



西安交通大学研究生教育系列教材

数值分析

李乃成 梅立泉 编著



科学出版社

西安交通大学研究生教育系列教材

数值分析

李乃成 梅立泉 编著

科学出版社

北 京

内 容 简 介

本书介绍了科学与工程计算中常用的数值计算方法及相关理论. 内容包括解线性方程组的直接法和迭代法、插值法、函数最优逼近、数值微积分、非线性方程(组)的迭代解法、矩阵特征值和特征向量的计算、常微分与偏微分方程数值解法等. 其中包含了一些在实际中有重要应用的新方法, 如求解超定方程组的最小二乘法、求解线性方程组的基于伽辽金原理的迭代法、奇异值分解、广义特征值问题的求解方法等. 同时, 对数值计算方法的计算效率、稳定性、收敛性、误差估计、适用范围及优缺点也进行了分析和介绍.

本书可作为高等院校数学系各专业本科生和各类工科专业研究生的教材或教学参考书, 也可供从事科学与工程计算的科研工作者阅读参考.

图书在版编目(CIP)数据

数值分析/李乃成, 梅立泉编著. —北京: 科学出版社, 2011

(西安交通大学研究生教育系列教材)

ISBN 978-7-03-032192-3

I. ①数… II. ①李… ②梅… III. ①数值分析—研究生—教材 IV. ①O241

中国版本图书馆 CIP 数据核字(2011) 第 174452 号

责任编辑: 赵彦超 徐园园 房 阳 / 责任校对: 何艳萍

责任印制: 钱玉芬 / 封面设计: 王 浩

科学出版社 出版

北京东黄城根北街 16 号

邮政编码: 100717

<http://www.sciencep.com>

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2011 年 9 月第 一 版 开本: B5(720 × 1000)

2011 年 9 月第一次印刷 印张: 21 1/2

印数: 1—2 500 字数: 421 000

定价: 65.00 元

(如有印装质量问题, 我社负责调换)

前 言

在现代科学研究与工程设计中, 计算机已成为不可或缺的有力工具. 科学计算与理论、实验三足鼎立, 成为科学实践的三大手段. 科教兴国, 要求现代理工科大学学生掌握计算机常用的数值计算方法, 具有使用计算机解决实际问题的能力. “数值计算方法”已成为各高等学校本科生、研究生的必修课.

本书是为理工科大学各专业研究生 72 学时“数值计算方法”课程编写的教材. 以介绍常用、实用的数值计算方法为主, 在介绍经典计算方法的基础上, 本着推陈出新、与时俱进的精神, 增加了一些在实际中有着重要应用的、近年来发展的新方法. 如超定方程组的求解方法、求解线性方程组的基于伽辽金原理的迭代法、奇异值分解、广义特征值问题的求解方法等.

本书重点突出、层次分明、推导详细、清晰易懂, 内容由浅入深、循序渐进, 便于自学与教学, 可供从事科学与工程计算的科技工作者参考, 也适宜于相关专业人员自学.

承蒙邓建中教授在百忙中审阅本书, 并提出了不少宝贵的意见和建议. 本书的出版得到西安交通大学研究生院的支持、鼓励和资助. 科学出版社的有关同志对本书的出版给予了极大的帮助. 编者在此向他们表示衷心的感谢.

由于编者水平所限, 书中不妥之处在所难免, 恳请读者批评指正.

编 者

2010 年 10 月

目 录

前言

第 1 章 绪论	1
1.1 数值分析研究的内容与特点	1
1.2 误差	2
1.2.1 误差的来源与分类	2
1.2.2 绝对误差、相对误差与准确数字	2
1.2.3 计算机中数的表示与舍入误差	4
1.2.4 数据误差影响的估计	6
1.3 算法的数值稳定性	8
小结	9
习题	10
第 2 章 解线性方程组的直接法	12
2.1 高斯消去法	13
2.1.1 高斯消去法	13
2.1.2 高斯消去法中乘除法的运算量	18
2.1.3 高斯消去法顺利进行的条件	18
2.1.4 高斯消去法的算法组织	19
2.1.5 列主元高斯消去法	20
2.2 矩阵的三角分解	22
2.2.1 高斯消去法的矩阵形式	22
2.2.2 矩阵的 LU 分解	25
2.2.3 平方根法和改进平方根法	30
2.2.4 求解三对角方程组的追赶法	35
2.3 舍入误差对解的影响	37
2.3.1 向量范数与矩阵范数	37
2.3.2 舍入误差对解的影响	44
2.4 正交变换与矩阵的 QR 分解	49
2.4.1 吉文斯变换与豪斯霍尔德变换	49
2.4.2 矩阵的 QR 分解	52
*2.5 超定方程组	65

2.5.1 线性最小二乘问题	65
2.5.2 最小二乘问题的求解	67
小结	71
习题	72
计算实习	75
第 3 章 解线性方程组的迭代法	77
3.1 向量序列和矩阵序列的极限	77
3.2 解线性方程组的基本迭代法	78
3.2.1 迭代法的一般格式	78
3.2.2 三种基本迭代法	78
3.3 迭代法的收敛性	83
3.3.1 迭代法的矩阵表示	83
3.3.2 迭代法的收敛性	84
3.4 共轭梯度法	92
3.4.1 求解线性方程组与求解二次函数极小点的等价性	92
3.4.2 共轭梯度法	93
*3.5 基于伽辽金原理的迭代法	100
3.5.1 伽辽金原理和克雷洛夫子空间	100
3.5.2 阿诺尔迪过程	101
3.5.3 阿诺尔迪算法	103
3.5.4 广义极小残余算法	106
小结	110
习题	111
计算实习	113
第 4 章 插值法	115
4.1 多项式插值问题	115
4.2 拉格朗日插值多项式	118
4.3 牛顿插值多项式	120
4.3.1 差商的定义	121
4.3.2 牛顿插值多项式	121
4.3.3 差商的性质	124
4.4 埃尔米特插值多项式	125
4.5 分段低次插值多项式	129
4.5.1 高次插值多项式的缺陷	129
4.5.2 分段低次插值法	130

4.6 三次样条插值函数	132
4.6.1 三次样条插值函数的定义	132
4.6.2 三次样条插值函数的导出	132
4.6.3 三次样条插值函数的收敛性与误差估计	138
小结	138
习题	139
计算实习	141
第 5 章 函数最优逼近	142
5.1 函数的内积、范数和正交多项式	142
5.1.1 函数的内积和范数	142
5.1.2 正交多项式	144
5.2 最优平方逼近	151
5.2.1 最优平方逼近	151
5.2.2 正规方程组	152
5.3 最优一致逼近	162
5.3.1 最优一致逼近多项式	162
5.3.2 近似最优一致逼近多项式	165
小结	173
习题	174
计算实习	176
第 6 章 数值积分与数值微分	177
6.1 牛顿-科茨求积公式	177
6.1.1 数值积分的基本思想	177
6.1.2 牛顿-科茨求积公式	178
6.1.3 复化求积公式	180
6.1.4 变步长积分法	183
6.1.5 龙贝格积分法	184
6.2 待定系数法与高斯型求积公式	187
6.2.1 代数精度与待定系数法	187
6.2.2 广义佩亚诺定理	189
6.2.3 高斯型求积公式	191
6.2.4 常用的 4 种高斯型求积公式	197
6.3 数值积分的稳定性	201
6.4 数值微分	201
6.4.1 插值型数值微分公式	202

6.4.2 待定系数法	204
6.4.3 外推求导法	205
6.4.4 利用三次样条插值函数求导法	208
小结	208
习题	209
计算实习	210
第 7 章 非线性方程 (组) 的迭代解法	212
7.1 求解非线性方程的迭代法	212
7.1.1 几种基本迭代法	212
7.1.2 迭代法的收敛性	219
7.1.3 迭代法的收敛速度	224
7.1.4 加速收敛技术	226
7.2 求解非线性代数方程组的迭代法	228
7.2.1 简单迭代法	229
7.2.2 牛顿法	231
7.2.3 弦割法	234
7.2.4 布洛依登法	236
小结	237
习题	238
计算实习	239
第 8 章 矩阵特征值与特征向量的计算	241
8.1 基本性质	241
8.2 求一般矩阵特征值的计算方法	242
8.2.1 乘幂法及反幂法	242
8.2.2 求矩阵全部特征值与特征向量的 QR 方法	245
8.2.3 阿诺尔迪方法	251
8.3 求实对称矩阵特征值的计算方法	253
8.3.1 雅可比方法	253
8.3.2 吉文斯方法	256
8.3.3 兰乔斯方法	258
8.4 奇异值 (SVD) 的计算	259
8.5 广义特征值问题	261
8.5.1 广义 Schur 分解	261
8.5.2 对称正定矩阵的广义 Schur 分解	262
小结	262

习题	263
计算实习	263
第 9 章 常微分方程数值解法	265
9.1 初值问题常用数值解法的建立与使用	265
9.1.1 基本数值解法的建立与隐式法的求解	265
9.1.2 龙格-库塔法	273
9.1.3 待定系数法、预测—校正公式	278
9.2 数值解中误差的积累、数值方法的收敛性和绝对稳定性	282
9.2.1 数值解中误差的积累和数值方法的收敛性	282
9.2.2 绝对稳定性	286
9.3 一阶微分方程组与高阶方程的数值解法	289
9.3.1 一阶微分方程组	289
9.3.2 高阶常微分方程	291
9.4 边值问题的数值解法	293
9.4.1 有限差分法	293
9.4.2 打靶法	300
小结	302
习题	303
计算实习	304
第 10 章 偏微分方程的数值解法	305
10.1 椭圆型边值问题	305
10.1.1 差分方程的建立	305
10.1.2 差分解的误差估计与收敛性	307
10.1.3 一般二阶椭圆型方程边值问题	310
10.2 抛物型方程初、边值问题	310
10.2.1 差分方程的建立与求解	311
10.2.2 差分格式的稳定性	313
10.2.3 差分解的误差估计与收敛性	315
10.3 双曲型方程混合问题	316
10.3.1 一阶双曲型方程	316
10.3.2 一阶常系数双曲型方程组	317
10.3.3 二阶双曲型方程	318
10.4 有限元法	320
10.4.1 变分原理	320
10.4.2 伽辽金逼近解	323

10.4.3 单元及形状函数	324
10.4.4 有限元求解步骤	327
小结	329
习题	329
计算实习	332
参考文献	333

第1章 绪 论

本章简要介绍数值分析研究的内容与特点、误差的来源与分类、数据误差的影响、计算机浮点数集和数值算法的稳定性等数值分析的基本概念。

1.1 数值分析研究的内容与特点

随着科学技术的发展和计算机的广泛应用, 解决科学研究和工程技术中的问题通常采用定量分析方法, 其过程分为两个阶段. 第一阶段, 利用有关科学知识和数学理论把实际问题提炼为数学问题 (即建立数学模型). 这一阶段属于应用数学范畴. 所提炼的数学问题一般比较复杂, 无法求出准确解, 只好采用数值计算方法求其近似解. 第二阶段, 针对数学问题提出用计算机求解的数值计算方法, 编写算法程序, 上机求解并分析计算结果. 这一阶段的工作属于计算数学范畴. 数值分析是计算数学的一个主要组成部分, 它研究用计算机求解各类数学问题的数值计算方法及其相关理论. 其内容包括线性方程组的求解、插值法、函数最优逼近法、数值积分与微分、非线性方程 (组) 的求解、矩阵特征值问题、常微分方程与偏微分方程的数值解法等.

可用于计算机求解的可行且有效的数值计算方法应具备以下特点:

(1) 可执行性. 设计算法要面向计算机, 计算机只能作加、减、乘、除和逻辑运算, 设计的算法中只能包含加、减、乘、除四则运算和逻辑运算.

(2) 理论可靠性. 算法要有可靠的理论分析, 即设计的算法要保证其收敛性、稳定性, 并给出其误差分析.

(3) 计算复杂性. 算法要有良好的计算复杂性, 计算复杂性包括时间复杂性 (算法的运算次数, 运算次数决定计算机的计算时间) 和空间复杂性 (占用计算机存储空间的大小).

最后还需通过数值试验验证算法的有效性.

由于数值算法的复杂性, 有些算法虽然在理论上不够严密, 但通过大量的数值试验和对比分析, 证明是行之有效的方法, 在实际中有着广泛的应用, 人们在应用中逐步完善它的理论基础.

鉴于数值计算方法的特点, 学习“数值分析”课程, 首先应掌握设计算法的思想和原理, 重视误差分析、收敛性、稳定性等理论基础; 其次要重视实践, 需要做

一些习题和在计算机上的数值试验, 掌握数值计算方法的计算过程, 加深对算法的理解.

1.2 误 差

1.2.1 误差的来源与分类

科学与工程计算中的数通常是近似数, 近似值与真正值之差称为误差. 误差的来源或分类有 4 种. 从实际问题提炼数学问题时, 由于实际问题比较复杂, 为使问题简化, 往往忽略了许多次要因素, 所提炼的数学问题是对实际问题的近似描述. 即使数学问题能求出准确解, 也与实际问题的真正解不同, 二者之差称为模型误差. 数学问题中含有若干参量, 如质量、温度、电压等, 它们的值是通过观测得到的, 难免存在着误差, 这种误差称为观测误差、数据误差或参量误差. 数学问题一般都比较复杂, 难以求其准确解. 通常用近似公式代替准确公式使之容易求解, 由此求出问题的近似解. 准确解与近似解之差称为截断误差或方法误差. 例如, 为计算函数值 $f(x)$, 取它的 n 次泰勒多项式

$$p_n(x) = f(0) + f'(0)x + \frac{1}{2!}f''(0)x^2 + \cdots + \frac{1}{n!}f^{(n)}(0)x^n$$

近似代替 $f(x)$, 即取 $f(x) \approx p_n(x)$, 则截断误差

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}x^n,$$

其中 ξ 介于 0 与 x 之间.

求解数学问题时, 由于计算机的数位有限, 一般对数要进行舍入, 由此产生的误差称为舍入误差或计算误差.

“数值分析”课程只研究截断误差和舍入误差的估计、在计算过程中的传播及控制和对计算结果的影响.

1.2.2 绝对误差、相对误差与准确数字

定义 1.2.1 设 x 为准确值, $\tilde{x}$ 是 x 的一个近似值, 则

$$\Delta x = x - \tilde{x} \quad \text{或} \quad |\Delta x| = |x - \tilde{x}|$$

称为近似值 $\tilde{x}$ 的绝对误差.

通常, 准确值是不知道的, 因而绝对误差也是不知道的, 但可以估计出绝对误差的一个上界 ε , 即 $|\Delta x| \leq \varepsilon$. ε 称为绝对误差界或绝对误差限. 绝对误差或绝对误差界简称为误差.

虽然不知道准确值, 但由绝对误差界可知准确值所在的范围, 即 $\tilde{x} - \varepsilon \leq x \leq \tilde{x} + \varepsilon$, 并记

$$x = \tilde{x} \pm \varepsilon.$$

例如, 用千分尺测量工件的直径 d 时, 只能测得工件直径的近似值 $\tilde{d}$, 此时绝对误差

$$|\Delta d| = |d - \tilde{d}| \leq \frac{1}{2} \times 10^{-2} \text{ mm}.$$

所测工件的直径的范围为

$$\tilde{d} - \frac{1}{2} \times 10^{-2} \leq d \leq \tilde{d} + \frac{1}{2} \times 10^{-2}.$$

对于绝对值较大或较小的近似值, 绝对误差不足以刻画近似数的准确程度, 于是引进相对误差, 用相对误差的大小来衡量其准确程度.

定义 1.2.2 设 x 为准确值, $\tilde{x}$ 是 x 的一个近似值, 则

$$\delta x = \frac{x - \tilde{x}}{x} \quad \text{或} \quad |\delta x|$$

称为近似值 $\tilde{x}$ 的相对误差.

在实际计算中, 准确值 x 是不知道的, 但由于

$$\frac{x - \tilde{x}}{\tilde{x}} - \frac{x - \tilde{x}}{x} = \Delta x \frac{x - \tilde{x}}{x\tilde{x}} = \frac{(\Delta x)^2}{\tilde{x}(\tilde{x} + \Delta x)} = \frac{(\Delta x/\tilde{x})^2}{1 + \Delta x/\tilde{x}},$$

两者之差是 $\Delta x/\tilde{x}$ 的平方级, 当 $|\Delta x|/|\tilde{x}|$ 较小时, 可以忽略不计. 这时, 相对误差可取为

$$|\delta x| = \frac{|x - \tilde{x}|}{|\tilde{x}|}.$$

若 $|\delta x| \leq \varepsilon_r$, 则 ε_r 称为相对误差界或相对误差限, 常用百分数表示, 并记为

$$x = \tilde{x}(1 \pm \varepsilon_r).$$

由上述定义可知, 绝对误差和绝对误差界是有量纲量, 而相对误差和相对误差界是无量纲量.

近似数 $\tilde{x}$ 的准确数字与误差界有关. 设

$$x = \pm x_1 x_2 \cdots x_m . x_{m+1} x_{m+2} \cdots x_{m+n} x_{m+n+1} \cdots,$$

若 x 的近似值 $\tilde{x}$ 取到小数后第 n 位, 则

$$\tilde{x} = \pm x_1 x_2 \cdots x_m . x_{m+1} \cdots \tilde{x}_{m+n},$$

其中最后一位

$$\tilde{x}_{m+n} = \begin{cases} x_{m+n}, & x_{m+n+1} \leq 4, \\ x_{m+n} + 1, & x_{m+n+1} \geq 5. \end{cases}$$

这时, $|x - \tilde{x}| \leq \underbrace{0.00 \cdots 0}_n 5 = 0.5 \times 10^{-n} = \frac{1}{2} \times 10^{-n}$, 则称近似值 $\tilde{x}$ 准确到 n 位小

数, 并将从该位起直到最左端的非零数字之间的一切数字称为近似值 $\tilde{x}$ 的准确数字 (有效数字), 它的多少反映了 $\tilde{x}$ 近似 x 的准确程度.

有了准确数字的概念后, 数 1.23 和 1.230 是有区别的. 前者准确到两位小数, 后者准确到三位小数, 所以不能随意给数的末尾添零.

又如, $\pi = 3.14159265 \cdots$, 若取 π 的近似值分别为

$$\pi_1 = 3.14, \quad \pi_2 = 3.142,$$

则

$$|\pi - \pi_1| = 0.00159265 \cdots < 0.005 = 0.5 \times 10^{-2} = \frac{1}{2} \times 10^{-2},$$

$$|\pi - \pi_2| = 0.000407346 \cdots < 0.0005 = 0.5 \times 10^{-3} = \frac{1}{2} \times 10^{-3}.$$

由此可知近似值 π_1 准确到两位小数, 具有三位准确数字; 近似值 π_2 准确到三位小数, 具有四位准确数字.

若把 x 写成标浮点形式

$$x = \pm 10^m \times 0.x_1x_2 \cdots, \quad x_1 \neq 0,$$

当 x 的近似值 $\tilde{x}$ 的绝对误差界为 $\frac{1}{2} \times 10^{-n}$ 时, $\tilde{x}$ 准确到 n 位小数, 具有 $n+m$ 位准确数字.

若已知 $\tilde{x}$ 的相对误差界为 $\frac{1}{2} \times 10^{-t}$, 即

$$\frac{|x - \tilde{x}|}{|\tilde{x}|} \leq \frac{1}{2} \times 10^{-t},$$

则 $\tilde{x}$ 的绝对误差为

$$|x - \tilde{x}| \leq |\tilde{x}| \times \frac{1}{2} \times 10^{-t} = 10^m \times 0.x_1x_2 \cdots \times \frac{1}{2} \times 10^{-t} \leq \frac{1}{2} \times 10^{-(t-m)}.$$

由此可知, $\tilde{x}$ 至少具有 $t-m+m=t$ 位准确数字.

1.2.3 计算机中数的表示与舍入误差

计算机中的实数采用浮点表示法, 即将一个数分为指数和尾数两部分来表示. 设计算机采用 β 进制 (二进制、八进制、十六进制或十进制), 字长为 t 位. 按舍入原则, 将非零实数 x 在计算机上表示为

$$fl(x) = \tilde{x} = \pm \left\{ \frac{x_1}{\beta} + \frac{x_2}{\beta^2} + \cdots + \frac{x_t}{\beta^t} \right\} \times \beta^l = \pm 0.x_1x_2 \cdots x_t \times \beta^l,$$

其中, $x_1 \in \{1, 2, \cdots, \beta - 1\}$, $x_i \in \{0, 1, \cdots, \beta - 1\}$ ($i = 2, 3, \cdots, t$).

$0.x_1x_2 \cdots x_t$ 称为尾数部分, β^l 称为指数部分. 指数 l 是整数, 也称为阶码. 阶码 l 的取值范围为 $L \leq l \leq U$ ($L < 0, U > 0$). L, U 分别称为指数的下、上界. $fl(x)$ 称为规格化浮点数. 计算机所能表示的全部浮点数的集合称为计算机的浮点数集, 记为

$$F(\beta, t, L, U) = \{0\} \cup \{fl(x) = \pm 0.x_1x_2 \cdots x_t \times \beta^l\}.$$

显然, 浮点数集 $F(\beta, t, L, U)$ 中共有 $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$ 个数, 它所能表示的数的范围为

$$fl_{\min}(x) = \beta^{L-1} \leq |fl(x)| \leq \beta^U \times (1 - \beta^{-t}) = fl_{\max}(x).$$

由此可知, 当 $|x| > fl_{\max}(x)$ 时, 计算机无法表示 (这种情况称为“上溢”), 计算机中断运行; 当 $x \neq 0$ 但 $|x| < fl_{\min}(x)$ 时, 计算机令 $fl(x) = 0$ (这种情况称为“下溢”), 这时计算机虽然继续运算, 但可能会得出难以预料的结果. 因此, 在计算过程中应尽量调整计算顺序, 避免出现“上溢”、“下溢”的情况. 由此还可知, 计算机字长有限, 一个实数 x 一般只能用最接近的浮点数 $fl(x)$ 代替它, 如

$$x = \pm 0.x_1x_2 \cdots x_tx_{t+1} \cdots \times \beta^l, \quad fl(x) = \pm 0.x_1x_2 \cdots \tilde{x}_t \times \beta^l,$$

$fl(x)$ 尾数部分最末位可能有半个单位的误差, 由此将产生舍入误差. 浮点数 $fl(x)$ 的绝对误差与相对误差分别为

$$|x - fl(x)| \leq \frac{1}{2}\beta^{-t} \times \beta^l = \frac{1}{2}\beta^{l-t},$$

$$\frac{|x - fl(x)|}{|x|} \leq \frac{\frac{1}{2}\beta^{l-t}}{\beta^{l-1}} = \frac{1}{2}\beta^{-(t-1)}.$$

相对误差限 $\frac{1}{2}\beta^{-(t-1)}$ 只与计算机数的进制和字长有关, 称为计算机的相对精度.

在计算机的每步运算中, 得到的数都可能有舍入误差, 如两个数加、减运算, 计算机先对阶, 使得阶小的数与阶大的数的阶码相同, 然后尾数相加减, 最后对计算结果规格化. 例如, 在浮点数集 $F(10, 4, -10, 10)$ 中, 两个数 $a = 0.1234 \times 10^3$ 与 $b = 0.4567 \times 10^{-1}$ 相加,

$$\begin{aligned} a + b &= 0.1234 \times 10^3 + 0.4567 \times 10^{-1} = 0.1234 \times 10^3 + 0.00004567 \times 10^3 \\ &= 0.1234 \times 10^3 + 0.0000 \times 10^3 = 0.1234 \times 10^3 = a, \end{aligned}$$

得到了 $a + b$ ($b \neq 0$) $= a$ 的结果, 即大数 a “吃掉” 了小数 b .

一般地, 为避免大数 “吃掉” 小数, 若个数相加时, 宜采取绝对值较小的数先加的原则.

为了减少舍入误差和节省计算时间, 在计算中应尽量简化计算步骤, 减少运算次数. 下面以两个例子予以说明.

例如, 计算多项式 $p_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ 的值, 若先计算各项 a_kx^k , 再逐项相加, 则需进行 $1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$ 次乘法和 n 次加法; 若用秦九韶算法^①, 将 $p_n(x)$ 改写为

$$p_n(x) = (\cdots((a_nx + a_{n-1})x + a_{n-2})x + \cdots + a_1)x + a_0,$$

则只需 n 次乘法和 n 次加法.

又如, 计算 ABx , 其中, A, B 为 n 阶矩阵, x 为 n 维向量. 若按 $(AB)x$ 的次序计算, 则算出 AB 的 n^2 个数各需 n 次乘法和 $n-1$ 次加法; 再算出 $(AB)x$ 的 n 个数各需 n 次乘法和 $n-1$ 次加法, 因此, 共需 $n^2 \cdot n + n \cdot n = n^3 + n^2$ 次乘法和 $n^2 \cdot (n-1) + n \cdot (n-1) = n^3 - n$ 次加法. 同理, 若按 $A(Bx)$ 的次序计算, 则只需 $2n^2$ 次乘法和 $2n^2 - 2n$ 次加法.

1.2.4 数据误差影响的估计

前面已述数学问题中的参数 $x_1, x_2, \cdots, x_n$ 存在着数据误差, 设其近似值为 $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n$. 数学问题的解与这些参数有关, 记为

$$y = \phi(x_1, x_2, \cdots, x_n).$$

设其近似解为

$$\tilde{y} = \phi(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n).$$

由泰勒展开式知, 近似解 $\tilde{y}$ 的绝对误差和相对误差分别为

$$\begin{aligned} \Delta y &= y - \tilde{y} = \phi(x_1, x_2, \cdots, x_n) - \phi(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n) \\ &\approx \sum_{i=1}^n \frac{\partial \phi(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n)}{\partial x_i} (x_i - \tilde{x}_i) = \sum_{i=1}^n \frac{\partial \phi(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n)}{\partial x_i} \Delta x_i, \\ \delta y &= \frac{\Delta y}{y} \approx \sum_{i=1}^n \frac{\partial \phi(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n)}{\partial x_i} \frac{x_i}{y} \frac{\Delta x_i}{x_i} = \sum_{i=1}^n \frac{\partial \phi(\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n)}{\partial x_i} \frac{x_i}{y} \delta x_i. \end{aligned}$$

^① 秦九韶算法也称为霍纳 (Horner) 算法, 秦九韶于 1247 年提出此法, 比霍纳 1819 年提出此算法早 500 多年.

以上两式中系数的绝对值 $\left| \frac{\partial \phi}{\partial x_i} \right|$ 和 $\left| \frac{\partial \phi}{\partial x_i} \frac{x_i}{y} \right|$ 是数学问题的近似解 $\tilde{y}$ 的绝对误差和相对误差关于参数 x_i 绝对误差和相对误差的放大或缩小倍数. 若其值较大, 则 $\tilde{x}_i$ 的小误差就会导致近似解 $\tilde{y}$ 的大误差. $\left| \frac{\partial \phi}{\partial x_i} \right|$ 和 $\left| \frac{\partial \phi}{\partial x_i} \frac{x_i}{y} \right|$ 称为数学问题的条件数, 其值很大的问题称为病态问题. 换句话说, 病态问题就是数据发生微小的变化将引起其解发生剧烈变化的问题. 这是数学问题本身的固有特征. 对病态问题必须采取相应的处理措施和特殊的计算方法, 以减少误差影响.

注意到当 $y = \phi(x_1, x_2, \dots, x_n) \approx 0$ 时, 相对误差的条件数 $\left| \frac{\partial \phi}{\partial x_i} \frac{x_i}{y} \right|$ 很大. 因此, 凡是计算结果接近于零的问题往往是病态问题. 例如, 相近的两个数相减, 因差接近于零, 就是病态问题. 因此, 在设计算法时, 应尽量避免两个相近的数相减. 常用的做法是变换算式. 这是因为数学上等价的公式在数值计算时并不总是等效的. 例如, 当 $|x|$ 的绝对值充分大时,

$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}, \quad \frac{1}{x} - \frac{1}{x+1} = \frac{1}{x(x+1)},$$

$$\ln\left(x + \frac{1}{x}\right) - \ln\left(x - \frac{1}{x}\right) = \ln\left(\frac{x^2+1}{x^2-1}\right).$$

当 $|x|$ 的绝对值充分小时,

$$1 - \cos x = 2 \sin^2 \frac{x}{2},$$

$$\sin x - x = -\frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

又如, 二次方程 $ax^2 + bx + c = 0$ 的求根公式为

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

当 $b^2 \gg 4|ac|$ 时, $b \approx \sqrt{b^2 - 4ac}$, 在求根公式中总有一式存在相近数相减的情况, 使得计算结果不可靠. 为避免两个相近数相减, 在计算机上求解宜取

$$x_1 = \frac{-b - \operatorname{sgn}(b)\sqrt{b^2 - 4ac}}{2a},$$

其中 $\operatorname{sgn}(x)$ 为符号函数, 即

$$\operatorname{sgn}(x) = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$$

由韦达定理取 $x_2 = \frac{c}{ax_1}$.

由绝对误差估计式可知, 两数乘、除运算的绝对误差分别为

$$\Delta(x_1 x_2) \approx x_2 \Delta x_1 + x_1 \Delta x_2,$$

$$\Delta\left(\frac{x_1}{x_2}\right) \approx \frac{\Delta x_1}{x_2} - \frac{x_1}{x_2^2} \Delta x_2.$$

可见, 在前一式中, 当 $|x_1|$ 或 $|x_2|$ 较大时, 在后式中当 $|x_2|$ 很小时, 计算结果的绝对误差可能很大. 因此, 在实际计算时, 应尽量避免绝对值很大的数作乘数、绝对值较小的数作除数.

1.3 算法的数值稳定性

用计算机求解数学问题时, 参数存在数据误差, 在计算过程中的每一步又存在舍入误差, 这些误差积累起来可能对计算结果会产生很大的影响, 导致计算结果不可靠. 人们用算法的“数值稳定性”来评价一个算法舍入误差对计算结果的影响.

定义 1.3.1 如果一个算法在运算过程中舍入误差能得到控制, 或者舍入误差的积累不影响产生可靠的计算结果, 则称该算法是数值稳定的; 否则, 称其是数值不稳定的.

