偶函数, 那么$a_n\equiv 0(\forall n)$, 且如果$n$是偶数, 则$p_n$是偶函数; 如果$n$是奇数, 则$p_n$是奇函数.
3. 用Lagrange插值多项式导出积分公式
$$\int_0^1 f(x)dx\approx Af\left(\frac{1}{3}\right)+Bf\left(\frac{2}{3}\right),$$
并把这条公式变为区间$[a,b]$的积分公式.
4. 下一节要讲的一条Gauss求积公式为
$$\int_{-1}^{1}f(x)dx\approx f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right),$$
写出$\int_a^b f(x)dx$的积分公式.
5. 用中点矩形公式
$$\int_{-1}^{1}f(x)dx\approx 2f(0)$$
推导出$\int_a^b f(x)dx$的复合求积公式, 可以选取等距结点$x_0,x_1,\cdots,x_n$.
6. **【2012Individual, 1】**数值积分公式
$$\int_{-1}^{1}f(x)\mathrm{d}x\approx af(-1)+bf(c).$$
若a,b,c可以任意选取, 求k的最大值, 使得这个积分公式可以对不超过k的多项式精确成立? 并找出常数a,b,c的表达式使得求积公式对不超过这个k的多项式精确成立.

§6.4 Gauss积分

前面小节讨论了Newton-Cotes型求积公式对某些次数较低的多项式精确成立, 而这些公式中, **结点**$x_0,x_1,\cdots,x_n$**事先给定**. 给定了结点以后, 对于能使不超过n次的多项式精确成立的那个多项式, Newton-Cotes型积分的系数A_i就可以唯一确定.

如何选取结点“更好”? 前一节指出带权的Newton-Cotes求积公式是
$$\int_a^b f(x)W(x)dx\approx\sum_{i=0}^{n}A_if(x_i),\qquad(*)$$
其中
$$A_i=\int_a^b W(x)l_i(x)dx,\qquad(**)$$
且$W(x)$是给定的正的**权函数**.

如果我们不预先限制结点, 把A_i, x_i都看作待定系数的变量, 那么这里有$n+1$个系数A_i和$n+1$个结点x_i, 一共有$2n+2$个待定变量, 那我们就猜测可以找到形如$(*)$且次数不超过$2n+1$的多项式是精确成立的求积公式. 下面证明结果确实如此.

Gauss利用结点的可变性使得求积公式对所有$2n+1$次多项式精确成立, 下面定理也给出了结点所处的位置. (用$P^n[a,b]$表示$[a,b]$上不超过n次的多项式空间)

定理 6.4.1

设W是正的权函数, q是$n+1$次非零多项式, 且$q \perp P_W^n[a,b]$, 即

$$\int_a^b q(x)p(x)W(x)dx = 0, \forall p \in P^n[a,b].$$

若$x_0, x_1, \cdots, x_n$是$q(x)$的**零点**, 则具有$(**)$式给定系数的求积公式$(*)$对所有$f \in P^{2n+1}[a,b]$精确成立.

证明: 设$f \in P^{2n+1}[a,b]$, 用q除f, 得到

$$f = qp + r, p, r \in P^n[a,b].$$

由于$q(x_i) = 0$, 则$f(x_i) = r(x_i)$. 又由$(*)$对$P^n[a,b]$精确成立, 因此

$$\begin{aligned}\int_a^b f(x)W(x)dx &= \int_a^b q(x)p(x)W(x) + r(x)W(x)dx \\ &= \int_a^b r(x)W(x)dx \\ &= \sum_{i=0}^n A_i r(x_i) = \sum_{i=0}^n A_i f(x_i).\end{aligned}$$

□

注: q的根都是单根, 可以由下面定理得到:

定理 6.4.2

设$W \in C[a,b]$是正的权函数, $f \in C[a,b]$非零元, 且满足$f \perp P_W^n[a,b]$. 则f在(a,b)上至少变号$n+1$次.

证明: 由于$1 \in P^n[a,b]$, 则$\int_a^b f(x)W(x)dx = 0$, 从而f至少变号1次.

假设f只变号r次, $r \le n$, 选t使得

$$a = t_0 < t_1 < t_2 < \cdots < t_r < t_{r+1} = b,$$

满足在每个区间(t_i, t_{i+1})上不变号$(i = 0, 1, \cdots, r)$, 则$f(x)$与$p(x) = \prod_{i=1}^r (x - t_i)$的符号相同, 从而

$$\int_a^b f(x)p(x)W(x)dx \neq 0,$$

由于$p \in P^n[a,b]$, 这与$f \perp P_W^n[a,b]$矛盾. □

如何寻找$q(x)$使得$q \perp P_W^n[a,b]$成为了关心的问题. 根据前一节的理论, 我们可以找q是$n+1$次正交多项式, 这样前面的Chebyshev多项式、Legendre多项式就派上了用场.

例 6.4.1 *设$W(x) = 1$, $n = 1$(即q是2次正交多项式), 积分区间是$[-1,1]$, 那么此时$q(x)$可以取Legendre多项式:*

$$q(x) = x^2 - \frac{1}{3}.$$

$q(x)$的零点是$\pm\frac{1}{\sqrt{3}}$, 这样求积公式$(*)$就变为

$$\int_{-1}^{1} f(x)dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

这里系数1可以用待定系数法求出来. (作业)

例 6.4.2 设$W(x) = 1$, $n = 2$, 积分区间是$[-1, 1]$, 那么此时$q(x)$可以取3次*Legendre*多项式:

$$q(x) = x^3 - \frac{3}{5}x.$$

$q(x)$的三个零点是$\pm\sqrt{\frac{3}{5}}, 0$, 这样求积公式$(*)$就变为

$$\int_{-1}^{1} f(x)dx \approx \frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f(0) + \frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right)$$

这里$\frac{5}{9}, \frac{8}{9}$可以用待定系数法求出来. (作业)

例 6.4.3 求出下列形式的公式

$$\int_{-1}^{1} f(x)dx \approx c\sum_{i=0}^{2} f(x_i).$$

使得这个公式对所有二次多项式精确成立.

解: 代入$f(x) = 1, x, x^2$可得

$$\begin{aligned} 2 &= c(1+1+1) \\ 0 &= c(x_0 + x_1 + x_2) \\ \frac{2}{3} &= c(x_0^1 + x_1^2 + x_2^2). \end{aligned}$$

解得$c = \frac{2}{3}$. 如果令$x_1 = 0$, 那么$x_{0,2} = \pm\frac{\sqrt{2}}{2}$. 这样就得到了

$$\int_{-1}^{1} f(x)dx \approx \frac{2}{3}\left(f(0) + f\left(\frac{\sqrt{2}}{2}\right) + f\left(-\frac{\sqrt{2}}{2}\right)\right).$$

6.4.1 收敛性分析

引理 6.4.3

在Gauss求积公式中, 它的系数A_i都是正的, 且它们的和是$\int_a^b W(x)dx$.

证明: 固定n, 令$\deg q(x) = n+1$且与$P_W^n[a,b]$正交, q的零点为$x_0, \cdots, x_n$, 作为$(*)$的结点. 对某固定j, 设$p(x) = \frac{q(x)}{x - x_j}$. 由于$p^2$次数最多是$2n$, 则Gauss公式对它是精确成立的, 从而

$$0 < \int_a^b p^2(x)W(x)dx = \sum_{i=0}^{n} A_i p^2(x_i) = A_j p^2(x_j).$$

因此$A_j > 0$, 即系数是正的. 又由于Gauss公式对$f(x) \equiv 1$精确成立, 则$\int_a^b W(x)dx = \sum_{i=0}^{n} A_i$. □

定理 6.4.4

设$f \in C^{2n}[a,b]$, 则

$$\int_a^b f(x)W(x)dx = \sum_{i=0}^{n-1} A_i f(x_i) + \frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b q^2(x)W(x)dx.$$

其中$a < \xi < b, q(x) = \prod_{i=0}^{n-1}(x - x_i)$.

证明: 由Hermite插值, 存在次数最多是$2n-1$的多项式p使得

$$p(x_i) = f(x_i), p'(x_i) = f'(x_i), 0 \le i \le n-1.$$

这个插值公式的误差公式是

$$f(x) - p(x) = \frac{1}{(2n)!}f^{(2n)}(\zeta_x)q^2(x).$$

两边乘$W(x)$再积分可得

$$\int_a^b f(x)W(x)dx - \int_a^b p(x)W(x)dx = \frac{1}{(2n)!}\int_a^b f^{(2n)}(\zeta_x)q^2(x)W(x)dx.$$

由于p的次数最多是$2n-1$, 则Gauss求积公式精确成立, 则

$$\int_a^b p(x)W(x)dx = \sum_{i=0}^{n-1} A_i p(x_i) = \sum_{i=0}^{n-1} f(x_i).$$

又由积分中值定理可得

$$\int_a^b f^{(2n)}(\zeta_x)q^2(x)W(x)dx = f^{(2n)}(\xi)\int_a^b q^2(x)W(x)dx.$$

整理一下即可得最终结论. □

作业:

1. 补全例6.4.1和例6.4.2.
2. 把例6.4.1和例6.4.2的权函数改为$W(x) = \frac{1}{\sqrt{1-x^2}}$, 写出求积公式.
3. 用前一题的求积公式计算积分:
   (1) $\int_{-1}^{1}(1-x^2)^{1/2}dx$.
   (2) $\int_1^3 x\sqrt{4x-x^2-3}dx$.
   (3) $\int_0^{1/3}\frac{6x}{\sqrt{x(1-3x)}}dx$.
4. **【2019Team, 1】**证明求积公式$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x = \frac{\pi}{n}\sum_{k=0}^{n-1} f\left(\cos\pi\frac{2k+1}{2n}\right)$对所有不超过$2n-1$次的多项式精确成立.

§6.5 自适应积分

自适应积分指自动用被积函数的性质计算定积分. 为了计算积分$\int_a^b f(x)dx$, 用户只需提供被积函数f, 区间$[a,b]$与精度ε, 然后把区间划分为各种不同长度的小段, 使得数值积分产生满足精度要去i的结果.

以Simpson公式为例,

$$\int_a^b f(x)dx = S(a,b) - \frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi),$$

其中$\xi \in (a,b)$且

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

思想: 如果在给定小区间上的Simpson公式不符合精度要求, 那么该区间被等分为两部分, 在每个**长度减半**的区间用Simpson公式, 重复过程得到积分的近似. 这样在所有子区间上都有相同的精度.

这部分的思想比较简单, 误差分析和算法设计会比较复杂, 不过算法设计过程可以用到递归, 可以减小代码量.

§6.6 Bernoulli多项式与Euler-Maclaurin公式

在介绍Romberg积分之前, 首先介绍Euler-Maclaurin公式, 因为Romberg积分的正确性取决于Euler-Maclaurin公式. 首先给出Bernoulli多项式的性质, 然后推导Euler-Maclaurin公式.

Bernoulli多项式的定义如下:

$$\sum_{k=0}^{n} C_{n+1}^k B_k(t) = (n+1)t^n.$$

其中$C_n^m = \dfrac{n!}{m!(n-m)!}$是组合数.

当$n=0$时, 立即得$B_0(t)=1$; 当$n=1$时, $B_0(t)+2B_1(t)=2t$, 则$B_1(t)=t-\dfrac{1}{2}$. 以此可以用$B_0, B_1, \cdots, B_{n-1}$求出$B_n$. 前四个Bernoulli多项式是

$$\begin{aligned}
B_0(t) &= 1,\\
B_1(t) &= t - \frac{1}{2},\\
B_2(t) &= t^2 - t + \frac{1}{6},\\
B_3(t) &= t^3 - \frac{3}{2}t^2 + \frac{1}{2}t.
\end{aligned}$$

定理 6.6.1

Bernoulli多项式有如下性质:

(1)$B_n'(t) = nB_{n-1}(t), n \geq 1.$

(2)$B_n(t+1) - B_n(t) = nt^{n-1}, n \geq 2.$

(3)$B_n(t) = \sum_{k=0}^{n} C_n^k B_k(0) t^{n-k}.$

(4)$B_n(1-t) = (-1)^n B_n(t).$

引理 6.6.2

函数$G(t) = B_{2n}(t) - B_{2n}(0)$在开区间$(0,1)$中没有零点.

证明: 由前一定理的(2)(4)得

$$B_n(0) = B_n(1) = (-1)^n B_n(0),$$

因此

$$B_3(0) = B_5(0) = B_7(0) = \cdots = 0.$$

(反证)设$G(t)$在区间$(0,1)$有一个零点, 由于$G(0) = G(1) = 0$, 由Rolle定理, G'在区间$(0,1)$有两个零点. 由于

$$G'(t) = B_{2n}'(t) = 2nB_{2n-1}(t),$$

则B_{2n-1}在区间$(0,1)$中有两个零点. 由于$B_{2n-1}(0)=B_{2n-1}(1)=0$, 则$B'_{2n-1}$在区间$(0,1)$有三个零点, 故$B_{2n-2}$在区间$(0,1)$中有三个零点. 这样, 对于奇数指标$k<2n$, B_k在$(0,1)$中至少有两个零点, 因此B_3除了两个零点0,1以外, 在区间$(0,1)$中还有两个零点, 但B_3是三次多项式, 这是不可能的. □

定理 6.6.3: Euler-Maclaurin公式

若$f\in C^{2n}[0,1]$, 则

$$\int_0^1 f(t)dt=\frac{1}{2}[f(0)+f(1)]-\sum_{k=0}^{n-1}\frac{b_{2k}}{(2k)!}[f^{(2k-1)}(1)-f^{(2k-1)}(0)]+R,$$

其中$b_k=B_k(0), R_k=-\frac{b_{2n}}{(2n)!}f^{(2n)}(\xi), 0<\xi<1.$

证明: 由前一定理(1)可知

$$B_n(t)=\frac{1}{n+1}B'_{n+1}(t),$$

由该公式与分部积分,

$$\int_0^1 f(t)dt=\int_0^1 f(t)B_0(t)dt=B_1(t)f(t)|_0^1-\int_0^1 B_1(t)f'(t)dt.$$

由于$B_1(1)=\frac{1}{2}, B_1(0)=-\frac{1}{2}$, 则

$$\int_0^1 f(t)dt=\frac{1}{2}[f(0)+f(1)]-\int_0^1 B_1(t)f'(t)dt.$$

进一步用分部积分, 可得

$$\int_0^1 f(t)dt=\frac{1}{2}[f(0)+f(1)]-\frac{b_2}{2}[f'(1)-f'(0)]+\frac{1}{2}\int_0^1 B_2(t)f''(t)dt.$$

继续这个过程, 由于

$$B_n(0)=B_n(1)=b_n, b_3=b_5=b_7=\cdots=b_0,$$

重复$2n$步分部积分可得

$$\int_0^1 f(t)dt=\frac{1}{2}[f(0)+f(1)]-\sum_{k=1}^{n}\frac{b_{2k}}{(2k)!}[f^{(2k-1)}(1)-f^{(2k-1)}(0)]+\frac{1}{(2n)!}\int_0^1 B_{2n}(t)f^{(2n)}(t)dt.$$

和式最后一项可以用newton-Leibniz公式表示为

$$\frac{b_{2n}}{(2n)!}[f^{(2n-1)}(1)-f^{(2n-1)}(0)]=\frac{b_{2n}}{(2n)!}\int_0^1 f^{(2n)}(t)dt.$$

现在, 除了余项是

$$R=\frac{1}{(2n)!}\int_0^1[B_{2n}(t)-b_{2n}]f^{(2n)}(t)dt$$

以外, 可以把公式写成定理中叙述的形式, 由前一引理, $B_{2n}(t)-b_{2n}$在$[0,1]$中不改变符号, 用积分中值定理可得

$$R=\frac{1}{(2n)!}f^{(2n)}(\xi)\int_0^1[B_{2n}(t)-b_{2n}]dt.$$

又由$B_{2n}=\dfrac{B'_{2n+1}}{2n+1}$, 且$B_{2n+1}(0)=B_{2n+1}(1)=0$, 最后得到

$$R=-\frac{b_{2n}}{(2n)!}f^{(2n)}(\xi).$$

□

作业:

1. 证明第一个定理中的四条Bernoulli多项式性质.
2. 用第一个定理的性质(2), 证明

$$\sum_{k=1}^{n}k^p=\frac{B_{p+1}(n+1)-B_{p+1}(0)}{p+1}.$$

3. 证明$B_n(t)$的首项是t^n, 而t^{n-1}项的系数是$-\dfrac{n}{2}$.
4. 证明: 当n是奇数时, $B_n(x)-B_n(0)$在$\dfrac{1}{2}$处有单个零点.
5. 证明: $B_0(0),B_2(0),B_4(0),\cdots$的符号交替变化.

§ 6.7 Romberg积分

Romberg给出了求数值积分$\displaystyle\int_a^b f(x)dx$的递推形式.

由复合梯形公式

$$T(n)=\frac{h}{2}\left[f(a)+2\sum_{i=1}^{n-1}f(a+ih)+f(b)\right]$$

其中$h=\dfrac{b-a}{n}$.

当区间是[0, 1]时,$T(1),T(2),T(4),T(8)$的公式是

$$\begin{aligned}
T(1)&=\frac{1}{2}f(0)+\frac{1}{2}f(1)\\
T(2)&=\frac{1}{4}f(0)+\frac{1}{2}\left[f\left(\frac{1}{2}\right)\right]+\frac{1}{4}f(1)\\
T(4)&=\frac{1}{8}f(0)+\frac{1}{4}\left[f\left(\frac{1}{4}\right)+f\left(\frac{1}{2}\right)+f\left(\frac{3}{4}\right)\right]+\frac{1}{8}f(1)\\
T(8)&=\frac{1}{17}f(0)+\frac{1}{8}\left[f\left(\frac{1}{8}\right)+f\left(\frac{1}{4}\right)+f\left(\frac{3}{8}\right)+f\left(\frac{1}{2}\right)+f\left(\frac{5}{8}\right)+f\left(\frac{3}{4}\right)+f\left(\frac{7}{8}\right)\right]+\frac{1}{16}f(1)
\end{aligned}$$

如果要计算$T(2n)$, 可以利用$T(n)$计算已有结果, 只需计算未计算的结果. 根据上式可得

$$\begin{aligned}
T(2)&=\frac{1}{2}T(1)+\frac{1}{2}\left[f\left(\frac{1}{2}\right)\right]\\
T(4)&=\frac{1}{2}T(2)+\frac{1}{4}\left[f\left(\frac{1}{4}\right)+f\left(\frac{3}{4}\right)\right]\\
T(8)&=\frac{1}{2}T(4)+\frac{1}{8}\left[f\left(\frac{1}{8}\right)+f\left(\frac{3}{8}\right)+f\left(\frac{5}{8}\right)+f\left(\frac{7}{8}\right)\right]
\end{aligned}$$

而对于一般的$h=\dfrac{b-a}{2n}$与任意区间$[a,b]$, 有如下公式:

$$\boxed{T(2n)=\frac{1}{2}T(n)+h\sum_{i=1}^{n}f(a+(2i-1)h).}$$

证明略.