研究算法的数值稳定性, 理应考虑计算过程中每一步计算所产生的舍入误差, 由于计算过程的复杂性, 这是难以做到的. 为简单起见, 往往只考虑某一步计算产生的舍入误差的影响. 例如, 运算开始时舍入误差的影响. 这实质上是把算法的稳定性研究简化为初始舍入误差对算法影响的分析. 显然, 若一步的误差逐步增大, 则多步的误差更加增大; 反之, 若一步的误差逐步减小, 则多步误差也势必逐步减小. 因此, 研究初始一步误差传播的影响得出的结论与全程误差影响的结果应是一致的, 这对于实际误差分析具有指导意义.

例如, 计算积分

$$I_k = e^{-1} \int_0^1 x^k e^x dx, \quad k = 0, 1, \dots, 7.$$

利用分部积分法可得计算积分 I_k 的递推关系

$$I_k = 1 - k I_{k-1}, \quad k = 1, 2, \dots, 7.$$

算法 1

$$I_0 = e^{-1} \int_0^1 e^x dx = 1 - \frac{1}{e},$$

$$I_k = 1 - k I_{k-1}, \quad k = 1, 2, \dots, 7.$$

算法 2 若用某种方法 (如数值积分法) 算得 I_7 的近似值 $\tilde{I}_7$, 则

$$I_{k-1} = \frac{1 - I_k}{k}, \quad k = 7, 6, \dots, 1.$$

现在分析这两种算法的数值稳定性.

对于算法 1, 只能得到 I_0 的近似值 $\tilde{I}_0$. 现在只考虑这一步产生误差的影响, 假定以后各步计算无误差. 记 $\tilde{I}_k$ 为 I_k 的近似值, 则

$$I_k = 1 - kI_{k-1}, \quad k = 1, 2, \dots, 7,$$

$$\tilde{I}_k = 1 - k\tilde{I}_{k-1}, \quad k = 1, 2, \dots, 7.$$

以上两式相减得

$$I_k - \tilde{I}_k = (-k)(I_{k-1} - \tilde{I}_{k-1}), \quad k = 1, 2, \dots, 7.$$

由此递推可得

$$|I_7 - \tilde{I}_7| = 7!|I_0 - \tilde{I}_0| = 5040|I_0 - \tilde{I}_0|.$$

这表明计算 I_7 产生的误差是第一步计算 I_0 产生的误差的 5040 倍. 由此可知, 算法 1 是数值不稳定的方法, 不可用.

对于算法 2, 同理可得

$$|I_0 - \tilde{I}_0| = \frac{|I_7 - \tilde{I}_7|}{5040}.$$

这表明计算 I_0 产生的误差是计算 I_7 产生的误差的 5040 分之一, 所以算法 2 是数值稳定的方法. 两种积分方法计算结果见表 1.1.

表 1.1 两种积分方法计算结果的比较

I_k	I_0	I_1	I_2	I_3	I_4	I_5	I_6	I_7
准确值	0.632 1	0.367 9	0.264 2	0.207 3	0.170 9	0.145 5	0.126 8	0.112 4
算法 1	0.632 1	0.367 9	0.264 2	0.207 4	0.170 4	0.148 0	0.112 0	0.216 0
算法 2	0.632 1	0.367 9	0.264 2	0.207 3	0.170 9	0.145 5	0.126 8	0.112 4

注意到积分 $I_k = e^{-1} \int_0^1 x^k e^x dx > e^{-1} \int_0^1 x^{k+1} e^x dx = I_{k+1}$, 即序列 I_k 单调减. 而由算法 1 计算的 $I_7 = 0.2160 > I_6 = 0.1120$, 显然, 计算结果是错误的. 此例说明, 不稳定的算法不能用, 实际计算时应采用数值稳定性好的算法.

小 结

本章简要介绍了数值分析研究的内容与特点、误差的来源与分类、计算机的浮点数表示、数据误差影响的估计、病态问题及算法的数值稳定性等数值分析的基本概念.

在计算中应注意采用以下原则:

- (1) 在设计算法时尽量避免两个相近的数相减;
- (2) 合理安排数量级相差很大的数之间的运算次序, 防止大数“吃”小数; 若个数相加, 宜绝对值较小者先相加;
- (3) 避免绝对值很小的数作除数、绝对值很大的数作乘数;
- (4) 简化计算步骤, 尽量减少乘除法的运算次数;
- (5) 选用数值稳定性好的算法.

习 题

1.1 已知 $\sqrt{2} = 1.414\,213\,562\,373\cdots$, 分别写出准确到 3~5 位小数的近似值, 指出它们的绝对误差界、相对误差界以及有效数字的位数.

1.2 求下列近似数的绝对误差界、相对误差界以及有效数字的位数:

- (1) 3 450; (2) 0.005 67; (3) 0.2340×10^5 ; (4) $4\,567 \times 10^{-3}$; (5) 123.45×10^1 .

1.3 在浮点数集 $F(10, 4, -5, 5)$ 中将下列实数写成规格化浮点数:

- (1) 303.35; (2) -0.004 97; (3) 21 346; (4) 1.344 5;
 (5) $48\frac{1}{3}$; (6) $48\frac{3}{5}$; (7) $\frac{2}{3\,000}$; (8) $0.009\,9 \times 10^{-4}$.

1.4 证明浮点数集 $F(\beta, t, L, U)$ 中共有 $2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$ 个数.

1.5 (1) 正方形的边长约为 10cm, 测量边长的误差界多大时, 才能保证面积的误差不超过 0.1cm^2 ?

(2) 计算球体允许相对误差限为 1%, 测量球体半径 r 时允许的相对误差限是多少?

1.6 设 $|x| \ll 1$, 如何计算下列公式, 使得到的结果比较准确:

(1) $\frac{1}{1+2x} - \frac{1-x}{1+x}$; (2) $\frac{\sqrt{1+x^2} - \sqrt{1-x^2}}{\sqrt{x}}$;

(3) $\frac{1 - \cos(2x)}{x}$; (4) $\ln \frac{1 - \sqrt{1-x^2}}{|x|}$.

1.7 证明

(1) 设函数 $y = f(x)$ 可导, $f(x^*) = 0$, $\tilde{x}$ 是 x^* 的近似值, 则按相对误差计算函数值 $f(\tilde{x})$ 是病态问题;

(2) 设 x^* 是方程 $p_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = 0$ 的根, 则当 x^* 接近重根 (即 $p'_n(x^*) \approx 0$) 时, 按绝对误差求根 x^* 的问题是病态问题.

1.8 化简或改写下列算式, 以减少运算次数:

(1) $(x-5)^4 + 9(x-5)^3 + 7(x-5)^2 + 6(x-5) + 4$;

(2) $1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$;

(3) $\frac{1}{1 \times 3} + \frac{1}{2 \times 4} + \frac{1}{3 \times 5} + \cdots + \frac{1}{99 \times 101}$.

1.9 在计算机上怎样计算 $y = 29^{71}/71!$, 才能避免溢出.

1.10 已知积分 $I_k = \int_0^1 \frac{x^k}{x+6} dx$ 具有如下递推关系:

$$I_k = \int_0^1 \frac{x+6-6}{x+6} x^{k-1} dx = \int_0^1 x^{k-1} dx - 6I_{k-1} = \frac{1}{k} - 6I_{k-1}.$$

取 4 位小数, 用以下两种递推方法计算积分 $I_0, I_1, \dots, I_7$ 的近似值.

算法 1 令 $I_0 = 0.1542$ ($\approx \ln \frac{7}{6}$), 计算

$$I_k = \frac{1}{k} - 6I_{k-1}, \quad k = 1, 2, \dots, 7.$$

算法 2 令 $I_{11} = 0$, 计算

$$I_{k-1} = \frac{1 - k I_k}{6k}, \quad k = 11, 10, \dots, 1.$$

哪种算法计算准确, 原因为何?

第2章 解线性方程组的直接法

在科学研究和工程技术中,许多问题都可以归结为求解如下的线性方程组:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \dots\dots\dots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{cases} \quad (2.0.1)$$

方程组 (2.0.1) 常写成矩阵形式, 即

$$Ax = b, \quad (2.0.1')$$

其中

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

$$x = (x_1, x_2, \cdots, x_n)^T, \quad b = (b_1, b_2, \cdots, b_n)^T.$$

由线性代数知, 若系数矩阵的行列式 $|A| \neq 0$, 则方程组的解存在且唯一. 本章在线性方程组的解存在且唯一的条件下, 介绍各种求解方法.

根据求解线性方程组的克拉默 (Cramer) 法则, 方程组 (2.0.1) 的解为

$$x_i = \frac{|A_i|}{|A|}, \quad i = 1, 2, \cdots, n,$$

其中 A_i ($i = 1, 2, \cdots, n$) 为把系数矩阵 A 的第 i 列用方程组的右端项 b 替换后所得到的矩阵. 用克拉默方法求解 n 阶线性方程组, 乘除法的运算量为 $(n+1)!(n-1)+n$. 当 $n = 20$ 时, 此数约为 9.7×10^{20} . 用每秒钟作 10^8 次乘除法运算的计算机求解这个方程组, 大约需要 30.8 万年. 这说明, 即使求解阶数较小的方程组, 该法也是不能使用的. 鉴于这个原因, 人们提出了许多求解线性方程组的数值解法, 而且至今仍在不断地探索研究高效实用的新算法.

本章介绍求解线性方程组 (2.0.1) 的直接法. 所谓直接法, 即在无舍入误差的假定下, 经过有限次运算就可以求得方程组的准确解.

实际上, 由于计算过程中舍入误差的存在和影响, 直接法也只能求得方程组的近似解.

2.1 高斯消去法

2.1.1 高斯消去法

高斯消去法, 其本质就是初等数学中的消元法, 只不过是求解步骤规范化, 便于编写计算机程序而已.

高斯消去法首先是将方程组 (2.0.1) 进行消元运算, 将其化为一个等价同解的上三角方程组, 这个过程称为消元过程. 然后通过求解上三角方程组得到原方程组的解, 后一过程称为回代过程.

下面以求解一个三阶方程组为例来说明高斯消去法的思想.

例 2.1.1 用高斯消去法求解线性方程组

$$\begin{cases} x_1 + 2x_2 + x_3 = 0, \\ 2x_1 + 2x_2 + 3x_3 = 3, \\ -x_1 - 3x_2 = 2. \end{cases}$$

解 消元过程:

第 1 步 保留第 1 个方程不动, 将第 2, 3 个方程中 x_1 前的系数消为零. 为此, 第 2, 3 个方程分别减去第 1 个方程的 2, -1 倍得

$$\begin{cases} x_1 + 2x_2 + x_3 = 0, \\ -2x_2 + x_3 = 3, \\ -x_2 + x_3 = 2. \end{cases}$$

第 2 步 保留第 1, 2 个方程不动, 将第 3 个方程中 x_2 前的系数消为零. 第 3 个方程减去第 2 个方程的 $\frac{1}{2}$ 倍, 则得与原方程同解的上三角方程组

$$\begin{cases} x_1 + 2x_2 + x_3 = 0, \\ -2x_2 + x_3 = 3, \\ \frac{1}{2}x_3 = \frac{1}{2}. \end{cases}$$

回代过程:

求解以上上三角方程组. 从第 3 个方程组求得 $x_3 = 1$, 代入第 2 个方程可得 $x_2 = -1$, 再将 x_2, x_3 代入第 1 个方程组得 $x_1 = 1$. 解毕.

下面介绍求解 n 阶线性方程组 (2.0.1) 的高斯消去法.

为了书写方便起见, 用符号 “ $\triangleq$ ” 表示 “记为”,

$$A^{(0)} \triangleq A, \quad a_{ij}^{(0)} \triangleq a_{ij}, \quad b^{(0)} \triangleq b, \quad b_i^{(0)} \triangleq b_i.$$

1. 消元过程

第 1 步 假定 $a_{11}^{(0)} \neq 0$ (若 $a_{11}^{(0)} = 0$, 则由 $|A| \neq 0$ 知, 矩阵 A 的第 1 列元素不全为零, 交换方程位置, 使新的 $a_{11}^{(0)} \neq 0$). 第 1 个方程保留不动, 将第 $2 \sim n$ 个方程中 x_1 的系数消为零, 即第 i 个方程减去第 1 个方程乘以 $\frac{a_{i1}^{(0)}}{a_{11}^{(0)}} (i = 2, 3, \dots, n)$ 得

$$\left\{ \begin{array}{l} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + a_{13}^{(0)} x_3 + \cdots + a_{1n}^{(0)} x_n = b_1^{(0)}, \\ a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 + \cdots + a_{2n}^{(1)} x_n = b_2^{(1)}, \\ a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 + \cdots + a_{3n}^{(1)} x_n = b_3^{(1)}, \\ \dots\dots\dots \\ a_{n2}^{(1)} x_2 + a_{n3}^{(1)} x_3 + \cdots + a_{nn}^{(1)} x_n = b_n^{(1)}, \end{array} \right. \quad (2.1.1)$$

其中

$$\begin{aligned} a_{ij}^{(1)} &= a_{ij}^{(0)} - \frac{a_{i1}^{(0)}}{a_{11}^{(0)}} a_{1j}^{(0)} = a_{ij}^{(0)} - l_{i1} a_{1j}^{(0)}, \quad i = 2, 3, \dots, n; j = 2, 3, \dots, n, \\ b_i^{(1)} &= b_i^{(0)} - \frac{a_{i1}^{(0)}}{a_{11}^{(0)}} b_1^{(0)} = b_i^{(0)} - l_{i1} b_1^{(0)}, \quad i = 2, 3, \dots, n, \\ l_{i1} &\triangleq \frac{a_{i1}^{(0)}}{a_{11}^{(0)}}, \quad i = 2, 3, \dots, n. \end{aligned}$$

将方程组 (2.1.1) 写为矩阵形式, 即

$$A^{(1)} x = b^{(1)},$$

其中

$$A^{(1)} = \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3n}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}, \quad b^{(1)} = \begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ b_3^{(1)} \\ \vdots \\ b_n^{(1)} \end{pmatrix}.$$

第 2 步 假定 $a_{22}^{(1)} \neq 0$, 保留方程组 (2.1.1) 中第 1, 2 个方程不动, 将第 $3 \sim n$ 个方程中 x_2 的系数消为零, 即第 i 个方程减去第 2 个方程乘以 $\frac{a_{i2}^{(1)}}{a_{22}^{(1)}} (i = 3, 4, \dots, n)$ 得

$$\left\{ \begin{array}{l} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + a_{13}^{(0)} x_3 + \cdots + a_{1n}^{(0)} x_n = b_1^{(0)}, \\ a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 + \cdots + a_{2n}^{(1)} x_n = b_2^{(1)}, \\ a_{33}^{(2)} x_3 + \cdots + a_{3n}^{(2)} x_n = b_3^{(2)}, \\ \dots\dots\dots \\ a_{n3}^{(2)} x_3 + \cdots + a_{nn}^{(2)} x_n = b_n^{(2)}, \end{array} \right. \quad (2.1.2)$$

其中

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} a_{2j}^{(1)} = a_{ij}^{(1)} - l_{i2} a_{2j}^{(1)}, \quad i = 3, 4, \dots, n; j = 3, 4, \dots, n,$$

$$b_i^{(2)} = b_i^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} b_2^{(1)} = b_i^{(1)} - l_{i2} b_2^{(1)}, \quad i = 3, 4, \dots, n,$$

$$l_{i2} \triangleq \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = 3, 4, \dots, n.$$

方程组 (2.1.2) 用矩阵表示为

$$A^{(2)} x = b^{(2)},$$

其中

$$A^{(2)} = \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ b_3^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix}.$$

重复以上过程, 进行 $k-1$ 次消元后可得

$$\left\{ \begin{array}{l} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + \cdots + a_{1k}^{(0)} x_k + \cdots + a_{1n}^{(0)} x_n = b_1^{(0)}, \\ a_{22}^{(1)} x_2 + \cdots + a_{2k}^{(1)} x_k + \cdots + a_{2n}^{(1)} x_n = b_2^{(1)}, \\ \dots\dots\dots \\ a_{kk}^{(k-1)} x_k + \cdots + a_{kn}^{(k-1)} x_n = b_k^{(k-1)}, \\ a_{k+1,k}^{(k-1)} x_k + \cdots + a_{k+1,n}^{(k-1)} x_n = b_{k+1}^{(k-1)}, \\ \dots\dots\dots \\ a_{nk}^{(k-1)} x_k + \cdots + a_{nn}^{(k-1)} x_n = b_n^{(k-1)}. \end{array} \right. \quad (2.1.3)$$

方程组 (2.1.3) 用矩阵表示为

$$A^{(k-1)} x = b^{(k-1)},$$

其中

$$A^{(k-1)} = \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1k}^{(0)} & \cdots & a_{1n}^{(0)} \\ & a_{22}^{(1)} & \cdots & a_{2k}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \ddots & \vdots & & \vdots \\ & & & a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ & & & a_{k+1,k}^{(k-1)} & \cdots & a_{k+1,n}^{(k-1)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{pmatrix}, \quad b^{(k-1)} = \begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ \vdots \\ b_k^{(k-1)} \\ b_{k+1}^{(k-1)} \\ \vdots \\ b_n^{(k-1)} \end{pmatrix}.$$

第 k 步 假定 $a_{kk}^{(k-1)} \neq 0$, 保留第 $1 \sim k$ 个方程不动, 将第 $k+1 \sim n$ 个方程中 x_k 的系数消为零, 即第 i 个方程减去第 k 个方程乘以 $\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} (i = k+1, k+2, \dots, n)$ 得

$$\left\{ \begin{array}{l} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + \cdots + a_{1k}^{(0)} x_k + a_{1,k+1}^{(0)} x_{k+1} + \cdots + a_{1n}^{(0)} x_n = b_1^{(0)}, \\ a_{22}^{(1)} x_2 + \cdots + a_{2k}^{(1)} x_k + a_{2,k+1}^{(1)} x_{k+1} + \cdots + a_{2n}^{(1)} x_n = b_2^{(1)}, \\ \dots\dots\dots \\ a_{kk}^{(k-1)} x_k + a_{k,k+1}^{(k-1)} x_{k+1} + \cdots + a_{kn}^{(k-1)} x_n = b_k^{(k-1)}, \\ a_{k+1,k+1}^{(k)} x_{k+1} + \cdots + a_{k+1,n}^{(k)} x_n = b_{k+1}^{(k)}, \\ \dots\dots\dots \\ a_{n,k+1}^{(k)} x_{k+1} + \cdots + a_{nn}^{(k)} x_n = b_n^{(k)}, \end{array} \right.$$

其中

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)} = a_{ij}^{(k-1)} - l_{ik} a_{kj}^{(k-1)},$$

$$i = k+1, k+2, \dots, n; j = k+1, k+2, \dots, n,$$

$$b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)} = b_i^{(k-1)} - l_{ik} b_k^{(k-1)}, \quad i = k+1, k+2, \dots, n,$$

$$l_{ik} \triangleq \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad i = k+1, k+2, \dots, n.$$

同理, 进行 $n-1$ 次消元后, 原方程组 (2.0.1) 化为如下等价同解上三角方程组:

$$\left\{ \begin{array}{l} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + \cdots + a_{1k}^{(0)} x_k + a_{1,k+1}^{(0)} x_{k+1} + \cdots + a_{1n}^{(0)} x_n = b_1^{(0)}, \\ a_{22}^{(1)} x_2 + \cdots + a_{2k}^{(1)} x_k + a_{2,k+1}^{(1)} x_{k+1} + \cdots + a_{2n}^{(1)} x_n = b_2^{(1)}, \\ \dots\dots\dots \\ a_{kk}^{(k-1)} x_k + a_{k,k+1}^{(k-1)} x_{k+1} + \cdots + a_{kn}^{(k-1)} x_n = b_k^{(k-1)}, \\ a_{k+1,k+1}^{(k)} x_{k+1} + \cdots + a_{k+1,n}^{(k)} x_n = b_{k+1}^{(k)}, \\ \dots\dots\dots \\ a_{nn}^{(n-1)} x_n = b_n^{(n-1)}. \end{array} \right. \quad (2.1.4)$$

用矩阵表示为

$$A^{(n-1)} x = b^{(n-1)},$$

其中

$$A^{(n-1)} = \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1k}^{(0)} & a_{1,k+1}^{(0)} & \cdots & a_{1n}^{(0)} \\ & a_{22}^{(1)} & \cdots & a_{2k}^{(1)} & a_{2,k+1}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \ddots & \vdots & \vdots & & \vdots \\ & & & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ & & & & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ & & & & & \ddots & \vdots \\ & & & & & & a_{nn}^{(n-1)} \end{pmatrix}, \quad b^{(n-1)} = \begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ \vdots \\ b_k^{(k-1)} \\ b_{k+1}^{(k)} \\ \vdots \\ b_n^{(n-1)} \end{pmatrix}.$$

2. 回代过程

求解上三角方程组 (2.1.4).

假定 $a_{nn}^{(n-1)} \neq 0$, 则从 (2.1.4) 的最后一个方程解出 $x_n = b_n^{(n-1)} / a_{nn}^{(n-1)}$. 将 x_n 代入倒数第 2 个方程, 则可求得 x_{n-1} , 同理依次求得 x_k . 回代过程的计算公式如下:

$$\begin{cases} x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}, \\ b_k^{(k-1)} - \sum_{j=k+1}^n a_{kj}^{(k-1)} x_j \\ x_k = \frac{\quad}{a_{kk}^{(k-1)}}, \quad k = n-1, n-2, \dots, 2, 1. \end{cases} \quad \text{)} \quad (2.1.1)$$

2.1.2 高斯消去法中乘除法的运算量

在消去过程中, 计算 $l_{ik} = a_{ik}^{(k-1)} / a_{kk}^{(k-1)} (i = k+1, k+2, \dots, n)$ 共有 $n-k$ 个除法, $a_{ij}^{(k)} = a_{ij}^{(k-1)} - l_{ik} a_{kj}^{(k-1)} (i = k+1, k+2, \dots, n; j = k+1, k+2, \dots, n)$ 共有 $(n-k)^2$ 个乘法, $b_i^{(k)} = b_i^{(k-1)} - l_{ik} b_k^{(k-1)} (i = k+1, k+2, \dots, n)$ 共有 $n-k$ 个乘法. 消元过程共进行 $n-1$ 步, 即 $k = 1, 2, \dots, n-1$.

于是消元过程中乘除法的运算量为

$$N_1 = \sum_{k=1}^{n-1} [(n-k)^2 + 2(n-k)] = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6},$$

回代过程乘除法的运算量为

$$N_2 = \sum_{k=1}^{n-1} (n-k) + n = \frac{n^2}{2} + \frac{n}{2},$$

所以用高斯消去法解 n 阶线性方程组 (2.0.1) 的乘法运算量为

$$N = N_1 + N_2 = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} + \frac{n^2}{2} + \frac{n}{2} = \frac{1}{3}n^3 + n^2 - \frac{1}{3}n.$$

2.1.3 高斯消去法顺利进行的条件

从消元过程知, 当且仅当 $a_{kk}^{(k-1)} \neq 0 (k = 1, 2, \dots, n)$ 时, 高斯消去法才能进行下去, 下面的定理给出了高斯消去法顺利进行的条件.

定理 2.1.1 $a_{kk}^{(k-1)} \neq 0 (k = 1, 2, \dots, n)$ 的充分必要条件是线性方程组的系数矩阵 A 的各阶顺序主子式不等于零, 即

$$a_{kk}^{(k-1)} \neq 0, \quad k = 1, 2, \dots, n$$

$$\Leftrightarrow D_1 = a_{11} \neq 0, \quad D_k = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix} \neq 0, \quad k = 2, \dots, n.$$

证明 略.

定义 2.1.1 若矩阵 A 的元素满足

$$\sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| < |a_{ii}|, \quad i = 1, 2, \dots, n,$$

则称矩阵 A 为严格对角占优矩阵.

由定理 2.1.1 和线性代数理论知, 高斯消去法顺利进行的几个条件分别为

条件 1 系数矩阵 A 的各阶顺序主子式均不等于零.

条件 2 系数矩阵 A 是对称正定矩阵.

条件 3 系数矩阵 A 是严格对角占优矩阵.

2.1.4 高斯消去法的算法组织

为了便于高斯消去法在计算机上的实现以及减少计算机的存储量, 在开始计算时, 将方程组的系数矩阵存放在二维数组 A 中, 方程的右端项存放在一维数组 b 中. 第 1 步消元后产生的元素 $a_{ij}^{(1)}$ 存放在元素 $a_{ij}^{(0)}$ 的对应位置上, $b_i^{(1)}$ 放在 $b_i^{(0)}$ 的对应位置上, 这是因为在以后的运算中 $a_{ij}^{(0)}$, $b_i^{(0)}$ 不再有用. 矩阵 A 的第 1 列 $a_{11}^{(0)}$ 以下的 $n-1$ 个元素消为零, 没有必要将这些零存入, 这些位置上存放数据 l_{i1} ($i = 2, 3, \dots, n$). 对于以后的各步消元, 同理存放相应数据. 消元过程结束后, 二维数组 A 和一维数组 b 的内容如下:

$$A = \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1,n-1}^{(0)} & a_{1n}^{(0)} \\ l_{21} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2,n-1}^{(1)} & a_{2n}^{(1)} \\ l_{31} & l_{32} & a_{33}^{(2)} & \cdots & a_{3,n-1}^{(2)} & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{n,n-1} & a_{nn}^{(n-1)} \end{pmatrix}, \quad b = \begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ b_3^{(2)} \\ \vdots \\ b_n^{(n-1)} \end{pmatrix}.$$

高斯消去法的计算过程如下:

```

for k = 1 to n - 1 do
  if  $|a_{kk}| \leq \varepsilon$ , then stop
  for i = k + 1 to n do
     $a_{ik} := a_{ik} / a_{kk}$ 
    for j = k + 1 to n do
       $a_{ij} := a_{ij} - a_{ik} a_{kj}$ 
    end do
     $b_i := b_i - a_{ik} b_k$ 
  end do
end do

```

```

end do
 $x_n := b_n / a_{nn}$ 
for  $k = n - 1$  to 1 step  $-1$  do
     $s := b_k$ 
    for  $j = k + 1$  to  $n$  do
         $s := s - a_{kj}x_j$ 
    end do
     $x_k := s / a_{kk}$ 
end do

```

2.1.5 列主元高斯消去法

在高斯消去法的消元过程中要求主元 $a_{kk}^{(k-1)} \neq 0$ ($k = 1, 2, \dots, n$), 但方程组的条件 $|A| \neq 0$ 不能保证 $a_{kk}^{(k-1)} \neq 0$ ($k = 1, 2, \dots, n$). 即使 $a_{kk}^{(k-1)} \neq 0$ ($k = 1, 2, \dots, n$), 但 $|a_{kk}^{(k-1)}|$ 很小, 这就会引起其他元素数量级的剧增和舍入误差的增长, 导致计算结果不可靠, 甚至计算不能进行下去.

为了避免出现小主元, 在高斯消去法中引进所谓的选主元技术, 构成主元素 (即绝对值最大的元素) 高斯消去法. 由于是按列选主元, 故称为列主元高斯消去法.

列主元高斯消去法在消去的第 1 步, 从矩阵 A 的第 1 列选取绝对值最大的元 a_{p1} , 交换第 p 个方程与第 1 个方程的位置, 交换后的方程组 a_{11} 的绝对值最大, 然后进行消元的第 1 步.

一般地, 在第 k 步消元之前, 从 $A^{(k-1)}$ 的第 k 列元素 $a_{kk}^{(k-1)}, \dots, a_{nk}^{(k-1)}$ 中选取绝对值最大的元 $a_{pk}^{(k-1)}$, 交换第 p 个方程与第 k 个方程的位置, 使新的 $a_{kk}^{(k-1)}$ 的绝对值最大, 然后进行第 k 步消元. 同理, 进行 $n - 1$ 步消元后, 将原方程组化为一个等价同解上三角方程组.

由于列主元高斯消去法中 $|a_{ik}^{(k-1)}| / |a_{kk}^{(k-1)}| \leq 1$ ($i = k + 1, k + 2, \dots, n$), 有利于控制误差的传播, 故具有较好的数值稳定性.

在高斯消去法的计算过程中每一步消元之前, 增加按列选主元素和交换方程位置两点, 便构成了列主元素高斯消去法的计算过程.

增加的按列选主元素和交换方程位置的计算过程如下:

选主元素 $\left\{ \begin{array}{l} \max := 0.0 \\ \text{for } i = k \text{ to } n \text{ do} \\ \quad \text{if } |a_{ik}| \leq |\max| \text{ then go to 标号 10} \\ \quad \text{else } \max := a_{ik} \\ \quad q := i \\ \text{标号 10 end do} \end{array} \right.$

if $|\max| \leq \varepsilon$, then stop

if $q = k$ then go to 消元部分

交换方程位置 $\left\{ \begin{array}{l} \text{for } j = k \text{ to } n \text{ do} \\ \quad c := a_{kj} \\ \quad a_{kj} := a_{qj} \\ \quad a_{qj} := c \\ \text{end do} \\ c := b_k \\ b_k := b_q \\ b_q := c \end{array} \right.$

例 2.1.2 用列主元素高斯消去法求解例 2.1.1 的方程组

$$\begin{cases} x_1 + 2x_2 + x_3 = 0, \\ 2x_1 + 2x_2 + 3x_3 = 3, \\ -x_1 - 3x_2 = 2. \end{cases}$$

解 消元过程.

第 1 步 在系数矩阵 A 的第 1 列元素中选取绝对值最大的元 $a_{21} = 2$, 交换第 1 个方程与第 2 个方程的位置得

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 3, & (1) \\ x_1 + 2x_2 + x_3 = 0, & (2) \\ -x_1 - 3x_2 = 2. & (3) \end{cases}$$

进行第 1 步消元. 方程 (2) - 方程 (1) $\times \frac{1}{2}$, 方程 (3) - 方程 (1) $\times \left(-\frac{1}{2}\right)$ 得

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 3, & (1') \\ x_2 - \frac{1}{2}x_3 = -\frac{3}{2}, & (2') \\ -2x_2 + \frac{3}{2}x_3 = \frac{7}{2}. & (3') \end{cases}$$

第 2 步 在第 1 步消元后的矩阵 $A^{(1)}$ 的元素 $a_{22}^{(1)} = 1$, $a_{32}^{(1)} = -2$ 中选取绝对值最大的元 $a_{32}^{(1)} = -2$ 作为主元素, 交换第 2 个方程与第 3 个方程的位置得

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 3, & (1'') \\ -2x_2 + \frac{3}{2}x_3 = \frac{7}{2}, & (2'') \\ x_2 - \frac{1}{2}x_3 = -\frac{3}{2}. & (3'') \end{cases}$$

进行第 2 步消元. 方程 (3'') - 方程 (2'') $\times \left(-\frac{1}{2}\right)$ 得

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 3, \\ -2x_2 + \frac{3}{2}x_3 = \frac{7}{2}, \\ \frac{1}{4}x_3 = \frac{1}{4}. \end{cases}$$

由回代过程求得原方程组的解 $x = (1, -1, 1)^T$. 解毕.

2.2 矩阵的三角分解

在实际问题中, 经常遇到系数矩阵 A 相同, 但右端项不同的多个线性方程组 $Ax = b^{(i)}$ ($i = 1, 2, \dots, m$). 对高斯消去法作进一步的深入分析, 由矩阵运算导出矩阵的三角分解法. 利用矩阵的三角分解求解上述的一类线性方程组, 以达到减少运算量的目的. 其次, 针对具有特殊系数矩阵的线性方程组, 由矩阵的三角分解建立一些特殊的求解方法.

2.2.1 高斯消去法的矩阵形式

从矩阵运算的观点来看, 高斯消去法的消元过程实质上是将增广矩阵 $(A^{(0)}, b^{(0)})$ 通过逐步左乘一系列的初等下三角矩阵 L_k ($k = 1, \dots, n-1$), 最终变为矩阵 $(A^{(n-1)}, b^{(n-1)})$, 即将原方程组化为一个同解的上三角方程组 $A^{(n-1)}x = b^{(n-1)}$. 事实上,

$$A^{(0)} \triangleq A,$$

$$A^{(1)} \triangleq \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}, \quad L_1 \triangleq \begin{pmatrix} 1 & & & \\ -l_{21} & 1 & & \\ -l_{31} & & 1 & \\ \vdots & & & \ddots \\ -l_{n1} & & & & 1 \end{pmatrix},$$

则 $A^{(1)} = L_1 A^{(0)}$.

$$A^{(2)} \triangleq \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}, \quad L_2 \triangleq \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & -l_{32} & 1 & & \\ & -l_{33} & & 1 & \\ & \vdots & & & \ddots \\ & -l_{nn} & & & & 1 \end{pmatrix},$$

则 $A^{(2)} = L_2 A^{(1)} = L_2 L_1 A^{(0)}$.

一般地,

$$A^{(k-1)} \triangleq \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & \cdots & a_{1n}^{(0)} \\ & a_{22}^{(1)} & \cdots & \cdots & a_{2n}^{(1)} \\ & & \ddots & & \vdots \\ & & & a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{pmatrix},$$

$$L_k \triangleq \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{k+1,k} & 1 & & \\ & & -l_{k+2,k} & & 1 & \\ & & \vdots & & & \ddots \\ & & -l_{nk} & & & & 1 \end{pmatrix},$$

$$A^{(k)} \triangleq \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1,k+1}^{(0)} & \cdots & a_{1n}^{(0)} \\ & a_{22}^{(1)} & \cdots & a_{2,k+1}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \ddots & \vdots & & \vdots \\ & & & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ & & & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{n,k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}.$$

则 $A^{(k)} = L_k A^{(k-1)} = L_k L_{k-1} \cdots L_2 L_1 A^{(0)}$.

由于消元过程进行了 $n-1$ 步, 故

$$A^{(n-1)} = L_{n-1} A^{(n-2)} = \cdots = L_{n-1} L_{n-2} \cdots L_2 L_1 A^{(0)}, \quad (2.2.1)$$

$$A^{(n-1)} \triangleq \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1n}^{(0)} \\ & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n-1)} \end{pmatrix}.$$

同理,

$$b^{(n-1)} = L_{n-1}L_{n-2}\cdots L_2L_1b^{(0)}.$$

注意到矩阵 L_k 的逆矩阵为

$$L_k^{-1} \triangleq \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & l_{nk} & & & 1 \end{pmatrix},$$

则由式 (2.2.1) 有

$$A = A^{(0)} = L_1^{-1}L_2^{-1}\cdots L_{n-1}^{-1}A^{(n-1)} = LU, \quad b = b^{(0)} = Lb^{(n-1)},$$

其中

$$L \triangleq L_1^{-1}L_2^{-1}\cdots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1 \end{pmatrix}.$$

$$U \triangleq A^{(n-1)}.$$

为了以后使用方便, 将上三角矩阵 U 的元记为 u_{ij} , 则

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix}.$$

由于 L 的对角线元均为 1, 故将 L 称为单位下三角矩阵. 由以上讨论可知, 若高斯消去法能顺利进行下去, 则线性方程组的系数矩阵 A 可以分解成一个单位下三角矩阵 L 与一个上三角矩阵 U 的乘积, 即

$$A = LU.$$

由以上讨论并结合定理 2.1.1 可得下述定理:

定理 2.2.1 设 A 为 n 阶矩阵, 如果 A 的顺序主子式 $D_k \neq 0$ ($k = 1, 2, \dots, n$), 则 A 可以唯一地分解成一个单位下三角矩阵 L 与一个上三角矩阵 U 的乘积, 即

$$A = LU. \quad (2.2.2)$$

证明 由高斯消去法的矩阵形式可知 $A = LU$ 分解的存在性. 下面证明分解的唯一性. 设矩阵 A 有两种分解:

$$A = LU = L_1 U_1,$$

其中 L, L_1 为单位下三角矩阵, U, U_1 为上三角矩阵. 由 A 可逆知, L 与 U 均可逆, 故有 $L_1^{-1}L = U_1U^{-1}$. 该式左端为单位下三角矩阵, 右端为上三角矩阵. 因此, 该式左右两端必是单位矩阵. 于是 $L_1 = L, U_1 = U$, 唯一性得证. 证毕.

式 (2.2.2) 称为矩阵 A 的 LU 分解.

2.2.2 矩阵的 LU 分解

下面根据定理 2.2.1, 利用矩阵乘法, 直接导出矩阵 LU 分解的计算公式. 设

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad L = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1 \end{pmatrix},$$

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix}.$$

由 L 是单位下三角矩阵和 U 是上三角矩阵知 $l_{ik} = 0$ ($k > i$), $l_{ii} = 1$ ($i = 1, 2, \dots, n$), $u_{kj} = 0$ ($k > j$).

根据 $A = LU$, 由矩阵乘法可得

$$a_{ij} = \sum_{k=1}^n l_{ik} u_{kj} = \sum_{k=1}^{\min\{i,j\}} l_{ik} u_{kj}, \quad i, j = 1, 2, \dots, n. \quad (2.2.3)$$

先计算矩阵 U 的元 u_{ij} ($i \leq j$). 由式 (2.2.3) 有

$$a_{ij} = \sum_{k=1}^i l_{ik} u_{kj} = \sum_{k=1}^{i-1} l_{ik} u_{kj} + u_{ij}, \quad l_{ii} = 1,$$

故

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \quad i = 1, 2, \dots, n; j = i, i+1, \dots, n.$$

特别地, 当 $i = 1$ 时, $u_{1j} = a_{1j}$ ($j = 1, 2, \dots, n$).

再算矩阵 L 的元 l_{ij} ($i > j$). 由式 (2.2.3) 有

$$a_{ij} = \sum_{k=1}^j l_{ik} u_{kj} = \sum_{k=1}^{j-1} l_{ik} u_{kj} + l_{ij} u_{jj},$$

由此可得

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj}}{u_{jj}}, \quad j = 1, 2, \dots, n-1; i = j+1, j+2, \dots, n.$$

特别地, 当 $j = 1$ 时, $l_{i1} = \frac{a_{i1}}{u_{11}}$ ($i = 2, 3, \dots, n$).

综上所述, $A = LU$ 分解公式如下:

$$\left\{ \begin{array}{ll} u_{1j} = a_{1j}, & j = 1, 2, \dots, n, \\ l_{i1} = \frac{a_{i1}}{u_{11}}, & i = 2, 3, \dots, n, \\ u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, & i = 2, 3, \dots, n; j = i, i+1, \dots, n, \\ l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj}}{u_{jj}}, & j = 2, 3, \dots, n-1; i = j+1, j+2, \dots, n. \end{array} \right. \quad (2.2.4)$$

为了编程方便, 将 l_{ij} 计算式中的下标 i, j 互换, 即

$$l_{ji} = \frac{a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki}}{u_{ii}}, \quad i = 2, 3, \dots, n-1; j = i+1, i+2, \dots, n.$$

矩阵 $A = LU$ 分解的计算过程如下:

$$u_{1j} = a_{1j}, \quad j = 1, 2, \dots, n.$$

$$l_{i1} = a_{i1}/u_{11}, \quad i = 2, 3, \dots, n.$$

for $i = 2$ to $n-1$ do

$$\begin{aligned}
& u_{ii} := a_{ii} - \sum_{k=1}^{i-1} l_{ik} u_{ki} \\
& \text{for } j = i+1 \text{ to } n \text{ do} \\
& \quad u_{ij} := a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \\
& \quad l_{ji} := \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki} \right) / u_{ii} \\
& \text{end do} \\
& \text{end do} \\
& u_{nn} := a_{nn} - \sum_{k=1}^{n-1} l_{nk} u_{kn}
\end{aligned}$$

在计算机上实现矩阵 $A = LU$ 分解时, 为了减少存储量, 将 L 的下三角部分的元素存放在 A 的下三角部分相应的位置上 (L 的对角元均等于 1, 不用存储). 将矩阵 U 的上三角部分的元素放于 A 的上三角部分位置, 其编程时只需要将上述分解过程中 l_{ij}, u_{ij} 改为相应的 a_{ij} 即可.

上面介绍的将矩阵分解为单位下三角矩阵 L 与上三角矩阵 U 的乘积 $A = LU$, 称为杜利特尔 (Doolittle) 分解. 同理, 矩阵 A 也可分解成下三角矩阵 L 与单位上三角矩阵 U 之积. 这种分解称为克劳特 (Crout) 分解. 读者仿照杜利特尔分解法, 不难导出克劳特分解法的计算公式, 故此处从略.

$A = LU$ 分解的紧凑格式 从矩阵的 LU 分解公式 (2.2.4) 知, 除了矩阵 U 的第 1 列元素 $u_{1j} = a_{1j}$ ($j = 1, 2, \dots, n$), 矩阵 L 的第 1 列元素 $l_{i1} = a_{i1}/a_{11}$ ($i = 2, 3, \dots, n$) 外, 其余元素 u_{ij} 和 l_{ij} 都是从矩阵 A 的对应元素 a_{ij} 减去同行左边 L 的元素与同列上边 U 的元素的乘积. 然后 L 的元素除以 U 的对角元 u_{ii} , 按先算矩阵 U 的第 i 行, 再算矩阵 L 的第 i 列的顺序计算, 计算流程按图 2.1 所示逐框进行.

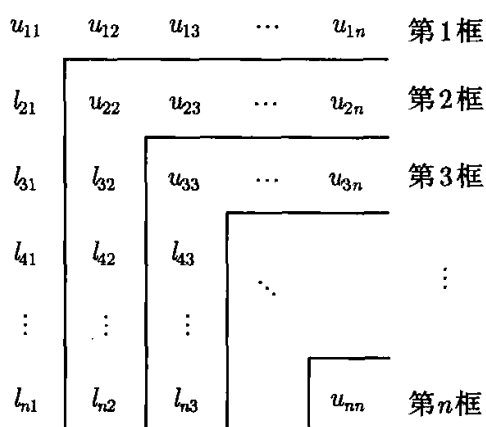


图 2.1

例 2.2.1 分解矩阵 $A = LU$, 其中

$$A = \begin{pmatrix} 4 & -2 & 0 & 4 \\ -2 & 2 & -3 & 1 \\ 0 & -3 & 13 & -7 \\ 4 & 1 & -7 & 23 \end{pmatrix}.$$

解 按计算公式 (2.2.4), 由紧凑格式可得图 2.2.

所以

$$A = LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ 1 & 3 & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 4 & -2 & 0 & 4 \\ 0 & 1 & -3 & 3 \\ 0 & 0 & 4 & 2 \\ 0 & 0 & 0 & 9 \end{pmatrix}.$$

图 2.2

解毕.

下面介绍利用矩阵 $A = LU$ 分解求解线性方程组 $Ax = b$.

设对 A 已作 LU 分解, 即有 $A = LU$. 将其代入 $Ax = b$ 得

$$LUx = b.$$

显然, 它等价于下面两个三角方程组:

$$Ly = b,$$

$$Ux = y.$$

先从 $Ly = b$ 解出 y , 再从 $Ux = y$ 解出 x .

解方程组 $Ly = b$. $Ly = b$ 的第一个方程为 $y_1 = b_1$, $Ly = b$ 的第 i 个方程为

$$\sum_{j=1}^{i-1} l_{ij}y_j + y_i = b_i, \quad i = 2, 3, \dots, n,$$

所以

$$\begin{cases} y_1 = b_1, \\ y_i = b_i - \sum_{j=1}^{i-1} l_{ij}y_j, \quad i = 2, 3, \dots, n. \end{cases}$$

从 y 的计算公式可以看到, 可不必先将 A 分解后再求解 $Ly = b$, 计算 y 与 A 的分解可以同时进行. 具体做法是对 A 的增广矩阵 $\bar{A} = (A, b)$ 进行 LU 分解, 分解结束后, $\bar{A}$ 的第 $n+1$ 列位置上的元即是 y 的元.

求得 y 后, 再求解上三角方程组 $Ux = y$, 同高斯消去法的回代过程可得

$$\begin{cases} x_n = \frac{y_n}{u_{nn}}, \\ x_i = \frac{y_i - \sum_{j=i+1}^n u_{ij}x_j}{u_{ii}}, \quad i = n-1, n-2, \dots, 2, 1. \end{cases}$$

用 $A = LU$ 分解解方程 $Ax = b$ 乘除法的运算量如下:

由矩阵分解公式 (2.2.4) 知, 计算 u_{ij} 和 l_{ij} 各有 $\frac{1}{6}n^3 - \frac{1}{6}n$ 个乘除法, 故矩阵分解共有 $\frac{1}{3}n^3 - \frac{1}{3}n$ 个乘除法. 求解下三角方程 $Ly = b$ 有 $\frac{1}{2}n(n-1)$ 个乘除法, 求解上三角方程组 $Ux = y$ 有 $\frac{1}{2}n(n-1) + n$ 个乘除法, 所以用 $A = LU$ 分解解方程 $Ax = b$ 共有 $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ 个乘除法, 与高斯消去法的乘除法运算次数相同.

矩阵的三角分解主要应用于求解具有相同的系数矩阵 A , 不同右端项的多个线性方程组

$$Ax = b^{(i)}, \quad i = 1, 2, \dots, m.$$

先将 $A = LU$ 分解后, 再求解 $2m$ 个三角方程组

$$Ly^{(i)} = b^{(i)}, \quad Ux = y^{(i)}, \quad i = 1, 2, \dots, m.$$

这样求解大大地减少了计算量.

例 2.2.2 用 $A = LU$ 分解求解线性方程组

$Ax = b$, 其中

$$A = \begin{pmatrix} 9 & 18 & 9 & -27 \\ 18 & 45 & 0 & -45 \\ 9 & 0 & 126 & 9 \\ -27 & -45 & 9 & 135 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 16 \\ 8 \end{pmatrix}.$$

$$\begin{array}{c} 9 \\ 2 \\ 1 \\ -3 \end{array} \left[\begin{array}{cccc|c} 18 & 9 & -27 & & 1 \\ \hline 9 & -18 & 9 & & 0 \\ \hline -2 & 81 & 54 & & 15 \\ \hline 1 & \frac{2}{3} & 9 & & 1 \end{array} \right]$$

解 将向量 b 作为 $\bar{A}$ 的第 5 列, 按紧凑格式计算得图 2.3. 则

图 2.3

$$A = LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ -3 & 1 & \frac{2}{3} & 1 \end{pmatrix} \begin{pmatrix} 9 & 18 & 9 & -27 \\ 0 & 9 & -18 & 9 \\ 0 & 0 & 81 & 54 \\ 0 & 0 & 0 & 9 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 0 \\ 15 \\ 1 \end{pmatrix}.$$

由

$$Ax = b \Rightarrow \begin{cases} Ly = b, \\ Ux = y. \end{cases}$$

解方程组 $Ly = b$, 从矩阵分解得 $y = (1, 0, 15, 1)^T$. 从 $Ux = y$ 解得 $x = \left(\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\right)^T$. 解毕.

2.2.3 平方根法和改进平方根法

在用有限元法求解结构力学和最优化方法等诸多问题中, 都需要求解一个系数矩阵是对称正定的线性方程组 $Ax = b$. 针对这种方程组的特点, 本小节建立一种更有效的求解方法——平方根法. 所谓平方根法, 就是先将矩阵 A 分解成一个下三角矩阵与该下三角矩阵转置的乘积, 然后利用这种分解求解线性方程组 $Ax = b$.

定理 2.2.2 (对称矩阵的三角分解) 设 A 是 n 阶对称矩阵, 若 A 的各阶顺序主子式均不等于零, 则 A 可以唯一地分解为

$$A = LDL^T.$$

证明 由定理的条件, 据定理 2.2.1 知, A 可以唯一地分解为 $A = LU$, 其中 L 为单位下三角矩阵, U 为上三角矩阵. 记 D 是由 U 的对角线元构成的对角矩阵, 即

$$D = \text{diag}(d_1, d_2, \dots, d_n) = \text{diag}(u_{11}, u_{22}, \dots, u_{nn}),$$

则 $A = LU = LDD^{-1}U = LDM^T$, 其中 $M^T = D^{-1}U$, M 为单位下三角矩阵. 又 $LU = A = A^T = (LDM^T)^T = M(LD)^T$. 注意到 $(LD)^T$ 是上三角矩阵, 两边都是 A 的杜利特尔分解, 由三角分解的唯一性知

$$L = M,$$

于是 $A = LDL^T$. 证毕.

再进一步假定 A 是对称正定矩阵, 则有以下定理:

定理 2.2.3 (对称正定矩阵的楚列斯基 (Cholesky) 分解) 设 A 是 n 阶对称正定矩阵, 则存在一个可逆的下三角矩阵 G , 使得

$$A = GG^T.$$

当限定 G 的对角元为正时, 这种分解是唯一的.

证明 由线性代数知, 正定矩阵的各阶顺序主子式均大于零, 由定理 2.2.2 知 $A = LDL^T$. 可证 D 的对角元均大于零.

事实上, $\forall x \neq 0$,

$$0 < x^T Ax = x^T LDL^T x = (L^T x)^T D (L^T x).$$

分别取 x 为方程 $L^T x = e_i$ ($i = 1, 2, \dots, n$) 的解, 其中 $e_i = (0, \dots, 0, \underset{i}{1}, 0, \dots, 0)^T \in \mathbb{R}^n$ 为各坐标轴上的单位向量, 则有

$$0 < (L^T x)^T D (L^T x) = e_i^T D e_i = d_i = u_{ii}, \quad i = 1, 2, \dots, n.$$

记 $D^{\frac{1}{2}} = \text{diag}(\sqrt{u_{11}}, \sqrt{u_{22}}, \dots, \sqrt{u_{nn}})$, 则

$$A = LDL^T = LD^{\frac{1}{2}}D^{\frac{1}{2}}L^T = (LD^{\frac{1}{2}})(LD^{\frac{1}{2}})^T \triangleq GG^T,$$

其中 $G = LD^{\frac{1}{2}}$ 为下三角矩阵, 即

$$G \triangleq \begin{pmatrix} g_{11} & & & & \\ g_{21} & g_{22} & & & \\ g_{31} & g_{32} & g_{33} & & \\ \vdots & \vdots & & \ddots & \\ g_{n1} & g_{n2} & \cdots & g_{n,n-1} & g_{nn} \end{pmatrix}.$$

证毕.

下面导出矩阵楚列斯基分解的计算公式.

由 G 是下三角矩阵 ($g_{ij} = 0, j > i$) 及 A 的对称性, 故只考虑 A 的下三角部分的元素, 即 $i \geq j$ 的情况. 据 $A = GG^T$, 由矩阵乘法得

$$a_{ij} = \sum_{k=1}^n g_{ik}g_{jk} = \sum_{k=1}^{\min\{i,j\}} g_{ik}g_{jk} = \sum_{k=1}^j g_{ik}g_{jk} = \sum_{k=1}^{j-1} g_{ik}g_{jk} + g_{ij}g_{jj}.$$

当 $i = j$ 时, $a_{jj} = \sum_{k=1}^{j-1} g_{jk}^2 + g_{jj}^2$, 所以,

$$g_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2 \right)^{\frac{1}{2}}, \quad j = 1, 2, \dots, n.$$

当 $j = 1$ 时, $g_{11} = \sqrt{a_{11}}$.

当 $i > j$ 时,

$$g_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} g_{ik}g_{jk}}{g_{jj}}, \quad j = 1, 2, \dots, n-1; i = j+1, j+2, \dots, n.$$

当 $j = 1$ 时, $g_{i1} = \frac{a_{i1}}{g_{11}} \quad (i = 2, 3, \dots, n)$.

于是得分解公式

$$\left\{ \begin{array}{l} g_{11} = \sqrt{a_{11}}, \\ g_{i1} = \frac{a_{i1}}{g_{11}}, \quad i = 2, 3, \dots, n. \\ g_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2 \right)^{\frac{1}{2}}, \quad j = 2, 3, \dots, n. \\ g_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} g_{ik}g_{jk}}{g_{jj}}, \quad j = 2, 3, \dots, n-1; i = j+1, j+2, \dots, n. \end{array} \right. \quad (2.2.5)$$

由分解公式知

$$a_{jj} = \sum_{k=1}^j g_{jk}^2 \Rightarrow |g_{jk}| \leq \sqrt{a_{jj}}, \quad j = 1, 2, \dots, n; \quad k = 1, 2, \dots, j.$$

这说明矩阵 G 的元素是可以控制的, 故舍入误差的增长是可以控制的, 因此, 分解过程是稳定的.

$A = GG^T$ 分解的算法组织如下:

$$g_{11} := \sqrt{a_{11}}$$

$$g_{i1} := a_{i1}/g_{11}, \quad i = 2, 3, \dots, n.$$

for $j = 2$ to $n - 1$ do

$$g_{jj} := \left(a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2 \right)^{\frac{1}{2}}$$

for $i = j + 1$ to n do

$$g_{ij} := \left(a_{ij} - \sum_{k=1}^{j-1} g_{ik}g_{jk} \right) / g_{jj}$$

end do

end do

$$g_{nn} := \left(a_{nn} - \sum_{k=1}^{n-1} g_{nk}^2 \right)^{\frac{1}{2}}$$

在上述算法中, 因为 A 是对称正定矩阵, 故只需要储存 A 的下三角部分的元素, 将算得的 g_{ij} 存放于 A 的相应位置上.

用 $A = GG^T$ 分解求解方程组 $Ax = b$ 的方法称为平方根法.

$$Ax = b \Rightarrow GG^Tx = b \Rightarrow \begin{cases} Gy = b \Rightarrow \text{解出 } y, \\ G^Tx = y \Rightarrow \text{解出 } x. \end{cases}$$

由 $Gy = b$ 知

$$\begin{cases} g_{11}y_1 = b_1, \\ g_{i1}y_1 + g_{i2}y_2 + \dots + g_{ii}y_i = b_i, \quad i = 2, 3, \dots, n. \end{cases}$$

由此可得

$$\begin{cases} y_1 = \frac{b_1}{g_{11}}, \\ y_i = \frac{b_i - \sum_{k=1}^{i-1} g_{ik}y_k}{g_{ii}}, \quad i = 2, 3, \dots, n. \end{cases}$$

然后从 $G^T x = y$, 解出 x .

类似于用杜利特尔分解法解方程组, 求解下三角方程组 $Gy = b$ 可在分解 A 的同时进行, 其紧凑格式如图 2.4 所示.

$$\begin{array}{cccccc}
 g_{11} & g_{21} & g_{31} & \cdots & g_{n1} & y_1 & \text{第1框} \\
 g_{21} & g_{22} & g_{32} & \cdots & g_{n2} & y_2 & \text{第2框} \\
 g_{31} & g_{32} & g_{33} & \cdots & g_{n3} & y_3 & \text{第3框} \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
 g_{n1} & g_{n2} & g_{n3} & & g_{nn} & y_n & \text{第}n\text{框}
 \end{array}$$

图 2.4

例 2.2.3 用平方根法求解例 2.2.2 的方程组.

解

$$\bar{A} = \begin{pmatrix} 9 & 18 & 9 & -27 & 1 \\ 18 & 45 & 0 & -45 & 2 \\ 9 & 0 & 126 & 9 & 16 \\ -27 & -45 & 9 & 135 & 8 \end{pmatrix},$$

按图 2.4 计算有

$$\begin{array}{cccc|c}
 3 & 6 & 3 & -9 & \frac{1}{3} \\
 6 & 3 & -6 & 3 & 0 \\
 3 & -6 & 9 & 6 & \frac{5}{3} \\
 -9 & 3 & 6 & 3 & \frac{1}{3}
 \end{array}$$

由此可得 $A = GG^T$, 其中

$$G = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 6 & 3 & 0 & 0 \\ 3 & -6 & 9 & 0 \\ -9 & 3 & 6 & 3 \end{pmatrix}, \quad y = \begin{pmatrix} \frac{1}{3} \\ 0 \\ \frac{5}{3} \\ \frac{1}{3} \end{pmatrix}.$$

因此,

$$Ax = b \Rightarrow GG^T x = b \Rightarrow \begin{cases} Gy = b \Rightarrow y = \left(\frac{1}{3}, 0, \frac{5}{3}, \frac{1}{3} \right)^T, \\ G^T x = y \Rightarrow x = \left(\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9} \right)^T. \end{cases}$$

解毕.

用平方根法求解方程组 $Ax = b$ 共需要作 $\frac{1}{6}n^3 + n^2 - \frac{1}{6}n$ 个乘除法和 n 个开方运算, 比高斯消去法和用杜利特尔分解法解方程减少了近一半的运算量. 但 n 个开方运算需要耗费较多的机器时间. 为了避免平方根法的开方运算, 利用定理 2.2.2 对 A 作 $A = LDL^T$ 分解. 下面导出 $A = LDL^T$ 的分解公式. 由

$$A = LU = LDD^{-1}U = LDL^T$$

知 $L^T = D^{-1}U$. 据此,

$$l_{ji} = \frac{u_{ij}}{d_i} = \frac{u_{ij}}{u_{ii}}, \quad i = 1, 2, \dots, n-1; j = i+1, i+2, \dots, n.$$

根据杜利特尔分解式 (2.2.4) 有

$$\begin{cases} u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, & i = 1, 2, \dots, n; j = i, i+1, \dots, n, \\ l_{ji} = \frac{u_{ij}}{u_{ii}}, & i = 1, 2, \dots, n-1; j = i+1, i+2, \dots, n. \end{cases} \quad (2.2.6)$$

算法组织如下:

```

 $u_{1j} := a_{1j}, \quad j = 1, 2, \dots, n.$ 
 $l_{j1} := u_{1j}/u_{11}, \quad j = 2, 3, \dots, n.$ 
for  $i = 2$  to  $n-1$  do
     $u_{ii} := a_{ii} - \sum_{k=1}^{i-1} l_{ik}u_{ki}$ 
    for  $j = i+1$  to  $n$  do
         $u_{ij} := a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}$ 
         $l_{ji} := u_{ij}/u_{ii}$ 
    end do
end do
 $u_{nn} := a_{nn} - \sum_{k=1}^{n-1} l_{nk}u_{kn}$ 

```

用 $A = LDL^T$ 分解求解方程组 $Ax = b$, 称为改进平方根法. 下面导出其计算公式.

$$Ax = b \Rightarrow LDL^T x = b \Rightarrow \begin{cases} Ly = b \Rightarrow \text{解出 } y, \\ Dz = y \Rightarrow \text{解出 } z, \\ L^T x = z \Rightarrow \text{解出 } x. \end{cases}$$

由 $Ly = b$ 得

$$y_1 = b_1,$$

$$y_i = b_i - \sum_{k=1}^{i-1} l_{ik} y_k, \quad i = 2, 3, \dots, n.$$

由 $Dz = y$ 得

$$z_i = \frac{y_i}{u_{ii}}, \quad i = 1, 2, \dots, n.$$

由 $L^T x = z$ 得

$$\begin{cases} x_n = z_n, \\ x_i = z_i - \sum_{k=i+1}^n l_{ki} x_k, \quad i = n-1, n-2, \dots, 2, 1. \end{cases}$$

将 z_i 和 $l_{ki} = \frac{u_{ik}}{u_{ii}}$ 代入 x_i 的表达式, 则得

$$\begin{cases} x_n = \frac{y_n}{u_{nn}}, \\ x_i = \frac{y_i - \sum_{k=i+1}^n u_{ik} x_k}{u_{ii}}, \quad i = n-1, n-2, \dots, 2, 1. \end{cases}$$

而用改进平方根法解方程组 $Ax = b$ 共用 $\frac{1}{6}n^3 + \frac{3}{2}n^2 - \frac{2}{3}n$ 个乘除法, 它比平方根法多了 $\frac{1}{2}n^2 - \frac{1}{2}n$ 个乘除法, 少了 n 个开方运算, 而比高斯消去法和利用杜利特尔分解解方程几乎少了一半的乘除法. 因此, 改进平方根法是求解系数矩阵为对称正定的线性方程组最有效的方法之一.