设$T_{n,0}$是具有2^n个子区间的梯形估计, 则

$$T_{0,0}=\frac{1}{2}(b-a)[f(a)+f(b)],$$
$$T_{n,0}=\frac{1}{2}T_{n-1,0}+h_n\sum_{i=1}^{2^{n-1}}f(a+(2i-1)h_n).$$

其中$h_0=b-a, h_n=\dfrac{h_{n-1}}{2}$. 这样不断计算$T_{0,0},T_{1,0},\cdots,T_{M,0}$即可得到积分的近似值.

此外还可以作加速处理. 从Euler-Maclaurin公式入手,

$$\int_0^1 f(t)\mathrm{d}t=\frac{1}{2}[f(0)+f(1)]+\sum_{k=1}^{m-1}A_{2k}[f^{(2k-1)}(0)-f^{(2k-1)}(1)]-A_{2m}f^{(2m)}(\xi_0),$$

其中ξ_0介于0与1, 常数$k!A_k$是Bernoulli数. Bernoulli数可以如下定义:

$$\frac{x}{e^x-1}=\sum_{k=0}^{\infty}A_kx^k.$$

定义函数$g(t)=f(x_i+ht)$, 其中$h=x_{i+1}-x_i$, 把Euler-Maclaurin公式用于g, 并作变量代换$t=(x-x_i)/h$, 得到

$$\int_{x_i}^{x_{i+1}}f(x)\mathrm{d}x=\frac{h}{2}[f(x_i)+f(x_{i+1})]+\sum_{k=1}^{m-1}A_{2k}h^{2k}[f^{(2k-1)}(x_i)-f^{(2k-1)}(x_{i+1})]-A_{2m}h^{2m+1}f^{(2m)}(\xi_i).$$

上式两端对$i=0,1,\cdots,2^n-1$求和,$x_i=a+ih, h=\dfrac{b-a}{2^n}$, 可得

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x&=\frac{h}{2}\sum_{i=0}^{2^n-1}[f(x_i)+f(x_{i+1})]+\sum_{k=1}^{m-1}A_{2k}h^{2k}[f^{(2k-1)}(a)-f^{(2k-1)}(b)]-A_{2m}(b-a)h^{2m}f^{(2m)}(\xi)\\&=T_{n,0}+c_2h^2+c_4h^4+c_6h^6+\cdots+c_{2m-2}h^{2m-2}+c_{2m}h^{2m}f^{(2m)}(\xi).\end{aligned}$$

其中$\xi\in(a,b)$, $c_2,\cdots,c_{2m}$与h无关. 这就是前面提到的**Richardson外推**的特殊情形. 最终可以得到如下的递推式:

$$T_{n,m}=T_{n,m-1}+\frac{1}{4^m-1}[T_{n,m-1}-T_{n-1,m-1}],$$

可以构造积分近似值的**Romberg积分阵**:

$$\begin{array}{cccccccc}T_{0,0}&&&&&&&\\T_{1,0}&T_{1,1}&&&&&&\\T_{2,0}&T_{2,1}&T_{2,2}&&&&&\\T_{3,0}&T_{3,1}&T_{3,2}&T_{3,3}&&&&\\T_{4,0}&T_{4,1}&T_{4,2}&T_{4,3}&T_{4,4}&&&\\\vdots&\vdots&\vdots&\vdots&\vdots&\ddots&&\\T_{M,0}&T_{M,1}&T_{M,2}&T_{M,3}&T_{M,4}&\cdots&T_{M,M}\end{array}$$

定理 6.7.1

若$f \in C[a,b]$, 则Romberg积分阵每一列都收敛于f的积分, 即

$$\lim_{n\to\infty} T_{n,m} = \int_a^b f(x)\mathrm{d}x, \forall m.$$

CHAPTER 7

函数逼近理论

所谓的函数逼近问题, 就是用一个构造简单的函数$\varphi(x)$来代替一个复杂的函数$f(x)$的问题. 通常, $\varphi(x)$可以写成一组线性无关函数系的线性组合:

$$\varphi(x)=\sum_{j=0}^{n}c_j\varphi_j(x),$$

其中$c_0,\cdots,c_n$都是实数. 例如, 我们可以取φ_j为幂函数:

$$1,x,\cdots,x^n,$$

或者三角函数:

$$1,\cos x,\sin x,\cdots,\cos nx,\sin nx.$$

在讨论函数逼近的时候, 希望$\varphi(x)$和$f(x)$误差尽可能小, 那误差度量标准就需要指出来了, 常用的度量有:

(1)一致度量: $d_\infty(f,\varphi)=\max\limits_{a\le x\le b}|f(x)-\varphi(x)|$;

(2)带权函数W的L^p度量: $d_p(f,\varphi)=\displaystyle\int_a^b|f(x)-\varphi(x)|^pW(x)\mathrm{d}x$, 其中$p\ge1,W(x)\ge0$.

对于给定的函数系$\{\varphi_j(x)\}$, 若函数$\varphi^{(n)}(x)=\displaystyle\sum_{j=0}^{n}c_j\varphi_j(x)$满足

$$\lim_{n\to\infty}d_\infty(f,\varphi^{(n)})=0,$$

我们称这样的函数逼近是一致逼近. 若函数$\varphi^{(n)}(x)$满足

$$\lim_{n\to\infty}d_p(f,\varphi^{(n)})=0,$$

我们称这样的函数逼近是L^p-逼近.

§7.1 最佳一致逼近

基本问题: 设$f(x)$是定义在$[a,b]$上的连续函数, $H_n=\mathrm{span}\{1,x,\cdots,x^n\}$, 求$n$次多项式$p_n(x)$使得

$$E_n=\min_{p_n(x)\in H_n}\max_{a\le x\le b}|f(x)-p_n(x)|.$$

尽可能小. 使得E_n最小的那个$p_n(x)$叫做$f(x)$的**最佳一致逼近多项式**. 根据Weierstrass定理, E_n单调下降趋于0.

设$\|g(x)\|_{C[a,b]} = \max\limits_{x\in[a,b]} |g(x)|$.

定义 7.1.1: 交错点组

若$x_1, x_2, \cdots, x_n \in [a,b]$满足$|g(x_i)| = \|g(x)\|_{C[a,b]}$, 且$g(x_i) = (-1)^i\sigma\|g(x)\|_{C[a,b]}$, 其中$\sigma = \pm 1$代表起点的符号$\mathrm{sgn}(-g(x_1))$. 即$g(x_i)$的符号交替变化, 则把$x_1, x_2, \cdots, \cdots, x_n$称作$g(x)$在$[a,b]$上的一个交错点组.

换言之, 这n个点x_i需要同时满足:

- $a \le x_1 < x_2 < \cdots < x_n \le b$;
- $|g(x_i)| = \max\limits_{x\in[a,b]} |g(x)|, \qquad i = 1, 2, \cdots, n$;
- $g(x_i) = -g(x_{i+1}), \qquad i = 1, 2, \cdots, n-1$.

那么我们就称$x_1, x_2, \cdots, x_n$是$g(x)$在$[a,b]$上的一个交错点组.

定理 7.1.1: Chebyshev

$p_n(x)$是$f(x)$的最佳一致逼近多项式的充分必要条件是: $e(x) = f(x) - p_n(x)$在$[a,b]$上存在$n+2$个点形成的交错点组.

推论 7.1.2

相同次数的最佳一致逼近多项式是唯一的.

证明: 设$p(x), q(x) \in H_n$是最佳一致逼近, 则

$$\|f(x) - p(x)\|_{C[a,b]} = \|f(x) - q(x)\|_{C[a,b]} = E_n.$$

构造$h(x) = \dfrac{1}{2}(p(x) + q(x)) \in H_n$, 则

$$E_n \le \|f(x) - h(x)\|_{C[a,b]} \le \frac{1}{2}\|f(x) - p(x)\|_{C[a,b]} + \frac{1}{2}\|f(x) - p(x)\|_{C[a,b]} = E_n.$$

因此$h(x)$也是最佳一致逼近多项式, 根据Chebyshev定理, 有$f(x) - h(x)$有$n+2$个交错点$x_1, \cdots, x_{n+2}$.

$$f(x_i) - h(x_i) = \frac{1}{2}(f(x_i) - p(x_i)) + \frac{1}{2}(f(x_i) - q(x_i)), \forall i.$$

则

$$(-1)^i\sigma E_n = f(x_i) - h(x_i) = f(x_i) - p(x_i) = f(x_i) - q(x_i) \forall i.$$

从而$p(x_i) = q(x_i), i = 1, 2, \cdots, n+2$. 而$\deg(p) \le n, \deg(q) \le n$, 则必有$p(x) = q(x)$. □

注: n次多项式在H_n中的最佳一致逼近是本身.

7.1.1 特殊的最佳一致逼近多项式计算

1. 高阶导数符号不变的情形

定理 7.1.3

若$f^{(n+1)}(x)$在(a,b)内不变号(恒正或恒负), 则最佳一致逼近多项式的交错点组必含端点a, b.

证明: 多次使用Rolle定理. 设a或b不属于$f(x) - p(x)$的交错点组. 根据Chebyshev定理, 能找到$n+2$个点的函数值正负交替变换, 于是$r(x) = f(x) - p(x)$在区间(a,b)内至少有$n+1$个点

$$a < x_1 < x_2 < \cdots < x_{n+1} < b$$

使得

$$r'(x_i)=0,\qquad i=1,2,\cdots,n+1.$$

不断使用Rolle定理可知$r^{(n+1)}(x)$在(a,b)中至少有一个零点η, 即

$$r^{(n+1)}(\eta)=0.$$

但是,

$$r^{(n+1)}(x)=f^{(n+1)}(x)-p^{(n+1)}(x)=f^{(n+1)}(x),$$

所以$f^{(n+1)}(\eta)=0$. 这与$f^{(n+1)}(x)$在$[a,b]$中恒正或恒负的假设矛盾, 定理得证. □

例 7.1.1 *设$f(x)\in C^2[a,b]$为凸函数, 求它的最佳一致逼近**线性**多项式.*

解: 根据前一定理, 交错点组是a,ξ,b, 多项式形如$p(x)=kx+h$. 记$\tilde{E}_1=\sigma E_1$, 其中$\sigma=\operatorname{sgn}(-g(x_1))$.

$$\begin{aligned}f(a)-(ka+h)&=(-1)^1\tilde{E}_1\\ f(\xi)-(k\xi+h)&=(-1)^2\tilde{E}_1\\ f(b)-(kb+h)&=(-1)^3\tilde{E}_1\end{aligned}$$

注意ξ是驻点, 从而还有

$$f'(\xi)-k=0.$$

联立四个方程可以解$\xi,k,h,\tilde{E}_1$. □

2. 用低阶多项式来逼近高阶多项式

复习: 考虑区间$[-1,1]$上的函数的最佳逼近. 对于n次首一多项式$p(x)\in H_n$, 满足$\min\limits_{p(x)\in H_n}\|p(x)\|$的$p(x)$是$n$次首一Chebyshev多项式$\dfrac{T_n(x)}{2^{n-1}}$. (注意(高于1次的)Chebyshev多项式不是首一的!)

根据这个理论, 如果我们要找$q(x)\in H_{n-1}[-1,1]$来逼近$f(x)\in H_n[-1,1]$, 使得$\|f(x)-q(x)\|_{C[-1,1]}$最小的那个$q(x)$满足

$$\frac{f(x)-q(x)}{a_n}=\frac{T_n(x)}{2^{n-1}},$$

即

$$\boxed{q(x)=f(x)-\frac{a_n}{2^{n-1}}T_n(x).}$$

其中a_n代表$f(x)$的最高次项系数.

注: 若$f(x)$定义在$[a,b]$上, 先把$f(x)$作个平移伸缩变成$[-1,1]$.

例 7.1.2 *$f(x)=2x^4+3x^3-x^2+1$在区间$[-1,1]$上的三次最佳一致逼近多项式为*

$$p(x)=f(x)-\frac{2}{2^3}T_4(x)=3x^3+x^2+\frac{3}{4}.$$

注: Chebyshev多项式递推关系式是$T_{n+1}(x)=2xT_n(x)-T_{n-1}(x)$. 可以表示成$T_n(x)=\cos(n\arccos x)$.

$$\begin{aligned}T_0(x)&=1\\ T_1(x)&=x\\ T_2(x)&=2x^2-1\\ T_3(x)&=4x^3-3x\\ T_4(x)&=8x^4-8x^2+1\end{aligned}$$

3. 近似最佳一致逼近多项式

设$T_{n+1}(x)=0$的根是$x_1,\cdots,x_{n+1}$, 以这$n+1$个点作为插值基点, 作$f(x)\in C^{\infty}[-1,1]$的n次Lagrange插值多项式

$$p(x)=\sum_{j=1}^{n+1}f(x_j)\prod_{j\neq i}\frac{x-x_i}{x_j-x_i}.$$

此时$f(x)-p(x)=\dfrac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)$的余项达到极小. 这样的Lagrange插值多项式叫**Chebyshev插值多项式**.

注: 若$f(x)$定义在$[a,b]$上, 先把$f(x)$作个平移伸缩变成$[-1,1]$.

§7.2 最佳平方逼近

基本问题: 设$H_n=\mathrm{span}\{\phi_0(x),\cdots,\phi_n(x)\}$, 求$p_n(x)=\sum_{j=0}^{n}a_j\phi_j(x)\in H_n$, 使得

$$\|f(x)-p_n(x)\|_{L^2_W[a,b]}:=\int_a^b(f(x)-p_n(x))^2W(x)\mathrm{d}x:=F(a_0,\cdots,a_n)$$

最小, 其中$W(x)\geq 0$为权函数, 至多有限个点为0.

我们记

$$(f,g)_{L^2_W[a,b]}=\int_a^b f(x)g(x)W(x)\mathrm{d}x,$$

或者简记为(f,g).

定理 7.2.1

$p_n(x)$是$f(x)$的最佳平方逼近多项式等价于$\dfrac{\partial F}{\partial a_i}=0,i=0,\cdots,n$, 而这可以写成一个线性方程组.

证明: $p_n(x)$是$f(x)$的最佳平方逼近多项式等价于$F(a_0,a_1,\cdots,a_n)$取到最小值. 为了使得F达到最小, 根据极值存在的必要条件, 需要

$$\frac{\partial F}{\partial a_k}=\frac{\partial}{\partial a_k}\left[\int_a^b(f(x)-p_n(x))^2W(x)\mathrm{d}x\right]=0,k=0,1,\cdots,n,$$

求导可得

$$\int_a^b[f(x)-p_n(x)]\varphi_k(x)W(x)\mathrm{d}x=0,\qquad k=0,1,\cdots,n.$$

或者写

$$\int_a^b p_n(x)\varphi_k(x)W(x)\mathrm{d}x=\int_a^b f(x)\varphi_k(x)W(x)\mathrm{d}x,\qquad k=0,1,\cdots,n.$$

即

$$\sum_{j=0}^{n}a_j(\varphi_j,\varphi_k)=(p_n,\varphi_k)=(f,\varphi_k).$$

于是我们得到了由$n+1$个线性方程构成的方程组:

$$\begin{cases}(\varphi_0,\varphi_0)a_0+(\varphi_0,\varphi_1)a_1+\cdots+(\varphi_0,\varphi_n)a_n=(\varphi_0,f),\\(\varphi_1,\varphi_0)a_0+(\varphi_1,\varphi_1)a_1+\cdots+(\varphi_1,\varphi_n)a_n=(\varphi_1,f),\\\cdots\cdots\\(\varphi_n,\varphi_0)a_0+(\varphi_n,\varphi_1)a_1+\cdots+(\varphi_n,\varphi_n)a_n=(\varphi_n,f),\end{cases}\tag{7.1}$$

系数矩阵是线性无关系$\{\varphi_j(x)\}_{j=0}^n$的Gram矩阵, 因而非奇异, 故上述方程组有唯一的解$a_0,a_1,\cdots,a_n$.

下面再来证明前面所构造的$p_n(x)=\sum_{j=0}^{n}a_j\phi_j(x)$是最佳平方逼近函数.

对任意的H_n中元素

$$Q(x)=\sum_{j=0}^{n}b_j\varphi_j(x),$$

我们证明$\|Q-f\|_{L^2_W[a,b]}\geq\|\varphi-f\|_{L^2_W[a,b]}$, 从而上述构造的$\varphi$是最佳平方逼近. 事实上, 注意到

$$\begin{aligned}D&:=\|Q-f\|^2-\|\varphi-f\|^2\\&=\|Q\|^2-\|\varphi\|^2-2(Q,f)+2(\varphi,f)\\&=\|Q\|^2-2(Q,\varphi)+\|\varphi\|^2-2(Q-\varphi,f)-2(\varphi,\varphi)+2(Q,\varphi)\\&=\|Q-\varphi\|^2+2(Q-\varphi,\varphi-f)\end{aligned}$$

注意到

$$\begin{aligned}(Q-\varphi,\varphi-f)&=\int_a^b[Q(x)-\varphi(x)][\varphi(x)-f(x)]W(x)\mathrm{d}x\\&=\sum_{k=0}^{n}(b_k-a_k)\int_a^b[\varphi(x)-f(x)]\varphi_k(x)W(x)\mathrm{d}x=0,\end{aligned}$$

所以$D\geq 0$. □

我们把(7.1)叫**法方程组**. 系数矩阵也叫**质量矩阵**.

例 7.2.1 取$H_{n+1}=\mathrm{span}\{1,x,\cdots,x^n\},W(x)=1,[a,b]=[0,1]$, 则

$$(x^i,x^j)=\frac{1}{i+j+1},\qquad i,j=0,1,\cdots,n,$$

此时质量矩阵是*Hilbert*矩阵, 条件数很可怕, 由于计算过程中舍入误差的影响, 得到的近似解的精确度是很差的.