2.2.4 求解三对角方程组的追赶法

在样条插值函数、常微分方程边值问题和热传导方程的有限差分法等问题中, 常遇到求解三对角方程组 $Ax = d$, 即

$$\begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}.$$

通常 A 是严格对角占优矩阵. 严格对角占优矩阵存在唯一的杜利特尔分解.

对于三对角矩阵 A , 则有如下形式的杜利特尔分解:

$$A = LU = \begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{n-1} & 1 & \\ & & & l_n & 1 \end{pmatrix} \begin{pmatrix} u_1 & c_1 & & & \\ & u_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & u_{n-1} & c_{n-1} \\ & & & & u_n \end{pmatrix}.$$

根据矩阵乘法易得其分解公式

$$\begin{cases} u_1 = b_1, \\ l_i = \frac{a_i}{u_{i-1}}, & i = 2, 3, \dots, n, \\ u_i = b_i - l_i c_{i-1}, & i = 2, 3, \dots, n. \end{cases}$$

用矩阵 $A = LU$ 分解求解三对角方程组称为追赶法, 即

$$\text{求解 } Ax = d \Rightarrow LUx = d \Rightarrow \begin{cases} Ly = d, \\ Ux = y. \end{cases}$$

由 $Ly = d$ 得

$$\begin{cases} y_1 = d_1, \\ y_i = d_i - l_i y_{i-1}, & i = 2, 3, \dots, n. \end{cases}$$

由 $Ux = y$ 得

$$\begin{cases} x_n = \frac{y_n}{u_n}, \\ x_i = \frac{y_i - c_i x_{i+1}}{u_i}, & i = n-1, n-2, \dots, 1. \end{cases}$$

求解三对角方程 $Ax = d$ 的追赶法的算法组织如下:

```

 $u_1 := b_1, \quad y_1 := d_1$ 
for  $i = 2$  to  $n$  do
     $l_i := a_i / u_{i-1}$ 
     $u_i := b_i - l_i c_{i-1}$ 
     $y_i := d_i - l_i y_{i-1}$ 
end do
 $x_n := y_n / u_n$ 
for  $i = n-1$  to  $1$  step  $-1$  do
     $x_i := (y_i - c_i x_{i+1}) / u_i$ 
end do

```

计算 l_i, u_i, y_i 的过程称为“追”的过程, 计算 x_i 的过程称为“赶”的过程. 追赶法的乘除法的运算次数仅为 $5n - 4$, 比高斯消去法的运算次数少得多. 由于方程组的系数矩阵 A 是严格对角占优矩阵, 保证了追赶法能顺利进行, 并且计算过程稳定.

例 2.2.4 用追赶法求解线性方程组 $Ax = d$, 其中

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & 4 & 2 & 0 \\ 0 & 0 & 4 & 7 & 1 \\ 0 & 0 & 0 & -5 & 6 \end{pmatrix}, \quad d = \begin{pmatrix} 5 \\ 9 \\ 2 \\ 19 \\ -4 \end{pmatrix}.$$

$$\begin{array}{c|ccccc|c} 1 & 2 & 0 & 0 & 0 & 5 \\ 2 & -1 & 1 & 0 & 0 & -1 \\ 0 & 3 & 1 & 2 & 0 & 5 \\ 0 & 0 & 4 & -1 & 1 & -1 \\ 0 & 0 & 0 & 5 & 1 & 1 \end{array}$$

解 追赶法是矩阵 $A = LU$ 分解用于求解三对

图 2.5

角方程组的特殊情形. 参照用杜利特尔分解求解方程组的方法有图 2.5. 所以

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 4 & 1 & 0 \\ 0 & 0 & 0 & 5 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 2 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} 5 \\ -1 \\ 5 \\ -1 \\ 1 \end{pmatrix}.$$

从 $Ly = d$ 解得 $y = (5, -1, 5, -1, 1)^T$. 从 $Ux = y$ 解得 $x = (1, 2, 1, 2, 1)^T$. 解毕.

2.3 舍入误差对解的影响

设线性方程组 $Ax = b$ 的准确解为 x^* , 近似解为 $\tilde{x}$. 为了衡量近似解的准确程度以及以后研究迭代法的收敛性问题, 则需推广绝对值的概念. 本节引入向量范数与矩阵范数, 利用这些范数进行误差分析和评判线性方程组的性态.

2.3.1 向量范数与矩阵范数

定义 2.3.1 (向量范数) 设 $f(x) = \|x\|$ 是定义在 $\mathbf{R}^n$ 上的非负实值函数, 若 $\|x\|$ 满足以下三条:

- (1) $\|x\| \geq 0$, $\|x\| = 0 \Leftrightarrow x = 0$;
- (2) $\forall \alpha \in \mathbf{R}$, $\|\alpha x\| = |\alpha| \|x\|$;
- (3) $\|x + y\| \leq \|x\| + \|y\|$ (三角不等式),

则称 $\|x\|$ 为 $\mathbf{R}^n$ 上的向量范数.

由三角不等式可以推得

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

三种常用的向量范数如下:

$$\|x\|_1 = \sum_{i=1}^n |x_i| \quad (\text{向量的 1 范数});$$

$$\|x\|_2 = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}} \quad (\text{向量的 2 范数});$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \quad (\text{向量的 } \infty \text{ 范数}).$$

易证, 上述三种向量范数满足定义 2.3.1.

例如, 向量 $x = (2, -5, 3)^T$ 的三种范数分别为

$$\|x\|_1 = 10, \quad \|x\|_2 = \sqrt{38}, \quad \|x\|_\infty = 5.$$

利用向量范数定义 x 的近似向量 $\tilde{x}$ 的绝对误差与相对误差如下:

绝对误差: $\|\Delta x\| = \|x - \tilde{x}\|;$

相对误差: $\frac{\|\Delta x\|}{\|x\|} = \frac{\|x - \tilde{x}\|}{\|x\|}.$

关于向量范数有以下两个定理:

定理 2.3.1 (向量范数的连续性) 设 $\|x\|$ 是 $\mathbf{R}^n$ 上的任一种向量范数, 则 $\|x\|$ 是 x 的连续函数.

证明 只需证明当 $x \rightarrow y$ 时 $\|x\| \rightarrow \|y\|$ 即可. 设

$$x = \sum_{i=1}^n x_i e_i, \quad y = \sum_{i=1}^n y_i e_i,$$

其中 $e_i = (0, \dots, 0, 1, 0, \dots, 0)^T \in \mathbf{R}^n$ ($i = 1, 2, \dots, n$), 则

$$\begin{aligned} |\|x\| - \|y\|| &\leq \|x - y\| = \left\| \sum_{i=1}^n (x_i - y_i) e_i \right\| \leq \sum_{i=1}^n |x_i - y_i| \|e_i\| \\ &\leq \max_i |x_i - y_i| \sum_{i=1}^n \|e_i\| = \|x - y\|_\infty \sum_{i=1}^n \|e_i\|. \end{aligned}$$

若记 $c \triangleq \sum_{i=1}^n \|e_i\|$, 则有

$$|\|x\| - \|y\|| \leq c \|x - y\|_\infty \rightarrow 0, \quad x \rightarrow y.$$

证毕.

定理 2.3.2 (向量范数的等价性) 设 $\|x\|_p$ 和 $\|x\|_q$ 是 $\mathbf{R}^n$ 上的任意两种向量范数, 则存在常数 $c_1 > 0$, $c_2 > 0$, 使得对任意的 $x \in \mathbf{R}^n$ 有

$$c_1 \|x\|_q \leq \|x\|_p \leq c_2 \|x\|_q.$$

证明 略.

常用的几个等价关系如下:

$$\begin{aligned} \|x\|_\infty &\leq \|x\|_1 \leq n \|x\|_\infty, \\ \frac{1}{\sqrt{n}} \|x\|_1 &\leq \|x\|_2 \leq \|x\|_1, \\ \frac{1}{\sqrt{n}} \|x\|_2 &\leq \|x\|_\infty \leq \|x\|_2. \end{aligned} \quad (2.3.1)$$

下面给出矩阵范数的定义.

定义 2.3.2 (矩阵范数) 设 $f(A) = \|A\|$ 是定义在 $\mathbf{R}^{n \times n}$ 上的非负实值函数, 若 $\|A\|$ 满足以下 4 条:

- (1) $\|A\| \geq 0$, $\|A\| = 0 \Leftrightarrow A = O$;
- (2) $\forall \alpha \in \mathbf{R}$, $\|\alpha A\| = |\alpha| \|A\|$;
- (3) $\|A + B\| \leq \|A\| + \|B\|$ (三角不等式);
- (4) $\|AB\| \leq \|A\| \|B\|$,

则称 $\|A\|$ 为 $\mathbf{R}^{n \times n}$ 上的矩阵范数.

由定义 2.3.2 中的三角不等式可得

$$|\|A\| - \|B\|| \leq \|A - B\|.$$

在误差分析中, 常有矩阵与向量相乘, 希望下式:

$$\|Ax\| \leq \|A\| \|x\|, \quad \forall x \in \mathbf{R}, A \in \mathbf{R}^{n \times n} \quad (2.3.2)$$

成立, 但式 (2.3.2) 不是对任意的矩阵范数和向量范数总是成立的, 于是引入矩阵范数与向量范数相容性的概念.

定义 2.3.3 若矩阵范数 $\|A\|$ 与向量范数 $\|x\|$ 满足式 (2.3.2), 则称矩阵范数 $\|A\|$ 与向量范数 $\|x\|$ 为相容的或协调的.

下面导出与向量范数 $\|x\|_p$ ($p = 1, 2, \infty$) 是相容的矩阵范数.

定理 2.3.3 (矩阵的算子范数) 设 $x \in \mathbf{R}^n$, $A \in \mathbf{R}^{n \times n}$, $\|x\|_p$ ($p = 1, 2, \infty$) 是给定的—种向量范数, 相应地定义一个矩阵的非负函数

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p=1} \|Ax\|_p, \quad (2.3.3)$$

则 $\|A\|_p$ 是一种矩阵范数, 称为矩阵 A 的算子范数, 并且满足相容性条件

$$\|Ax\|_p \leq \|A\|_p \|x\|_p.$$

显然, 由式 (2.3.3) 定义的矩阵范数, 对于任何的 p , n 阶单位矩阵 I 的范数 $\|I\|_p = 1$.

证明 (1) 显然,

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p \geq 0,$$

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = 0 \Leftrightarrow \|Ax\|_p = 0, \forall x \neq 0 \Leftrightarrow Ax = 0, \forall x \neq 0 \Leftrightarrow A = O,$$

所以, $\|A\|_p = 0 \Leftrightarrow A = O$.

(2) $\forall \alpha \in \mathbf{R}$,

$$\|\alpha A\|_p = \max_{\|x\|_p=1} \|\alpha Ax\|_p = |\alpha| \max_{\|x\|_p=1} \|Ax\|_p = |\alpha| \|A\|_p.$$

(3) 在证明三角不等式前先证相容性. $\forall x \neq 0$,

$$\frac{\|A\|_p}{\|x\|_p} \leq \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \|A\|_p \Rightarrow \|Ax\|_p \leq \|A\|_p \|x\|_p.$$

利用相容性证三角不等式,

$$\begin{aligned} \|A+B\|_p &= \max_{\|x\|_p=1} \|(A+B)x\|_p = \max_{\|x\|_p=1} \|Ax+Bx\|_p \\ &\leq \max_{\|x\|_p=1} \|Ax\|_p + \max_{\|x\|_p=1} \|Bx\|_p \\ &\leq \max_{\|x\|_p=1} \|A\|_p \|x\|_p + \max_{\|x\|_p=1} \|B\|_p \|x\|_p = \|A\|_p + \|B\|_p. \end{aligned}$$

(4) 由相容性有

$$\|ABx\|_p \leq \|A\|_p \|Bx\|_p \leq \|A\|_p \|B\|_p \|x\|_p.$$

$\forall x \neq 0$, 则

$$\frac{\|ABx\|_p}{\|x\|_p} \leq \|A\|_p \|B\|_p,$$

故

$$\|AB\|_p = \max_{x \neq 0} \frac{\|ABx\|_p}{\|x\|_p} \leq \|A\|_p \|B\|_p.$$

证毕.

由式 (2.3.3) 定义的矩阵范数 $\|A\|_p$ 与向量范数 $\|x\|_p$ 有关, 当给定一种具体的向量范数 $\|x\|_p$ ($p = 1, 2, \infty$) 时, 就得到了相应的一种矩阵范数 $\|A\|_p$, 它们的具体表达式如下述定理所示:

定理 2.3.4 设 $x \in \mathbf{R}^n, A \in \mathbf{R}^{n \times n}$, 按式 (2.3.3) 确定的对应于向量的三种范数 $\|x\|_p$ ($p = 1, 2, \infty$) 的矩阵范数分别为

$$(1) \|A\|_1 = \max_{1 \leq j \leq n} \left\{ \sum_{i=1}^n |a_{ij}| \right\} \quad (1 \text{ 范数或列范数});$$

$$(2) \|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} \quad (2 \text{ 范数或谱范数});$$

$$(3) \|A\|_\infty = \max_{1 \leq i \leq n} \left\{ \sum_{j=1}^n |a_{ij}| \right\} \quad (\infty \text{ 范数或行范数}),$$

其中 $\lambda_{\max}(A^T A)$ 表示矩阵 $A^T A$ 最大的特征值.

证明 (1) $A = (a_1, a_2, \dots, a_n)$, 其中 $a_j = (a_{1j}, a_{2j}, \dots, a_{nj})^T$ ($j = 1, 2, \dots, n$). $\forall x \in \mathbf{R}$, 则

$$\begin{aligned} \|Ax\|_1 &= \|x_1 a_1 + x_2 a_2 + \dots + x_n a_n\|_1 \leq |x_1| \|a_1\|_1 + |x_2| \|a_2\|_1 + \dots + |x_n| \|a_n\|_1 \\ &\leq (|x_1| + |x_2| + \dots + |x_n|) \max_{1 \leq j \leq n} \{\|a_j\|_1\} \\ &= \|x\|_1 \max_{1 \leq j \leq n} \{\|a_j\|_1\}. \end{aligned}$$

而

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 \leq \max_{1 \leq j \leq n} \{\|a_j\|_1\} = \max_{1 \leq j \leq n} \left\{ \sum_{i=1}^n |a_{ij}| \right\}.$$

下面说明存在一个向量 x , $\|x\|_1 = 1$, 使得 $\|Ax\|_1 = \max_{1 \leq j \leq n} \{\|a_j\|_1\}$.

不妨假定 $\max_{1 \leq j \leq n} \{\|a_j\|_1\} = \|a_k\|_1$. 取 $x = e_k$, 显然, $\|e_k\|_1 = 1$, 并且

$$\|Ae_k\|_1 = \|a_k\|_1 = \max_{1 \leq j \leq n} \{\|a_j\|_1\} = \max_{1 \leq j \leq n} \left\{ \sum_{i=1}^n |a_{ij}| \right\},$$

于是

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{1 \leq j \leq n} \left\{ \sum_{i=1}^n |a_{ij}| \right\}.$$

(2) 注意到 $A^T A$ 是实对称的且至少是半正定矩阵, 其特征值是非负实数, 设为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$, u_i 是对应于 λ_i 的特征向量.

由线性代数理论知, $A^T A$ 有完全的标准正交特征向量系 u_i ($i = 1, 2, \dots, n$). 取 u_i ($i = 1, 2, \dots, n$) 作为 $\mathbf{R}^n$ 的一组基底. $\forall x \in \mathbf{R}, x \neq 0$, 则

$$x = \sum_{i=1}^n c_i u_i.$$

于是

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{(Ax, Ax)}{(x, x)} = \frac{(A^T Ax, x)}{(x, x)} = \frac{\sum_{i=1}^n c_i^2 \lambda_i}{\sum_{i=1}^n c_i^2} \leq \lambda_1.$$

另一方面, 取 $x = u_1$, 则上式等号成立, 故

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_1} = \sqrt{\lambda_{\max}(A^T A)}.$$

(3) 与 (1) 同理可证. 证毕.

例 2.3.1 计算矩阵

$$A = \begin{pmatrix} 1 & -2 \\ 3 & 4 \end{pmatrix}$$

的三种范数.

解

$$\|A\|_1 = \max\{4, 6\} = 6, \quad \|A\|_\infty = \max\{3, 7\} = 7.$$

以下求 $\|A\|_2$.

$$A^T A = \begin{pmatrix} 1 & 3 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 10 & 10 \\ 10 & 20 \end{pmatrix},$$

$$|\lambda I - A^T A| = \begin{vmatrix} \lambda - 10 & -10 \\ -10 & \lambda - 20 \end{vmatrix} = \lambda^2 - 30\lambda + 100 = 0.$$

解特征方程得 $\lambda = 15 \pm 5\sqrt{5}$, 所以

$$\lambda_{\max}(A^T A) = 15 + 5\sqrt{5},$$

故 $\|A\|_2 = \sqrt{15 + 5\sqrt{5}}$. 解毕.

除了以上介绍的三种常用的矩阵范数之外, 还有一种常用的矩阵范数是 F (Frobenius) 范数:

$$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2}.$$

易证, 它满足矩阵范数定义 2.3.2 中的 4 条. 矩阵的 F 范数与向量的 2 范数 $\|x\|_2$ 是相容的, 即

$$\|Ax\|_2 \leq \|A\|_F \|x\|_2.$$

事实上,

$$\begin{aligned}\|Ax\|_2^2 &= \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij} x_j \right)^2 \leq \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij}^2 \sum_{k=1}^n x_k^2 \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 \cdot \sum_{k=1}^n x_k^2 = \|A\|_F^2 \|x\|_2^2.\end{aligned}$$

上式两边开方即得结论.

注意: 对 n 阶单位矩阵 I , $\|I\|_F = \sqrt{n}$.

与向量范数的等价性相似, 矩阵范数也具有其等价性, 即 $\|A\|_p$, $\|A\|_q$ 是 $\mathbf{R}^{n \times n}$ 中任意两种矩阵范数, 则存在常数 $c_1 > 0$, $c_2 > 0$, 使得

$$c_1 \|A\|_q \leq \|A\|_p \leq c_2 \|A\|_q, \quad \forall A \in \mathbf{R}^{n \times n}$$

成立.

定义 2.3.4 设 $A \in \mathbf{R}^{n \times n}$, λ_i ($i = 1, 2, \dots, n$) 是 A 的特征值, 则

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

称为矩阵 A 的谱半径.

定理 2.3.5 设 $A \in \mathbf{R}^{n \times n}$, 则 $\rho(A) \leq \|A\|$, 即谱半径不超过 A 的任一种范数.

证明 设 λ 是 A 的任一特征值, x 是对应的特征向量, 则

$$Ax = \lambda x.$$

于是

$$|\lambda| \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\| \|x\|.$$

注意到 $\|x\| \neq 0$, 则得

$$|\lambda| \leq \|A\|,$$

因而 $\rho(A) \leq \|A\|$. 证毕.

对于矩阵的算子范数有下述定理:

定理 2.3.6 设 $\|B\| < 1$, 则 $I - B$ 是可逆矩阵, 并且有

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

证明 用反证法证明 $I - B$ 可逆. 假定 $I - B$ 不可逆, 则方程组

$$(I - B)x = 0$$

有非零解, 于是

$$\begin{aligned} 0 = \|0\| &= \|(I - B)x\| = \|x - Bx\| \geq \|x\| - \|Bx\| \geq \|x\| - \|B\| \|x\| \\ &= (1 - \|B\|)\|x\| > 0. \end{aligned}$$

得到矛盾.

由 $I - B$ 可逆, 则有

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

从而有

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

证毕.

2.3.2 舍入误差对解的影响

下面讨论舍入误差对线性方程组解的影响.

求解线性方程组 $Ax = b$, 由于系数矩阵 A 和右端项 b 的元素存在着数据误差. 其次, 在解方程过程中还存在舍入误差, 只能求得方程组的近似解 $\tilde{x}$. 近似解 $\tilde{x}$ 也与方程本身的固有性态有关.

设方程组 $Ax = b$ 的准确解为 x , 近似解 $\tilde{x}$ 可以看成是原方程组的系数矩阵 A 和右端项 b 有微小扰动 (误差) $\Delta A, \Delta b$ ($\|\Delta A\|$ 和 $\|\Delta b\|$ 的大小与原始数据的精度和计算机的字长有关) 后的方程组

$$(A - \Delta A)x = b - \Delta b$$

的准确解, 即 $\tilde{x} = x - \Delta x$ 满足方程

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

利用向量范数和矩阵范数可以导出方程组近似解 $\tilde{x}$ 的相对误差界.

定理 2.3.7 设线性方程组 $Ax = b$ ($|A| \neq 0, b \neq 0$) 的系数矩阵 A 和右端项 b 有微小的扰动 $\Delta A, \Delta b$, 扰动后的方程组为

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

当 $\|A^{-1}\| \|\Delta A\| < 1$ 时, 则方程组 $Ax = b$ 近似解 $\tilde{x}$ 的相对误差估计式为

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A\| \|A^{-1}\|}{1 - \|A\| \|A^{-1}\| \frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right). \quad (2.3.4)$$

显然, 只要 $\|\Delta A\|$ 足够小, 就能使定理条件 $\|A^{-1}\| \|\Delta A\| < 1$ 成立.

证明 由 $(A - \Delta A)(x - \Delta x) = b - \Delta b$ 和 $Ax = b$ 得

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

即

$$(I - A^{-1}\Delta A)\Delta x = A^{-1}(\Delta b - \Delta A x). \quad (2.3.5)$$

由定理 2.3.6 知, $I - A^{-1}\Delta A$ 可逆, 并且

$$\|(I - A^{-1}\Delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\Delta A\|} \leq \frac{1}{1 - \|A^{-1}\| \|\Delta A\|}.$$

由式 (2.3.5) 知

$$\Delta x = (I - A^{-1}\Delta A)^{-1}A^{-1}(\Delta b - \Delta A x).$$

两边取范数有

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

上式两端除以 $\|x\|$ 得

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

注意到 $\|b\| = \|Ax\| \leq \|A\| \|x\|$, 则得

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

证毕.

式 (2.3.4) 表明, 当 $\frac{\|\Delta A\|}{\|A\|}$ 较小时, 近似解的相对误差 $\frac{\|\Delta x\|}{\|x\|}$ 约为 $\frac{\|\Delta A\|}{\|A\|}$ 与 $\frac{\|\Delta b\|}{\|b\|}$ 之和的 $\|A\| \|A^{-1}\|$ 倍. 由于 $\|A\| \|A^{-1}\| \geq \|I\| = 1$, 可认为 $\|A\| \|A^{-1}\|$ 是误差的一个放大倍数. $\|A\| \|A^{-1}\|$ 越小, 近似解的相对误差就越小; 反之, 近似解的相对误差就可能越大. $\|A\| \|A^{-1}\|$ 与方程组的系数矩阵有关, 它反映了原始数据对解的影响.

定义 2.3.5 设 A 是非奇异矩阵, 数

$$\text{Cond}(A) \triangleq \|A\| \|A^{-1}\| \quad (2.3.6)$$

称为矩阵 A 的条件数. 条件数与矩阵的范数有关, 也将条件数记为

$$\text{Cond}_p(A) \triangleq \|A\|_p \|A^{-1}\|_p, \quad p = 1, 2, \infty. \quad (2.3.6')$$

常用的矩阵条件数为

$$(1) \text{Cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty;$$

$$(2) \text{Cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 = \sqrt{\frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)}}.$$

当 A 是对称矩阵时, $\text{Cond}_2(A) = \frac{|\lambda_1|}{|\lambda_n|}$, 其中 λ_1, λ_n 分别为矩阵 A 的绝对值最大和绝对值最小的特征值.

矩阵的条件数刻画了方程组的性态, 条件数大的矩阵称为“病态”矩阵, 相应的方程组称为“病态”方程组; 条件数小的矩阵称为“良态”矩阵, 相应的方程组称为“良态”方程组.

下面给出几个病态矩阵的例子.

例 2.3.2 设 $\varepsilon > 0$,

$$A = \begin{pmatrix} 1 & 1 + \varepsilon \\ 1 - \varepsilon & 1 \end{pmatrix},$$

则

$$A^{-1} = \varepsilon^{-2} \begin{pmatrix} 1 & -1 - \varepsilon \\ -1 + \varepsilon & 1 \end{pmatrix},$$

$\|A\|_\infty = 2 + \varepsilon$, $\|A^{-1}\|_\infty = \varepsilon^{-2}(2 + \varepsilon)$, $\text{Cond}_\infty(A) = \left(\frac{2 + \varepsilon}{\varepsilon}\right)^2 > \frac{4}{\varepsilon^2}$. 若 $\varepsilon \leq 0.01$, 则 $\text{Cond}_\infty(A) \geq 40000$.

例 2.3.3 设有线性方程组

$$\begin{pmatrix} 1 & 10^5 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10^5 \\ 2 \end{pmatrix},$$

则

$$A^{-1} = \frac{1}{10^5 - 1} \begin{pmatrix} -1 & 10^5 \\ 1 & -1 \end{pmatrix},$$

$$\|A\|_\infty = 10^5 + 1, \quad \|A^{-1}\|_\infty = \frac{10^5 + 1}{10^5 - 1}, \quad \text{Cond}_\infty(A) > 10^5 \gg 1.$$

例 2.3.4 一个著名的病态矩阵的例子是 n 阶希尔伯特矩阵

$$H_n = \begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n-1} \end{pmatrix}.$$

随着矩阵 H_n 阶数 n 的增高, H_n 的条件数急剧增大, 此时, 方程组 $H_n x = b$ 是严重的病态方程组. 下面以三阶希尔伯特方程组为例说明其病态性.

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{11}{6} \\ \frac{13}{12} \\ \frac{47}{60} \end{pmatrix},$$

$$H_3^{-1} = \begin{pmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{pmatrix},$$

$$\|H_3\|_{\infty} = \frac{11}{6}, \quad \|H_3^{-1}\|_{\infty} = 408, \quad \text{Cond}_{\infty}(H_3) = \|H_3\|_{\infty} \|H_3^{-1}\|_{\infty} = 748.$$

若将方程组中的分数舍入到两位小数后求解, 则求得的近似解为 $\tilde{x} = (-6.22, 38.25, -33.65)^T$, 与真解 $x = (1, 1, 1)^T$ 相差甚远.

病态方程组的特征/判别 条件数的计算很复杂, 在实际中难以应用, 往往具有下列现象之一的方程组可能是病态方程组:

- (1) 系数矩阵的行或列近似线性相关 (如例 2.3.2);
- (2) 系数矩阵的元素数量级相差悬殊 (如例 2.3.3, 例 2.3.4);
- (3) 解对系数矩阵元素的变化比较敏感;
- (4) 在主元素消去法求解过程中, 出现量级很小的主元素;
- (5) 求出的解与预期的解相差较大.

求解病态方程组应采取的措施如下:

- (1) 采用高精度 (如双精度) 计算, 以减少舍入误差的影响;
- (2) 采用稳定性好的算法 (如全主元高斯消去法);
- (3) 采取平衡措施;
- (4) 采用迭代改善技术.

所谓平衡措施是将系数矩阵的各行除以该行绝对值最大的元素.

例如, 对例 2.3.3 的系数矩阵平衡后变为

$$A = \begin{pmatrix} 10^{-5} & 1 \\ 1 & 1 \end{pmatrix}, \quad A^{-1} = \frac{1}{1-10^{-5}} \begin{pmatrix} -1 & 1 \\ 1 & -10^{-5} \end{pmatrix},$$

$$\|A\|_{\infty} = 2, \quad \|A^{-1}\|_{\infty} = \frac{2}{1-10^{-5}}, \quad \text{Cond}_{\infty}(A) \approx 4.$$

经平衡后方程组变为良态方程组. 取 4 位小数计算, 可得解 $x_1 = 1$, $x_2 = 1$, 接近于真解.

实际上, 方程组经一次平衡后未必变为良态, 可进行多次平衡, 即将平衡措施实施于算法的每一步执行之中. 例如, 用列主元高斯消去法解方程组, 将原来每一步消元之前的直接选主元素, 改进为先将各行除以其绝对值最大的元素后, 再选主元素.

迭代改善技术适用于不是十分病态的线性方程组, 其实质是采用迭代的思想, 对近似解不断修正, 使得近似解逐渐逼近方程组的准确解.

设 $\tilde{x}$ 是方程组 $Ax = b$ 的一个近似解, 求修正量 $\Delta\tilde{x}$, 使得 $\tilde{x} + \Delta\tilde{x}$ 满足方程组 $Ax = b$, 即求解方程组

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

将上面方程组改写为

$$A\Delta\tilde{x} = b - A\tilde{x}. \quad (2.3.7)$$

记 $r = b - A\tilde{x}$, r 称为残向量, 则方程组 (2.3.7) 可以写为

$$A\Delta\tilde{x} = r. \quad (2.3.8)$$

从方程组 (2.3.8) 求得 $\Delta\tilde{x}$ 后, $\tilde{x} + \Delta\tilde{x}$ 则是方程组 $Ax = b$ 更准确的近似解. 令 $\tilde{\tilde{x}} = \tilde{x} + \Delta\tilde{x}$, 用上述方法对近似解 $\tilde{\tilde{x}}$ 再进行修正. 重复以上过程, 直到求得满足精度要求的近似解.

为了减少计算量, 应事先将 A 进行 LU 分解, 用 LU 分解解方程组 (2.3.8), 即将求解 $A\Delta\tilde{x} = r$ 改为求解两个三角方程组 $Ly = r$, $U\Delta\tilde{x} = y$. 为了提高计算精度, 应采用双精度计算残量 r .

迭代改善技术计算过程如下:

- (1) 给定误差限 $\varepsilon > 0$, 将 A 进行三角分解 $A = LU$, 用某种方法 (如列主元高斯消去法或三角分解法) 求得方程组 $Ax = b$ 的一个近似解 $x^{(0)}$, $0 \Rightarrow k$;
- (2) 用双精度计算 $r^{(k)} = b - Ax^{(k)}$;
- (3) 解方程组 $Ly = r^{(k)}$, $U\Delta x = y$, 设其解为 $\Delta x^{(k)}$;

(4) 令 $x^{(k+1)} = x^{(k)} + \Delta x^{(k)}$;

(5) 如果 $\|\Delta x^{(k)}\| \leq \varepsilon$, 或者 $\frac{\|\Delta x^{(k)}\|}{\|x^{(k+1)}\|} \leq \varepsilon$, 则取 $x^{(k+1)}$ 作为方程组 $Ax = b$ 的近似解, 停止; 否则, $k+1 \Rightarrow k$, 转步骤 (2).