我们自然要问, 怎样作函数系$\{\varphi_j(x)\}_{j=0}^n$才是最合适的? 显然, 如果$\{\varphi_j(x)\}_{j=0}^n$是$[a,b]$上关于权函数$W(x)$的正交函数系:

$$(\varphi_i,\varphi_j)_{L^2_W[a,b]}=\int_a^b\varphi_i(x)\varphi_j(x)W(x)\mathrm{d}x=0,\qquad i\neq j,$$

那么质量矩阵变成对角阵, 可以很轻松解出系数a_j为

$$a_j=\frac{(\varphi_j,f)}{(\varphi_j,\varphi_j)}=\frac{\int_a^b\varphi_j(x)f(x)W(x)\mathrm{d}x}{\int_a^b[\varphi_j(x)]^2W(x)\mathrm{d}x}.$$

此时我们把a_j叫做$f(x)$关于正交函数系$\{\varphi_j(x)\}_{j=0}^n$的**Fourier系数**, 把级数$\sum_{j=0}^{\infty}a_j\varphi_j(x)$ 叫做$f(x)$的**广义Fourier级数**, 它是Fourier级数的直接推广.

例 7.2.2 区间$[-1,1]$、权重$W(x)=1$的正交多项式系是*Legendre*多项式.

注: Legendre多项式P满足

$$\begin{aligned}P_0(x)&=1\\P_1(x)&=x\\P_2(x)&=\frac{1}{2}(3x^2-1)\\P_3(x)&=\frac{1}{2}(5x^3-3x)\end{aligned}$$

例 7.2.3 权重$W(x)=1$的正交三角函数系

$$\mathrm{span}\{1,\cos x,\sin x,\cos 2x,\sin 2x,\cdots,\cos nx,\sin nx\}$$

是$[-\pi,\pi]$上的正交函数系. $f(x)$在$[-\pi,\pi]$上的最佳平方逼近函数是

$$\varphi(x)=\frac{1}{2}a_0+\sum_{j=1}^{n}(a_j\cos jx+b_j\sin jx),$$

并且

$$a_j=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\cos jx\mathrm{d}x,\qquad j=0,1,\cdots,n,$$
$$b_j=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\sin jx\mathrm{d}x,\qquad j=1,2,\cdots,n.$$

注: 让$\varphi(x)$的右端$n\to\infty$, 可得关于$f(x)$的**经典Fourier级数**.

例 7.2.4 区间$[-1,1]$、权重$W(x)=\dfrac{1}{\sqrt{1-x^2}}$的正交多项式系是*Chebyshev*多项式:

$$T_n(x)=\cos(n\arccos x),\qquad n=0,1,2\cdots.$$

注意到

$$\int_{-1}^{1}T_i(x)T_j(x)W(x)\mathrm{d}x=\begin{cases}0, & i\neq j,\\ \dfrac{\pi}{2}, & i=j\neq 0,\\ \pi, & i=j=0,\end{cases}$$

所以$f(x)$的最佳平方逼近是

$$p_n(x)=\frac{1}{2}a_0+\sum_{j=1}^{n}a_jT_j(x).$$

其中,

$$a_j=\frac{2}{\pi}\int_{-1}^{1}T_j(x)f(x)\frac{\mathrm{d}x}{\sqrt{1-x^2}},\qquad j=0,1,\cdots,n.$$

注: 令$p_n(x)$的右端让$n\to\infty$, 可得级数

$$\frac{1}{2}a_0+\sum_{j=1}^{\infty}a_jT_j(x),$$

把这个级数叫做$f(x)$的**Chebyshev级数**. 可以证明, $f(x)$的Chebyshev级数是$g(\theta)=f(\cos\theta)$的经典Fourier级数.

§ 7.3 离散Fourier变换

假设已知函数$f(x)$在区间$[a,b]$上的点$x_1,\cdots,x_m$处的值分别为$f(x_1),\cdots,f(x_m)$, 而$\{\varphi_j(x)\}_{j=0}^{n}$是区间$[a,b]$上的一个线性无关函数系, $n<m$. 令

$$\varphi(x)=\sum_{j=0}^{n}a_j\varphi_j(x),$$

记**残量**为

$$r_i=\sum_{j=0}^{n}a_j\varphi_j(x_i)-f(x_i),\qquad i=1:m.$$

一般来说不可能选择出实数$\{a_j\}_{j=0}^{n}$使得残量都为零, 但是我们可以选实数$\{a_j\}_{j=0}^{n}$使得残量的平方和达到极小:

$$\min\sum_{i=1}^{m}r_i^2=\min\sum_{i=1}^{m}\left(\sum_{j=0}^{n}a_j\varphi_j(x_i)-f(x_i)\right)^2.$$

这种函数逼近问题叫做**离散的最佳平方逼近问题**, 这样求得的$\varphi(x)$称为$f(x)$在点集$\{x_i\}_{i=1}^{m}$上的最佳平方逼近.

假设给定了数据集$S=\{(x_j,y_j)\}_{j=0}^{2m-1}$, 其中$\{x_j\}_{j=0}^{2m-1}$是区间$[-\pi,\pi]$上的等距点:

$$x_j=-\pi+\frac{j}{m}\pi,\qquad j=0,1,\cdots,2m-1.$$

我们取三角函数系$\{\varphi_k(x)\}_{k=0}^{2n-1}$:

$$\begin{aligned}\varphi_0(x)&=\frac{1}{2},\\ \varphi_k(x)&=\cos kx,\qquad k=1,2,\cdots,n,\\ \varphi_{n+k}(x)&=\sin kx,\qquad k=1,2,\cdots,n-1.\end{aligned}$$

考虑数据点集S的形如

$$\begin{aligned}\varphi(x)&=a_0\varphi_0(x)+a_n\varphi_n(x)+\sum_{k=1}^{n-1}(a_k\varphi_k(x)+b_k\varphi_{n+k}(x))\\ &=\frac{a_0}{2}+a_n\cos nx+\sum_{k=1}^{n-1}(a_k\cos kx+b_k\sin kx)\end{aligned}$$

的最小二乘拟合, 记向量

$$\boldsymbol{\varphi}_j=[\varphi_j(x_0),\varphi_j(x_1),\cdots,\varphi_j(x_{2m-1})]^T\in\mathbb{R}^{2m},\qquad j=0,1,\cdots,2n-1.$$

可以证明, $\{\boldsymbol{\varphi}_k\}_{k=0}^{2n-1}$是正交向量系, 即

$$(\boldsymbol{\varphi}_k,\boldsymbol{\varphi}_l)=\sum_{j=0}^{2m-1}\varphi_k(x_j)\varphi_l(x_j)=0,\qquad k\neq l.$$

定理 7.3.1

设$x_j=-\pi+\frac{j}{m}\pi, j=0,1,\cdots,2m-1$. 则数据$S=\{(x_j,y_j)\}_{j=0}^{2m-1}$的最小二乘拟合

$$\varphi(x)=\frac{a_0}{2}+a_n\cos nx+\sum_{k=1}^{n-1}(a_k\cos kx+b_k\sin kx)\tag{7.2}$$

的系数为

$$\begin{aligned}a_k&=\frac{1}{m}\sum_{j=0}^{2m-1}y_j\cos kx_j,\qquad k=0,1,\cdots,n,\\ b_k&=\frac{1}{m}\sum_{j=0}^{2m-1}y_j\sin kx_j,\qquad k=1,2,\cdots,n-1.\end{aligned}\tag{7.3}$$

证明:

注: 如果我们假设$y_j=f(x_j), j=0,1,\cdots,2m-1$, 那么具有系数表达式(7.3)的三角多项式(7.2)称为函数$f(x)$的**离散Fourier展开式**. 它是用三角多项式作为数据S的最小二乘拟合. 由(7.3)式所确定的a_k,b_k叫做**离散Fourier系数**.

命题 7.3.2

如果前一定理中$n=m$, $f(x)$是一个函数, 并且$y_j=f(x_j)$, 把前一定理的(7.2)改写成

$$\varphi(x)=\frac{a_0+a_m\cos mx}{2}+\sum_{k=1}^{m-1}(a_k\cos kx+b_k\sin kx) \tag{7.4}$$

其中系数为

$$\begin{aligned} a_k&=\frac{1}{m}\sum_{j=0}^{2m-1}f(x_j)\cos kx_j, \qquad k=0,1,\cdots,m,\\ b_k&=\frac{1}{m}\sum_{j=0}^{2m-1}f(x_j)\sin kx_j, \qquad k=1,2,\cdots,m-1. \end{aligned} \tag{7.5}$$

那么$\varphi(x)$满足插值原则:

$$\varphi(x_j)=f(x_j), \qquad j=0,1,\cdots,2m-1,$$

此时三角多项式$\varphi(x)$是$f(x)$的经过点集$S=\{(x_j,f(x_j))\}_{j=0}^{2m-1}$的插值函数, 把(7.4)式叫做**三角插值多项式**. 并且此时系数a_j,b_j可以使得残量平方和

$$E(a_0,a_1,\cdots,a_m,b_1,\cdots,b_{m-1})=\sum_{j=0}^{2m-1}(f(x_j)-\varphi(x_j))^2$$

达到极小.

下面来考虑计算

$$S(x)=\frac{1}{m}\sum_{k=0}^{2m-1}c_ke^{ikx},$$

中的复系数

$$c_k=\sum_{j=0}^{2m-1}y_je^{\frac{ik\pi j}{m}}, \qquad k=0,1,\cdots,2m-1.$$

由$\{f(x_j)\}_{j=0}^{2m-1}$计算$\{c_k\}_{k=0}^{2m-1}$的过程称为**离散的Fourier变换(DFT)**. 计算$\{c_k\}$相当于计算前一命题中的$\{a_k\},\{b_k\}$, 这是因为

$$\begin{aligned} \frac{1}{m}c_k(-1)^k&=\frac{1}{m}c_ke^{-i\pi k}=\frac{1}{m}\sum_{j=0}^{2m-1}y_je^{\frac{ik\pi j}{m}}\cdot e^{-i\pi k}=\frac{1}{m}\sum_{j=0}^{2m-1}y_je^{ik(-\pi+\frac{\pi j}{m})}\\ &=\frac{1}{m}\sum_{j=0}^{2m-1}y_j\left(\cos k\left(-\pi+\frac{j}{m}\pi\right)+i\sin k\left(-\pi+\frac{j}{m}\pi\right)\right)\\ &=\frac{1}{m}\sum_{j=0}^{2m-1}y_j(\cos kx_j+i\sin kx_j). \end{aligned}$$

于是

$$a_k+ib_k=\frac{1}{m}(-1)^kc_k.$$

§7.4 (*)快速Fourier变换

本节带(*)表示是课程要求之外的补充内容.

设$c_0,c_1,\cdots,c_{n-1}$如下定义:

$$c_k=\frac{1}{n}\sum_{j=0}^{n-1}f(x_j)e^{\frac{-2i\pi kj}{n}},$$

直接计算c_k需要大约用n次乘法和n次加法. 因为需要算n个系数c_k, 所以计算

$$S(x) = \frac{1}{n}\sum_{k=0}^{n-1} c_k e^{ikx}$$

总共需要$O(n^2)$次运算. 快速Fourier变换(FFT)可以把这个计算成本降低到$O(n\log_2 n)$.

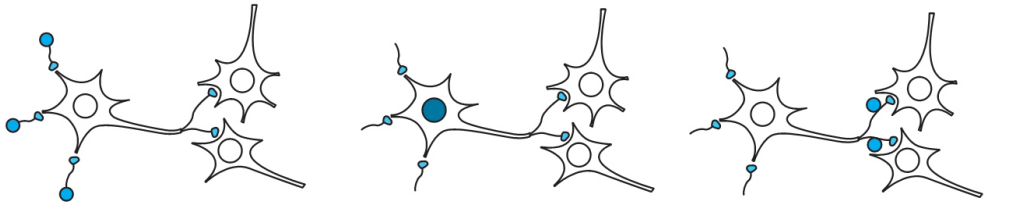
§ 7.5 (*)神经网络的逼近性质

本节带(*)表示是课程要求之外的补充内容.

7.5.1 什么是神经网络

神经网络(neural network)是个相当大的、多学科交叉的学科领域, 不同学科对神经网络的定义有所差别. [①]

神经网络中最基本的成分是**神经元(neuron)**. 在生物神经网络中, 每个神经元与其他神经元相连, 当它“兴奋”时, 就会向相连的神经元发送化学物质, 从而改变这些神经元内的电位; 如果神经元的电位超过了一个“阈值”, 那么它就会被激活, 即“兴奋”起来, 向其他神经元发送化学物质. [②]



信号被输入到神经元中　　细胞体判断输入信号之和　　当信号之和大于阈值时, 点火, 向相邻神经元传递信号.

Figure 7.1: 兴奋在神经元之间的传递

将神经元的工作在数学上抽象化, 并以其为单位人工地形成网络, 这样的人工网络就是**神经网络(neural network)**. 将构成大脑的神经元的集合体抽象为数学模型, 这就是神经网络的出发点.

1943年, W. McCulloch和W. Pitts将上述情形抽象为“McCulloch-Pitts(M-P)神经元模型”, 简单模拟了神经元的反应流程, 包括:

- 多个带有权重的输入$w_i \cdot x_i$, 相当于“突触”. 其中,
  x_i是**输入值(input)**, 表示外界的刺激;
  w_i表示**权重(weight)**, 表示刺激的不同强度.
- 转换函数Σ, 相当于“汇聚电信号的细胞膜”.
  $\sum_{i=1}^{n} w_i x_i$也就是对所有带权重的输入进行简单的求和, 将多个值合并为一个值.
- **阈值(threshold)**或者**偏置(bias)**b与**激活函数(activation function)**σ, 神经元接受到的总输入值将与阈值进行比较, 然后通过激活函数处理以产生神经元的输出.

于是我们得到一个基本的神经元表示:

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i - b\right).$$

[①]周志华, 机器学习, 清华大学出版社, 2016. (ISBN 978-7-302-42328-7)
[②]图片来源: https://www.ituring.com.cn/book/tupubarticle/28234

在生物神经网络中, 激活函数σ取为Heaviside函数:

$$\sigma(x) = \operatorname{sgn}(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

对应于神经细胞中神经元的电位超过阈值之后, 就会被激活, 并向下一个神经元传递信号.

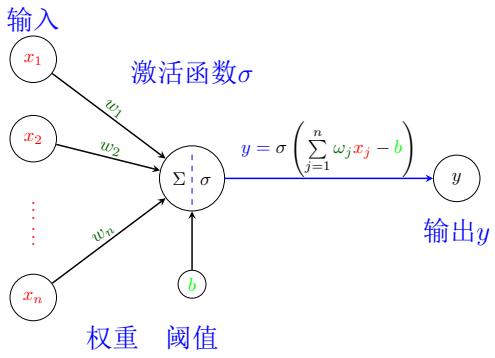


Figure 7.2: M-P神经元模型

感知机模型(perceptron)只包含一个M-P神经元, 所以其表达能力非常有限, 只能用于二分类, 无法学习比较复杂的非线性模型.

神经网络在感知机的模型上作了扩展, 主要有三点:

- 在输入层和输出层之间加入了隐藏层, 增强模型的表达能力.
- 输出层的神经元也可以不止一个输出, 可以有多个输出. 这样模型可以灵活应用于分类或回归, 以及其他的学习任务.
- 对激活函数进行扩展. 感知机的激活函数是$\operatorname{sgn}(x)$. 虽然简单但是存在不连续、不光滑等缺点, 因此神经网络中一般使用其他激活函数.

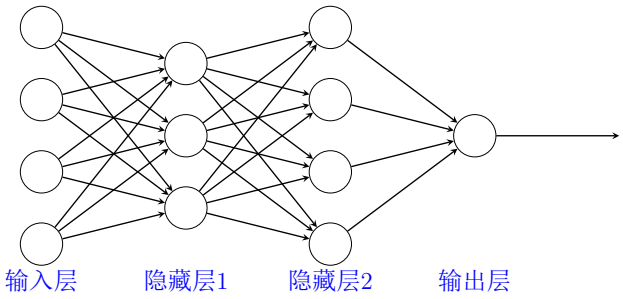


Figure 7.3: 有隐藏层的推广

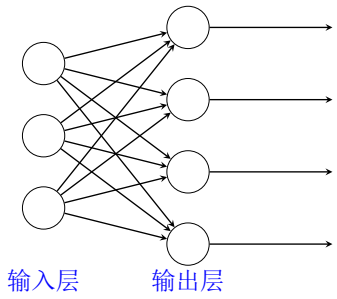
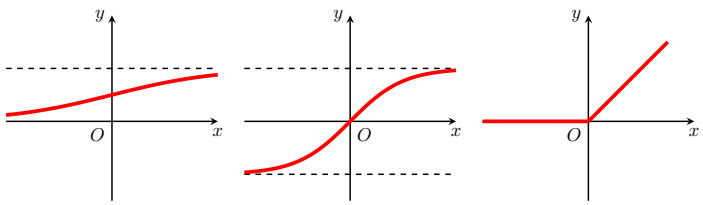


Figure 7.4: 有多个输出的推广

一些激活函数的例子:

- Logistic sigmoid函数: $\sigma(t) = \dfrac{1}{1+e^{-t}}$.
- tanh函数: $\sigma(t) = \tanh t$或者$\sigma(t) = \tanh \dfrac{t}{2}$, 它相当于sigmoid函数作了平移和伸缩.
- ReLU函数: $\sigma(t) = \max\{0, t\}$.



$y = \dfrac{1}{1+e^{-x}}$ $y = \tanh x$ $y = \max\{0, x\}$

神经网络是基于感知机的扩展, 而深度神经网络可以理解为有很多隐藏层的神经网络, 有时也把深度神经网络叫做**多层感知机**. 常见的深度神经网络具有如图所示的层级结构, 第一层为**输入层(input layer)**, 最后一层为**输出层(output layer)**, 而中间的层都是**隐藏层(hidden layer)**. 每层神经元与下一层神经元全互连, 神经元之间不存在同层连接, 也不存在跨层连接. 这样的网络结构通常称为**多层前馈神经网络(multi-layer feedforward neural network)**.