2.4 正交变换与矩阵的 QR 分解

本节介绍两个最基本的正交变换——吉文斯 (Givens) 变换和豪斯霍尔德 (Householder) 变换, 并由此导出矩阵的 QR 分解. 它们在数值线性代数中有着广泛的应用, 是许多重要算法的基础.

2.4.1 吉文斯变换与豪斯霍尔德变换

1. 吉文斯变换

定义 2.4.1 对某个 θ , 记 $s = \sin \theta$, $c = \cos \theta$, 设 $1 \leq i < j \leq n$, 矩阵

$$P(i, j, \theta) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & c & & & s \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ & & & -s & & & c \\ & & & & & & & 1 \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix}_{n \times n}$$

称为吉文斯变换、吉文斯变换阵或平面旋转阵, θ 称为转角.

它是把 n 阶单位矩阵 I 的第 i, j 行的对角元改为 c , 将第 i 行第 j 列的元素改为 s , 第 j 行第 i 列的元素改为 $-s$ 后形成的矩阵. 从几何上来看, $P(i, j, \theta)x$ 是在 (i, j) 坐标平面内将向量 x 按顺时针方向作了 θ 度的旋转, 所以吉文斯变换也称为平面旋转变换.

吉文斯矩阵具有以下性质:

性质 2.4.1 $P(i, j, \theta)$ 是正交矩阵, 即 $P(i, j, \theta)^{-1} = P(i, j, \theta)^T$.

为了书写简便, 以下将 $P(i, j, \theta)$ 记为 P_{ij} .

性质 2.4.2 设 $x, y \in \mathbf{R}^n$, 则吉文斯变换 $y = P_{ij}x$ 只改变 x 的第 i 个和第 j 个分量.

事实上, 设 $x = (x_1, x_2, \dots, x_n)^T$, $y = (y_1, y_2, \dots, y_n)^T$, 由 $y = P_{ij}x$ 可得

$$y_i = cx_i + sx_j, \quad y_j = -sx_i + cx_j, \quad y_k = x_k, \quad k \neq i, j.$$

吉文斯变换常用于将给定向量 x 的某个 (如第 j 个) 分量化为零. 由性质 2.4.2 知, 当 $x_i^2 + x_j^2 \neq 0$ 时, 欲使 $y_j = 0$, 只要选取

$$c = \frac{x_i}{\sqrt{x_i^2 + x_j^2}}, \quad s = \frac{x_j}{\sqrt{x_i^2 + x_j^2}},$$

就有 $y_i = \sqrt{x_i^2 + x_j^2} > 0$, $y_j = 0$.

性质 2.4.3 设 A 是 n 阶矩阵, 若用 P_{ij} 左乘矩阵 A , 则只需计算第 i 行和第 j 行的元素.

令 $B = P_{ij}A$, 则有 $b_{lk} = a_{lk}$ ($l \neq i, j$),

$$b_{ik} = ca_{ik} + sa_{jk}, \quad b_{jk} = -sa_{ik} + ca_{jk}, \quad k = 1, 2, \dots, n,$$

即

$$\begin{pmatrix} b_{ik} \\ b_{jk} \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a_{ik} \\ a_{jk} \end{pmatrix}, \quad k = 1, 2, \dots, n.$$

性质 2.4.4 若用 P_{ij} 右乘 n 阶矩阵 A , 则只需计算第 i 列和第 j 列的元素.

令 $B = AP_{ij}$, 则有 $b_{kl} = a_{kl}$ ($l \neq i, j$),

$$b_{ki} = ca_{ki} - sa_{kj}, \quad b_{kj} = sa_{ki} + ca_{kj}, \quad k = 1, 2, \dots, n,$$

即

$$(b_{ki}, b_{kj}) = (a_{ki}, a_{kj}) \begin{pmatrix} c & s \\ -s & c \end{pmatrix}, \quad k = 1, 2, \dots, n.$$

定理 2.4.1 设 $x \in \mathbf{R}^n$ 是任意给定的非零向量, 则存在有限个吉文斯矩阵的乘积 P , 使得 $Px = \|x\|_2 e_1$, 其中 $e_1 = (1, 0, \dots, 0)^T \in \mathbf{R}^n$.

证明 (1) 若 $x_1 \neq 0$, 根据性质 2.4.2, 依次构造

$$P_{12}: \quad c = \frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \quad s = \frac{x_2}{\sqrt{x_1^2 + x_2^2}},$$

则

$$P_{12}x = \left(\sqrt{x_1^2 + x_2^2}, 0, x_3, \dots, x_n \right)^T;$$

$$P_{13}: \quad c = \frac{\sqrt{x_1^2 + x_2^2}}{\sqrt{x_1^2 + x_2^2 + x_3^2}}, \quad s = \frac{x_3}{\sqrt{x_1^2 + x_2^2 + x_3^2}},$$

则

$$P_{13}(P_{12}x) = \left(\sqrt{x_1^2 + x_2^2 + x_3^2}, 0, 0, x_4, \dots, x_n \right)^T;$$

.....

$$P_{1n}: \quad c = \frac{\sqrt{x_1^2 + x_2^2 + \dots + x_{n-1}^2}}{\sqrt{x_1^2 + x_2^2 + \dots + x_n^2}}, \quad s = \frac{x_n}{\sqrt{x_1^2 + x_2^2 + \dots + x_n^2}},$$

则

$$P_{1n}(P_{1,n-1} \cdots P_{12}x) = \left(\sqrt{x_1^2 + x_2^2 + \dots + x_n^2}, 0, \dots, 0 \right)^T.$$

记 $P = P_{1n}P_{1,n-1} \cdots P_{12}$, 则

$$Px = \left(\sqrt{x_1^2 + x_2^2 + \dots + x_n^2}, 0, \dots, 0 \right)^T = \|x\|e_1.$$

(2) 若 $x_1 = \dots = x_{k-1} = 0, x_k \neq 0$ ($1 < k \leq n$), 则由 $P_{1,k}$ 开始即可. 证毕.

2. 豪斯霍尔德变换

定义 2.4.2 设 $u \in \mathbf{R}^n$ 且 $\|u\|_2 = 1$, 矩阵 $H = I - 2uu^T$ 称为初等反射阵、豪斯霍尔德变换或豪斯霍尔德矩阵.

显然, H 是一个对称的正交矩阵. 由 H 是正交矩阵知, 对任意的 $x \in \mathbf{R}^n$, 令 $y = Hx$, 则 $\|y\|_2 = \|Hx\|_2 = \|x\|_2$, 这说明豪斯霍尔德变换不改变向量的长度.

豪斯霍尔德变换的几何意义. 当 $n = 3$ 时, 如图 2.6 所示. 设 $u \in \mathbf{R}^3$ 且 $\|u\|_2 = 1$, π 是过原点以 u 为法向量的平面. 对任意的 $x \in \mathbf{R}^3$, x 可分解为 $x = x^{(1)} + x^{(2)}$, 其中 $x^{(1)} \in \pi, x^{(2)} \perp \pi$. 注意到 $u^T x^{(1)} = 0, x^{(2)} = ku$, 则有 $Hx^{(1)} = x^{(1)}, Hx^{(2)} = -x^{(2)}$, 故 $Hx = H(x^{(1)} + x^{(2)}) = x^{(1)} - x^{(2)}$. 这说明豪斯霍尔德变换后的向量 Hx 是 x 关于平面 π 对称的向量. 几何学上称 Hx 为关于平面 π 的镜面反射, 故称 H 为初等反射阵.

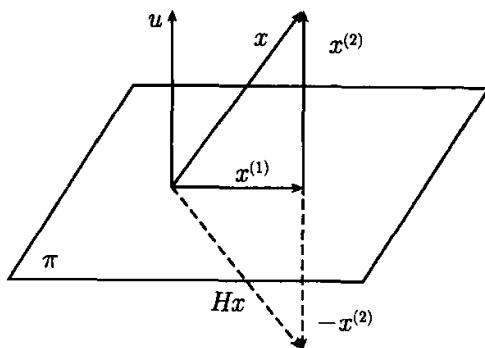


图 2.6

定理 2.4.2 设 $x \in \mathbf{R}^n$ 是任意给定的非零向量, $v \in \mathbf{R}^n$ 是任意给定的单位向量, 则存在初等反射阵 $H = I - 2uu^T$, 使得 $Hx = \sigma v$, 其中 σ 为常数.

证明 $\sigma^2 v^T v = x^T H^T H x = x^T x = \|x\|_2^2$. 由于 v 是单位向量, 所以 $\sigma^2 = \|x\|_2^2$, 故 $\sigma = \pm \|x\|_2$. 又 $Hx = (I - 2uu^T)x = x - 2(u^T x)u = \sigma v$, 所以 $2(u^T x)u = x - \sigma v$. 这说明向量 u 与 $x - \sigma v$ 的方向相同, 注意到 u 是单位向量, 故取

$$u = \frac{x - \sigma v}{\|x - \sigma v\|_2}, \quad (2.4.1)$$

则由 u 确定的矩阵 H 必然满足 $Hx = \sigma v$. 证毕.

设 $x = (x_1, x_2, \dots, x_n)^T$, 为了避免 $x - \sigma v$ 时损失有效数字, 宜取 $\sigma = -\text{sgn}(x_1)\|x\|_2$.

从以上两种变换可知, 若要将向量中的某个分量化为零, 则利用吉文斯变换; 若要将若干个相邻分量化为零, 则用豪斯霍尔德变换.

2.4.2 矩阵的 QR 分解

定义 2.4.3 设 $A \in \mathbf{R}^{m \times n} (m \geq n)$, 若存在正交阵 $Q \in \mathbf{R}^{m \times m}$ 和从第 $n+1$ 行到 m 行均是零元素的上梯形矩阵 $R \in \mathbf{R}^{m \times n}$, 使得 $A = QR$, 则称 QR 为矩阵 A 的 QR 分解.

下面分别介绍利用吉文斯变换和豪斯霍尔德变换对矩阵 A 进行 QR 分解.

1. 基于吉文斯变换的 QR 分解

设 A 是一个 $m \times n$ ($m \geq n$) 阶矩阵且它的列向量线性无关, 则利用吉文斯变换可以把 A 逐步化为上梯形矩阵. 设

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = (a_1, a_2, \dots, a_n) \quad (\text{按列分块}).$$

为了符号统一起见, 记

$$A^{(0)} = (a_1^{(0)}, a_2^{(0)}, \dots, a_n^{(0)}) = A.$$

具体变换过程如下:

第 1 步 由 A 的列向量线性无关知 $a_1^{(0)} \neq 0$, 根据定理 2.4.1, 可构造有限个 m 阶吉文斯矩阵的乘积 P_1 , 使得

$$P_1 a_1^{(0)} = \|a_1^{(0)}\|_2 e_1, \quad e_1 \in \mathbf{R}^m.$$

记 $a_{11}^{(1)} = \|a_1^{(0)}\|_2$, 用 P_1 左乘 $A^{(0)}$ 得

$$A^{(1)} = P_1 A^{(0)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a_{m2}^{(1)} & a_{m3}^{(1)} & \cdots & a_{mn}^{(1)} \end{pmatrix}.$$

第 2 步 记 $a_2^{(1)} = (a_{12}^{(1)}, a_{22}^{(1)}, \dots, a_{m2}^{(1)})^T = (a_{12}^{(1)}, \tilde{a}_2^{(1)T})^T$. 由 A 的列向量线性无关易知 $\tilde{a}_2^{(1)} \neq 0$, 构造有限个 $m-1$ 阶吉文斯矩阵的乘积 $\tilde{P}_2$, 使得

$$\tilde{P}_2 \tilde{a}_2^{(1)} = \|\tilde{a}_2^{(1)}\|_2 e_1, \quad e_1 \in \mathbf{R}^{m-1}.$$

令

$$P_2 = \begin{pmatrix} 1 & 0^T \\ 0 & \tilde{P}_2 \end{pmatrix},$$

用 P_2 左乘 $A^{(1)}$ 得

$$A^{(2)} = P_2 A^{(1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{m3}^{(2)} & a_{m4}^{(2)} & \cdots & a_{mn}^{(2)} \end{pmatrix}.$$

重复以上过程, 假定已经对矩阵 A 进行了 $k-1$ 次变换, 即构造 $P_1, P_2, \dots, P_{k-1}$, 则

$$\begin{aligned} A^{(k-1)} &= P_{k-1} P_{k-2} \cdots P_1 A^{(0)} \\ &= \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & a_{2,k+1}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots & \vdots & \vdots & & \vdots \\ & & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & a_{k-1,k+1}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ & & & & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ & & & & \vdots & \vdots & & \vdots \\ & & & & a_{mk}^{(k-1)} & a_{m,k+1}^{(k-1)} & \cdots & a_{mn}^{(k-1)} \end{pmatrix}. \end{aligned}$$

第 k 步 记

$$a_k^{(k-1)} = (a_{1k}^{(1)}, a_{2k}^{(2)}, \dots, a_{k-1,k}^{(k-1)}, a_{kk}^{(k-1)}, \dots, a_{mk}^{(k-1)})^T = (a_{1k}^{(1)}, a_{2k}^{(2)}, \dots, a_{k-1,k}^{(k-1)}, \tilde{a}_k^{(k-1)T})^T.$$

显然, $\tilde{a}_k^{(k-1)} \neq 0$, 构造有限个 $(m-k+1)$ 阶吉文斯矩阵的乘积 $\tilde{P}_k$, 使得

$$\tilde{P}_k \tilde{a}_k^{(k-1)} = \|\tilde{a}_k^{(k-1)}\|_2 e_1, \quad e_1 \in \mathbf{R}^{m-k+1}.$$

令

$$P_k = \begin{pmatrix} I_{k-1} & O^T \\ O & \tilde{P}_k \end{pmatrix},$$

用 P_k 左乘 $A^{(k-1)}$ 得

$$\begin{aligned} A^{(k)} &= P_k P_{k-1} \cdots P_1 A^{(0)} \\ &= \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & a_{1,k+2}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & a_{2,k+1}^{(2)} & a_{2,k+2}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ & & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & a_{k-1,k+1}^{(k-1)} & a_{k-1,k+2}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ & & & & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & a_{k,k+2}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & & & a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ & & & & & \vdots & \vdots & & \vdots \\ & & & & & & a_{m,k+1}^{(k)} & a_{m,k+2}^{(k)} & \cdots & a_{mn}^{(k)} \end{pmatrix}. \end{aligned}$$

继续上述变换过程, 直至 $s = \min\{m-1, n\}$ 步, 最后得到

$$A^{(s)} = P_s P_{s-1} \cdots P_1 A = R. \quad (2.4.2)$$

由于每个 P_i 都是正交矩阵, 所以 $P = P_s P_{s-1} \cdots P_1$ 是 m 阶正交矩阵, 故式 (2.4.2) 可以写为

$$PA = R. \quad (2.4.3)$$

当 $m = n$ 时, R 是 n 阶上三角矩阵, 这时式 (2.4.3) 可以写为

$$A = QR, \quad (2.4.4)$$

其中 $Q = P^T$ 为 n 阶正交矩阵, 即把 n 阶矩阵 A 分解成一个 n 阶正交矩阵 Q 与一个可逆的 n 阶上三角矩阵 R 的乘积.

当 $m > n$ 时, R 是 $m \times n$ 阶的上梯形矩阵. 这时, 式 (2.4.3) 为

$$PA = R = \begin{pmatrix} R_1 \\ O \end{pmatrix},$$

其中 R_1 为 n 阶可逆的上三角矩阵. 若令 $Q = P^T = (Q_1, Q_2)$, 则 Q 是 m 阶正交矩阵, 其中 Q_1 具有 n 个正交列, 于是有

$$A = (Q_1, Q_2) \begin{pmatrix} R_1 \\ O \end{pmatrix} = Q_1 R_1. \quad (2.4.5)$$

式 (2.4.5) 实现了 $m \times n$ 阶矩阵 A 的一个正交三角分解.

例 2.4.1 用吉文斯变换作矩阵

$$A = \begin{pmatrix} 3 & 5 & 5 \\ 0 & 3 & 4 \\ 4 & 0 & 5 \end{pmatrix}$$

的 QR 分解.

解 第 1 步 为将第 1 列中的元素 $a_{31} = 4$ 消为零, 取 $c = \frac{3}{5}$, $s = \frac{4}{5}$. 构造

$$P_{13} = \begin{pmatrix} \frac{3}{5} & 0 & \frac{4}{5} \\ 0 & 1 & 0 \\ -\frac{4}{5} & 0 & \frac{3}{5} \end{pmatrix},$$

取 $P_1 = P_{13}$, 则

$$A^{(1)} = P_1 A = \frac{1}{5} \begin{pmatrix} 3 & 0 & 4 \\ 0 & 5 & 0 \\ -4 & 0 & 3 \end{pmatrix} \begin{pmatrix} 3 & 5 & 5 \\ 0 & 3 & 4 \\ 4 & 0 & 5 \end{pmatrix} = \begin{pmatrix} 5 & 3 & 7 \\ 0 & 3 & 4 \\ 0 & -4 & -1 \end{pmatrix}.$$

第 2 步 为将矩阵 $A^{(1)}$ 第 2 列中的元素 $a_{32}^{(1)} = -4$ 消为零, 针对子阵

$$\begin{pmatrix} 3 & 4 \\ -4 & -1 \end{pmatrix}$$

构造 2 阶吉文斯矩阵

$$\tilde{P}_{12} = \begin{pmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix},$$

取

$$P_2 = \begin{pmatrix} 1 & 0^T \\ 0 & \tilde{P}_{12} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{5} & -\frac{4}{5} \\ 0 & \frac{4}{5} & \frac{3}{5} \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & -4 \\ 0 & 4 & 3 \end{pmatrix},$$

则

$$\begin{aligned}
 A^{(2)} &= P_2 A^{(1)} = P_2 P_1 A \\
 &= \frac{1}{5} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & -4 \\ 0 & 4 & 3 \end{pmatrix} \begin{pmatrix} 5 & 3 & 7 \\ 0 & 3 & 4 \\ 0 & -4 & -1 \end{pmatrix} = \begin{pmatrix} 5 & 3 & 7 \\ 0 & 5 & \frac{16}{5} \\ 0 & 0 & \frac{13}{5} \end{pmatrix} = R.
 \end{aligned}$$

令 $P = P_2 P_1$, 则

$$P = \frac{1}{25} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & -4 \\ 0 & 4 & 3 \end{pmatrix} \begin{pmatrix} 3 & 0 & 4 \\ 0 & 5 & 0 \\ -4 & 0 & 3 \end{pmatrix} = \frac{1}{25} \begin{pmatrix} 15 & 0 & 20 \\ 16 & 15 & -12 \\ -12 & 20 & 9 \end{pmatrix},$$

则有

$$Q = P^T = \frac{1}{25} \begin{pmatrix} 15 & 16 & -12 \\ 0 & 15 & 20 \\ 20 & -12 & 9 \end{pmatrix}, \quad R = \begin{pmatrix} 5 & 3 & 7 \\ 0 & 5 & \frac{16}{5} \\ 0 & 0 & \frac{13}{5} \end{pmatrix}, \quad A = QR.$$

解毕.

2. 基于豪斯霍尔德变换的 QR 分解

设 A 是一个 $m \times n$ ($m \geq n$) 阶矩阵且它的列向量线性无关, 则利用豪斯霍尔德变换可以把 A 逐步化为上梯形矩阵. 设

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = (a_1, a_2, \cdots, a_n).$$

记

$$A^{(0)} = (a_1^{(0)}, a_2^{(0)}, \cdots, a_n^{(0)}) = A.$$

具体变换过程如下:

设 e_i ($i = 1, 2, \cdots, n$) 是 m 维单位坐标向量.

第 1 步 为使把矩阵 A 的第 1 列 $a_1^{(0)}$ 化为 $(\sigma_1, 0, \cdots, 0)^T$, 取 $x = a_1^{(0)}, v = e_1 = (1, 0, \cdots, 0)^T \in \mathbf{R}^m$ (或取 $v = -e_1$). 根据式 (2.4.1), 取

$$u_1 = \frac{a_1^{(0)} - \sigma_1 e_1}{\|a_1^{(0)} - \sigma_1 e_1\|_2} = \frac{w_1}{\|w_1\|_2},$$

其中

$$\sigma_1 = -\operatorname{sgn}(a_{11}^{(0)}) \|a_1^{(0)}\|_2, \quad w_1 = a_1^{(0)} - \sigma_1 e_1.$$

而

$$\begin{aligned} \|w_1\|_2 &= \|a_1^{(0)} - \sigma_1 e_1\|_2 = \sqrt{2(\sigma_1^2 - \sigma_1 a_{11}^{(0)})}, \\ 2u_1 u_1^T &= \frac{2w_1 w_1^T}{\|w_1\|_2^2} = \frac{w_1 w_1^T}{\sigma_1(\sigma_1 - a_{11}^{(0)})} = \frac{w_1 w_1^T}{\alpha_1}, \quad \alpha_1 = \sigma_1(\sigma_1 - a_{11}^{(0)}). \end{aligned}$$

令 $H_1 = I_m - 2u_1 u_1^T = I_m - \alpha_1^{-1} w_1 w_1^T$, 用 H_1 左乘 $A^{(0)}$ 得

$$\begin{aligned} A^{(1)} &= H_1 A^{(0)} = (H_1 a_1^{(0)}, H_1 a_2^{(0)}, \dots, H_1 a_n^{(0)}) \\ &= (\sigma_1 e_1, a_2^{(1)}, a_3^{(1)}, \dots, a_n^{(1)}) \\ &= \begin{pmatrix} \sigma_1 & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a_{m2}^{(1)} & a_{m3}^{(1)} & \cdots & a_{mn}^{(1)} \end{pmatrix}, \end{aligned}$$

其中

$$\begin{aligned} a_j^{(1)} &= H_1 a_j^{(0)} = (I_m - \alpha_1^{-1} w_1 w_1^T) a_j^{(0)} = a_j^{(0)} - \alpha_1^{-1} w_1^T a_j^{(0)} w_1 \\ &= a_j^{(0)} - \beta_{1j} w_1 = a_j^{(0)} - \beta_{1j} (a_1^{(0)} - \sigma_1 e_1) \\ &= \begin{pmatrix} a_{1j}^{(0)} - \beta_{1j} (a_{11}^{(0)} - \sigma_1) \\ a_{2j}^{(0)} - \beta_{1j} a_{21}^{(0)} \\ \vdots \\ a_{mj}^{(0)} - \beta_{1j} a_{m1}^{(0)} \end{pmatrix}, \quad j = 2, 3, \dots, n. \end{aligned}$$

由此可知

$$\begin{aligned} a_{1j}^{(1)} &= a_{1j}^{(0)} - \beta_{1j} (a_{11}^{(0)} - \sigma_1), \quad j = 2, 3, \dots, n, \\ a_{ij}^{(1)} &= a_{ij}^{(0)} - \beta_{1j} a_{i1}^{(0)}, \quad i = 2, 3, \dots, m; j = 2, 3, \dots, n. \end{aligned}$$

而

$$\begin{aligned} \beta_{1j} &= \alpha_1^{-1} w_1^T a_j^{(0)} = \alpha_1^{-1} (a_1^{(0)} - \sigma_1 e_1)^T a_j^{(0)} \\ &= \alpha_1^{-1} (a_1^{(0)T} a_j^{(0)} - \sigma_1 a_{1j}^{(0)}) = \alpha_1^{-1} \left(\sum_{i=1}^m a_{i1}^{(0)} a_{ij}^{(0)} - \sigma_1 a_{1j}^{(0)} \right) \\ &= \alpha_1^{-1} \left((a_{11}^{(0)} - \sigma_1) a_{1j}^{(0)} + \sum_{i=2}^m a_{i1}^{(0)} a_{ij}^{(0)} \right), \quad j = 2, 3, \dots, n. \end{aligned}$$

第 2 步 记 $a_2^{(1)} = (a_{12}^{(1)}, a_{22}^{(1)}, \dots, a_{m2}^{(1)})^T = (a_{12}^{(1)}, \tilde{a}_2^{(1)T})^T$, 显然, $\tilde{a}_2^{(2)} \neq 0$. 取 $x = \tilde{a}_2^{(1)}, v = e_1^{(1)}$ (或取 $v = -e_1^{(1)}$), 其中 $e_1^{(1)}$ 为 $e_2 = (0, e_1^{(1)T})^T$ 中的 $m-1$ 维单位坐标向量. 这时, 取

$$u_2 = \frac{\tilde{a}_2^{(1)} - \sigma_2 e_1^{(1)}}{\|\tilde{a}_2^{(1)} - \sigma_2 e_1^{(1)}\|_2} = \frac{w_2}{\|w_2\|_2},$$

其中

$$\sigma_2 = -\operatorname{sgn}(a_{22}^{(1)}) \|\tilde{a}_2^{(1)}\|_2, \quad w_2 = \tilde{a}_2^{(1)} - \sigma_2 e_1^{(1)}.$$

而

$$\begin{aligned} \|w_2\|_2 &= \|\tilde{a}_2^{(1)} - \sigma_2 e_1^{(1)}\|_2 = \sqrt{2(\sigma_2^2 - \sigma_2 a_{22}^{(1)})}, \\ 2u_2 u_2^T &= \frac{2w_2 w_2^T}{\|w_2\|_2^2} = \frac{w_2 w_2^T}{\sigma_2(\sigma_2 - a_{22}^{(1)})} = \frac{w_2 w_2^T}{\alpha_2}, \end{aligned}$$

其中

$$\alpha_2 = \sigma_2(\sigma_2 - a_{22}^{(1)}).$$

构造

$$\tilde{H}_2 = I_{m-1} - 2u_2 u_2^T = I_{m-1} - \alpha_2^{-1} w_2 w_2^T,$$

令

$$H_2 = \begin{pmatrix} 1 & 0^T \\ 0 & \tilde{H}_2 \end{pmatrix}.$$

用 H_2 左乘 $A^{(1)}$ 得

$$\begin{aligned} A^{(2)} &= H_2 A^{(1)} = (\sigma_1 e_1, a_{12}^{(1)} e_1 + \sigma_2 e_2, a_3^{(2)}, \dots, a_n^{(2)}) \\ &= \begin{pmatrix} \sigma_1 & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & \sigma_2 & a_{23}^{(2)} & a_{24}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{m3}^{(2)} & a_{m4}^{(2)} & \cdots & a_{mn}^{(2)} \end{pmatrix}. \end{aligned}$$

重复以上过程, 假定已经对矩阵 A 进行了 $k-1$ 次变换, 即存在初等反射阵 $H_1, H_2, \dots, H_{k-1}$, 使得

$$\begin{aligned} A^{(k-1)} &= H_{k-1} H_{k-2} \cdots H_1 A \\ &= \left(\sigma_1 e_1, a_{12}^{(2)} e_1 + \sigma_2 e_2, \dots, \sum_{j=1}^{k-2} a_{j,k-1}^{(j)} e_j + \sigma_{k-1} e_{k-1}, a_k^{(k-1)}, \dots, a_n^{(k-1)} \right) \end{aligned}$$

$$= \begin{pmatrix} \sigma_1 & a_{12}^{(1)} & \cdots & a_{1,k-2}^{(1)} & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1n}^{(1)} \\ & \sigma_2 & \cdots & a_{2,k-2}^{(2)} & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & a_{2,k+1}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ & & & \sigma_{k-2} & a_{k-2,k-1}^{(k-2)} & a_{k-2,k}^{(k-2)} & a_{k-2,k+1}^{(k-2)} & \cdots & a_{k-2,n}^{(k-2)} \\ & & & & \sigma_{k-1} & a_{k-1,k}^{(k-1)} & a_{k-1,k+1}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ & & & & & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ & & & & & \vdots & \vdots & & \vdots \\ & & & & & & a_{mk}^{(k-1)} & a_{m,k+1}^{(k-1)} & \cdots & a_{mn}^{(k-1)} \end{pmatrix}.$$

第 k 步 记

$$\begin{aligned} a_k^{(k-1)} &= (a_{1k}^{(1)}, a_{2k}^{(2)}, \cdots, a_{k-1,k}^{(k-1)}, a_{kk}^{(k-1)}, \cdots, a_{mk}^{(k-1)})^T \\ &= (a_{1k}^{(1)}, a_{2k}^{(2)}, \cdots, a_{k-1,k}^{(k-1)}, \tilde{a}_k^{(k-1)})^T, \\ \tilde{a}_j^{(k-1)} &= (a_{kj}^{(k-1)}, a_{k+1,j}^{(k-1)}, \cdots, a_{mj}^{(k-1)})^T, \quad j = k, k+1, \cdots, n. \end{aligned}$$

显然, $\tilde{a}_k^{(k-1)} \neq 0$, 取 $x = \tilde{a}_k^{(k-1)}$, $v = e_1^{(k-1)}$ (或取 $v = -e_1^{(k-1)}$), 其中 $e_1^{(k-1)}$ 为 $e_k = (\underbrace{0, \cdots, 0}_{k-1 \text{ 个}}, e_1^{(k-1)})^T$ 中的 $m-k+1$ 维单位坐标向量. 这时取

$$u_k = \frac{\tilde{a}_k^{(k-1)} - \sigma_k e_1^{(k-1)}}{\|\tilde{a}_k^{(k-1)} - \sigma_k e_1^{(k-1)}\|_2} = \frac{w_k}{\|w_k\|_2},$$

其中

$$\sigma_k = -\operatorname{sgn}(a_{kk}^{(k-1)}) \|\tilde{a}_k^{(k-1)}\|_2, \quad w_k = \tilde{a}_k^{(k-1)} - \sigma_k e_1^{(k-1)}.$$

而

$$\begin{aligned} \|w_k\|_2 &= \|\tilde{a}_k^{(k-1)} - \sigma_k e_1^{(k-1)}\|_2 = \sqrt{2(\sigma_k^2 - \sigma_k a_{kk}^{(k-1)})}, \\ 2u_k u_k^T &= \frac{2w_k w_k^T}{\|w_k\|_2^2} = \frac{w_k w_k^T}{\sigma_k(\sigma_k - a_{kk}^{(k-1)})} = \frac{w_k w_k^T}{\alpha_k}, \end{aligned}$$

其中

$$\alpha_k = \sigma_k(\sigma_k - a_{kk}^{(k-1)}).$$

构造

$$\tilde{H}_k = I_{m-k+1} - 2u_k u_k^T = I_{m-k+1} - \alpha_k^{-1} w_k w_k^T,$$

令

$$H_k = \begin{pmatrix} I_{k-1} & O^T \\ O & \tilde{H}_k \end{pmatrix},$$

用 H_k 左乘 $A^{(k-1)}$ 得

$$\begin{aligned} A^{(k)} &= H_k A^{(k-1)} = H_k H_{k-1} \cdots H_1 A \\ &= \begin{pmatrix} \sigma_1 e_1, a_{12}^{(2)} e_1 + \sigma_2 e_2, \cdots, \sum_{j=1}^{k-1} a_{j,k}^{(j)} e_j + \sigma_k e_k, a_{k+1}^{(k)}, \cdots, a_n^{(k)} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1 & a_{12}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & a_{1,k+2}^{(1)} & \cdots & a_{1n}^{(1)} \\ & \sigma_2 & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & a_{2,k+1}^{(2)} & a_{2,k+2}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ & & & \sigma_{k-1} & a_{k-1,k}^{(k-1)} & a_{k-1,k+1}^{(k-1)} & a_{k-1,k+2}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ & & & & \sigma_k & a_{k,k+1}^{(k)} & a_{k,k+2}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & & & a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ & & & & & \vdots & \vdots & & \vdots \\ & & & & & & a_{m,k+1}^{(k)} & a_{m,k+2}^{(k)} & \cdots & a_{mn}^{(k)} \end{pmatrix}, \end{aligned}$$

其中

$$\begin{aligned} (a_{kj}^{(k)}, a_{k+1,j}^{(k)}, \cdots, a_{mj}^{(k)})^T &= \tilde{H}_k \tilde{a}_j^{(k-1)} = (I_{m-k+1} - \alpha_k^{-1} w_k w_k^T) \tilde{a}_j^{(k-1)} \\ &= \tilde{a}_j^{(k-1)} - \beta_{kj} w_k \\ &= \tilde{a}_j^{(k-1)} - \beta_{kj} (\tilde{a}_k^{(k-1)} - \sigma_k e_1^{(k-1)}), \quad j = k+1, \cdots, n. \end{aligned}$$

由此即得

$$\begin{aligned} a_{kj}^{(k)} &= a_{kj}^{(k-1)} - \beta_{kj} (a_{kk}^{(k-1)} - \sigma_k), \quad j = k+1, \cdots, n, \\ a_{ij}^{(k)} &= a_{ij}^{(k-1)} - \beta_{kj} a_{ik}^{(k-1)}, \quad i = k+1, \cdots, m; j = k+1, \cdots, n, \end{aligned}$$

其中

$$\beta_{kj} = \alpha_k^{-1} w_k^T \tilde{a}_j^{(k-1)} = \alpha_k^{-1} \left((a_{kk}^{(k-1)} - \sigma_k) a_{kj}^{(k-1)} + \sum_{i=k+1}^m a_{ik}^{(k-1)} a_{ij}^{(k-1)} \right), \quad j = k+1, \cdots, n.$$

继续上述变换过程, 直至 $s = \min\{m-1, n\}$ 步, 最后得到

$$A^{(s)} = H_s H_{s-1} \cdots H_1 A = R.$$

由于每个 H_i 都是正交矩阵, 所以 $H = H_s H_{s-1} \cdots H_1$ 也是正交矩阵. 上式可以写为

$$HA = R. \quad (2.4.6)$$

当 $m = n$ 时, R 是 n 阶上三角矩阵. 这时, 式 (2.4.6) 可以写为

$$A = QR, \quad (2.4.7)$$

其中 $Q = H^T$ 为 n 阶正交矩阵, 即把 n 阶矩阵 A 分解为一个 n 阶正交矩阵与一个可逆的 n 阶上三角矩阵 R 的乘积.