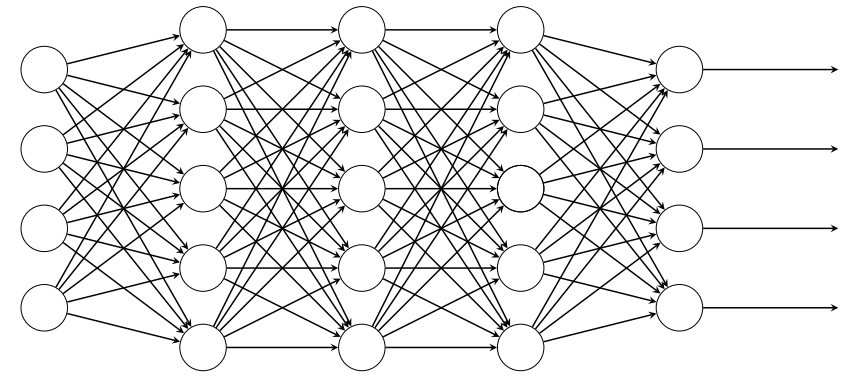


Figure 7.5: 深度神经网络

7.5.2 神经网络的数学表示

虽然深度神经网络看起来很复杂, 但是从小的局部模型来说, 还是和感知机一样的, 即一个线性变换$z = \boldsymbol{W} \cdot x + b$加上一个激活函数$\sigma(\cdot)$.

由于我们的神经网络推广到了多个输出, 那么$\sigma(\boldsymbol{x})$这一记号需要加以说明. 如果$\boldsymbol{x} = (x_1, \cdots, x_n)^T$是个向量, 我们记$\sigma(\boldsymbol{x})$为逐分量作用:

$$\sigma(\boldsymbol{x}) = (\sigma(x_1), \cdots, \sigma(x_n))^T.$$

我们考虑相邻两层, 第$l-1$层的输出$\boldsymbol{z}^{l-1}$即为第l层的输入, 从而第l层的输出可以表示为

$$\boldsymbol{z}^l = \sigma(\boldsymbol{W}^l \boldsymbol{z}^{l-1} + b^l), \tag{7.6}$$

其中$\boldsymbol{W}^l$表示第$l-1$层到第l层的连接权重矩阵, b^l表示第l层的偏置向量, 激活函数作用于向量视作作用于该向量的每一个分量. 最后, 整个多层前馈神经网络可以表示为

$$y = A^L \circ \sigma \circ A^{L-1} \circ \sigma \circ \cdots \circ \sigma \circ A^0(x),$$

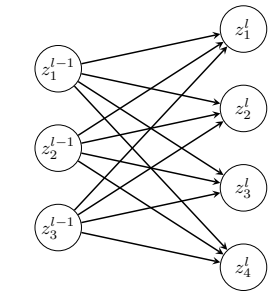
其中, $A^l(z) = \boldsymbol{W}^l z + b^l (l = 0, 1, \cdots, L)$. 这样的神经网络是一个$L+1$层神经网络, 共有$L$个隐藏层、1个输出层.

例 7.5.1 在下图中，第$l-1$层共有3个神经元，第l层共有4个神经元，记

$$\boldsymbol{W}^l = \begin{pmatrix} w_{11}^l & w_{12}^l & w_{13}^l \\ w_{21}^l & w_{22}^l & w_{23}^l \\ w_{31}^l & w_{32}^l & w_{33}^l \\ w_{41}^l & w_{42}^l & w_{43}^l \end{pmatrix}, \boldsymbol{b}^l = \begin{pmatrix} b_1^l \\ b_2^l \\ b_3^l \\ b_4^l \end{pmatrix}$$

则对于$j=1,2,3,4$，可以把前一页的(7.6)式写成分量形式为

$$z_j^l = \sigma\left(z_1^{l-1}w_{j1}^l + z_2^{l-1}w_{j2}^l + z_3^{l-1}w_{j3}^l + b_j^l\right).$$



$\boldsymbol{W}^l$: 4×3矩阵

$\boldsymbol{b}^l$: 4维向量

$\boldsymbol{z}^{l-1}=(z_1^{l-1},z_2^{l-1},z_3^{l-1})^T$: 3维向量

$\boldsymbol{z}^l=(z_1^l,z_2^l,z_3^l,z_4^l)^T$: 4维向量

Figure 7.6: 深度神经网络的局部

事实上，从计算科学的角度看，我们可以先不考虑神经网络是否真的模拟了生物神经网络，只需将一个神经网络视为**包含了许多参数的数学模型**，这个模型是**若干个函数相互组合嵌套**而得.

神经网络的**学习过程**，就是根据训练数据来调整神经元之间的连接权重以及每个神经元的偏置；换言之，神经网络“学”到的东西，蕴含在连接权重和偏置中.

7.5.3 神经网络的损失函数

神经网络可以说是由其结构和参数控制. 确定了网络结构后，神经网络由其参数唯一确定，因此我们可以把神经网络模型记为$\phi(x;\theta)$，其中，θ表示神经网络的参数(权重与偏置). 那么我们如何确定神经网络的参数θ从而得到最佳的模型$\phi(x;\theta)$呢？

首先我们要确立一个评价模型好坏的标准，通常我们使用**损失函数(loss function)** $J(\theta)$. 通常针对不同的问题，以及不同的网络结构，我们会使用不同的损失函数. 我们以最简单的**均方误差(mean-square loss)**为例:

$$J(\theta) := \|\phi(x;\theta) - y\|_2^2,$$

其中，对于$f\in L^2(\Omega)$，记$\|f\|_2 = \left(\int_\Omega |f(x)|^2 \mathrm{d}x\right)^{\frac{1}{2}}$表示$f$的$L^2$**-范数**.

这个损失函数的直观意义相当明确，模型的预测值和真值的欧氏距离越大，损失就越大，反之就越小. 因此，最优的参数θ^*应使得损失函数$J(\theta)$达到最小，即

$$\theta^* = \arg\min_\theta J(\theta) \tag{7.7}$$

这样，模型的训练就变成了一个优化问题.

由于计算机只能处理离散数据，没有办法精确计算

$$J(\theta) := \|\phi(x;\theta) - y\|_2^2,$$

并且在一般情况下, 我们只有部分数据集$\{(x_i, y_i)\}_{i=1}^n$, 所以我们通常使用**经验损失函数(empirical loss)**

$$J_e(\theta) := \frac{1}{n}\|\phi(x_i;\theta) - y_i\|_2^2,$$

其中, 对于$\boldsymbol{z} = (z_1, \cdots, z_n)$, 记$\|\boldsymbol{z}\|_2 = \left(\sum_{i=1}^n |z_i|^2\right)^{\frac{1}{2}}$表示$\boldsymbol{z}$的**欧几里得范数(Euclidean norm)**.

神经网络的训练过程就是寻找最优参数θ^*使得

$$\theta^* = \arg\min_{\theta} J_e(\theta) \tag{7.8}$$

7.5.4 梯度下降法简介

一般地, 设f是$\mathbb{R}^n$中连续可微函数, 现在需要求它的最小值. 把f的梯度记为

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right),$$

梯度下降法(gradient descent method)是下面的迭代算法: [①]

$$x_{k+1} = x_k - \alpha_k g_k,$$

其中, $g_k = g(x_k) = \nabla f(x_k)^T$, 非负实数$\alpha_k$是**步长(stepsize)**.

梯度下降法的计算过程就是沿梯度下降的方向迭代求解极小值, 换言之, 从x_k这一点开始, 我们沿着负梯度方向$-g_k$来进行搜索, 找到一点x_{k+1}, 使得$f(x_{k+1}) \leq f(x_k)$.

对于优化问题(7.7), 我们把步长设成常数, 迭代公式为

$$\theta_{k+1} = \theta_k - \eta\nabla J(\theta).$$

其中η在机器学习中也叫做**学习率(learning rate)**.

在梯度下降法的求解过程中, 只需求解**损失函数的一阶导数, 计算代价比较小**, 可以在很多大规模数据集上应用.

但是，梯度下降法由于方向选择的问题, 得到的结果不一定是全局最优, 即参数寻优陷入了**局部极小**, 这显然不是我们所希望的.

除此之外, 梯度下降法每次迭代更新都需要遍历所有的数据, 当样本数据很多时, **计算量开销大, 计算速度慢**.

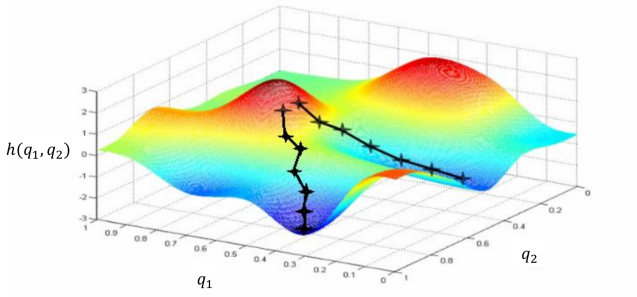


Figure 7.7: 梯度下降法陷入局部极小

一般来说学习率需要合理选取. 如果太低, 会导致收敛速度太慢; 如果太高, 会导致“跨过”了极小值点:

[①] 关于梯度下降法的参考书: Luenberger, David & Ye, Yinyu. (1984). Linear and Nonlinear Programming. doi:10.2307/1240727.

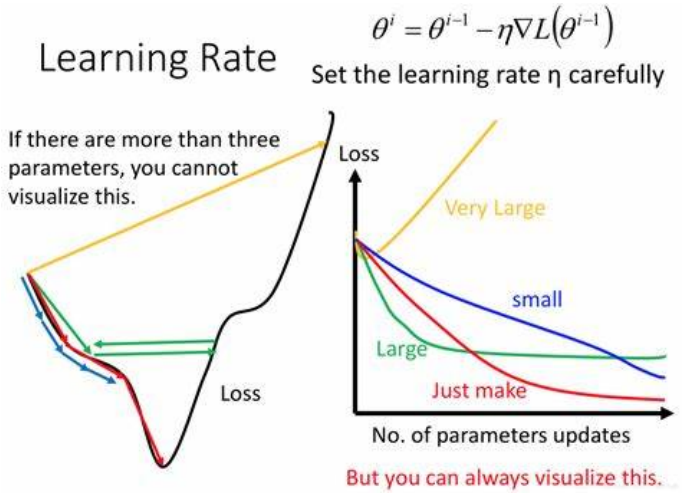


Figure 7.8: 学习率对收敛性的影响

因此，在梯度下降法的基础上，设计出了**小批量梯度下降法(mini-batch gradient descent)**, 把数据分为若干个批次, 按批次来更新参数, 这样, 一个批次中的一组数据共同决定了本次梯度的方向, 从而减少了计算量. 然而这种方法并不能解决陷入局部极小的问题, 且不恰当的分组可能导致更新效果出现“抵消”的现象.

在此基础上, **随机梯度下降法(stochastic gradient descent)**被提了出来, 每个批次中的数据都是随机选取的, 这样就解决了人为分批次可能出现的问题, 并且即便陷入局部极小值点, 它计算出来梯度仍可能不为零, 这样就有机会跳出局部极小继续搜索迭代.

7.5.5 函数逼近性质

我们已经指出, L个隐藏层的神经网络可以表示成

$$y = A^L \circ \sigma \circ A^{L-1} \circ \sigma \circ \cdots \circ \sigma \circ A^0(\boldsymbol{x}), \tag{7.9}$$

其中, $A^l(z) = \boldsymbol{W}^l z + b^l (l = 0, 1, \cdots, L)$. 如果输入层是$d$维, 即$\boldsymbol{x} \in \mathbb{R}^d$, 输出层是$p$维, 即$y \in \mathbb{R}^p$, 那么我们就得到了映射关系$\boldsymbol{x} \mapsto y$的函数表达式.

把所有形如(7.9)的神经网络函数放在一起, 构成一个函数族, 记为$\mathcal{F}$. 这样的函数族中, 所有权重和偏置都是可调的, 这样神经网络函数(7.9)可以作为函数逼近空间. [①]

虽然大部分函数不能精确地用(7.9)表示, 但是随着隐藏层数、神经元数的增加, 网络的非线性建模能力(即表达能力)会快速增长, 这是一般的逼近方法所不能比拟的. 另外, 简便统一的训练算法以及分部并行计算的易于实现, 使得神经网络逼近函数具有传统函数逼近方法所不具备的优势.

于是我们现在关心的是$\mathcal{F}$究竟对哪些函数有逼近能力, 逼近的阶和精度是怎样的, 与经典的函数逼近方法相比有哪些优缺点.

1989年, Hornik[②] 等人建立了下面的**通用逼近定理(universal approximation theorem)**.

记$C(X, Y)$表示从X到Y的连续函数全体, 激活函数$\sigma \in C(\mathbb{R}, \mathbb{R})$. 记

$$\Sigma^r(\sigma) = \left\{ f : \mathbb{R}^r \to \mathbb{R} \,\middle|\, \sum_{i=1}^{q} \beta_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} + b_i), \boldsymbol{w}_i, \boldsymbol{x} \in \mathbb{R}^r, b_i, \beta_i \in \mathbb{R}, q = 1, 2, \cdots \right\}.$$

[①] 虽然我们把$\mathcal{F}$称作“空间”, 但如果对每层的神经元数个数加以限制, 那么它就不是严格意义上的线性空间, 因为对加法不封闭.
[②] Hornik, K., Stinchcombe, M, & White, H. (1989). Multilayer feedforward networks are universal approximators.

定理 7.5.1: 通用逼近定理

若σ不是多项式、单调递增, 且满足$\lim_{x\to+\infty}\sigma(x)=1$, $\lim_{x\to-\infty}\sigma(x)=0$, 则对任意$r\in\mathbb{N}^+$、紧集$K\subset\mathbb{R}^r$、$\varepsilon>0$、连续函数$f\in C(\mathbb{R}^r,\mathbb{R})$, 存在$g\in\Sigma^r(\sigma)$, 使得

$$\sup_{x\in K}|f(x)-g(x)|<\varepsilon.$$

这个定理也有推广到多个隐藏层、多个输出的深度神经网络的版本.

ReLU激活函数$\mathrm{ReLU}(x)=\max\{0,x\}$不满足上面定理的条件, 但Yarotsky [①]给出了如下结果, 不仅证明了ReLU网络有逼近性质, 还给出了网络大小的估计.

记$W^{n,\infty}([0,1]^d)$是Sobolev空间, 其范数定义为

$$\|f\|_{W^{n,\infty}([0,1]^d)}=\max_{\alpha:|\alpha|\le n}\operatorname*{ess\,sup}_{\boldsymbol{x}\in[0,1]^d}|D^\alpha f(\boldsymbol{x})|.$$

并且记$W^{n,\infty}([0,1]^d)$中的单位球为$F_{n,d}=\{f\in W^{n,\infty}([0,1]^d):\|f\|_{W^{n,\infty}([0,1]^d)}\le 1\}$, 则有如下结论.

定理 7.5.2

对任意正整数d,n与$\varepsilon\in(0,1)$、任意$f\in F_{n,d}$, 存在ReLU-神经网络f_θ满足

$$\sup_{\theta\in\Theta}\|f_\theta(x)-f(x)\|_{W^{n,\infty}([0,1]^d)}<\varepsilon.$$

并且这个神经网络的隐藏层数至多为$c(\ln(1/\varepsilon)+1)$, 神经元总数是$c\varepsilon^{-d/n}(\ln(1/\varepsilon)+1)$, 其中常数$c=c(d,n)$.

Maiorov[②]给出了对1个隐藏层的神经网络而言的逼近结论, 但只是说存在激活函数, 并不是对所有激活函数都成立.

设

$$\mathcal{M}_r(\sigma)=\left\{\sum_{i=1}^r a_i\sigma(\boldsymbol{w}^i\cdot\boldsymbol{x}+b_i):a_i,b_i\in\mathbb{R},\boldsymbol{w}^i\in\mathbb{R}^d\right\},$$

以及

$$\|f\|_{W^{n,p}([0,1]^d)}=\Big(\sum_{0\le|\alpha|\le m}\|D^\alpha f\|_p^p\Big)^{\frac{1}{p}},\qquad 1\le p<\infty.$$

定理 7.5.3

存在一个激活函数σ, 满足单调递增且$\lim_{x\to+\infty}\sigma(x)=1$, $\lim_{x\to-\infty}\sigma(x)=0$, 使得对$1\le p\le\infty$, $d\ge 2$, $n\ge 1$, 有

$$\sup_{f\in F_{n,d}}\inf_{g\in\mathcal{M}_r(\sigma)}\|f-g\|_{W^{n,p}([0,1]^d)}\le Cr^{-\frac{n}{d-1}}.$$

上述的逼近定理都是考虑的神经网络的**最佳逼近误差**, 即神经网络能够表示的函数中对目标函数近似效果最好的情况. 类似的结论还有很多.

在实际应用中, 我们还要考虑由于对损失函数离散和采样带来的**泛化误差(generalization error)**, 以及优化算法带来的**优化误差(optimization error)**. 此外, 实际应用中还要考虑网络结构的选择以及初始参数的设置.

从数学上来看, 神经网络函数逼近代表了非线性函数的一类仿射展开, 其特殊优点是基函数的选取自由度大, 展开系数可由统一的训练算法获得. 不仅可以用来逼近我们通常意义上的函数, 还可以用来数值求解常微分方程与偏微分方程(组), 应用非常广泛. 可以说, 神经网络用于函数逼近为函数逼近理论的研究开辟了一个崭新的方向.

[①]Dmitry Yarotsky. Error bounds for approximations with deep relu networks. Neural Networks, 94:103 - 114, 2017.

[②]V. E. Maiorov (1999), 'On best approximation by ridge functions', to appear in J. Approx. Theory

7.5.6 用Python实现

我们考虑用1个隐藏层的神经网络逼近一元函数f. 考虑下面的函数族

$$\mathcal{M}_r(\sigma)=\left\{\sum_{i=1}^{r}a_i\sigma(w_i\cdot x+b_i):a_i,b_i\in\mathbb{R},w_i\in\mathbb{R}\right\}$$

对于有界区间K上的一个充分光滑的函数$f(x)$, 希望找$f_{\boldsymbol{\theta}}(x)\in\mathcal{M}_r(\sigma)$, 使得$\|f-f_{\boldsymbol{\theta}}\|$充分小. 程序中采取的损失函数为

$$J(\boldsymbol{\theta})=\frac{1}{N}\sum_{j=1}^{N}(f_{\boldsymbol{\theta}}(x_j)-y_j)^2,$$

其中$\boldsymbol{\theta}$是所有参数a_i,b_i,w_i构成的向量, $y_j=f(x_j)(j=1,2,\cdots,N)$是精确给出来的, 并且我们把$\{(x_j,j_i)\}_{j=1}^{N}$作为训练集. 而优化的算法采用梯度下降法.

附件的程序各部分细节可以参见程序中的注释, 激活函数$\sigma(x)=\dfrac{1}{1+e^{-x}}$是logistic sigmoid函数, 学习率取为0.2, 终止准则取为

$$|J(\boldsymbol{\theta}_{k+1})-J(\boldsymbol{\theta}_k)|<\varepsilon=10^{-9}.$$

其中, $\boldsymbol{\theta}_k$表示梯度下降法第k步所得到的参数.

该程序用$\mathcal{M}_r(\sigma)$中的函数来逼近一元函数$f(x)=\dfrac{1}{1+x^2},x\in[-5,5]$, 神经元数取为20, 训练集取为等距结点$\{x_i,f(x_i)\}_{i=1}^{21}$, 其中$x_i=-5+\dfrac{i-1}{2}$.

运行程序, 得到的$f_\theta(x)$函数图象如下, 一共迭代了25138次. (由于神经网络参数的初值是随机取的, 所以每次运行会得到不同的神经网络与不同的迭代次数).

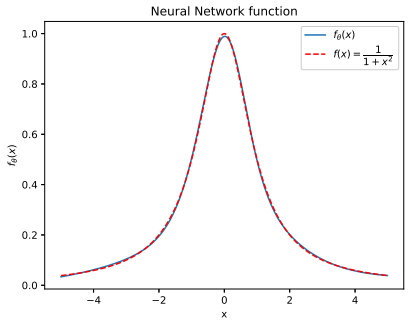


Figure 7.9: 程序运行结果

可以考虑:

- 探究不同的终止准则、学习率、神经元数、激活函数、训练集点的选取等因素对运行速度、收敛阶等的影响.
- 考虑把梯度下降法改成随机梯度下降法, 探究对运行速度的影响.
- 用神经网络求单调函数的反函数.
- 把这个程序中的单个隐藏层神经网络推广到多个隐藏层的神经网络并作相应的探究. (需要重写程序中核心的代码——`train`函数, 比较有挑战性)
- 用深度神经网络**求ODE数值解、求积分方程的解(目前研究热点!)**.

CHAPTER 8

常微分方程数值解基本理论

§ 8.1 几种基本方法

问题: 求解线性ODE方程(组)

$$\begin{cases} y' = f(t,y), \quad a \le t \le b, \\ y(a) = \eta. \end{cases}$$

其中f关于t,y是Lipschitz的.

8.1.1 Euler方法

设步长为h, 则对上式方程两边关于t积分, 可得

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_n+h} f(t,y(t))dt.$$

近似对右端积分取为左端点, 可得**Euler法**:

$$y_{n+1} = y_n + hf(t_n, y_n).$$

局部截断误差$R_n = O(h^2)$.

注: **局部截断误差**定义为迭代方法中把y_k换成$y(t_k)$之后多出来的余项. 比如上面的局部截断误差R_n满足

$$y(t_{n+1}) - y(t_n) = hf(t_n, y(t_n)) + R_n.$$

如果改用梯形方法估计右端积分, 可得**改进的Euler法**:

$$y_{n+1} = y_n + \frac{h}{2}[f(t_n,y_n) + f(t_{n+1}, y_{n+1})].$$

此时$R_n = O(h^3)$. 它是个**隐式方法**.

注: 局部阶段误差$R_n = O(h^k)$代表方法是$k-1$阶的.

定理 8.1.1

Euler方法的局部截断误差是$R_n = O(h^2)$, 从而迭代格式是一阶格式.

证明: 记

$$y(t_{n+1})-y(t_n)=hf(t_n,y(t_n))+R_n, e_n=y(t_n)-y_n,$$

作差可得**误差方程**

$$e_{n+1}-e_n=h[f(t_n,y(t_n))-f(t_n,y_n)]+R_n$$

从而

$$\begin{aligned}
|e_{n+1}| &\le |e_n|+hL|y(t_n)-y_n|+|R_n| \\
&\le (1+hL)|e_n|+Mh^2 \\
&\le (1+hL)\left((1+hL)|e_{n-1}|+Mh^2\right)+Mh^2 \\
&\le \cdots \le (1+hL)^{n+1}|e_0|+[(1+hL)^n+\cdots+(1+hL)^0]\cdot\frac{M}{h}^2 \\
&\le e^{hL(n+1)}|e_0|+\frac{(1+hL)^{n+1}-1}{hL}\cdot Mh^2 \\
&\le e^{hL(n+1)}|e_0|+\frac{e^{hL(n+1)}-1}{L}\cdot Mh \\
&\le \frac{e^{L(b-a)}-1}{L}\cdot Mh \triangleq Ch.
\end{aligned}$$

(注意$e_0=0, Nh=b-a$) □

注: 同理可证改进Euler法是二阶的.

8.1.2 Runge-Kutta方法

思想: 单步法的一般形式是

$$y_0=\eta, y_{n+1}=y_n+h\Phi(t_n,y_n,h), n=0,1,\cdots,N-1.$$

根据Taylor展开,

$$\begin{aligned}
y(t+h) &= y(t)+hf(t,y(t))+\frac{1}{2}h^2f'(t,y(t))+O(h^3) \\
&\triangleq y(t)+h\Phi(t,y,h)+O(h^3)
\end{aligned}$$

而

$$\begin{aligned}
f'(t,y(t)) &= \frac{d}{dt}f(t,y)=f'_t(t,y)+f'_y(t,y)\cdot y'(t), \\
y'(t) &= f(t,y),
\end{aligned}$$

于是上式可以写成

$$y(t+h)=y(t)+h[f(t,y)+\frac{1}{2}h(f'_t(t,y)+f'_y(t,y)f(t,y))]+O(h^3).$$

另一方面, 令

$$\begin{aligned}
&\Phi(t,y,h)=c_1K_1+c_2K_2, \\
&K_1=f(t,y) \\
&K_2=f(t+a_2h,y+b_{21}hK_1),
\end{aligned}$$

展开K_2得

$$K_2=f(t,y)+a_2hf'_t(t,y)+b_{21}hf(t,y)f'_y(t,y)+O(h^2).$$

把K_1, K_2代入, 并比较系数可得

$$\begin{aligned} c_1 + c_2 &= 1, \\ c_2 a_2 &= \frac{1}{2}, \\ c_2 b_{21} &= \frac{1}{2}. \end{aligned}$$

上面有4个未知数、3个方程, 所以有一个未知数可以自由选取.

- 选$a_2 = 1$, 则$c_1 = c_2 = \frac{1}{2}$, 此时得到**RK2**公式(又叫**Heun方法**:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{2}(K_1 + K_2), \\ K_1 &= f(t_n, y_n), \\ K_2 &= f(t_n + h, y_n + hK_1), \\ y_0 &= \eta. \end{aligned}$$

- 选$a_2 = \frac{1}{2}$, 则$c_1 = 0, c_2 = 1$, 此时得到**变形的Euler方法(中点方法)**:

$$\begin{aligned} y_{n+1} &= y_n + hK_2 = y_n hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right) \\ y_0 &= \eta. \end{aligned}$$

注: 展开到4阶并比较系数可以得到如下**RK4**方法:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4), \\ K_1 &= f(t_n, y_n), \\ K_2 &= f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hK_1\right) \\ K_3 &= f\left(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hK_2\right) \\ K_4 &= f\left(t_n + h, y_n + hK_3\right) \\ y_0 &= \eta. \end{aligned}$$

注: 也可以推隐式, 方法也是用Taylor展开. 隐式RK2方法为

$$\begin{aligned} y_{n+1} &= y_n + K_1, \\ K_1 &= hf(t_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1), \\ y_0 &= \eta. \end{aligned}$$

1. 用步长h和$h/2$对Euler方法执行Richardson外推, 推导前面的变形的Euler方法.
2. 推导RK3方法

$$y_{n+1} = y_n + \frac{1}{9}(2K_1 + 3K_2 + 4K_3),$$

其中

$$\begin{aligned} K_1 &= hf(t, y) \\ K_2 &= hf\left(t + \frac{1}{2}h, y + \frac{1}{2}K_1\right) \\ K_3 &= hf\left(t + \frac{3}{4}h, y + \frac{3}{4}K_2\right) \end{aligned}$$

3. 说明当RK4方法用于问题$y'=\lambda y$时, 一步推进的公式将会是

$$y_{n+1}=\left(1+h\lambda+\frac{1}{2}h^2\lambda^2+\frac{1}{6}h^3\lambda^3+\frac{1}{24}h^4\lambda^4\right)y_n.$$

并证明局部截断误差是$O(h^5)$的.

§8.2 单步法的相容性、稳定性、收敛性

下面要求解初值问题

$$\begin{cases}y'=f(t,y), & a\le t\le b,\\ y(a)=\eta.\end{cases}$$

它的显式单步法的一般形式是

$$\begin{aligned}&y_{n+1}=y_n+h\Phi(t_n,y_n,h),\\ &y_0=\eta.\\ &h=\frac{b-a}{N},t_n=a+nh.\end{aligned}$$

作差并让$h\to 0$, 我们希望

$$\lim_{h\to 0}\left[\frac{y(t+h)-y(t)}{h}-\Phi(t,y(t),h)\right]=0,$$

即$y'(t)=\Phi(t,y(t),0)$. 而$y'(t)=f(t,y)$.

定义 8.2.1: 相容性

若关系式$\Phi(t,y,0)=f(t,y)$成立, 则称单步法与微分方程初值问题**相容**. 或者把这个关系式叫做**相容条件**.

注: 相容的单步法至少是一阶的. 证明方法: $y(t+h)-y(t)=h\Phi(t,y(t),h)+R(t,h)$, 其中$R(t,h)=O(h^{p+1})$, 即

$$\left|\frac{R(t,h)}{h}\right|\le Mh^p$$

对迭代公式取极限$h\to 0$并用相容条件即可得$\lim\limits_{h\to 0}\dfrac{R(t,h)}{h}=0$, 因此$p\ge 1$, 方法至少为一阶的. □

定义 8.2.2: 收敛

设$f(t,y)$在区域$R=\{(t,y):a\le t\le b,-\infty<y<\infty\}$连续, 且关于$y$满足Lipschitz条件, 若对所有$t\in[a,b]$都有

$$\lim_{h\to 0}y_n=y(t),(t_n\text{固定})$$

则称单步法收敛.

注: 证明单步法收敛的方法: 把真解求出来可得$y(t)$, 再根据迭代公式把y_n求出来, 然后估计$y_n-y(t_n)$. 中间可能要用到Taylor公式(书P242例题)

定理 8.2.1

若$\Phi(t,y,h)$关于t,h,y满足Lipschitz条件, 则收敛⇔相容.

定理 8.2.2

在前一定理的条件下, 若$R(t,h)=O(h^{p+1})$(方法是p阶的), 则p阶单步法的整体离散误差$\varepsilon_n=y(t_n)-y_n=O(h^p)$.

证明: $|R(t,h)|\le Mh^{p+1}$, 然后证明$|\varepsilon_n|\le e^{L(b-a)}|\varepsilon_0|+h^p\dfrac{M}{L}(e^{L(b-a)}-1)$. □

定义 8.2.3: 稳定性

若存在正常数h_0,C使得对任意的初值$y_0,\tilde{y}_0$, 单步法对应的相应精确解$y_n,\tilde{y}_n$与所有的$0<h\le h_0$都有

$$|y_n-\tilde{y}_n|\le C|y_0-\tilde{y}_0|,$$

则称这个单步法是稳定的.

定理 8.2.3

若$\Phi(t,y,h)$关于y是Lipschitz的, 则单步法稳定.

证明: 分别写

$$y_{n+1}=y_n+h\Phi(t_n,y_n,h),$$
$$\tilde{y}_{n+1}=\tilde{y}_n+h\Phi(t_n,\tilde{y}_n,h),$$

作差然后作简单放缩即可. □

定理 8.2.4

在单步法中若相容、收敛、稳定之中有两个成立, 则第三个也成立.

注: 验证收敛的时候, 可以先验证相容与稳定性再去验证收敛性.

定义 8.2.4

对给定微分方程和给定步长h, 如果由单步法计算y_n时有大小为δ的误差, 即计算得$\tilde{y}_n=y_n+\delta$, 而引起其后值$y_m(m>n)$的变化小于δ, 即$|\tilde{y}_m-y_m|<|\delta|$, 则说该单步法**绝对稳定**.

注: 一般只限于典型微分方程

$$y'=\mu y$$

考虑数值方法的绝对稳定性, 其中μ是实数. **若对所有$\mu h\in(\alpha,\beta)$, 单步法都绝对稳定, 则称(α,β)是绝对温度区间.**

例 8.2.1 *求Euler方法$y_{n+1}=y_n+hf(t_n,y_n)$的绝对稳定区间.*

证明: 考虑方程$y'=\mu y$, 得$y_{m+1}=(1+\mu h)y_m$. 如果$\tilde{y}_n=y_n+\delta$, 则$\tilde{y}_{m+1}=(1+\mu h)\tilde{y}_m(m\ge n)$, 所以

$$\begin{aligned}|\tilde{y}_{m+1}-y_{m+1}|&=|1+\mu h||\tilde{y}_m-y_m|\\&=|1+\mu h|^2|\tilde{y}_{m-1}-y_{m-1}|\\&=|1+\mu h|^{m+1-n}|\tilde{y}_n-y_n|,m\ge n-1.\end{aligned}$$

若$|1+\mu h|<1$, 则$|\tilde{y}_{m+1}-y_{m+1}|<|\delta|$. 所以绝对稳定区间是$\mu h\in(-2,0)$. □

例 8.2.2 *求梯形方法$y_{n+1}=y_n+\dfrac{1}{2}[f(t_n,y_n)+f(t_{n+1},y_{n+1})$的绝对稳定区间.*

证明: 考虑方程$y'=\mu y$, 有

$$y_{n+1}=y_n+\frac{h}{2}(\mu y_n+\mu y_{n+1}),$$

如果要绝对稳定, 则

$$\left|\frac{1+\frac{\mu h}{2}}{1-\frac{\mu h}{2}}\right|<1.$$

解得$\mu h<0$, 绝对稳定区间是$(-\infty,0)$. □

例 8.2.3 *RK4方法的绝对稳定区间满足*

$$\left|1+\mu h+\frac{1}{2}h^2\mu^2+\frac{1}{6}h^3\mu^3+\frac{1}{24}h^4\mu^4\right|<1.$$

大概为$(-2.78,0)$.

8.2.1 附录：通常的迭代公式的稳定性分析方法

如果一个ODE可以用下面的公式数值求解：

$$y_{n+1}=T(y_n,t_n),$$

若对y_n施加一个扰动ε_n, 那么y_{n+1}也会有一个扰动ε_{n+1}, 于是

$$y_{n+1}+\varepsilon_{n+1}=T(y_n+\varepsilon_n,t_n).$$

我们作一下线性化, 当ε_n充分小的时候, 忽略高阶项, 我们有

$$y_{n+1}+\varepsilon_{n+1}=T(y_n,t_n)+\varepsilon_n\frac{\partial T(y_n,t_n)}{\partial y_n},$$

或者写

$$\varepsilon_{n+1}=g_n\varepsilon_n.$$

其中, g_n叫做**增长因子(growth factor)**, 如下给出:

$$g_n=\frac{\partial T(y_n,t_n)}{\partial y_n}.$$

如果格式是稳定的, 那么我们需要$|g_n|<1$, 或者记为$\left|\frac{\partial y_{n+1}}{\partial y_n}\right|<1$也行.

特别地, 对于Euler方法

$$y_{n+1}=y_n+hf(y_n,t_n),$$

我们有$T(y_n,t_n)=y_n+hf(y_n,t_n)$, 所以

$$g_n=\frac{\partial T(y_n,t_n)}{\partial y_n}=1+h\frac{\partial f(y_n,t_n)}{\partial y_n}.$$

下面分两类讨论:

(i)衰减问题(decay-type problem), 即$\frac{\partial f(y,t)}{\partial y}<0$, 此时

$$g_n=1-h\left|\frac{\partial f}{\partial y_n}\right|.$$

代入$|g_n|<1$解得

$$0<h<\frac{2}{|\frac{\partial f(y_n,t_n)}{\partial y_n}|}.$$

所以只有当步长充分小的时候才有稳定性.

(ii)增长问题(growth problem), 即$\dfrac{\partial f(y,t)}{\partial y}>0$, 此时

$$g_n=1+h\frac{\partial f}{\partial y_n}>1$$

恒成立, 所以Euler方法是无条件不稳定的.

对于我们熟悉的$f(y,t)=-\mu y(\mu>0)$的情形, (i)相当于$0<h<-\dfrac{2}{\mu}$, 即$-2<\mu h<0$, 这是我们熟悉的结果.

对于其他格式, 假设

$$\Delta=h\frac{\partial f(y,t)}{\partial y}.$$

可以证明:

格式名称	格式	g_n	稳定条件
RK2	$y_{n+1/2}=y_n+\dfrac{h}{2}f(y_n,t_n)$ $y_{n+1}=y_n+hf(y_{n+1/2},t_{n+1/2})$	$1+\Delta+\dfrac{\Delta^2}{2}$	$-2<\Delta<0$
反向Euler	$y_{n+1}=y_n+hf(y_{n+1},t_{n+1})$	$g_n=1+g_n\Delta$	$\Delta<0$或$\Delta>2$
Crank-Nicolson	$y_{n+1}=y_n+\dfrac{h}{2}[f(y_n,t_n)+f(y_{n+1},t_{n+1})]$	$g_n=\dfrac{2+\Delta}{2-\Delta}$	$\Delta<0$

§8.3 多步法

初值问题

$$y'=f(t,y),a\le t\le b,$$
$$y(a)=\eta$$

的线性k步法的一般公式是

$$\sum_{j=0}^{k}\alpha_j y_{n+j}=h\sum_{j=0}^{k}\beta_j f(t_{n+j},y_{n+j}),n=0,1,\cdots,N-k.$$

其中$h=\dfrac{b-a}{N}$, 且α_j,β_j是常数.

用它来算$\{y_n\}$需要k个初始值$y_0,y_1,\cdots,y_{k-1}$, 初值问题只能给出y_0, 而$y_1,\cdots,y_{n-1}$需要用单步法来算出来.