当 $m > n$ 时, R 是 $m \times n$ 阶的上梯形矩阵. 这时, 式 (2.4.6) 为

$$HA = R = \begin{pmatrix} R_1 \\ O \end{pmatrix},$$

其中 R_1 为可逆的 n 阶上三角矩阵. 若令 $Q = H^T = (Q_1, Q_2)$, 则 Q 是 m 阶正交矩阵, 其中 Q_1 具有 n 个正交列, 于是有

$$A = QR = (Q_1, Q_2) \begin{pmatrix} R_1 \\ O \end{pmatrix} = Q_1 R_1. \quad (2.4.8)$$

式 (2.4.8) 实现了 $m \times n$ 阶矩阵 A 的一个正交三角分解.

总结以上基于两种变换的矩阵 QR 分解的讨论, 可得下述定理:

定理 2.4.3 (1) 设 A 是 n 阶非奇异矩阵, 则存在 n 阶正交矩阵 Q 和 n 阶非奇异上三角矩阵 R , 使得

$$A = QR,$$

当限定 R 的对角元为正数时, 这种分解是唯一的;

(2) 设 A 是 $m \times n$ ($m > n$) 阶矩阵, $\text{rank}(A) = n$, 则存在 m 阶正交矩阵 Q 和 n 阶非奇异上三角矩阵 R_1 , 使得

$$A = QR = (Q_1, Q_2) \begin{pmatrix} R_1 \\ O \end{pmatrix} = Q_1 R_1,$$

其中 R 为 $m \times n$ 阶的上梯形矩阵, Q_1 表示由 Q 的前 n 列构成的 $m \times n$ 阶矩阵.

证明 (1) 前面的讨论已表明矩阵 A 能作 QR 分解, 现证其唯一性. 设矩阵 A 有两种分解法, 即

$$A = QR = \tilde{Q}\tilde{R}. \quad (2.4.9)$$

由 A 是非奇异矩阵知, $R, \tilde{R}$ 均是非奇异矩阵, 并且 $R, \tilde{R}$ 的对角元均为正数, 而 $Q, \tilde{Q}$ 是正交矩阵, 所以

$$A^T A = R^T Q^T Q R = R^T R,$$

$$A^T A = \tilde{R}^T \tilde{Q}^T \tilde{Q} \tilde{R} = \tilde{R}^T \tilde{R}.$$

由于 $A^T A$ 是对称正定矩阵, 由定理 2.2.3 知, $A^T A$ 的楚列斯基分解是唯一的. 由以上两式知 $R = \tilde{R}$, 再由 (2.4.9) 知 $Q = \tilde{Q}$.

(2) 类似前面的证明即得结论. 证毕.

在基于豪斯霍尔德变换的矩阵分解过程中, $\sigma = \pm \|x\|_2$, 即可取 $\sigma = \|x\|_2$, 也可取为 $\sigma = -\|x\|_2$. 而上三角矩阵 R_1 的对角元为 σ_k ($k = 1, 2, \dots, n$). 由此可得下述定理:

定理 2.4.4 设 A 是 $m \times n$ ($m \geq n$) 阶矩阵, $\text{rank}(A) = n$, 则基于豪斯霍尔德变换的矩阵 QR 分解

$$A = QR = (Q_1, Q_2) \begin{pmatrix} R_1 \\ O \end{pmatrix} = Q_1 R_1$$

几乎是唯一的 (即分解式的因子除去相差一个对角元模为 1 的对角矩阵外是唯一的).

用吉文斯变换实现矩阵 QR 分解的运算量大约是用豪斯霍尔德变换的 2 倍, 但若矩阵 A 含有大量的零元素, 则灵活地使用吉文斯变换会使运算量大为减少.

矩阵 QR 分解常用于求解病态线性方程组 $Ax = b$, 即求解

$$y = Q^T b, \quad Rx = y.$$

它的计算过程是非常稳定的, 得到的结果往往要比三角分解法好得多. QR 分解也常用于求解线性最小二乘问题、求矩阵全部特征值的 QR 方法等许多其他问题中, 它是数值代数中许多重要算法的基础.

下面给出基于豪斯霍尔德变换化矩阵 A 为上梯形矩阵的计算过程.

为了减少存储量, 可将上三角矩阵 R_1 的上三角部分的元素存放在矩阵 A 的相应位置上. 每一步计算中的豪斯霍尔德矩阵 H_k 是由向量 w_k 和数量 α_k 所生成, 故不必储存矩阵 H_k , 只需储存向量 w_k 和数量 α_k 即可. 将矩阵 A 消为零的元素的位置用来存放 w_k 的分量, 但在第 k 步计算中, w_k 是 $m - k + 1$ 维向量, 而矩阵 A 消为零的元素的个数为 $m - k$, 用来存放 w_k 还少一个单元. 因此, 或者将 w_k 的第一个分量存放在别的位置上, 或者将 R_1 的对角元 $\sigma_1, \sigma_2, \dots, \sigma_n$ 另行存放, 空出 A 的对角元的位置用来存放 w_k 的第一个分量. 下面的计算过程采用的是后一种储存方法, 将 R_1 的对角元 $\sigma_1, \sigma_2, \dots, \sigma_n$ 依次存放在 $d_1, d_2, \dots, d_n$ 中. 具体计算过程如下:

输出生成豪斯霍尔德矩阵的向量 $w_1, w_2, \dots, w_n$ 以及数量 $\alpha_1, \alpha_2, \dots, \alpha_n$ 和上三角矩阵 R_1 .

```

for k = 1 to n - 1 do
   $\sigma := \sqrt{\sum_{i=k}^m a_{ik}^2}$ 
  if  $a_{kk} \geq 0$ , then  $\sigma := -\sigma$  ( $\sigma_k = -\text{sgn}(a_{kk}^{(k-1)}) \|\tilde{a}_k^{(k-1)}\|_2$ )
   $d_k := \sigma$ 
   $\alpha_k := \sigma(\sigma - a_{kk})$ 
   $a_{kk} := a_{kk} - \sigma$  (存放  $w_k$  的第一个分量)
  for j = k + 1 to n do
     $\beta := \left( a_{kk}a_{kj} + \sum_{i=k+1}^m a_{ik}a_{ij} \right) / \alpha_k$ 
     $a_{kj} := a_{kj} - \beta a_{kk}$ 
    for i = k + 1 to m do
       $a_{ij} := a_{ij} - \beta a_{ik}$ 
    end do
  end do
end do
if m = n, then  $d_n := a_{nn}$ , 转标号 1
if m > n, then  $\sigma := \sqrt{\sum_{i=n}^m a_{in}^2}$ 
if  $a_{nn} \geq 0$ , then  $\sigma := -\sigma$ ;
 $d_n := \sigma$ 
 $\alpha_n := \sigma(\sigma - a_{nn})$ ,
 $a_{nn} := a_{nn} - \sigma$ ; (存放  $w_n$  的第一个分量)

```

标号 1: 输出矩阵 A , 向量 $d = (d_1, d_2, \dots, d_n)^T$, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$.

下面以矩阵 $A_{5 \times 4}$ 为例来说明上述算法执行完毕后 A 中元素存放数据的情况.

$$A = \begin{pmatrix} a_{11}^{(0)} - \sigma_1 & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} \\ a_{21}^{(0)} & a_{22}^{(1)} - \sigma_2 & a_{23}^{(2)} & a_{24}^{(2)} \\ a_{31}^{(0)} & a_{32}^{(1)} & a_{33}^{(2)} - \sigma_3 & a_{34}^{(3)} \\ a_{41}^{(0)} & a_{42}^{(1)} & a_{43}^{(2)} & a_{44}^{(3)} - \sigma_4 \\ a_{51}^{(0)} & a_{52}^{(1)} & a_{53}^{(2)} & a_{54}^{(3)} \end{pmatrix} = \begin{pmatrix} w_{11} & r_{12} & r_{13} & r_{14} \\ w_{21} & w_{22} & r_{23} & r_{24} \\ w_{31} & w_{32} & w_{33} & r_{34} \\ w_{41} & w_{42} & w_{34} & w_{44} \\ w_{51} & w_{52} & w_{35} & w_{45} \end{pmatrix}.$$

以上计算过程结束后, 可由向量 w_k 和数量 α_k 构造矩阵 $\tilde{H}_k$, 形成矩阵 H_k , 进而计算 $H = H_s H_{s-1} \cdots H_1$, 最后得到正交矩阵 $Q = H^T$.

例 2.4.2 用豪斯霍尔德变换作矩阵

$$A = \begin{pmatrix} 3 & 14 & 9 \\ 6 & 43 & 3 \\ 6 & 22 & 15 \end{pmatrix}$$

的 QR 分解.

解 在豪斯霍尔德变换中, $\sigma = \pm \|x\|_2$, 为计算简便, 在本例中取 $\sigma = \|x\|_2$.

第 1 步 取 $x = a_1 = (3, 6, 6)^T$, $e_1 = (1, 0, 0)^T$, 则

$$\sigma_1 = \|a_1\|_2 = 9,$$

$$\alpha_1 = \sigma_1(\sigma_1 - a_{11}) = 54, \quad w_1 = a_1 - \sigma_1 e_1 = 6(-1, 1, 1)^T,$$

$$H_1 = I_3 - 2u_1 u_1^T = I_3 - \frac{w_1 w_1^T}{\alpha_1} = \frac{1}{3} \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{pmatrix},$$

$$A^{(1)} = H_1 A = \begin{pmatrix} 9 & 48 & 15 \\ 0 & 9 & -3 \\ 0 & -12 & 9 \end{pmatrix}.$$

第 2 步 对于子矩阵

$$\begin{pmatrix} 9 & -3 \\ -12 & 9 \end{pmatrix},$$

取 $x = \tilde{a}_2^{(1)} = (9, -12)^T$, $e_1^{(1)} = (1, 0)^T$, 则

$$\sigma_2 = \|\tilde{a}_2^{(1)}\|_2 = 15, \quad \alpha_2 = \sigma_2(\sigma_2 - a_{22}^{(1)}) = 90,$$

$$w_2 = \tilde{a}_2^{(1)} - \sigma_2 e_1^{(1)} = -6(1, 2)^T,$$

$$\tilde{H}_2 = I_2 - 2u_2 u_2^T = I_2 - \frac{w_2 w_2^T}{\alpha_2} = \frac{1}{5} \begin{pmatrix} 3 & -4 \\ -4 & -3 \end{pmatrix},$$

$$H_2 = \begin{pmatrix} 1 & 0^T \\ 0 & \tilde{H}_2 \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & -4 \\ 0 & -4 & -3 \end{pmatrix},$$

$$A^{(2)} = H_2 A^{(1)} = \frac{1}{5} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & -4 \\ 0 & -4 & -3 \end{pmatrix} \begin{pmatrix} 9 & 48 & 15 \\ 0 & 9 & -3 \\ 0 & -12 & 9 \end{pmatrix} = \begin{pmatrix} 9 & 48 & 15 \\ 0 & 15 & -9 \\ 0 & 0 & -3 \end{pmatrix},$$

$$H = H_2 H_1 = \frac{1}{5} \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & -4 \\ 0 & -4 & -3 \end{pmatrix} \times \frac{1}{3} \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{pmatrix} = \frac{1}{15} \begin{pmatrix} 5 & 10 & 10 \\ -2 & 11 & -10 \\ -14 & 2 & 5 \end{pmatrix},$$

则有

$$Q = H^T = \frac{1}{15} \begin{pmatrix} 5 & -2 & -14 \\ 10 & 11 & 2 \\ 10 & -10 & 5 \end{pmatrix},$$

$$R = A^{(2)} = H_2 H_1 A = \begin{pmatrix} 9 & 48 & 15 \\ 0 & 15 & -9 \\ 0 & 0 & -3 \end{pmatrix},$$

$A = QR$. 解毕.

*2.5 超定方程组

通常求解的线性方程组 $Ax = b$ 中方程的个数与未知量的个数相等 (即 $A \in \mathbf{R}^{n \times n}$). 有时 (如在数据拟合问题中) 需要求解方程个数大于未知量个数的线性方程组 (即 $A \in \mathbf{R}^{m \times n}, m > n$), 本节介绍这种方程组的求解方法.

2.5.1 线性最小二乘问题

求解线性方程组

$$Ax = b, \quad (2.5.1)$$

其中 $A \in \mathbf{R}^{m \times n}, m > n, x \in \mathbf{R}^n, b \in \mathbf{R}^m, \text{rank}(A) = n$.

由线性代数知, 方程组 (2.5.1) 解存在的充分必要条件是系数矩阵的秩等于增广矩阵的秩, 即

$$\text{rank}(A) = \text{rank}(A, b).$$

若上述条件不成立, 则方程组 (2.5.1) 无解. 这时, 方程组 (2.5.1) 称为超定方程组或矛盾方程组.

当方程组 (2.5.1) 无解时, 对任意的 $x \in \mathbf{R}^n$, 残向量 $r = b - Ax \neq 0$. 这时, 可考虑取使残向量 r 在某种范数意义下达到最小的 x 作为方程组 (2.5.1) 的一种解. 最小二乘问题就是求残向量 r 在 2 范数意义下达到最小的 x .

定义 2.5.1 给定矩阵 $A \in \mathbf{R}^{m \times n} (m > n)$, 向量 $b \in \mathbf{R}^m$, 求 $\tilde{x} \in \mathbf{R}^n$, 使得 $\|Ax - b\|_2$ 达到最小, 即

$$\|A\tilde{x} - b\|_2 = \min_{x \in \mathbf{R}^n} \|Ax - b\|_2.$$

该问题称为线性最小二乘问题. $\tilde{x}$ 称为最小二乘问题的解, 也称为超定方程组 (2.5.1) 的最小二乘解.

为了便于求解, 记 $f(x) = \|Ax - b\|_2^2$, 则最小二乘问题等价于求解一个无条件极值问题

$$\min_{x \in \mathbf{R}^n} f(x) = \|Ax - b\|_2^2. \quad (2.5.2)$$

而

$$f(x) = \|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) = x^T A^T A x - 2(A^T b)^T x + b^T b.$$

由多元函数求极值的必要条件, 令

$$\nabla f(x) = 0,$$

由此可得

$$A^T A x = A^T b. \quad (2.5.3)$$

方程组 (2.5.3) 称为超定方程组 (2.5.1) 的正则方程组或法方程组.

下面的定理给出了超定方程组 (2.5.1) 的最小二乘解与法方程组 (2.5.3) 的解的关系.

定理 2.5.1 $\tilde{x} \in \mathbf{R}^n$ 是超定方程组 (2.5.1) 的最小二乘解的充分必要条件是 $\tilde{x}$ 是其法方程组 (2.5.3) 的解.

证明 必要性. 由前多元函数求极值的必要条件即可证明结论.

充分性. 设 $\tilde{x}$ 是法方程组 (2.5.3) 的解, 则对任意的 $y \in \mathbf{R}^n$, 令 $x = \tilde{x} + y$, 则

$$\begin{aligned} \|Ax - b\|_2^2 &= \|A\tilde{x} - b + Ay\|_2^2 = \|A\tilde{x} - b\|_2^2 + 2(A\tilde{x} - b)^T Ay + \|Ay\|_2^2 \\ &= \|A\tilde{x} - b\|_2^2 + 2y^T A^T (A\tilde{x} - b) + \|Ay\|_2^2 = \|A\tilde{x} - b\|_2^2 + \|Ay\|_2^2 \geq \|A\tilde{x} - b\|_2^2, \end{aligned}$$

即 $\tilde{x}$ 是方程组 (2.5.1) 的一个最小二乘解. 证毕.

定理 2.5.1 将求解超定方程组 (2.5.1) 的最小二乘解的问题转化为求解其法方程组 (2.5.3) 的问题. 由定理 2.5.1 可得以下推论:

推论 2.5.1 设 $\text{rank}(A) = n$, 则超定方程组 (2.5.1) 有唯一的最小二乘解

$$\tilde{x} = (A^T A)^{-1} A^T b.$$

证明 由 $\text{rank}(A^T A) = \text{rank}(A) = n$ 知, 法方程组 (2.5.3) 的系数矩阵 $A^T A$ 非奇异, 所以结论成立. 证毕.

2.5.2 最小二乘问题的求解

定理 2.5.1 给出了一个求超定方程组的最小二乘解的方法, 即通过求解法方程组 (2.5.3):

$$A^T A x = A^T b$$

来得到超定方程组 $Ax = b$ 的最小二乘解.

当 $\text{rank}(A) = n$ 时, 法方程组的系数矩阵 $A^T A$ 是 n 阶对称正定矩阵. 然而, 遗憾的是 $A^T A$ 的条件数往往很大, 所以法方程组 (2.5.3) 是病态方程组, 而且方程组的阶数 n 越大, 方程组越病态. 当 n 较小或者 $A^T A$ 的条件数不大时, 可用前面介绍的有关线性方程组的解法 (如改进平方根法) 求解. 但当 n 较大时, 不仅法方程组严重病态, 而且形成法方程组的系数矩阵 $A^T A$ 大约需要 $O(mn^2)$ 浮点运算, 其计算工作量较大, 并且计算出的 $A^T A$ 的精度较差, 甚至不正定, 致使楚列斯基分解无法进行. 一般来说, 应避免采用解法方程组的方法求最小二乘解. 一个常用的方法是利用矩阵的 QR 分解求解超定方程组 (2.5.1). 下面介绍用基于豪斯霍尔德变换的 QR 分解求解超定方程组.

设对矩阵 A 已作 QR 分解, 则有

$$A = QR = (Q_1, Q_2) \begin{pmatrix} R_1 \\ O \end{pmatrix} = Q_1 R_1,$$

其中 Q 为 $m \times m$ 阶正交矩阵, Q_1, Q_2 分别为 $m \times n, m \times (m - n)$ 阶列正交矩阵, R_1 是 n 阶非奇异上三角矩阵. 将上式代入超定方程组 $Ax = b$ 中得 $Q_1 R_1 x = b$, 则将求解超定方程组的问题转化为求解一个 n 阶上三角方程组

$$R_1 x = Q_1^T b$$

的问题.

利用 QR 分解求解超定方程组, 在实际计算时并非先求出矩阵 Q 和 R , 而是采取对超定方程组的系数矩阵 A 和右端向量 b 逐次进行豪斯霍尔德变换的方法.

由矩阵的 QR 分解知 $s = \min\{m - 1, n\}$,

$$H_s H_{s-1} \cdots H_1 A = R = \begin{pmatrix} R_1 \\ O \end{pmatrix},$$

$$H_s H_{s-1} \cdots H_1 b = Q^T b = (Q_1, Q_2)^T b = \begin{pmatrix} Q_1^T b \\ Q_2^T b \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

其中 $c_1 \in \mathbf{R}^n$, $c_2 \in \mathbf{R}^{m-n}$.

由此可知, 在实际计算中, 不必计算矩阵 $H = H_s H_{s-1} \cdots H_1$, 而是构造它的因子 H_k . 对矩阵 A 和方程组的右端向量 b 同时作 s 次豪斯霍尔德变换 $H_1, H_2, \cdots, H_s$ 后, 即可获得 R_1 和 c_1 . 此时,

$$H(Ax - b) = \begin{pmatrix} R_1 \\ O \end{pmatrix} x - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} R_1 x - c_1 \\ c_2 \end{pmatrix}.$$

由于 H 是正交矩阵, 则

$$\|Ax - b\|_2^2 = \|H(Ax - b)\|_2^2 = \left\| \begin{pmatrix} R_1 x - c_1 \\ c_2 \end{pmatrix} \right\|_2^2 = \|R_1 x - c_1\|_2^2 + \|c_2\|_2^2.$$

由此可知, $\tilde{x}$ 是超定方程组的最小二乘解, 当且仅当 $\tilde{x}$ 是方程组 $R_1 x = c_1$ 的解. 所以, 通常由求解上三角方程组

$$R_1 x = c_1$$

获得最小二乘问题的解 $\tilde{x}$. 这时, 最小二乘解 $\tilde{x}$ 的残向量 r 的模为 $\|A\tilde{x} - b\|_2 = \|c_2\|_2$.

矩阵的 QR 分解可以用于超定方程组的求解, 也常用于病态线性方程组的求解, 故将两种方程组统一写为

$$Ax = b, \quad (2.5.4)$$

其中 $A \in \mathbf{R}^{m \times n}$ ($m \geq n$), $\text{rank}(A) = n$.

对矩阵的 QR 分解过程稍作修改即可得到求解方程组 (2.5.4) 的 QR 方法. 为此, 将方程组的右端向量 b 记为

$$b = (a_{1,n+1}, a_{2,n+1}, \cdots, a_{m,n+1})^T.$$

下面是用基于豪斯霍尔德变换的 QR 分解求解方程组 (2.5.4) 的计算过程.

for $k = 1$ to $n - 1$ do

$$\sigma := \sqrt{\sum_{i=k}^m a_{ik}^2}$$

$$\text{if } a_{kk} \geq 0, \text{ then } \sigma := -\sigma \quad (\sigma_k = -\text{sgn}(a_{kk}^{(k-1)}) \|\tilde{a}_k^{(k-1)}\|_2)$$

$$d_k := \sigma$$

$$\alpha_k := \sigma(\sigma - a_{kk})$$

$$a_{kk} := a_{kk} - \sigma$$

```

for  $j = k + 1$  to  $n + 1$  do
   $\beta := \left( a_{kk}a_{kj} + \sum_{i=k+1}^m a_{ik}a_{ij} \right) / \alpha_k$ 
   $a_{kj} := a_{kj} - \beta a_{kk}$ 
  for  $i = k + 1$  to  $m$  do
     $a_{ij} := a_{ij} - \beta a_{ik}$ 
  end do
end do
end do
if  $m = n$ , then  $d_n := a_{nn}$ , 转标号 1
if  $m > n$ , then  $\sigma := \sqrt{\sum_{i=n}^m a_{in}^2}$ 
if  $a_{nn} \geq 0$ , then  $\sigma := -\sigma$ ;
 $d_n := \sigma$ ;
 $\alpha_n := \sigma(\sigma - a_{nn})$ 
 $a_{nn} := a_{nn} - \sigma$ 
标号 1  $\beta := \left( a_{nn}a_{n,n+1} + \sum_{i=n+1}^m a_{in}a_{i,n+1} \right) / \alpha_n$ 
 $a_{n,n+1} := a_{n,n+1} - \beta a_{nn}$ 
for  $i = n + 1$  to  $m$  do
   $a_{i,n+1} := a_{i,n+1} - \beta a_{i,n}$ 
end do
 $x_n = a_{n,n+1} / d_n$ ,
for  $k = n - 1$  to 1 step  $-1$  do
   $x_k = \left( a_{k,n+1} - \sum_{j=k+1}^n a_{kj}x_j \right) / d_k$ ,
end do

```

最小二乘解 $\tilde{x}$ 的残向量的 2 范数为 $\|A\tilde{x} - b\|_2 = \sqrt{\sum_{i=n+1}^m a_{i,n+1}^2}$.

例 2.5.1 求下面超定方程组的最小二乘解:

$$\begin{cases} x_1 + 3x_2 = 4, \\ x_1 + 3x_2 = 3, \\ x_1 + x_2 = 3, \\ x_1 + x_2 = 0. \end{cases}$$

解法 1 用求解法方程的方法. 超定方程组的系数矩阵和右端向量分别为

$$A = \begin{pmatrix} 1 & 3 \\ 1 & 3 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 3 \\ 3 \\ 0 \end{pmatrix},$$

其法方程组 $A^T A x = A^T b$ 为

$$\begin{pmatrix} 4 & 8 \\ 8 & 20 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 24 \end{pmatrix}.$$

解之得 $x_1 = \frac{1}{2}$, $x_2 = 1$.

解法 2 用豪斯霍尔德变换法. 超定方程组的增广矩阵为

$$(A, b) = \begin{pmatrix} 1 & 3 & 4 \\ 1 & 3 & 3 \\ 1 & 1 & 3 \\ 1 & 1 & 0 \end{pmatrix}.$$

第 1 步

$$\sigma_1 = -\operatorname{sgn}(a_{11})\|a_1\|_2 = -\|a_1\|_2 = -\sqrt{1^2 + 1^2 + 1^2 + 1^2} = -2,$$

$$\alpha_1 = \sigma_1(\sigma_1 - a_{11}) = 6, \quad w_1 = a_1 - \sigma_1 e_1 = (3, 1, 1, 1)^T,$$

$$H_1 = I_4 - 2u_1 u_1^T = I_4 - \frac{w_1 w_1^T}{\alpha_1} = -\frac{1}{6} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 3 & -5 & 1 & 1 \\ 3 & 1 & -5 & 1 \\ 3 & 1 & 1 & -5 \end{pmatrix},$$

$$H_1(A, b) = \begin{pmatrix} -2 & -4 & -5 \\ 0 & \frac{2}{3} & 0 \\ 0 & -\frac{4}{3} & 0 \\ 0 & -\frac{4}{3} & -3 \end{pmatrix}.$$

第 2 步

$$\sigma_2 = -\operatorname{sgn}(a_{22}^{(1)})\|\tilde{a}_2^{(1)}\|_2 = -\|\tilde{a}_2^{(1)}\|_2 = -\sqrt{\left(\frac{2}{3}\right)^2 + \left(-\frac{4}{3}\right)^2 + \left(-\frac{4}{3}\right)^2} = -2,$$

$$\alpha_2 = \sigma_2(\sigma_2 - a_{22}^{(1)}) = \frac{16}{3},$$

$$w_2 = \tilde{a}_2^{(1)} - \sigma_2 e_1^{(1)} = \left(\frac{2}{3}, -\frac{4}{3}, -\frac{4}{3}\right)^T + 2(1, 0, 0)^T = \frac{4}{3}(2, -1, -1)^T,$$

$$\tilde{H}_2 = I_3 - 2u_2 u_2^T = I_3 - \frac{w_2 w_2^T}{\alpha_2} = \begin{pmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix},$$

$$H_2 = \begin{pmatrix} 1 & 0^T \\ 0 & \tilde{H}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ 0 & \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \\ 0 & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix},$$

$$H_2 H_1(A, b) = \begin{pmatrix} -2 & -4 & -5 \\ 0 & -2 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & -2 \end{pmatrix} = \begin{pmatrix} R_1 & c_1 \\ O & c_2 \end{pmatrix}.$$

最后, 解上三角方程组 $R_1 x = c_1$, 即解方程组

$$\begin{pmatrix} -2 & -4 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -5 \\ -2 \end{pmatrix}.$$

解之得超定方程组的最小二乘解为 $x_1 = \frac{1}{2}$, $x_2 = 1$, 其残向量 r 的 2 范数为 $\|A\tilde{x} - b\|_2 = \|c_2\|_2 = \sqrt{1^2 + (-2)^2} = \sqrt{5}$. 解毕.