若$\beta_k=0$, 则y_{n+k}可直接计算, 此时这个格式是**显式的**, 否则是**隐式的**.

8.3.1 Adams-Bashford公式

设$y(t)$是初值问题的解, 对$y'(t)=f(t,y(t))$ 在区间$[t_n,t_{n+1}]$两端积分可得

$$y(t_{n+1})=y(t_n)=\int_{t_n}^{t_{n+1}}f(t,y(t))\mathrm{d}t.$$

记

$$f_m\triangleq f(t_m,y_m)$$

为$f(t_m,y(t_m))$的近似值. 作$k+1$个点

$$(t_n,f_n),(t_{n-1},f_{n-1}),\cdots,(t_{n-k},f_{n-k})$$

的Newton外插多项式

$$p_k(t)=\sum_{m=0}^{k}(-1)^m\binom{-s}{m}\nabla^m f_n,$$

其中, $t = t_n + sh$, $\binom{-s}{m} = \frac{s(s-1)\cdots(s-m+1)}{m!}$., 得到

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} p_k(t)\mathrm{d}t,$$

对多项式作积分以后可得

$$y_{n+1} = y_n + h\sum_{m=0}^{k}\gamma_m\nabla^m f_n,$$

其中

$$\gamma_m = (-1)^m\frac{1}{h}\int_{t_n}^{t_{n+1}}\binom{-s}{m}\mathrm{d}t = (-1)^m\frac{1}{h}\int_0^1\binom{-s}{m}\mathrm{d}s, m = 0, 1, \cdots, k.$$

这公式叫**显式Adams外插公式**或者**Adams-Bashford公式**.

例如五阶Adams-Bashford公式为

$$y_{n+1} = y_n + \frac{h}{720}[1901f_n - 2774f_{n-1} + 2616f_{n-2} - 1274f_{n-3} + 251f_{n-4},$$

四阶Adams-Bashford公式为

$$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$$

二阶Adams-Bashford公式为

$$y_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1}).$$

8.3.2 Adams-Moulton公式

取插值多项式的结点为$t_{n+1}, t_n, t_{n-1}, \cdots, t_{n-k+1}$, 则Newton后插多项式$p_k(t)$是

$$p_k(t) = \sum_{m=0}^{k}(-1)^m\binom{-s}{m}\nabla^m f_{n+1},$$

其中$t_{n+1} + sh = t$. 则

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} p_k(t)\mathrm{d}t = y_n + h\int_{-1}^{0}\sum_{m=0}^{k}(-1)^m\binom{-s}{m}\nabla^m f_{n+1}\mathrm{d}s.$$

即

$$y_{n+1} = y_n + h\sum_{m=0}^{k}\gamma_m^*\nabla^m f_{n+1},$$

其中

$$\gamma_m^* = (-1)^n\int_{-1}^{0}\binom{-s}{m}d\mathrm{s}, m = 0, 1, \cdots, k.$$

这样的公式叫**隐式Adams公式**或者**Adams-Moulton公式**.

例如$k = 1$时

$$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n)$$

当$k = 2$时,

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}),$$

当$k=3$时,

$$y_{n+1}=y_n+\frac{h}{24}[9f_{n+1}+19f_n-5f_{n-1}+f_{n-2}].$$

根据Newton差商公式的结论, k步Adams公式的局部截断误差阶数至少为k.

8.3.3 预测-校正法

§8.4 多步法的相容性、稳定性、收敛性

8.4.1 相容性

线性k步法公式为

$$\sum_{j=0}^{k}a_jy_{n+j}=h\sum_{j=0}^{k}b_jf(t_{n+j},y_{n+j})\qquad(*)$$

其中$a_k\neq 0,a_0,b_0$不同时为0. 我们需要考虑这个差分格式是否逼近原问题$y'=f(t,y(t))$.

定义 8.4.1: 相容性

若求解初值问题的线性k步法$(*)$至少是一阶方法, 则称它是**相容的**.

考虑

$$\rho(\lambda)=a_k\lambda^k+a_{k-1}\lambda^{k-1}+\cdots+a_1\lambda+a_0,$$
$$\sigma(\lambda)=b_k\lambda^k+b_{k-1}\lambda^{k-1}+\cdots+b_1\lambda+b_0,$$

定理 8.4.1: 相容条件

线性k步法$(*)$相容的充要条件是

$$\rho(1)=0,\rho'(1)=\sigma(1).$$

证明: 把$(*)$改写为

$$\sum_{j=0}^{k}[a_jy(t+jh)-hb_jy'(t+jh)]=0.$$

作Taylor展开得到

$$c_0y(t)+c_1hy'(t)+\cdots+c_qh^qy^{(q)}(t)+\cdots=0.$$

其中,

$$c_0=a_0+a_1+\cdots+a_k,$$
$$c_1=a_1+2a_2+\cdots+ka_k-(b_0+b_1+\cdots+b_k),$$
$$\cdots\cdots$$
$$c_q=\frac{1}{q!}(a_1+2^qa_2+\cdots+k^qa_k)-\frac{1}{(q-1)!}(b_1+2^{q-1}b_2+\cdots+k^{q-1}b_k),q=2,3,\cdots.$$

k步公式为q阶的充要条件是$c_0=c_1=\cdots=c_q=0$且$c_{q+1}\neq 0$. □

8.4.2 稳定性

定义 8.4.2: 稳定性

假设$f(t,y)$在R中连续, 且关于y满足Lipschitz条件, 若存在$C>0$与h_0使得当$0<h\leq h_0$时, $(*)$的任何两个解y_n与$\tilde{y}_n$满足

$$\max_{nh\leq b-a}|y_n-\tilde{y}_n|\leq C\max_{0\leq j\leq k-1}|y_j-\tilde{y}_j|,$$

则称$(*)$稳定.

定理 8.4.2: 根条件

线性k步法$(*)$稳定的充要条件是: $\rho(\lambda)$满足特征根条件: 即$\rho(\lambda)$的所有根都在单位圆中, 且在单位圆周上的根只能是单根.

注: 这个条件又称为**弱根条件**. 满足弱根条件的稳定性是弱稳定性.

注: 还有**强根条件:** 特征多项式$\rho(\lambda)$的全部根除了$\lambda=1$以外都落在单位圆内. 满足强根条件的稳定性叫强稳定性.

例 8.4.1 *Adams显式公式*

$$y_{n+1}=y_n+h\sum_{j=0}^{k}\beta_{kj}f_{n-j}$$

的特征多项式是$\rho(\lambda)=\lambda^{k+1}-\lambda^k$, 只有一个根1, 其余根都是0, 则这个格式是强稳定的.

例 8.4.2 *Milne方法校正公式*

$$y_{n+1}=y_{n-1}+\frac{h}{3}(f_{n+1}+4f_n+f_{n-1})$$

的特征多项式是$\rho(\lambda)=\lambda^2-1$, 它的特征根为$\pm1$, 所以是弱稳定但不是强稳定的.

8.4.3 绝对稳定

讨论典型微分方程

$$y'=\mu y$$

那么线性k步法为

$$\sum_{j=0}^{k}a_jy_{n+j}=\mu h\sum_{j=0}^{k}b_jy_{n+j}. \qquad (**)$$

它是齐次常系数线性差分方程, 特征方程是

$$\rho(\lambda)-\mu h\sigma(\lambda)=0.$$

把差分方程的精确解记为y_n.

在求解的时候, 由于舍入误差的影响, 只能得到近似解$\tilde{y}_n$, 满足

$$\sum_{j=0}^{k}a_j\tilde{y}_{n+j}=\mu h\sum_{j=0}^{k}b_j\tilde{y}_{n+j}+\eta_n.$$

其中η_n是舍入误差. 令$e_n=\tilde{y}_n-y_n$, 则它满足

$$\sum_{j=0}^{k}a_je_{n+j}=\mu h\sum_{j=0}^{k}b_je_{n+j}+\eta_n. \qquad (\heartsuit)$$

差分方程$(\heartsuit)$的解可以表示为

$$e_n=\sum_{r=1}^{p}\sum_{l=1}^{s_r}C_{l,r}n^{l-1}\lambda_r^n+\sum_{i=0}^{n-k}g_{n,i}\eta_i.$$

其中λ_r是方程$(\heartsuit)$的互异根, 重数分别为s_r, 这里$r=1,2,\cdots,p$. **当$|\lambda_r|<1$时, 可以希望计算过程的舍入误差对以后计算结果的影响不会步步增长.**

定义 8.4.3

对给定的μ, h, 若特征方程

$$\rho(\lambda) - \mu h\sigma(\lambda) = 0$$

的所有根λ_r的模都小于1, 则称k步法$(**)$关于μh**绝对稳定**. 若对所有$\mu h \in (\alpha, \beta)$, $(**)$都绝对稳定, 称(α, β)为绝对稳定区间.

8.4.4 收敛性

线性k步法的数值启动需要有$y_0, y_1, \cdots, y_{k-1}$, 但是初值只有$y_0 = \eta$, 我们假设$y_1, \cdots, y_{k-1}$已经给定, 满足

$$\lim_{h\to 0} y_i = \eta_i(h).$$

定义 8.4.4: 收敛性

设$f(t, y)$在$R = \{(t, y) : a \le t \le b, y \in \mathbb{R}\}$连续, 关于$y$满足Lipschitz条件, 若对任意$t \in [a, b]$, 当$h \to 0$而$a + nh = t_n = t$固定时, $(*)$的解y_n收敛于原问题的真解$y(t)$, 则称k步法$(*)$收敛.

相容性与稳定性可以推出收敛性.

定理 8.4.3: 收敛⇒相容

若$(*)$收敛, 则$(*)$相容.

证明: 利用连续性与收敛性验证相容条件即可. □

定理 8.4.4: 收敛⇒稳定

若$(*)$收敛, 则$(*)$稳定.

证明: 验证稳定性的根条件. □

定理 8.4.5

若k步法$(*)$相容且稳定, 则$(*)$收敛.

8.4.5 多步法的绝对稳定区间

§8.5 (*)常微分方程组和高阶微分方程的数值解法

§8.6 (*)用深度神经网络数值求解常微分方程

CHAPTER 9

(*)积分方程的解法

本章带(*)表示是课程要求之外的补充内容. 本节需要一些泛函分析的基础. 参考书为[R. Kress, Numerical Analysis, GTM181].

我们要考虑求解下面的线性积分方程:

$$\int_a^b K(x,y)\varphi(y) = f(x), \qquad x \in [a,b]$$

或者

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\mathrm{d}y = f(x), \qquad x \in [a,b].$$

其中, φ是未知函数, K叫做**核(kernel)**, 右端项f是给定函数. 上面两个方程分别叫做**第一类Fredholm积分方程**和**第二类Fredholm积分方程**. 由于第一类Fredholm积分方程的理论和算法都比第二类复杂得多, 所以我们着重考虑第二类.

§9.1 Riesz-Fredholm理论

关于第二类Fredholm积分方程, 在1902年Fredholm建立了第二类积分方程解的存在性理论, 我们现在通常把它叫做Fredholm alternative.

考虑下面的非齐次问题

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\mathrm{d}y = f(x), \qquad x \in [a,b]. \tag{9.1}$$

对应的齐次问题为

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\mathrm{d}y = 0, \qquad x \in [a,b]. \tag{9.2}$$

其中, K是连续的核函数. 右端$f \in C[a,b]$. 可以证明, 非齐次问题(9.1)有唯一解$\varphi \in C[a,b]$当且仅当齐次问题(9.2)只有零解. 这个结论给出了非齐次问题的存在性的刻画, 即齐次问题是否有零解.

定义 9.1.1

设X, Y是赋范线性空间, 线性算子$A: X \to Y$称为**紧算子(compact)**, 若对任意的有界序列$\{\varphi_n\} \subset X$, 序列$\{A\varphi_n\}$在Y中有收敛子列(即$\{A\varphi_n\}$是序列紧的).

换言之, A是紧算子指的是集合$\{A\varphi : \varphi \in X, \|\varphi\| \le 1\}$包含收敛子列.

一些基本结论, 证明可以在许多泛函分析课本找到.

- 紧算子是有界算子,
- 紧算子的线性组合也是紧算子.
- 两个有界算子的乘积是紧算子的充分条件是其中一个算子是紧算子.
- 由Bolzano-Weierstrass定理, 如果有界算子$A : X \to X$的像集$A(X) := \{A\varphi : \varphi \in X\}$是有限维的, 则$A$是紧算子.
- 恒等算子$I : X \to X$是紧算子当且仅当X是有限维的. 而这个也说明了方程$A\varphi = f$与$\varphi - A\varphi = f$之间的区别. 若A是紧算子, 那么A和$I - A$有完全不一样的性质.

定理 9.1.1

设X是赋范线性空间, $A : X \to X$是紧算子, 则$I - A$是满的当且仅当$I - A$是单的. 如果其逆$(I - A)^{-1} : X \to X$存在, 那么$(I - A)^{-1}$是有界的.

为了说明Fredholm的存在性分析可以看做是定理9.1.1的一个特殊情形, 我们需要证明线性积分算子$A : C[a, b] \to C[a, b]$是紧算子, 其中

$$(A\varphi)(x) := \int_a^b K(x, y)\varphi(y)\mathrm{d}y, \qquad x \in [a, b]. \tag{9.3}$$

我们需要用到Ascoli-Arzelà定理.

称由定义在$[a, b]$上的函数构成的集合U是**一致有界的(uniformly bounded)**, 若存在常数C使得

$$|\varphi(x)| \le C, \qquad \forall x \in [a, b], \forall \varphi \in U.$$

称U是**等度连续的(equicontinuous)**, 若对任意$\varepsilon > 0$, 存在$\delta > 0$, 使得

$$|\varphi(x) - \varphi(y)| < \varepsilon$$

对任意的$x, y \in [a, b], |x - y| < \delta, \varphi \in U$都成立.

定理 9.1.2: Ascoli-Arzelà

设$U \subset C[a, b]$. U包含一致收敛子列(即U是序列紧集)的充分必要条件是U一致有界且等度连续.

定理 9.1.3

如果K是连续核, 那么$C[a, b]$上的积分算子(9.3)是紧算子.

证明: 对任意$\varphi \in C[a, b]$满足$\|\varphi\|_\infty \le 1$, 以及对任意$x \in [a, b]$, 我们有

$$|(A\varphi)(x)| \le (b - a) \max_{x,y\in[a,b]} |K(x, y)|.$$

所以集合$U := \{A\varphi : \varphi \in C[a, b], \|\varphi\|_\infty \le 1\} \subset C[a, b]$是一致有界的.

由于K在$[a, b] \times [a, b]$上一致连续, 所以对任意$\varepsilon > 0$, 存在$\delta > 0$, 使得

$$|K(x, z) - K(y, z)| < \frac{\varepsilon}{b - a}, \qquad \forall x, y, z \in [a, b], |x - y| < \delta.$$

于是

$$|(A\varphi)(x) - (A\varphi)(y)| = \left| \int_a^b [K(x, z) - K(y, z)]\varphi(z)\mathrm{d}z \right| < \varepsilon$$

对任意$x, y \in [a, b], |x - y| < \delta, \varphi \in C[a, b], \|\varphi\|_\infty \le 1$都成立. 所以$U$是等度连续的. 根据Ascoli-Arzelà定理, A是紧算子. □

定理 9.1.4

如果K是连续核, 则积分算子$A : C[a,b] \to C[a,b]$的算子范数为

$$\|A\|_\infty = \max_{a\le x\le b}\int_a^b |K(x,y)|\mathrm{d}y.$$

证明: 对任意$\varphi \in C[a,b]$满足$\|\varphi\|_\infty \le 1$, 我们有

$$|(A\varphi)(x)| \le \int_a^b |K(x,y)|\mathrm{d}y, \qquad x \in [a,b].$$

于是

$$\|A\|_\infty = \sup_{\|\varphi\|_\infty\le 1}\|A\varphi\|_\infty \le \max_{a\le x\le b}\int_a^b |K(x,y)|\mathrm{d}y.$$

另一方面, 由于K连续, 则存在$x_0 \in [a,b]$使得

$$\int_a^b |K(x_0,y)|\mathrm{d}y = \max_{a\le x\le b}\int_a^b |K(x,y)|\mathrm{d}y.$$

对$\varepsilon > 0$, 取$\psi \in C[a,b]$为

$$\psi(y) := \frac{K(x_0,y)}{|K(x_0,y)|+\varepsilon}, \qquad y \in [a,b],$$

则$\|\psi\|_\infty \le 1$, 并且

$$\begin{aligned}\|A\psi\|_\infty \ge |(A\psi)(x_0)| &= \int_a^b \frac{[K(x_0,y)]^2}{|K(x_0,y)|+\varepsilon}\mathrm{d}y\\ &\ge \int_a^b \frac{[K(x_0,y)]^2-\varepsilon^2}{|K(x_0,y)|+\varepsilon}\mathrm{d}y\\ &= \int_a^b |K(x_0,y)|\mathrm{d}y - \varepsilon(b-a).\end{aligned}$$

所以

$$\|A\|_\infty = \sup_{\|\varphi\|_\infty\le 1}\|A\varphi\|_\infty \ge \|A\psi\|_\infty \ge \int_a^b |K(x_0,y)|\mathrm{d}y - \varepsilon(b-a).$$

由ε的任意性, 我们有

$$\|A\|_\infty \ge \int_a^b |K(x_0,y)|\mathrm{d}y = \max_{a\le x\le b}\int_a^b |K(x,y)|\mathrm{d}y.$$

这就完成了证明. □

§9.2 算子逼近

求解积分方程

$$\varphi - A\varphi = f$$

的数值解的一个方法是考虑另外一个方程

$$\varphi_n - A_n\varphi_n = f_n,$$

其中$A_n \to A$且$f_n \to f$. 逼近序列对应的方程可以是一个线性方程组, 这样计算就很方便了. 在这一节我们提供这样逼近序列的收敛性和误差分析.

(1)$\{A_n\}$按算子范数收敛到A:

定理 9.2.1

设X是Banach空间, $A: X \to X$是紧的线性算子, $I-A$是单射. 若序列$\{A_n\}$是$X \to X$的有界线性算子, 满足按范数收敛:

$$\|A_n - A\| \to 0, \qquad n \to \infty,$$

那么对充分大的n, 逆算子$(I-A_n)^{-1}: X \to X$存在并且一致有界. 对于下面的方程

$$\varphi - A\varphi = f, \qquad \varphi_n - A_n\varphi_n = f_n,$$

有误差估计式

$$\|\varphi_n - \varphi\| \le C(\|(A_n - A)\varphi\| + \|f_n - f\|).$$

其中C是常数.

证明: 由定理9.1.1, 逆算子$(I-A)^{-1}: X \to X$存在且有界. 由于$\|A_n - A\| \to 0, n \to \infty$, 所以对充分大的$n$, 有

$$\|(I-A)^{-1}(A_n - A)\| \le q < 1.$$

根据Neumann定理,

$$I - (I-A)^{-1}(A_n - A) = (I-A)^{-1}(I - A_n) = \sum_{k=0}^{n}[(I-A)^{-1}(A_n - A)]^k \tag{9.4}$$

的逆存在且一致有界:

$$\|[I - (I-A)^{-1}(A_n - A)]^{-1}\| \le \frac{1}{1-q},$$

根据(9.4)可知$[I - (I-A)^{-1}(A_n - A)]^{-1}(I-A)^{-1}$是$I - A_n$的逆算子, 且一致有界.

误差估计很容易得到, 注意到

$$(I - A_n)(\varphi_n - \varphi) = (A - A_n)\varphi + f_n - f,$$

所以

$$\|\varphi_n - \varphi\| \le \|I - A_n\|^{-1}(\|(A_n - A)\varphi\| + \|f_n - f\|).$$

这就是欲证结论. □

(2)$\{A_n\}$逐点收敛到A, 即$A_n\varphi \to \varphi, n \to \infty$. 为了建立逐点收敛版本的误差估计, 我们需要建立按范数收敛和逐点收敛的关系, 需要用到共鸣定理(一致有界原理):

定理 9.2.2: 共鸣定理

设X是Banach空间, Y是赋范线性空间, $A_n: X \to Y$是有界线性算子$(n = 1, 2, \cdots)$, 并且A_n是逐点有界的, 即对任意$\varphi \in X$, 存在只依赖于φ的常数C_φ使得

$$\|A_n\varphi\| \le C_\varphi, \qquad \forall n \in \mathbb{N}.$$

则序列$\{A_n\}$是一致有界的, 即存在常数C使得

$$\|A_n\| \le C, \qquad \forall n \in \mathbb{N}.$$

定义 9.2.1: 整体紧

设X,Y是赋范线性空间, $A_n : X \to Y$是线性算子. 称算子序列$\{A_n\}$是**整体紧的(collectively compact)**, 若集合

$$\{A_n\varphi : \varphi \in X, \|\varphi\| \le 1, n \in \mathbb{N}\}$$

的每个序列包含收敛子列.

显然, 整体紧算子序列$\{A_n\}$的每个算子A_n都是紧算子.

引理 9.2.3

设X是Banach空间, $\{A_n : X \to X\}$是整体紧算子序列, $\{B_n : X \to X\}$是逐点收敛算子序列, 其极限为$B : X \to X$, 则

$$\|(B_n - B)A_n\| \to 0, \qquad n \to \infty. \tag{9.5}$$

证明: (反证)若(9.5)不成立, 则存在$\varepsilon_0 > 0$与一个数列$\{n_k\}_{\mathbb{N}}$, 满足$n_k \to \infty, k \to \infty$, 并且存在$(\varphi_k) \subset X$满足$\|\varphi_k\| \le 1$, 使得

$$\|(B_{n_k} - B)A_{n_k}\varphi_k\| \ge \varepsilon_0, \qquad k = 1, 2, \cdots. \tag{9.6}$$

由于算子序列$\{A_n\}$是整体紧的, 则存在子列使得

$$A_{n_{k(j)}}\varphi_{k(j)} \to \psi \in X, \qquad j \to \infty. \tag{9.7}$$

利用三角不等式, 我们有

$$\|(B_{n_{k(j)}} - B)A_{n_{k(j)}}\varphi_{k(j)}\| \le \|(B_{n_{k(j)}} - B)\psi\| + \|B_{n_{k(j)}} - B\|\|A_{n_{k(j)}}\varphi_{k(j)} - \psi\|.$$

根据$\{B_n\}$逐点收敛可知上式右端第一项随着$j \to \infty$会趋于0. 根据共鸣定理, $\{B_n\}$一致有界. 而由(9.7)式可知上式第二项也趋于零. 因此这与(9.6)式矛盾. □

定理 9.2.4

设$A : X \to X$是Banach空间X上的紧线性算子, $I - A$是单射, 线性算子序列$\{A_n : X \to X\}$是整体紧的, 并且逐点收敛, 即

$$A_n\varphi \to A\varphi (n \to \infty), \forall \varphi \in X.$$

则对充分大的n, 逆算子$(I - A_n)^{-1} : X \to X$存在且一致有界. 对于下面的方程

$$\varphi - A\varphi = f, \qquad \varphi_n - A_n\varphi_n = f_n,$$

有误差估计式

$$\|\varphi_n - \varphi\| \le C(\|(A_n - A)\varphi\| + \|f_n - f\|). \tag{9.8}$$

其中C是常数.

证明: 由定理9.1.1, 逆算子$(I - A)^{-1} : X \to X$存在且有界, 注意到恒等式

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

我们定义

$$M_n := I + (I - A)^{-1}A_n,$$

用$\{M_n\}$来逼近$I - A_n$的逆. 那么

$$M_n(I - A_n) = I - S_n, \tag{9.9}$$

其中

$$S_n := (I-A)^{-1}(A_n - A)A_n.$$

根据引理9.2.3, $\|S_n\| \to 0, n \to \infty$. 所以对充分大的$n$, 我们有$\|S_n\| \leq q < 1$. 由Neumann定理, $(I-S_n)^{-1}$存在且有界,

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

由于(9.9)式表明$I - A_n$是单的, 并且A_n是紧算子, 则根据定理9.1.1可知$(I-A_n)^{-1}$存在, 并且

$$(I-A_n)^{-1} = (I-S_n)^{-1}M_n,$$

根据共鸣定理, M_n是一致有界的, 再根据定理9.2.1可知欲证估计式成立. □

§ 9.3 Nyström方法

对于积分

$$Q(g) = \int_a^b g(x)\mathrm{d}x,$$

我们考虑收敛的数值积分公式

$$Q_n(g) = \sum_{k=0}^{n} a_k^{(n)} g(x_k^{(n)}),$$

其中数值积分的求积基点$x_0^{(n)}, \cdots, x_n^{(n)} \in [a,b]$, 权重系数$a_0^{(n)}, \cdots, a_n^{(n)}$是实数.

数值积分算子$Q_n : C[a,b] \to \mathbb{R}$是个有界线性泛函, 其范数是

$$\|Q_n\|_\infty = \sum_{k=0}^{n} |a_k|.$$

这是很容易证明的. 一方面, 我们有

$$|Q_n f| \leq \|f\|_\infty \sum_{k=0}^{n} |a_k|,$$

另一方面, 取f为分片线性连续函数, $\|f\|_\infty = 1$, $f(x_k)a_k = |a_k|, k = 0, 1, \cdots, n$(省略上标), 那么$\|Q_n\| \geq \sum_{k=0}^{n} |a_k|$.

下面为方便起见, 我们省略上标. 接下来我们要逼近积分算子

$$(A\varphi)(x) = \int_a^b K(x,y)\varphi(y)\mathrm{d}y, \qquad x \in [a,b],$$

其中K是连续核函数. 逼近的方式是采用下面的数值积分算子

$$(A_n\varphi)(x) := \sum_{k=0}^{n} a_k K(x, x_k)\varphi(x_k), \qquad x \in [a,b].$$

也就是说我们对函数$g(y) = K(x,y)\varphi(y)$用一下数值积分公式, 即

$$(A_n\varphi)(x) = (Q_n g)(x).$$

那么我们的第二类积分方程

$$\varphi - A\varphi = f$$

可以用下面的方程逼近:

$$\varphi_n - A_n\varphi_n = f.$$

这样就变成了一个有限维线性方程组.

定理 9.3.1

设φ_n是下面问题的解:

$$\varphi_n(x)-\sum_{k=0}^{n}a_kK(x,x_k)\varphi_n(x_k)=f(x),\qquad x\in[a,b], \tag{9.10}$$

则在数值积分的求积基点上的函数值$\varphi_j^{(n)}:=\varphi_n(x_j),j=0,1,\cdots,n$满足线性方程组

$$\varphi_j^{(n)}-\sum_{k=0}^{n}a_kK(x_j,x_k)\varphi_k^{(n)}=f(x_j),\qquad j=0,1,\cdots,n. \tag{9.11}$$

反之, 设$\varphi_j^{(n)},j=0,1,\cdots,n$是线性方程组(9.11)的解, 则函数$\varphi_n$定义为

$$\varphi_n(x):=f(x)+\sum_{k=0}^{n}a_kK(x,x_k)\varphi_k^{(n)},\qquad x\in[a,b], \tag{9.12}$$

是方程(9.10)的解.

证明: 第一部分的命题是平凡的. 反过来, 若$\varphi_j^{(n)},j=0,1,\cdots,n$是线性方程组(9.11)的解, 在由(9.12)定义的φ_n在x_j处的函数值是

$$\varphi_n(x_j)=f(x_j)+\sum_{k=0}^{n}a_kK(x_j,x_k)\varphi_k^{(n)}=\varphi_j^{(n)},\qquad j=0,1,\cdots,n.$$

代回到(9.12)可知φ_n满足(9.10). □

公式(9.12)可以看成在插值点$\{\varphi_j^{(n)}\}_{j=0}^n$上的插值函数. 这个公式是Nyström在1930年引入的.

定理 9.3.2

积分算子A_n的范数为

$$\|A_n\|_\infty=\max_{a\le x\le b}\sum_{k=0}^{n}|a_kK(x,x_k)|.$$

证明: 对任意$\varphi\in C[a,b]$满足$\|\varphi\|_\infty<1$, 我们有

$$\|A_n\varphi\|_\infty\le\max_{a\le x\le b}\sum_{k=0}^{n}|a_kK(x,x_k)|,$$

另一方面, 取$z\in[a,b]$使得

$$\sum_{k=0}^{n}|a_kK(z,x_k)|=\max_{a\le x\le b}\sum_{k=0}^{n}|a_kK(x,x_k)|,$$

取$\psi\in C[a,b]$满足$\|\psi\|_\infty=1$并且

$$a_kK(z,x_k)\psi(x_k)=|a_kK(z,x_k)|,\qquad k=0,1,\cdots,n.$$

则

$$\|A_n\|_\infty\ge\|A_n\psi\|_\infty\ge|(A_n\psi)(z)|=\sum_{k=0}^{n}|a_kK(z,x_k)|.$$

于是可得欲证结论. □

接下来我们分析收敛性.

定理 9.3.3

若数值积分公式的算子序列$\{Q_n\}$是收敛的, 其中则算子序列$\{A_n\}$是整体紧的, 并且逐点收敛, 即

$$A_n\varphi \to A\varphi, n \to \infty, \forall \varphi \in C[a,b].$$

但不是按范数收敛.

证明: 由于$\{Q_n\}$是收敛的, 根据$\{Q_n\}$泛函的定义以及共鸣定理, 存在常数C使得

$$\sum_{k=0}^{n} |a_k^{(n)}| \le C, \qquad \forall n \in \mathbb{N}.$$

注意到

$$\|A_n\varphi\|_\infty \le C \max_{x,y\in[a,b]} |K(x,y)| \|\varphi\|_\infty, \forall n \in \mathbb{N},$$

以及

$$|(A_n\varphi)(x_1) - (A_n\varphi)(x_2)| \le C \max_{a\le y\le b} |K(x_1,y) - K(x_2,y)| \|\varphi\|_\infty, \forall x_1, x_2 \in [a,b], \forall n \in \mathbb{N}, \tag{9.13}$$

所以

$$\{A_n\varphi : \varphi \in C[a,b], \|\varphi\|_\infty \le 1, n \in \mathbb{N}\}$$

是一致有界且等度连续的, 根据Ascoli-Arzelà定理, 算子序列$\{A_n\}$是整体紧的.

由于数值积分算子$\{Q_n\}$收敛, 并且我们已经知道A_n和Q_n的关系为

$$(A_n\varphi)(x) = (Q_n g)(x).$$

所以, 固定$\varphi \in C[a,b]$, 序列$\{A_n\varphi\}$是逐点收敛的, 即

$$(A_n\varphi)(x) \to (A\varphi)(x), n \to \infty, \forall x \in [a,b].$$

根据(9.13)式, $\{A_n\varphi\}$是等度连续的, 所以$(A_n\varphi)(x)$一致收敛于$(A\varphi)(x)$, 即

$$\|A_n\varphi - A\varphi\|_\infty \to 0, n \to \infty.$$

(需要给出证明, 参见本节习题1.) 也就是说, 算子序列$\{A_n\}$是逐点收敛, 即

$$A_n\varphi \to A\varphi, n \to \infty, \qquad \forall \varphi \in C[a,b].$$

对于$\varepsilon > 0$, 取函数$\psi_\varepsilon \in C[a,b]$满足$0 \le \psi_\varepsilon(x) \le 1, \forall x \in [a,b]$, 并且$\psi_\varepsilon(x_j) = 0 (j = 0,1,\cdots,n)$, 且当$\min\limits_{j=0,1,\cdots,n} |x - x_j| \ge \varepsilon$时, $\psi_\varepsilon(x) = 1$. 于是, 对任意$\varphi \in C[a,b]$满足$\|\varphi\|_\infty = 1$, 我们有

$$\|A(\varphi\psi_\varepsilon) - A\varphi\|_\infty \le \max_{x,y\in[a,b]} |K(x,y)| \int_a^b (1 - \psi_\varepsilon(y)) \mathrm{d}y \to 0, \qquad \varepsilon \to 0.$$

于是

$$\begin{aligned}\|A-A_n\|_\infty &= \sup_{\|\varphi\|_\infty=1}\|(A-A_n)\varphi\|_\infty\\ &\geq \sup_{\|\varphi\|_\infty=1}\sup_{\varepsilon>0}\|(A-A_n)(\varphi\psi_\varepsilon)\|_\infty\\ &= \sup_{\|\varphi\|_\infty=1}\sup_{\varepsilon>0}\|A(\varphi\psi_\varepsilon)\|_\infty \qquad (\text{因为}\psi_\varepsilon(x_j)=0)\\ &\geq \sup_{\|\varphi\|_\infty=1}\sup_{\varepsilon>0}(\|A\varphi\|_\infty-\|A(\varphi\psi_\varepsilon)-A\varphi\|_\infty)\\ &= \sup_{\|\varphi\|_\infty=1}\|A\varphi\|_\infty=\|A\|_\infty.\end{aligned}$$

所以$\{A_n\}$不是按范数收敛于A. □

注: 根据定理9.3.3, 我们可以利用定理9.2.4的结论. 注意在定理9.2.4的(9.8)式中, 我们需要估计一下$\|A\varphi-A_n\varphi\|_\infty$. 事实上, 我们可以写

$$\|A\varphi-A_n\varphi\|_\infty=\max_{a\leq x\leq b}\left|\int_a^b K(x,y)\varphi(y)\mathrm{d}y-\sum_{k=0}^{n}a_kK(x,x_k)\varphi(x_k)\right|,$$

所以我们需要得到关于函数$g(\cdot)=K(x,\cdot)\varphi(\cdot)$的数值积分的一致误差界. 所以, 如果我们知道核函数K和精确解φ的一些正则性(即光滑性), 那么从数值积分的收敛阶就可以知道求解积分方程的方法的收敛阶.

例如, 对于(复合)梯形公式, 我们假设$\varphi\in C^2[a,b]$, $K\in C^2([a,b]\times[a,b])$, 那么

$$\|A\varphi-A_n\varphi\|_\infty\leq\frac{1}{12}h^2(b-a)\max_{x,y\in[a,b]}\left|\frac{\partial^2}{\partial y^2}[K(x,y)\varphi(y)]\right|.$$

例 9.3.1 *考虑积分方程*

$$\varphi(x)-\frac{1}{2}\int_0^1(x+1)e^{-xy}\varphi(y)\mathrm{d}y=e^{-x}-\frac{1}{2}+\frac{1}{2}e^{-x-1},\qquad 0\leq x\leq 1,$$

它的精确解为$\varphi(x)=e^{-x}$.

注意它的核函数满足

$$\max_{0\leq x\leq 1}\int_0^1\frac{1}{2}(x+1)e^{-xy}\mathrm{d}y=\sup_{0\leq x\leq 1}\frac{x+1}{2x}(1-e^{-x})<1,$$

所以, 根据定理9.1.3, A的算子范数$\|A\|<1$, 所以$(I-A)^{-1}$存在且有界, 从而上述积分方程有唯一解.

用复合的梯形公式的Nyström方法来求解积分方程, 可以得到精确解和近似解在下面几个点处的误差如下表, 容易发现收敛阶是$O(h^2)$:

n	$x=0$	$x=0.25$	$x=0.5$	$x=0.75$	$x=1$
4	0.007146	0.008878	0.010816	0.013007	0.015479
8	0.001788	0.002224	0.002711	0.003261	0.003882
16	0.000447	0.000556	0.000678	0.000816	0.000971
32	0.000112	0.000139	0.000170	0.000204	0.000243

用复合Simpson公式的Nyström方法来求解积分方程, 可以得到精确解和近似解在下面几个点处的误差如下表, 容易发现收敛阶是$O(h^4)$:

n	$x=0$	$x=0.25$	$x=0.5$	$x=0.75$	$x=1$
4	0.00006652	0.00008311	0.00010905	0.00015046	0.00021416
8	0.00000422	0.00000527	0.00000692	0.00000956	0.00001366
16	0.00000026	0.00000033	0.00000043	0.00000060	0.00000086

习题:

1. 给定$[a,b]\to\mathbb{R}$的函数列$\{\varphi_n\}$, 如果$\{\varphi_n\}$是等度连续的, 并且在$[a,b]$上逐点收敛于$\varphi:[a,b]\to\mathbb{R}$, 则$\varphi_n$一致收敛于$\varphi$.
   提示: 利用Cauchy准则, 回顾数学分析.
2. 考虑第二类Volterra积分方程
$$\varphi(x)-\int_a^x K(x,y)\varphi(y)\mathrm{d}y=f(x),\qquad x\in[a,b],$$
   其中核函数K是连续函数, 证明对任意连续的右端函数f, 这个方程都存在唯一解φ.
   提示: 证明齐次方程只有平凡解, 用定理9.1.1.
3. 用逐次逼近法(successive approximations)求解Volterra积分方程:
$$\varphi(x)-\int_0^x e^{x-y}\varphi(y)\mathrm{d}y=f(x).$$
4. 设X,Y是赋范线性空间, $A_n:X\to Y$是紧线性算子, $n=1,2,\cdots$. 证明: $\{A_n\}$是整体紧的当且仅当对任意有界序列$\{\varphi_n\}\subset X$, 序列$\{A_n\varphi_n\}$有收敛子列.
5. 做数值试验验证Gauss求积公式版本的Nyström方法的收敛阶.