小 结

本章介绍了求解线性方程组 $Ax = b$ 的三类方法——高斯消去法、矩阵三角分解法和矩阵的 QR 分解法. 在无舍入误差的假定下, 这些方法经有限次运算就可以求得方程组的准确解.

列主元高斯消去法具有较好的数值稳定性, 算法组织比较容易, 计算复杂度为 $O(n^3)$, 对于阶数不高 ($n \leq 30$) 的线性方程组是一个常用的求解方法.

由高斯消去法的矩阵形式导出矩阵的 $A = LU$ 分解, 用 LU 分解解方程组等价于高斯消去法. 为了减少计算量, 矩阵 $A = LU$ 分解特别适宜于求解系数矩阵相同, 但右端项不同的多个方程组.

当系数矩阵 A 是对称正定矩阵时, A 可以分解成一个下三角矩阵与它的转置的乘积, 即 $A = GG^T$. 将其用于求解方程组即平方根法. 为了避免平方根法中的开方运算, 将 A 分解成 $A = LDL^T$, 由此得到改进平方根法. 平方根法和改进平方根法都是数值稳定的方法.

求解三对角方程组的追赶法则是矩阵 $A = LU$ 分解在解三对角方程组中的应用, 它具有计算量小、方法简单、计算过程稳定等优点.

舍入误差对解的影响取决于方程组系数矩阵的条件数, 条件数大的方程组称为“病态”方程组, 病态是问题本身固有的性态. 可根据病态问题的常见特征推测方程组是否病态.

求解病态方程组则必须事先采取预处理措施, 或采用高精度计算, 或用矩阵的 QR 分解求解.

由吉文斯变换和豪斯霍尔德变换分别导出了矩阵的 QR 分解. QR 分解常用于求解病态方程组、超定方程组和求矩阵特征值等数值代数中的许多其他问题.

当超定方程组 $Ax = b$ 的自变量个数 n 较小或条件数 $\text{Cond}(A^T A)$ 不大时, 用改进平方根法求解其正则方程组 $A^T Ax = A^T b$ 来得到超定方程组的最小二乘解; 当 n 较大时, 超定方程组则用基于豪斯霍尔德变换的 QR 分解来求解.

用基于豪斯霍尔德变换的 QR 分解求解线性方程组的计算过程是十分稳定的, 但由于它的计算量约是高斯消去法的 2 倍, 故一般不用它来求解良态方程组; 而且列主元高斯消去法也是一个稳定的方法, 在求解良态线性方程组时常用后者.

习 题

2.1 用列主元高斯消去法求解以下方程组:

$$(1) \begin{cases} 2x_1 - 3x_2 + 2x_3 = 5, \\ -4x_1 + 2x_2 - 6x_3 = 14, \\ 2x_1 + 2x_2 + 4x_3 = 8; \end{cases} \quad (2) \begin{cases} 12x_1 - 3x_2 + 3x_3 = 15, \\ -18x_1 + 3x_2 - x_3 = -15, \\ x_1 + x_2 + x_3 = 6. \end{cases}$$

2.2 在高斯消去法中方程组的系数矩阵 A 经过第 1 次消元后变为

$$\begin{pmatrix} a_{11} & a^T \\ 0 & A_1 \end{pmatrix},$$

其中 $a = (a_{12}, a_{13}, \dots, a_{1n})^T$, $0 = (0, 0, \dots, 0)^T$. 证明:

(1) 设 A 是严格对角占优矩阵, 则 A_1 仍是严格对角占优矩阵;

(2) 若 A 是对称正定矩阵, 则 A 的绝对值最大的元素必是对角元 (提示: $|a_{ij}| < \sqrt{a_{ii}a_{jj}}$);

(3) 若 A 是对称正定矩阵, 则 A_1 仍是对称正定矩阵 (提示: 利用等式 $\sum_{i=2}^n \sum_{j=2}^n a_{ij}^{(1)} x_i x_j = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j - a_{11} \left(x_1 + \sum_{i=2}^n \frac{a_{i1}}{a_{11}} x_i \right)^2$).

2.3 作以下矩阵的 $A = LU$ 分解:

$$(1) \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{pmatrix}; \quad (2) \begin{pmatrix} 3 & 0 & 1 \\ 0 & -1 & 3 \\ 1 & 3 & 0 \end{pmatrix}; \quad (3) \begin{pmatrix} 3 & 1 & 2 \\ 9 & 8 & 12 \\ 6 & 27 & 38 \end{pmatrix}.$$

2.4 试写出用 $A = LU$ 分解求解方程组

$$\begin{pmatrix} b_1 & c_1 & & & a_1 \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \vdots \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ c_n & & & a_n & b_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}$$

的计算公式.

2.5 试推导 n 阶矩阵 A 的克劳特分解 $A = LU$ 的计算公式, 其中 L 为下三角矩阵, U 为单位上三角矩阵.

2.6 用平方根法和改进平方根法求解以下方程组:

$$(1) \begin{pmatrix} 4 & -2 & -4 \\ -2 & 17 & 10 \\ -4 & 10 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 10 \\ 3 \\ -7 \end{pmatrix};$$

$$(2) \begin{pmatrix} 1 & 2 & 3 \\ 2 & 20 & 26 \\ 3 & 26 & 70 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 8 \\ -1 \end{pmatrix}.$$

2.7 用追赶法求解方程组 $Ax = b$, 其中

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

2.8 计算矩阵

$$A = \begin{pmatrix} 5 & -4 & -7 \\ -4 & 2 & -4 \\ -7 & -4 & 5 \end{pmatrix}$$

的 1 范数、2 范数、 ∞ 范数和相应的条件数.

2.9 证明

$$(1) \frac{1}{\sqrt{n}} \|x\|_1 \leq \|x\|_2 \leq \|x\|_1;$$

$$(2) \frac{1}{\sqrt{n}} \|A\|_F \leq \|A\|_2 \leq \|A\|_F.$$

2.10 证明对于任何的实数 $p \geq 1$, $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$ 定义了 $\mathbf{R}^n$ 上的一种向量范数.

2.11 设 A 是 n 阶对称正定矩阵, $x \in \mathbf{R}^n$. 证明

(1) $\|x\|_A = \sqrt{x^T A x}$ 是一种向量范数;

$$(2) \sqrt{\lambda_{\min}} \|x\|_2 \leq \|x\|_A \leq \sqrt{\lambda_{\max}} \|x\|_2.$$

2.12 (1) 设 $\|\cdot\|$ 是 $\mathbf{R}^n$ 上的一种向量范数, M 是 n 阶非奇异矩阵, 令 $\|x\|_M = \|Mx\|$ ($\forall x \in \mathbf{R}^n$), 证明 $\|x\|_M$ 是 $\mathbf{R}^n$ 上的向量范数;

(2) 将与 (1) 中向量范数 $\|\cdot\|$ 相容的矩阵范数记为 $\|\cdot\|$, 证明由 $\|A\|_M = \|MAM^{-1}\|$ 定义了一种矩阵范数且与向量范数 $\|x\|_M$ 相容.

2.13 证明矩阵的 F 范数满足矩阵范数定义的 4 个条件.

2.14 设 A 是非奇异矩阵, $\tilde{x}$ 是方程组 $Ax = b$ 的近似解, $r = b - A\tilde{x}$. 证明

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

2.15 设 A, B 是 n 阶非奇异矩阵, R 是 n 阶正交矩阵, c 是常数, 证明

$$(1) \text{Cond}(cA) = \text{Cond}(A);$$

$$(2) \text{Cond}(AB) \leq \text{Cond}(A)\text{Cond}(B);$$

$$(3) \text{Cond}_2(RA) = \text{Cond}_2(A) = \text{Cond}_2(AR);$$

$$(4) \text{Cond}_2(A) = \frac{\sqrt{\lambda_{\max}(A^T A)}}{\sqrt{\lambda_{\min}(A^T A)}};$$

$$(5) \text{当 } A \text{ 是对称矩阵时, } \text{Cond}_2(A) = \frac{|\lambda|_{\max}}{|\lambda|_{\min}}.$$

2.16 分别用吉文斯变换和豪斯霍尔德变换作矩阵 $A = QR$ 分解, 其中

$$(1) A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & -1 & -1 \\ 2 & -4 & 5 \end{pmatrix}; \quad (2) A = \begin{pmatrix} 0 & 4 & 1 \\ 1 & 1 & 1 \\ 0 & 3 & 2 \end{pmatrix}.$$

2.17 设 $A \in \mathbf{R}^{m \times n}$, $b_i \in \mathbf{R}^m$ ($i = 1, 2, \dots, k$), 证明 $x \in \mathbf{R}^n$ 使得

$$\sum_{i=1}^k \|Ax - b_i\|_2^2 = \min$$

的充分必要条件是 x 是方程组 $Ax = \frac{1}{k} \sum_{i=1}^k b_i$ 的最小二乘解.

2.18 求下列方程组的最小二乘解:

$$(1) \begin{cases} 2x_1 + 4x_2 = 11, \\ 3x_1 - 5x_2 = 3, \\ x_1 + 2x_2 = 6, \\ 2x_1 + x_2 = 7; \end{cases} \quad (2) \begin{cases} x_1 + x_2 - 3x_3 = -1, \\ 2x_1 + x_2 - 2x_3 = 1, \\ x_1 + x_2 + x_3 = 3, \\ x_1 + 2x_2 - 3x_3 = 1. \end{cases}$$

2.19 用豪斯霍尔德变换求如下方程组的最小二乘解:

$$\begin{cases} x_1 + 4x_2 = 7, \\ 2x_1 + 5x_2 = 8, \\ 3x_1 + 6x_2 = 9. \end{cases}$$

2.20 用豪斯霍尔德变换求方程组

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & -2 & 3 \\ 2 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}.$$

计算实习

2.1 用列主元高斯消去法求解线性方程组 $Ax = b$, 其中

$$A = \begin{pmatrix} 1.1348 & 3.8326 & 1.1651 & 3.4017 \\ 0.5301 & 1.7875 & 2.5330 & 1.5435 \\ 3.4129 & 4.9317 & 8.7643 & 1.3142 \\ 1.2371 & 4.9998 & 10.6721 & 0.0147 \end{pmatrix}, \quad b = \begin{pmatrix} 9.5342 \\ 6.3941 \\ 18.4231 \\ 16.9237 \end{pmatrix}.$$

2.2 作矩阵 $A = GG^T$ 和 $A = LDL^T$ 分解, 其中矩阵 $A = (a_{ij})_{20 \times 20}$,

$$a_{ij} = \begin{cases} i, & i = j, \\ \min\{i, j\}, & i \neq j. \end{cases}$$

2.3 用追赶法求解 20 阶三对角线性方程组 $Ax = d$ 的解, 其中

$$A = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}, \quad d = \begin{pmatrix} 3 \\ 2 \\ \vdots \\ 2 \\ 3 \end{pmatrix}.$$

2.4 用豪斯霍尔德变换求方程组 $Ax = b$, 其中

$$A = \begin{pmatrix} 5 & 4 & 7 & 5 & 6 & 7 & 5 \\ 4 & 12 & 8 & 7 & 8 & 8 & 6 \\ 7 & 8 & 10 & 9 & 8 & 7 & 7 \\ 5 & 7 & 9 & 11 & 9 & 7 & 5 \\ 6 & 8 & 8 & 9 & 10 & 8 & 9 \\ 7 & 8 & 7 & 7 & 8 & 10 & 10 \\ 5 & 6 & 7 & 5 & 9 & 10 & 10 \end{pmatrix}, \quad b = \begin{pmatrix} 39 \\ 53 \\ 56 \\ 53 \\ 58 \\ 57 \\ 52 \end{pmatrix}.$$

第3章 解线性方程组的迭代法

在科学研究和工程技术中的线性方程组 (如在偏微分方程数值解法中形成的线性方程组) 一般具有阶数高、稀疏性的特点 (稀疏性是指线性方程组的系数矩阵含有大量的零元素). 针对这类大型稀疏线性方程组的求解, 本节介绍的迭代法利用了方程组的稀疏性, 具有存储量小、算法简单、收敛速度快等特点. 迭代法是将原线性方程组作某种等价变形, 由此构造迭代格式, 给定初始点 $x^{(0)}$, 由迭代格式产生的迭代序列 $\{x^{(k)}\}$ 逐次逼近方程组的精确解 x^* .

3.1 向量序列和矩阵序列的极限

下面给出向量序列和矩阵序列收敛的定义.

定义 3.1.1 设 $\{x^{(k)}\}$ 是 $\mathbf{R}^n$ 中的向量序列, 其中 $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})^T$. 若 $\{x^{(k)}\}$ 的每一个分量 $x_i^{(k)}$ 都收敛于 x_i^* , 即

$$\lim_{k \rightarrow \infty} x_i^{(k)} = x_i^*, \quad i = 1, 2, \dots, n,$$

则称向量序列 $\{x^{(k)}\}$ 收敛于向量 $x^* = (x_1^*, x_2^*, \dots, x_n^*)^T$. 或者说, x^* 为向量序列 $\{x^{(k)}\}$ 的极限, 记作

$$\lim_{k \rightarrow \infty} x^{(k)} = x^*.$$

定义 3.1.2 设 $\{A^{(k)}\}$ 是 $\mathbf{R}^{n \times n}$ 中的矩阵序列, 其中 $A^{(k)} = (a_{ij}^{(k)})_{n \times n}$. 若 $A^{(k)}$ 的每个元素 $a_{ij}^{(k)}$ 都收敛于 a_{ij} , 即

$$\lim_{k \rightarrow \infty} a_{ij}^{(k)} = a_{ij}, \quad i, j = 1, 2, \dots, n,$$

则称矩阵序列 $\{A^{(k)}\}$ 收敛于 $A = (a_{ij})$, 或者说, 矩阵 A 是矩阵序列 $\{A^{(k)}\}$ 的极限, 记作

$$\lim_{k \rightarrow \infty} A^{(k)} = A.$$

显然, 向量序列和矩阵序列的收敛性可用范数描述为如下定理:

定理 3.1.1 (1) $\mathbf{R}^n$ 中的向量序列 $\{x^{(k)}\}$ 收敛于 x^* 的充分必要条件是

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

$$\lim_{k \rightarrow \infty} \|A - A^{(k)}\| = 0.$$

定理 3.1.2 设 $B \in \mathbb{R}^{n \times n}$, 则

$$\lim_{k \rightarrow \infty} B^k = 0$$

证明 略.

3.2 解线性方程组的基本迭代法

3.2.1 迭代法的一般格式

$$x = Bx + q.$$

构造迭代格式

$$x^{(k+1)} = Bx^{(k)} + g, \quad k = 0, 1, 2, \dots. \quad (3.2.1)$$

给定初始向量 $x^{(0)}$, 由迭代格式 (3.2.1) 产生迭代序列 $\{x^{(k)}\}$. 设 $\{x^{(k)}\}$ 收敛于 x^* , 对式 (3.2.1) 两端令 $k \rightarrow \infty$ 取极限. 由

$$\lim_{k \rightarrow \infty} x^{(k+1)} = \lim_{k \rightarrow \infty} (Bx^{(k)} + g)$$

得

$$x^* = Bx^* + g.$$

由此可知, x^* 满足 $Ax^* = b$, 即 x^* 是方程组 $Ax = b$ 的解.

迭代法的优点是在迭代过程中矩阵 B 不变, 从而利用了原方程组系数矩阵 A 含有大量零元素的特点, 具有计算量小的优点.

3.2.2 三种基本迭代法

1. 雅可比 (Jacobi) 迭代法

设求解的方程组为

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \qquad\qquad\qquad \dots\dots\dots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{cases}$$

不妨假定 $a_{ii} \neq 0$ ($i = 1, 2, \dots, n$) (交换方程的次序总会使 $a_{ii} \neq 0$).

把上述方程组改写为等价的同解方程组

$$\begin{cases} x_1 = (b_1 - a_{12}x_2 - a_{13}x_3 - \dots - a_{1n}x_n)/a_{11}, \\ x_2 = (b_2 - a_{21}x_1 - a_{23}x_3 - \dots - a_{2n}x_n)/a_{22}, \\ \dots\dots\dots \\ x_i = (b_i - a_{i1}x_1 - \dots - a_{i,i-1}x_{i-1} - a_{i,i+1}x_{i+1} - \dots - a_{in}x_n)/a_{ii}, \\ \dots\dots\dots \\ x_n = (b_n - a_{n1}x_1 - a_{n2}x_2 - \dots - a_{n,n-1}x_{n-1})/a_{nn}. \end{cases}$$

令

$$\begin{cases} x_1^{(k+1)} = (b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \dots - a_{1n}x_n^{(k)})/a_{11}, \\ x_2^{(k+1)} = (b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)} - \dots - a_{2n}x_n^{(k)})/a_{22}, \\ \dots\dots\dots \\ x_i^{(k+1)} = (b_i - a_{i1}x_1^{(k)} - \dots - a_{i,i-1}x_{i-1}^{(k)} - a_{i,i+1}x_{i+1}^{(k)} - \dots - a_{in}x_n^{(k)})/a_{ii}, \\ \dots\dots\dots \\ x_n^{(k+1)} = (b_n - a_{n1}x_1^{(k)} - a_{n2}x_2^{(k)} - \dots - a_{n,n-1}x_{n-1}^{(k)})/a_{nn}. \end{cases} \quad (3.2.2)$$

该式称为雅可比迭代格式或雅可比迭代法. 式 (3.2.2) 也写为

$$x_i^{(k+1)} = \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j^{(k)} \right) / a_{ii}, \quad i = 1, 2, \dots, n. \quad (3.2.2')$$

给定初始向量 $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})^T$ 后, 由式 (3.2.2) 产生迭代序列 $x^{(1)}, x^{(2)}, \dots, x^{(k)}, \dots$. 将式 (3.2.2) 写成矩阵形式

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

其中

$$B = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \dots & -\frac{a_{1,n-1}}{a_{11}} & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \dots & -\frac{a_{2,n-1}}{a_{22}} & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \dots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix},$$

$$g = \left(\frac{b_1}{a_{11}}, \frac{b_2}{a_{22}}, \dots, \frac{b_n}{a_{nn}} \right)^T.$$

例 3.2.1 给定初始向量 $x^{(0)} = (0, 0, 0)^T$, 用雅可比迭代法求解方程组

$$\begin{cases} 10x_1 - x_2 - 2x_3 = 72, \\ -x_1 + 10x_2 - 2x_3 = 83, \\ -x_1 - x_2 + 5x_3 = 42. \end{cases}$$

解 雅可比迭代格式为

$$\begin{cases} x_1^{(k+1)} = (72 + x_2^{(k)} + 2x_3^{(k)})/10, \\ x_2^{(k+1)} = (83 + x_1^{(k)} + 2x_3^{(k)})/10, \\ x_3^{(k+1)} = (42 + x_1^{(k)} + x_2^{(k)})/5. \end{cases}$$

将初始向量 $x^{(0)} = (0, 0, 0)^T$ 代入以上迭代格式得 $x^{(1)} = (7.2, 8.3, 8.4)^T$, $x_2 = (9.71, 10.70, 11.50)^T$. 继续迭代下去, 迭代结果如表 3.1 所示.

表 3.1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000 0	0.000 0	0.000 0
1	7.200 0	8.300 0	8.400 0
2	9.710 0	10.700 0	11.500 0
3	10.570 0	11.570 0	12.482 0
4	10.853 5	11.835 4	12.828 2
5	10.951 0	11.951 0	12.941 4
6	10.983 4	11.983 4	12.950 4
7	10.994 4	11.998 1	12.993 4
8	10.998 1	11.994 1	12.997 8
9	10.999 4	11.999 4	12.999 2

由表 3.1 可见, 迭代 9 次后可得近似解 $x^{(9)} = (10.999 4, 11.999 4, 12.999 2)^T$. 容易验证, 方程组的精确解为 $x^* = (11, 12, 13)^T$. 从表 3.1 可以看出, 随着迭代次数的增加, 迭代结果越来越接近精确解. 解毕.

2. 高斯-赛德尔 (Gauss-Seidel) 迭代法

在雅可比迭代法中计算 $x_2^{(k+1)}$ 时, $x_1^{(k+1)}$ 已算出, 将 $x_2^{(k+1)}$ 表达式中的 $x_1^{(k)}$ 用 $x_1^{(k+1)}$ 代替. 同理, 计算 $x_i^{(k+1)}$ 时, $x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_{i-1}^{(k+1)}$ 已算出, 用 $x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_{i-1}^{(k+1)}$ 代替 $x_i^{(k+1)}$ 表达式中的 $x_1^{(k)}, x_2^{(k)}, \dots, x_{i-1}^{(k)}$, 有望加速收敛. 由此可得下面的高斯-赛德尔迭代法或高斯-赛德尔迭代格式:

$$\begin{cases} x_1^{(k+1)} = (b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \dots - a_{1n}x_n^{(k)})/a_{11}, \\ x_2^{(k+1)} = (b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \dots - a_{2n}x_n^{(k)})/a_{22}, \\ \dots\dots\dots \\ x_i^{(k+1)} = (b_i - a_{i1}x_1^{(k+1)} - \dots - a_{i,i-1}x_{i-1}^{(k+1)} - a_{i,i+1}x_{i+1}^{(k)} - \dots - a_{in}x_n^{(k)})/a_{ii}, \\ \dots\dots\dots \\ x_n^{(k+1)} = (b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \dots - a_{n,n-1}x_{n-1}^{(k+1)})/a_{nn}. \end{cases} \quad (3.2.3)$$

式 (3.2.3) 也可以简记为

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right) / a_{ii}, \quad i = 1, 2, \dots, n. \quad (3.2.3')$$

例 3.2.2 给定初始向量 $x^{(0)} = (0, 0, 0)^T$, 用高斯-赛德尔迭代法求解例 3.2.1 中的方程组.

解 高斯-赛德尔迭代格式为

$$\begin{cases} x_1^{(k+1)} = (72 + x_2^{(k)} + 2x_3^{(k)})/10, \\ x_2^{(k+1)} = (83 + x_1^{(k+1)} + 2x_3^{(k)})/10, \\ x_3^{(k+1)} = (42 + x_1^{(k+1)} + x_2^{(k+1)})/5. \end{cases}$$

代入初始向量 $x^{(0)} = (0, 0, 0)^T$ 得

$$x^{(1)} = (7.2000, 9.0200, 11.6440)^T.$$

继续迭代下去, 计算结果如表 3.2 所示.

表 3.2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000 0	0.000 0	0.000 0
1	7.200 0	9.020 0	11.644 0
2	10.430 8	11.671 9	12.820 5
3	10.931 3	11.957 2	12.977 8
4	10.991 3	11.994 7	12.997 2
5	10.998 9	11.999 3	12.999 6
6	10.999 9	11.999 9	13.000 0

比较表 3.1 与表 3.2 的计算结果可见, 该题用高斯-赛德尔迭代法迭代 5 次的结果与雅可比迭代法迭代 8 次的结果差不多, 一般地, 当两种算法都收敛时, 高斯-赛德尔迭代法的收敛速度快于雅可比迭代法. 但也有雅可比迭代法收敛, 高斯-赛德尔迭代法发散的情况.

3. 逐次超松弛 (SOR) 迭代法

为了加快高斯-赛德尔迭代法的收敛速度, 对其进行改进, 将高斯-赛德尔迭代法改写为

$$x_i^{(k+1)} = x_i^{(k)} + \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^n a_{ij} x_j^{(k)} \right) / a_{ii}, \quad i = 1, 2, \dots, n, \quad (3.2.4)$$

$$r_i^{(k+1)} \triangleq b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^n a_{ij}x_j^{(k)}, \quad i = 1, 2, \dots, n.$$

$r^{(k+1)} \triangleq (r_1^{(k+1)}, r_2^{(k+1)}, \dots, r_n^{(k+1)})^T$ 称为第 $k+1$ 步迭代的残向量. 式 (3.2.4) 可写为

$$x_i^{(k+1)} = x_i^{(k)} + r_i^{(k+1)} / a_{ii}, \quad i = 1, 2, \dots, n. \quad (3.2.4')$$

显然, 当迭代序列 $x_i^{(k)}$ 收敛时, $r_i^{(k+1)} \rightarrow 0$ ($k \rightarrow \infty$) ($i = 1, 2, \dots, n$). 为了获得更快的收敛效果, 在式 (3.2.4) 修正量前乘以参数 ω 得

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

上式可写为

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \omega \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right) / a_{ii}, \quad i = 1, 2, \dots, n, \quad (3.2.5)$$

也即

$$\begin{cases} x_1^{(k+1)} = (1 - \omega)x_1^{(k)} + \omega(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \dots - a_{1n}x_n^{(k)}) / a_{11}, \\ x_2^{(k+1)} = (1 - \omega)x_2^{(k)} + \omega(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \dots - a_{2n}x_n^{(k)}) / a_{22}, \\ \dots\dots\dots \\ x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \omega(b_i - a_{i1}x_1^{(k+1)} - \dots - a_{i,i-1}x_{i-1}^{(k+1)} \\ \quad - a_{i,i+1}x_{i+1}^{(k)} - \dots - a_{in}x_n^{(k)}) / a_{ii}, \\ \dots\dots\dots \\ x_n^{(k+1)} = (1 - \omega)x_n^{(k)} + \omega(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \dots - a_{n,n-1}x_{n-1}^{(k+1)}) / a_{nn}. \end{cases} \quad (3.2.5')$$

式 (3.2.5) 或式 (3.2.5') 称为逐次超松弛 (successive over-relaxation, SOR) 迭代法, 简称为 SOR 方法, 参数 ω 称为松弛因子. 由式 (3.2.5) 可见, 高斯-赛德尔迭代法是超松弛迭代法中松弛因子 $\omega = 1$ 时的特例. 若将由高斯-赛德尔迭代法计算出的 $x^{(k+1)}$ 记为 $x_{G-S}^{(k+1)}$, 则超松弛迭代法可简记为

$$x_{SOR}^{(k+1)} = (1 - \omega)x^{(k)} + \omega x_{G-S}^{(k+1)},$$

故超松弛迭代法可以看成是 $x^{(k)}$ 与 $x_{G-S}^{(k+1)}$ 的一种加权平均. 若 ω 选取适当, 就能起到加速收敛的效果.

3.3 迭代法的收敛性

3.3.1 迭代法的矩阵表示

为了研究迭代法的收敛性, 先将三种迭代法改写为如下统一的迭代格式:

$$x^{(k+1)} = Bx^{(k)} + g.$$

为此, 将线性方程组的系数矩阵 A 分解为

$$A = D - E - F,$$

其中

$$D = \begin{pmatrix} a_{11} & & & & \\ & a_{22} & & & \\ & & \ddots & & \\ & & & a_{nn} & \end{pmatrix}, \quad E = \begin{pmatrix} 0 & & & & \\ -a_{21} & 0 & & & \\ -a_{31} & -a_{32} & 0 & & \\ \vdots & \vdots & & \ddots & \\ -a_{n1} & -a_{n2} & \cdots & -a_{n,n-1} & 0 \end{pmatrix},$$

$$F = \begin{pmatrix} 0 & -a_{12} & -a_{13} & \cdots & -a_{1n} \\ & 0 & -a_{23} & \cdots & -a_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & 0 & -a_{n-1,n} \\ & & & & 0 \end{pmatrix}.$$

1. 雅可比迭代法

雅可比迭代法 (3.2.2) 用矩阵-向量表示为

$$x^{(k+1)} = D^{-1}(E + F)x^{(k)} + D^{-1}b,$$

将上式记为

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

其中

$$B = D^{-1}(E + F), \quad g = D^{-1}b.$$

2. 高斯-赛德尔迭代法

高斯-赛德尔迭代法 (3.2.3) 可以表示为

$$x^{(k+1)} = D^{-1}Ex^{(k+1)} + D^{-1}Fx^{(k)} + D^{-1}b.$$

据此可得

$$x^{(k+1)} = (D - E)^{-1}Fx^{(k)} + (D - E)^{-1}b.$$

将上式记为

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

其中

$$B = (D - E)^{-1}F, \quad g = (D - E)^{-1}b.$$

3. 超松弛迭代法

超松弛迭代法 (3.2.5) 可表示为

$$x^{(k+1)} = (1 - \omega)x^{(k)} + \omega D^{-1}(Ex^{(k+1)} + Fx^{(k)} + b).$$

将上式改写为

$$x^{(k+1)} = (D - \omega E)^{-1}[(1 - \omega)D + \omega F]x^{(k)} + \omega(D - \omega E)^{-1}b.$$

同样, 将上式记为

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

其中

$$B = (D - \omega E)^{-1}[(1 - \omega)D + \omega F], \quad g = \omega(D - \omega E)^{-1}b.$$

因此, 三种迭代法可写成如下统一的迭代格式:

$$x^{(k+1)} = Bx^{(k)} + g, \quad k = 0, 1, 2, \dots, \quad (3.3.1)$$

B 称为迭代矩阵.

3.3.2 迭代法的收敛性

本小节针对迭代格式 (3.3.1) 讨论迭代法的收敛性. 首先介绍迭代法收敛的两个基本定理, 即一个充分条件和一个充分必要条件. 由此得到有关收敛性的几个推论.