9.3.1 部分习题解答

命题 9.3.4

给定$[a,b]\to\mathbb{R}$的函数列$\{\varphi_n\}$, 如果$\{\varphi_n\}$是等度连续的, 并且在$[a,b]$上逐点收敛于$\varphi:[a,b]\to\mathbb{R}$, 则$\varphi_n$一致收敛于$\varphi$.

证明: 对任意$\varepsilon>0$, 由于$\{\varphi_n\}$是等度连续的, 即对上述$\varepsilon>0$, 存在$\delta>0$, 使得当$x_1,x_2\in[a,b]$满足$|x_1-x_2|<\delta$时, 有
$$\|\varphi_n(x_1)-\varphi_n(x_2)\|<\frac{\varepsilon}{4},\forall n\in\mathbb{N}.$$
下面把区间$[a,b]$作$K=\left\lfloor\frac{1}{\delta}\right\rfloor+1$等分, 得到的点为
$$a=x_0<x_1<\cdots<x_K=b,$$
根据逐点收敛性, 对任意$j\in\{0,1,2,\cdots,K\}$, 存在N_j, 当$m,n>N_j$时,
$$|\varphi_n(x_j)-\varphi_m(x_j)|<\frac{\varepsilon}{2}.$$
我们取$N=\max\{N_0,\cdots,N_K\}$, 则当$m,n>N$时, 对任意$x\in[a,b]$, 总存在$j_0\in\{0,1,\cdots,K\}$使得$|x-j_0|<\delta$, 于是
$$\begin{aligned}|\varphi_n(x)-\varphi_m(x)|&\le|\varphi_n(x)-\varphi_n(x_{j_0})|+|\varphi_n(x_{j_0})-\varphi_m(x_{j_0})|+|\varphi_m(x_{j_0})-\varphi_m(x)|\\&\le\frac{\varepsilon}{4}+\frac{\varepsilon}{2}+\frac{\varepsilon}{4}=\varepsilon.\end{aligned}$$
所以$\{\varphi_n\}$一致收敛. □

§9.4 配点法

求解第二类积分方程问题
$$\varphi-A\varphi=f \tag{9.14}$$

的**配点法(collocation method)**是在一个有限维子空间中找近似解, 使得这个近似解只是在有限个点(叫做**配点法(collocation points)**上精确成立(9.14)式.

设$A:C[a,b]\to C[a,b]$是有界线性算子, $X_n=\mathrm{span}\{u_0^{(n)},\cdots,u_n^{(n)}\}\subset C[a,b]$, $\dim X_n=n+1$. 取$n+1$个点

$$a_0\le x_0^{(n)}<\cdots<x_n^{(n)}\le b$$

使得定义在子空间X_n上关于上述插值点的插值函数是唯一可解的. X_n的通常取法包括多项式、三角多项式、样条函数等等. 为了方便起见我们省略上标(n), 即插值点是$x_0,x_1,\cdots,x_n$, X_n的基函数是$u_0,u_1,\cdots,u_n$.

记$L_n:C[a,b]\to X_n$表示从$f\in C[a,b]$对应的插值函数$L_nf\in X_n$, 其性质满足

$$(L_nf)(x_j)=f(x_j),\qquad j=0,\cdots,n,$$

我们把L_n写成Lagrange基函数$l_0,\cdots,l_n\in X_n$的线性组合:

$$L_nf=\sum_{k=0}^{n}f(x_k)l_k,$$

其中Lagrange函数满足

$$l_k(x_j)=\delta_{jk},\qquad j,k=0,\cdots,n.$$

容易证明$L_n:C[a,b]\to X_n$是有界线性算子(范数取为最大模范数). 除此之外, 由于$L_nf=f,\forall f\in X_n$, 所以L_n是个**投影算子(projection operator)**, 即满足$L_n^2=L_n$.

配点法是利用X_n中的函数列$\{\varphi_n\}$来逼近问题(9.14)的解, 并且$\varphi_n\in X_n$满足

$$\varphi_n(x_j)-(A\varphi_n)(x_j)=f(x_j),\qquad j=0,1,\cdots,n. \tag{9.15}$$

我们把φ_n写成基函数的线性组合:

$$\varphi_n=\sum_{k=0}^{n}\gamma_ku_k,$$

这样我们可以写成关于系数$\{\gamma_k\}_{k=0}^n$的线性方程组

$$\sum_{k=0}^{n}\gamma_k\{u_k(x_j)-(Au_k)(x_j)\}=f(x_j),\qquad j=0,1,\cdots,n. \tag{9.16}$$

如果我们用Lagrange基函数, 即$X_n=\mathrm{span}\{l_0,\cdots,l_n\}$, 那么我们可以写

$$\varphi_n=\sum_{k=0}^{n}\gamma_kl_k,$$

此时$\gamma_j=\varphi_n(x_j)(j=0,1,\cdots,n)$, 这样我们可以把线性方程组(9.16)写成

$$\gamma_j-\sum_{k=0}^{n}\gamma_k(Al_k)(x_j)=f(x_j),\qquad j=0,1,\cdots,n. \tag{9.17}$$

注意线性方程组(9.16)和(9.17)并不是完全离散的, 因为$(Au_k)(x_j)$或$(Al_k)(x_j)$也需要作离散处理.

配点法又可以被称为**投影法(projection method)**, 因为插值函数被插值点唯一确定, 所以方程组(9.15)等价于

$$\varphi_n-L_nA\varphi_n=L_nf. \tag{9.18}$$

这个方程可以看作整个空间$C[a,b]$上的方程, 因为任意的解$\varphi_n=L_nA\varphi_n+L_nf$自动属于$X_n$. 于是我们可以把第二类积分方程的收敛性结果用在配点法上.

定理 9.4.1

设$A: C[a,b] \to C[a,b]$是紧线性算子, $I-A$是单射, 插值算子$L_n: C[a,b] \to X_n$满足

$$\|L_nA - A\|_\infty \to 0, \qquad n \to \infty,$$

则对充分大的n, 方程(9.18)对任意$f \in C[a,b]$存在唯一解, 且有误差估计式

$$\|\varphi_n - \varphi\|_\infty \le C\|L_n\varphi - \varphi\|_\infty,$$

其中常数C依赖于A.

证明: 由定理9.2.1, 令$A_n = L_nA$, 则对充分大的n, $(I-L_nA)^{-1}$存在且一致有界. 对(9.14)式作用插值算子L_n, 可得

$$\varphi - L_nA\varphi = L_nf + \varphi - L_n\varphi,$$

代入(9.18), 可得

$$(I-L_nA)(\varphi_n - \varphi) = L_n\varphi - \varphi,$$

于是可以得到欲证的误差估计式. □

推论 9.4.2

设$A: C[a,b] \to C[a,b]$是紧线性算子, $I-A$是单射, 插值算子$L_n: C[a,b] \to X_n$逐点收敛, 即

$$L_n\varphi \to \varphi, \qquad n \to \infty, \forall \varphi \in C[a,b]$$

则对充分大的n, 方程(9.18)对任意$f \in C[a,b]$存在唯一解, 且有误差估计式

$$\|\varphi_n - \varphi\|_\infty \le C\|L_n\varphi - \varphi\|_\infty,$$

证明: 由引理9.2.3, L_n的逐点收敛性和A的紧性可以推出$\|L_nA - A\|_\infty \to 0$. 用前一定理即可. □

注: 配点法除了$C[a,b]$, 也可以用在其他函数空间.

下面我们考虑用配点法求解第二类积分方程

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\mathrm{d}y = f(x), \qquad x \in [a,b]. \tag{9.19}$$

其中K是连续核函数. 利用插值算子, 我们可以把方程(9.17)写成

$$\varphi_n(x) - \int_a^b [L_nK(\cdot,y)](x)\varphi_n(y)\mathrm{d}y = (L_nf)(x), \qquad x \in [a,b]. \tag{9.20}$$

方程组(9.16)(9.17)分别写成

$$\sum_{k=0}^n \gamma_k \left\{ u_k(x_j) - \int_a^b K(x_j,y)u_k(y)\mathrm{d}y \right\} = f(x_j), j = 0,1,\cdots,n. \tag{9.21}$$

以及

$$\gamma_j - \sum_{k=0}^n \gamma_k \int_a^b K(x_j,y)l_k(y)\mathrm{d}y = f(x_j), \qquad j = 0,1,\cdots,n. \tag{9.22}$$

根据子空间X_n以及其基函数$u_0,\cdots,u_n$与配点$x_0,\cdots,x_n$的不同, 我们能得到不同种类的配点法. 我们只考虑两种配点法: 分别基于线性样条函数与三角多项式.

在这里由于篇幅限制, 我们只考虑线性样条插值. 令$x_j = a + jh$, $j = 0,1,\cdots,n$, 步长$h = \dfrac{b-a}{n}$, 子空

间$X_n \subset C[a,b]$满足在每个区间$[x_{j-1},x_j], j=1,\cdots,n$上都是线性函数. 我们定义Lagrange基函数为

$$l_k(x) := \begin{cases} \dfrac{1}{h}(x-x_{k-1}), & x \in [x_{k-1},x_k], \\ \dfrac{1}{h}(x_{k+1}-x), & x \in [x_k,x_{k+1}], \\ 0, & x \notin [x_{k-1},x_{k+1}]. \end{cases}$$

其中$k=0,1,\cdots,n$. 对于分片线性函数插值, 我们有

$$\|L_n f\|_\infty \le \max_{j=0,1,\cdots,n} |f(x_j)| \le \|f\|_\infty,$$

等号成立条件为f是常数, 于是$\|L_n\|_\infty = 1$.

可以证明, $L_n\varphi \to \varphi (n\to\infty)$, 所以利用推论9.4.2, 可得下面的结论:

定理 9.4.3

对于带有连续核函数的第二类积分方程, 线性样条插值的配点法收敛.

如果积分方程的精确解二次连续可微, 那么根据线性插值误差定理与推论9.4.2, 我们有

$$\|\varphi_n - \varphi\|_\infty \le C\|\varphi''\|_\infty h^2,$$

其中φ_n是线性样条配点法的逼近解, C只依赖于K.

注意在(9.22)式中, 我们需要用数值积分来计算$\int_a^b K(x_j,y)l_k(y)\mathrm{d}y$. 我们把$K(x_j,\cdot)$用它的分片线性插值函数来代替, 即

$$K(x_j,y) \approx (L_nK(x_j,\cdot))(y) = \sum_{i=0}^{n} K(x_j,x_i)l_i(y),$$

那么

$$\int_a^b K(x_j,y)l_k(y)\mathrm{d}y \approx \sum_{i=0}^{n} K(x_j,x_i)\int_a^b l_i(y)l_k(y)\mathrm{d}y.$$

其中$j,k=0,1,\cdots,n$. 令$w_{ik} = \int_a^b l_i(y)l_k(y)\mathrm{d}y$, 我们可以直接计算得到三对角阵$W=(w_{ik})_{i,k}$如下:

$$W = \frac{h}{6}\begin{pmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 2 \end{pmatrix}$$

因为计算积分也会带来误差, 所以我们也要作相应的误差分析. 我们把(9.22)用近似的数值积分代替原来的系数以后, 得到的解记为$\tilde{\varphi}_n$, 对应的方程为

$$\tilde{\varphi}_n - A_n\tilde{\varphi}_n = L_nf, \tag{9.23}$$

即配点方程

$$\tilde{\varphi}_n(x) - \int_a^b [L_nK_n(\cdot,y)](x)\tilde{\varphi}_n(y)\mathrm{d}y = (L_nf)(x), \qquad a \le x \le b.$$

其中,

$$K_n(x,y) := [L_nK_n(x,\cdot)](y) = \sum_{i=0}^{n} K(x,x_i)l_i(y).$$

假设核函数K是二次连续可微的, 那么我们有误差估计式

$$|K(x,y) - K_n(x,y)| \le \frac{h^2}{8}\left\|\frac{\partial^2 K}{\partial y^2}\right\|_\infty, \qquad \forall a \le x,y \le b.$$

我们写

$$K_n(x,y) - [L_nK_n(\cdot,y)](x) = L_n\{K(x,\cdot) - [L_nK_n(\cdot,\cdot)](x)\}(y),$$

然后根据$\|L_n\|_\infty = 1$, 我们有

$$|K_n(x,y) - [L_nK_n(\cdot,y)](x)| \le \frac{h^2}{8}\left\|\frac{\partial^2 K}{\partial x^2}\right\|_\infty, \qquad \forall a \le x,y \le b.$$

对于积分算子A_n, 对应的核函数为K_n, 我们有

$$\|A_n - A\|_\infty = O(h^2).$$

若f是二次连续可微的, 那么我们有

$$\|L_nf - f\|_\infty = O(h^2).$$

根据定理9.2.1, 方程(9.23)对充分大的n存在唯一解, 并且有误差估计

$$\|\tilde{\varphi}_n - \varphi\|_\infty = O(h^2).$$

所以总的离散误差也是$O(h^2)$阶.

例 9.4.1 考虑例9.3.1的问题, 下表给出了用线性样条的配点法逼近的误差, 误差阶是$O(h^2)$.

n	$x=0$	$x=0.25$	$x=0.5$	$x=0.75$	$x=1$
4	0.004808	0.005430	0.006178	0.007128	0.008331
8	0.001199	0.001354	0.001541	0.001778	0.002078
16	0.000300	0.000338	0.000385	0.000444	0.000519
32	0.000075	0.000085	0.000096	0.000111	0.000130

注: 如果用三次样条插值, 那么误差阶可以达到$O(h^4)$, 但是计算量会大大增加. 这也说明了Nyström方法更加实用(因为我们只需要改变求积公式就可以提升误差阶了).

习题:

1. 把线性样条插值改成三次样条插值, 并写代码实现. 随着n增大, 比较一下Simpson积分版本的Nyström方法和三次样条插值版本的配点法的运行时间.
2. 翻阅[R. Kress, Numerical Analysis, GTM181], 自行整理三角多项式版本的配点法, 并尝试给出实现程序.

§9.5 稳定性

由于积分方程$\varphi - A\varphi = f$的数值解法包括了许多算子以及线性方程组, 我们需要分析矩阵或者算子的条件数. 根据定理9.2.1和定理9.2.4, 条件数cond$(I - A_n)$是一致有界的, 所以我们只需要考虑逼近的方程$\varphi_n - A_n\varphi_n = f_n$涉及到的线性方程组的条件数.

(1)Nyström方法. 对应的方程组(9.11)中, 记$\tilde{A}_n$的各个分量为$a_kK(x_j,x_k)$, 引入算子$R_n : C[a,b] \to \mathbb{R}^{n+1}$为

$$R_n : f \mapsto (f(x_0),\cdots,f(x_n))^T, \qquad f \in C[a,b].$$

以及$M_n : \mathbb{R}^{n+1} \to C[a,b]$, 其中$M_n\Phi$是分片线性插值函数, 满足

$$(M_n\Phi)(x_j) = \Phi_j, \qquad j = 0,\cdots,n, \Phi = (\Phi_0,\cdots,\Phi_n)^T.$$

(若$a < x_0$, 记$(M_n\Phi)(x) = \Phi_0, a \le x \le x_0$; 若$x_n < b$, 记$(M_n\Phi)(x) = \Phi_n, x_n \le x \le b$.) 显然$\|R_n\|_\infty = \|M_n\|_\infty = 1$.

根据定理9.3.1, 我们有

$$(I - \tilde{A}_n) = R_n(I - A_n)M_n,$$

以及

$$(I - \tilde{A}_n)^{-1} = R_n(I - A_n)^{-1}M_n.$$

于是有如下定理:

定理 9.5.1

Nyström方法中线性方程组的矩阵的条件数一致有界.

因此, Nyström方法可以保持原来的积分方程的稳定性.

(2)配点法. 记E_n的各个分量为$u_k(x_j)$, $\tilde{A}_n$的各个分量为$(Au_k)(x_j)$. 由于$X_n = \mathrm{span}\{u_0,\cdots,u_n\}$对应于配点$x_0,\cdots,x_n$的插值问题存在唯一解, 所以$E_n$是可逆的. 记算子$W_n : \mathbb{R}^{n+1} \to C[a,b]$为

$$W_n : \gamma \mapsto \sum_{k=0}^{n} \gamma_k u_k,$$

其中$\gamma = (\gamma_0,\cdots,\gamma_n)^T$. 再记$R_n, M_n$如前, 则

$$W_n = L_nM_nE_n,$$

根据(9.16), 我们有

$$(E_n - \tilde{A}_n) = R_nL_n(I - A)W_n,$$

以及

$$(E_n - \tilde{A}_n)^{-1} = E_n^{-1}R_n(I - L_nA)^{-1}L_nM_n,$$

结合共鸣定理与定理9.4.1可知算子$(I - L_nA)^{-1}L_n$一致有界, 所以有如下结论:

定理 9.5.2

在定理9.4.1的条件下, 配点法的线性方程组的条件数满足

$$\mathrm{cond}(E_n - \tilde{A}_n) \le C\|L_n\|_\infty^2 \mathrm{cond}(E_n).$$

其中n充分大, C是常数.

定理标明基函数需要仔细选取. 如果我们取单项式(如$u_n = x^n$), 那么E_n的条件数增长很快. 而如果我们取Lagrange基函数, 即线性方程组(9.17), 那么E_n变成单位矩阵, 条件数是1. 另外, $\|L_n\|$也需要考虑进来. 对于线性样条插值, $\|L_n\| = 1$; 对于三角多项式插值, $\|L_n\| \to \infty (n \to \infty)$.

§9.6 用深度神经网络求解积分方程

*待补充.

CHAPTER 10

(*)最优控制

*待补充.

§10.1 最优控制简介

§10.2 Pontryagin极大值条件

§10.3 Hamilton-Jacobi-Bellman方程

§10.4 动态规划原理

§10.5 数值方法