定理 3.3.1 若迭代矩阵 B 的范数 $\|B\| < 1$, 则对任意给定的初始向量 $x^{(0)}$, 由迭代格式

$$x^{(k+1)} = Bx^{(k)} + g, \quad k = 0, 1, 2, \dots$$

产生的迭代序列 $\{x^{(k)}\}$ 都收敛于方程组 $x = Bx + g$ 的解 x^* , 并且有误差估计式

$$\|x^* - x^{(k)}\| \leq \frac{\|B\|}{1 - \|B\|} \|x^{(k)} - x^{(k-1)}\|, \quad (3.3.2)$$

$$\|x^* - x^{(k)}\| \leq \frac{\|B\|^k}{1 - \|B\|} \|x^{(1)} - x^{(0)}\|. \quad (3.3.3)$$

证明 (1) 证收敛性. 由题设知 $x^* = Bx^* + g$, 该式与 $x^{(k)} = Bx^{(k-1)} + g$ 相减得

$$x^* - x^{(k)} = B(x^* - x^{(k-1)}) = B^2(x^* - x^{(k-2)}) = \cdots = B^k(x^* - x^{(0)}).$$

由此可得

$$\|x^* - x^{(k)}\| \leq \|B\|^k \|x^* - x^{(0)}\|.$$

在上式两端令 $k \rightarrow \infty$ 取极限得

$$\lim_{k \rightarrow \infty} \|x^* - x^{(k)}\| \leq \lim_{k \rightarrow \infty} \|B\|^k \|x^* - x^{(0)}\| = 0,$$

即

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

所以

$$\lim_{k \rightarrow \infty} x^{(k)} = x^*.$$

由此可见, $\|B\|$ 越小, 则迭代序列收敛得越快.

(2) 证两个误差估计式.

$$\begin{aligned} \|x^{(k+n)} - x^{(k)}\| &= \|x^{(k+n)} - x^{(k+n-1)} + x^{(k+n-1)} - x^{(k+n-2)} + \cdots + x^{(k+1)} - x^{(k)}\| \\ &\leq \|x^{(k+n)} - x^{(k+n-1)}\| + \|x^{(k+n-1)} - x^{(k+n-2)}\| + \cdots + \|x^{(k+1)} - x^{(k)}\| \\ &= \sum_{i=1}^n \|x^{(k+i)} - x^{(k+i-1)}\|. \end{aligned} \quad (3.3.4)$$

而

$$\begin{aligned} \|x^{(k+i)} - x^{(k+i-1)}\| &= \|Bx^{(k+i-1)} + g - (Bx^{(k+i-2)} + g)\| \\ &\leq \|B\| \|x^{(k+i-1)} - x^{(k+i-2)}\| \leq \cdots \\ &\leq \|B\|^{i-1} \|x^{(k+1)} - x^{(k)}\|. \end{aligned} \quad (3.3.5)$$

将式 (3.3.5) 代入式 (3.3.4) 得

$$\|x^{(k+n)} - x^{(k)}\| \leq \sum_{i=1}^n \|B\|^{i-1} \|x^{(k+1)} - x^{(k)}\| = \frac{1 - \|B\|^n}{1 - \|B\|} \|x^{(k+1)} - x^{(k)}\|.$$

在上式两边令 $n \rightarrow \infty$ 取极限, 由迭代序列的收敛性及 $\|B\| < 1$ 知

$$x^{(k+n)} \rightarrow x^*, \quad \|B\|^n \rightarrow 0, \quad n \rightarrow \infty.$$

于是可得

$$\|x^* - x^{(k)}\| \leq \frac{\|x^{(k+1)} - x^{(k)}\|}{1 - \|B\|} \leq \frac{\|B\|}{1 - \|B\|} \|x^{(k)} - x^{(k-1)}\|.$$

又由式 (3.3.5) 知

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

将其代入式 (3.3.2) 得式 (3.3.3). 证毕.

由式 (3.3.2) 知, 当 $\|B\|$ 不太接近 1 时, 若事先给定精度 ε , 则可用 $\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon$ 作为迭代终止的条件, 即在迭代过程中, 若 $\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon$, 则取 $x^{(k)}$ 作为方程组的近似解.

也可用式 (3.3.3) 事先估计迭代次数. 令

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

上式两端取对数得

$$k \geq \ln \left(\frac{1 - \|B\|}{\|x^{(1)} - x^{(0)}\| \varepsilon} \right) / \ln \|B\|.$$

由此式确定的 k 可作为在给定精度 ε 下迭代次数的一个估计.

定理 3.3.1 是迭代序列收敛的一个充分条件, 也有对于任何的矩阵范数 $\|B\| < 1$ 不成立, 迭代序列仍可能收敛的情况.

定理 3.3.2 对任意给定的初始向量 $x^{(0)}$, 由迭代格式

$$x^{(k+1)} = Bx^{(k)} + g, \quad k = 0, 1, 2, \dots$$

产生的迭代序列 $\{x^{(k)}\}$ 都收敛的充分必要条件是以下两条件之一成立:

- (1) $B^k \rightarrow 0$ ($k \rightarrow \infty$);
- (2) 迭代矩阵 B 的谱半径 $\rho(B) < 1$.

证明 (1) 设 x^* 是方程组 $x = Bx + g$ 的解, 即 $x^* = Bx^* + g$, 该式与 $x^{(k)} = Bx^{(k-1)} + g$ 相减得

$$x^* - x^{(k)} = B(x^* - x^{(k-1)}) = B^2(x^* - x^{(k-2)}) = \dots = B^k(x^* - x^{(0)}).$$

必要性. 设 $x^{(k)} \rightarrow x^*$, 上式左端趋于零向量, 而右端由 $x^{(0)}$ 的任意性知 $x^* - x^{(0)} \neq 0$, 从而 $B^k \rightarrow 0$ ($k \rightarrow \infty$).

充分性. 若 $B^k \rightarrow 0$ ($k \rightarrow \infty$), 则上式右端趋于零向量, 从而左端趋于零向量, 即得 $x^{(k)} \rightarrow x^*$ ($k \rightarrow \infty$).

(2)

$$\lim_{k \rightarrow \infty} x^{(k)} = x^* \xLeftrightarrow{(1)} \lim_{k \rightarrow \infty} B^k = 0 \xLeftrightarrow{\text{定理 3.1.2}} \rho(B) < 1.$$

证毕.

定理 3.3.2 中的两个条件往往难以检验, 在实际应用中存在困难, 但由该定理可得几个应用方便的推论.

推论 3.3.1 超松弛迭代法收敛的必要条件是 $0 < \omega < 2$.

证明 超松弛迭代法的迭代矩阵

$$B = (D - \omega E)^{-1}[(1 - \omega)D + \omega F].$$

设 λ_i ($i = 1, 2, \dots, n$) 是矩阵 B 的特征值, 则

$$\begin{aligned} \lambda_1 \lambda_2 \cdots \lambda_n &= |B| = |(D - \omega E)^{-1}| |(1 - \omega)D + \omega F| \\ &= \frac{|(1 - \omega)D + \omega F|}{|D - \omega E|} = \frac{(1 - \omega)^n a_{11} a_{22} \cdots a_{nn}}{a_{11} a_{22} \cdots a_{nn}} = (1 - \omega)^n. \end{aligned}$$

若超松弛迭代法收敛, 则 $\rho(B) = \max_{1 \leq i \leq n} \{|\lambda_i|\} < 1$, 所以 $|1 - \omega|^n = |(1 - \omega)^n| = |\lambda_1 \lambda_2 \cdots \lambda_n| < 1$. 由此可得 $|1 - \omega| < 1 \Rightarrow 0 < \omega < 2$. 证毕.

引理 3.3.1 设 A 是严格对角占优矩阵, $0 < \omega \leq 1$, 当 $|\lambda| \geq 1$ 时, 矩阵

$$(\lambda + \omega - 1)D - \lambda\omega E - \omega F$$

是严格对角占优矩阵, 从而是非奇异矩阵.

证明 先证 $(\lambda + \omega - 1)D - \lambda\omega E - \omega F$ 是严格对角占优矩阵, 即证

$$|(\lambda + \omega - 1)a_{ii}| > \sum_{j=1}^{i-1} |\lambda\omega a_{ij}| + \sum_{j=i+1}^n |\omega a_{ij}|, \quad i = 1, 2, \dots, n.$$

而

$$\begin{aligned} & |(\lambda + \omega - 1)a_{ii}| - \sum_{j=1}^{i-1} |\lambda\omega a_{ij}| - \sum_{j=i+1}^n |\omega a_{ij}| \\ & \geq [|\lambda| - (1 - \omega)] |a_{ii}| - |\lambda|\omega \sum_{j=1}^{i-1} |a_{ij}| - \omega \sum_{j=i+1}^n |a_{ij}| \\ & > [|\lambda| - (1 - \omega)] \left(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^n |a_{ij}| \right) - |\lambda|\omega \sum_{j=1}^{i-1} |a_{ij}| - \omega \sum_{j=i+1}^n |a_{ij}| \\ & = (|\lambda| - 1)(1 - \omega) \sum_{j=1}^{i-1} |a_{ij}| + (|\lambda| - 1) \sum_{j=i+1}^n |a_{ij}| \geq 0, \quad i = 1, 2, \dots, n, \end{aligned}$$

从而 $(\lambda + \omega - 1)D - \lambda\omega E - \omega F$ 是严格对角占优矩阵. 根据线性代数理论知, 严格对角占优矩阵是非奇异矩阵. 证毕.

推论 3.3.2 设线性方程组的系数矩阵 A 是严格对角占优矩阵, 则对任意给定的初始向量 $x^{(0)}$, 雅可比迭代法、高斯-赛德尔迭代法和 $0 < \omega \leq 1$ 的超松弛迭代法收敛.

证明 (1) 雅可比迭代法的迭代矩阵为

$$B = D^{-1}(E + F),$$

则

$$|\lambda I - B| = |\lambda I - D^{-1}(E + F)| = |D^{-1}| |\lambda D - (E + F)|.$$

由 $A = D - E - F$ 是严格对角占优矩阵知, 当 $|\lambda| \geq 1$ 时, $\lambda D - (E + F)$ 是严格对角占优矩阵, 从而是非奇异矩阵. 因此, $|\lambda D - (E + F)| \neq 0$, 故当 $|\lambda| \geq 1$ 时, $|\lambda I - B| \neq 0$. 反之, $|\lambda I - B| = 0$, 必然 $|\lambda| < 1$. 据此, 矩阵 B 的特征值 $|\lambda| < 1 \Rightarrow \rho(B) < 1$. 由定理 3.3.2 知, 雅可比迭代法收敛.

(2) 超松弛迭代法 ($0 < \omega \leq 1$). 超松弛迭代法的迭代矩阵

$$B = (D - \omega E)^{-1}[(1 - \omega)D + \omega F],$$

则

$$\begin{aligned} |\lambda I - B| &= |\lambda I - (D - \omega E)^{-1}[(1 - \omega)D + \omega F]| \\ &= |(D - \omega E)^{-1}| |\lambda(D - \omega E) - [(1 - \omega)D + \omega F]| \\ &= |(D - \omega E)^{-1}| |(\lambda + \omega - 1)D - \lambda\omega E - \omega F|. \end{aligned}$$

由引理 3.3.1 知 $0 < \omega \leq 1$, 当 $|\lambda| \geq 1$ 时,

$$|(\lambda + \omega - 1)D - \lambda\omega E - \omega F| \neq 0.$$

又

$$|(D - \omega E)^{-1}| = \frac{1}{a_{11}a_{22} \cdots a_{nn}} \neq 0,$$

从而 $|\lambda I - B| \neq 0$. 据此可知, 对于 $0 < \omega \leq 1$, 只有当 $|\lambda| < 1$ 时, $|\lambda I - B| = 0$, 即矩阵 B 的特征值的模 $|\lambda| < 1 \Rightarrow \rho(B) < 1$. 据定理 3.3.2 知, $0 < \omega \leq 1$ 的超松弛迭代法收敛. 当 $\omega = 1$ 时的超松弛迭代法即高斯-赛德尔迭代法, 所以高斯-赛德尔迭代法收敛. 证毕.

推论 3.3.3 设线性方程组的系数矩阵 A 是对称正定矩阵, 则对任意给定的初始向量 $x^{(0)}$, 雅可比迭代法收敛的充分必要条件是 $2D - A$ 是对称正定矩阵.

证明 必要性. 设雅可比迭代法收敛. 由 A 是对称矩阵知 $A = D - E - E^T$, 则迭代矩阵 $B = D^{-1}(E + E^T) = I - D^{-1}A = D^{-1/2}(I - D^{-1/2}AD^{-1/2})D^{1/2}$. 设 $D^{-1/2}AD^{-1/2}$ 的特征值为 μ , 则 $I - D^{-1/2}AD^{-1/2}$ 的特征值为 $1 - \mu$. 由于 B 相似于 $I - D^{-1/2}AD^{-1/2}$, 故 B 的特征值为 $\lambda = 1 - \mu$. 由雅可比迭代法收敛知 $\rho(B) < 1$. 由此得 $|\lambda| = |1 - \mu| < 1 \Rightarrow 0 < \mu < 2$, 故 $2I - D^{-1/2}AD^{-1/2}$ 的特征值为 $2 - \mu \in (0, 2)$, 所以 $2I - D^{-1/2}AD^{-1/2}$ 是正定矩阵. 由此可知 $2D - A = D^{1/2}(2I - D^{-1/2}AD^{-1/2})D^{1/2}$ 是正定矩阵.

充分性. 设 $2D - A$ 是正定矩阵, 则由 A 是对称正定矩阵知, $D^{-1/2}AD^{-1/2}$ 也是对称正定矩阵, 设其特征值为 μ ($\mu > 0$). 同必要性的证明知, B 的特征值为 $\lambda = 1 - \mu$. 由 $\mu > 0$ 知

$$\lambda = 1 - \mu < 1. \quad (3.3.6)$$

另一方面,

$$\begin{aligned} -B &= -D^{-1/2}(I - D^{-1/2}AD^{-1/2})D^{1/2} \\ &= D^{-1/2}[I - D^{-1/2}(2D - A)D^{-1/2}]D^{1/2}. \end{aligned}$$

由此可知, $-B$ 的特征值与 $I - D^{-1/2}(2D - A)D^{-1/2}$ 的特征值相同. 由 $2D - A$ 是正定矩阵知, $D^{-1/2}(2D - A)D^{-1/2}$ 是正定矩阵, 设其特征值为 $\bar{\mu}$ ($\bar{\mu} > 0$), 而 $I - D^{-1/2}(2D - A)D^{-1/2}$ 的特征值为 $1 - \bar{\mu}$. B 的特征值为 λ , 则 $-B$ 的特征值为 $-\lambda$, 故有

$$-\lambda = 1 - \bar{\mu} < 1,$$

即

$$\lambda > -1. \quad (3.3.7)$$

由式 (3.3.6) 和 (3.3.7) 知 $|\lambda| < 1 \Rightarrow \rho(B) < 1$, 所以雅可比迭代法收敛. 证毕.

推论 3.3.4 设线性方程组系数矩阵 A 是对称正定矩阵, 则对任意给定的初始向量 $x^{(0)}$, 超松弛迭代法收敛的充分必要条件是 $0 < \omega < 2$.

*** 证明** 设 λ 是超松弛迭代法的迭代矩阵

$$B = (D - \omega E)^{-1}[(1 - \omega)D + \omega E^T]$$

的特征值, y 是相应的特征向量, 记 $\bar{y}$ 是 y 的共轭向量, 则有

$$[(1 - \omega)D + \omega E^T]y = \lambda(D - \omega E)y,$$

从而

$$(1 - \omega)\bar{y}^T Dy + \omega\bar{y}^T E^T y = \lambda[\bar{y}^T Dy - \omega\bar{y}^T E y]. \quad (3.3.8)$$

设 $\bar{y}^T E y = p + iq$, 则

$$\bar{y}^T E^T y = \overline{\bar{y}^T E y}^T = p - iq.$$

由 A 是正定矩阵知, 其对角元全大于零, 故 D 是正定矩阵, 所以 $\bar{y}^T D y = d > 0$.

由式 (3.3.8) 得

$$\lambda = \frac{(1-\omega)d + \omega(p-iq)}{d - \omega(p+iq)} \triangleq \frac{s}{t},$$

而

$$\begin{aligned} |s|^2 - |t|^2 &= (d - \omega d + \omega p)^2 + \omega^2 q^2 - [(d - \omega p)^2 + \omega^2 q^2] \\ &= d\omega(2 - \omega)(2p - d), \end{aligned}$$

$$2p - d = p + iq + p - iq - d = \bar{y}^T (E + E^T - D)y = -\bar{y}^T A y = -a,$$

故

$$|s|^2 - |t|^2 = ad\omega(\omega - 2).$$

由此可见, 当 $0 < \omega < 2$ 时, $|s|^2 - |t|^2 < 0$. 由此得到 $|\lambda| = \frac{|s|}{|t|} < 1$, 从而 $\rho(B) < 1$, 故超松弛迭代法收敛. 反之, 若超松弛迭代法收敛, 则 $|\lambda| < 1 \Rightarrow |s| < |t| \Rightarrow ad\omega(\omega - 2) < 0 \Rightarrow 0 < \omega < 2$. 证毕.

超松弛迭代法收敛的快慢取决于松弛因子 ω 的选取, 自然希望选取最佳松弛因子 ω , 使得超松弛迭代法具有较快的收敛速度, 但目前尚无可供计算最佳松弛因子的实用方法. 但当 A 是对称正定的三对角矩阵时, Young 于 1950 年从理论上建立了最佳松弛因子的计算公式

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(B_J)]^2}},$$

其中 B_J 为雅可比迭代矩阵.

一般来说, 确定最佳松弛因子比较困难. 通常在实际计算时, 根据线性方程组系数矩阵的特点及其实际计算经验, 采用试算的办法来确定一个合适的 ω .

例 3.3.1 设线性方程组 $Ax = b$ 的系数矩阵

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix},$$

若要雅可比迭代法、高斯-赛德尔迭代法收敛, 请给出数 a 的取值范围.

解法 1 显然, 当 $|a| < 1/2$ 时, A 是严格对角占优矩阵. 由推论 3.3.2 知, 雅可比迭代法、高斯-赛德尔迭代法均收敛.

解法 2 对于雅可比迭代法考察其迭代矩阵

$$B = \begin{pmatrix} 0 & -a & -a \\ -a & 0 & -a \\ -a & -a & 0 \end{pmatrix}$$

的 ∞ 范数或谱半径.

显然, 若 $\|B\|_{\infty} = |2a| < 1$, 即当 $|a| < 1/2$ 时, 据定理 3.3.1 知, 雅可比迭代法收敛. 或者

$$\det(\lambda I - B) = \lambda^3 - 3a^2\lambda + 2a^3 = (\lambda - a)^2(\lambda + 2a) = 0.$$

当 $\rho(B) = |2a| < 1$, 即 $|a| < 1/2$ 时, 据定理 3.3.2 知, 雅可比迭代法收敛.

对于高斯-赛德尔迭代法, 注意到 A 是对称矩阵, 故考察其正定性. 据线性代数知, 矩阵 A 正定的充分必要条件是 A 的顺序主子式全大于零. 据此,

$$\Delta_1 = 1 > 0, \quad \Delta_2 = \begin{vmatrix} 1 & a \\ a & 1 \end{vmatrix} = 1 - a^2 > 0, \quad \Delta_3 = |A| = (1 - a)^2(1 + 2a) > 0.$$

解不等式

$$\begin{cases} 1 - a^2 > 0, \\ (1 - a)^2(1 + 2a) > 0 \end{cases}$$

得 $-1/2 < a < 1$. 此时, A 是对称正定矩阵, 据推论 3.3.4 知, 高斯-赛德尔迭代法收敛. 解毕.

例 3.3.2 给定矩阵 A 如下, 讨论求解方程组 $Ax = b$ 的三种迭代法的收敛性:

$$(1) A = \begin{pmatrix} 1 & -1 & 2 \\ -1 & 3 & 0 \\ 2 & 0 & 7 \end{pmatrix}; \quad (2) A = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}.$$

解 (1) A 是对称矩阵, 其顺序主子式分别为

$$\Delta_1 = 1 > 0, \quad \Delta_2 = \begin{vmatrix} 1 & -1 \\ -1 & 3 \end{vmatrix} = 2 > 0, \quad \Delta_3 = |A| = 2 > 0,$$

故 A 是对称正定矩阵.

$$2D - A = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 3 & 0 \\ -2 & 0 & 7 \end{pmatrix},$$

其各阶顺序主子式与矩阵 A 的相同, 所以 $2D - A$ 也是对称正定矩阵, 故雅可比迭代法收敛. 高斯-赛德尔迭代法和 $0 < \omega < 2$ 的超松弛迭代法收敛.

(2) A 是对称矩阵, 其各阶顺序主子式分别为

$$\Delta_1 = 1 > 0, \quad \Delta_2 = \frac{3}{4} > 0, \quad \Delta_3 = |A| = \frac{1}{2} > 0,$$

即 A 是对称正定矩阵, 所以高斯-赛德尔迭代法和 $0 < \omega < 2$ 的超松弛迭代法收敛, 但 $|2D - A| = 0$, 即 $2D - A$ 不正定, 雅可比迭代法发散. 解毕.

3.4 共轭梯度法

本节介绍求解系数矩阵 A 是对称正定矩阵的线性方程组 $Ax = b$ 的共轭梯度法.

共轭梯度法是把求解线性方程组的问题转化为求解一个与之等价的二次函数极小化的问题. 从任意给定的初始点出发, 沿一组关于矩阵 A 的共轭方向进行线性搜索, 在无舍入误差的假定下, 最多迭代 n 次 (其中 n 为矩阵 A 的阶数), 就可求得二次函数的极小点, 也就求得了线性方程组 $Ax = b$ 的解.

对于良态大型稀疏线性方程组, 通常该法只需经过比方程阶数 n 小得多的迭代次数就能获得满足精度要求的近似解.

3.4.1 求解线性方程组与求解二次函数极小点的等价性

在高等数学中, n 元函数 $f(x)$ 的梯度定义为

$$\nabla f(x) = \text{grad } f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)^T$$

例 3.4.1 求 n 元二次函数 $f(x) = \frac{1}{2}x^T Ax - b^T x$ 的梯度 (其中 A 为 n 阶对称正定矩阵).

解 设 $A = (a_{ij})_{n \times n}$, $x = (x_1, x_2, \dots, x_n)^T$, $b = (b_1, b_2, \dots, b_n)^T$, 于是

$$f(x) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j - \sum_{i=1}^n b_i x_i.$$

由 A 是对称矩阵知 $a_{jk} = a_{kj}$, 对 $f(x)$ 关于 x_k 求偏导数得

$$\frac{\partial f(x)}{\partial x_k} = \frac{1}{2} \left(\sum_{j=1}^n a_{kj} x_j + \sum_{i=1}^n a_{ik} x_i \right) - b_k = \sum_{j=1}^n a_{kj} x_j - b_k, \quad k = 1, 2, \dots, n.$$

由此得

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^n a_{1j}x_j \\ \sum_{j=1}^n a_{2j}x_j \\ \vdots \\ \sum_{j=1}^n a_{nj}x_j \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = Ax - b.$$

解毕.

下述定理给出了求系数矩阵 A 是对称正定矩阵的线性方程组 $Ax = b$ 的解与求二次函数 $f(x) = \frac{1}{2}x^T Ax - b^T x$ 极小点的等价性:

定理 3.4.1 设 A 是 n 阶对称正定矩阵, 则 x^* 是方程组 $Ax = b$ 的解的充分必要条件是 x^* 是二次函数 $f(x) = \frac{1}{2}x^T Ax - b^T x$ 的极小点, 即

$$Ax^* = b \Leftrightarrow f(x^*) = \min_{x \in \mathbb{R}^n} f(x).$$

证明 充分性. 设 x^* 是 $f(x)$ 的极小点, 则由无约束最优化问题最优解的必要条件知

$$\nabla f(x^*) = Ax^* - b = 0,$$

即 x^* 是方程组 $Ax = b$ 的解.

必要性. 若 x^* 是方程组 $Ax = b$ 的解, 即 $Ax^* = b$. 注意到 A 是对称正定矩阵, 故 $\forall x \in \mathbb{R}^n$ 有

$$\begin{aligned} f(x) - f(x^*) &= \frac{1}{2}x^T Ax - b^T x - \frac{1}{2}x^{*T} Ax^* + b^T x^* \\ &= \frac{1}{2}(x^T Ax - 2b^T x + x^{*T} Ax^*) - x^{*T} Ax^* + b^T x^* \\ &= \frac{1}{2}(x^T Ax - 2(Ax^*)^T x + x^{*T} Ax^*) - (Ax^* - b)^T x^* \\ &= \frac{1}{2}(x - x^*)^T A(x - x^*) \\ &\geq 0, \end{aligned}$$

即 x^* 是 $f(x)$ 的极小点, 进而由 A 是正定矩阵知, x^* 是 $f(x)$ 的最小点. 证毕.

3.4.2 共轭梯度法

本小节介绍求解二次函数 $f(x) = \frac{1}{2}x^T Ax - b^T x$ (其中 A 为 n 阶对称正定矩阵) 极小点的共轭梯度法. 为此, 先介绍共轭向量的定义.

定义 3.4.1 (共轭向量) 设 A 是 n 阶对称正定矩阵, 若 $\mathbb{R}^n$ 中的一组非零向量 $d^{(0)}, d^{(1)}, \dots, d^{(m)}$ 满足

$$(d^{(i)}, Ad^{(j)}) = 0, \quad i \neq j,$$

则称 $d^{(i)}, d^{(j)}$ 是关于矩阵 A 的共轭向量, $d^{(0)}, d^{(1)}, \dots, d^{(m)}$ 为关于矩阵 A 的共轭向量组.

当 $A = I$ 时, 共轭向量组即正交向量组, 可以认为共轭是正交概念的推广.

定理 3.4.2 关于矩阵 A 的共轭向量 $d^{(0)}, d^{(1)}, \dots, d^{(m)}$ 线性无关.

证明 设

$$c_0 d^{(0)} + c_1 d^{(1)} + \dots + c_m d^{(m)} = 0,$$

对上式两边用 $Ad^{(i)} (i = 0, 1, \dots, m)$ 作内积, 则有

$$\left(Ad^{(i)}, \sum_{j=0}^m c_j d^{(j)} \right) = \sum_{j=0}^m c_j (Ad^{(i)}, d^{(j)}) = c_i (Ad^{(i)}, d^{(i)}) = 0, \quad i = 0, 1, \dots, m.$$

由 A 是对称正定矩阵及 $d^{(i)} \neq 0 (i = 0, 1, \dots, m)$ 知 $(Ad^{(i)}, d^{(i)}) > 0$, 从而 $c_i = 0 (i = 0, 1, \dots, m)$, 即向量 $d^{(0)}, d^{(1)}, \dots, d^{(m)}$ 线性无关. 证毕.

由定理 3.4.2 知, 关于矩阵 A 的共轭向量组最多含 n 个线性无关的向量.

共轭梯度法在形式上具有迭代法的特征, 即给定初始向量 $x^{(0)}$, 由迭代格式

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} \quad (3.4.1)$$

产生的迭代序列 $x^{(1)}, x^{(2)}, x^{(3)}, \dots$ 在无舍入误差的假定下, 最多经过 n 次迭代就可求得 $f(x)$ 的最小点, 也就是方程组 $Ax = b$ 的解, 从这点上来看, 它又是直接法.

共轭梯度法中关键的两点是, 确定迭代格式 (3.4.1) 中的搜索方向 $d^{(k)}$ 和最佳步长 $\alpha_k (\alpha_k \geq 0)$. 实际上, 搜索方向 $d^{(k)}$ 是关于矩阵 A 的共轭向量, 在迭代中逐步构造之. 步长 α_k 的确定原则是给定迭代点 $x^{(k)}$ 和搜索方向 $d^{(k)}$ 后, 要求选取非负实数 α_k , 使得 $f(x^{(k)} + \alpha_k d^{(k)})$ 达到最小, 即选择 α_k , 满足

$$f(x^{(k)} + \alpha_k d^{(k)}) = \min_{0 \leq \alpha} f(x^{(k)} + \alpha d^{(k)}). \quad (3.4.2)$$

下面首先导出最佳步长 α_k 的计算公式. 设迭代点 $x^{(k)}$ 和搜索方向 $d^{(k)}$ 已经给定, 由式 (3.4.2) 知, α_k 可以通过一元函数 $\phi(\alpha) = f(x^{(k)} + \alpha d^{(k)})$ 的极小化

$$\min_{0 \leq \alpha} \phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}) \quad (3.4.3)$$

来求得. 根据多元复合函数的求导法得

$$\phi'(\alpha) = \nabla f(x^{(k)} + \alpha d^{(k)})^T d^{(k)}.$$

令

$$\begin{aligned} 0 = \phi'(\alpha) &= \nabla f(x^{(k)} + \alpha d^{(k)})^T d^{(k)} = [A(x^{(k)} + \alpha d^{(k)}) - b]^T d^{(k)} \\ &= (Ax^{(k)} - b + \alpha Ad^{(k)})^T d^{(k)} = (-r^{(k)} + \alpha Ad^{(k)})^T d^{(k)}, \end{aligned}$$

其中 $r^{(k)} = b - Ax^{(k)}$. 由此解得最佳步长

$$\alpha_k = \frac{r^{(k)T} d^{(k)}}{d^{(k)T} A d^{(k)}}. \quad (3.4.4)$$

下面确定搜索方向 $d^{(k)}$. 给定初始向量 $x^{(0)}$ 后, 由于负梯度方向是函数下降最快的方向, 故第 1 次迭代取搜索方向 $d^{(0)} = r^{(0)} = -\nabla f(x^{(0)}) = b - Ax^{(0)}$. 令

$$x^{(1)} = x^{(0)} + \alpha_0 d^{(0)},$$

其中 $\alpha_0 = \frac{r^{(0)T} d^{(0)}}{d^{(0)T} A d^{(0)}}$. 第 2 次迭代时, 从 $x^{(1)}$ 出发的搜索方向不再取 $r^{(1)}$, 而是选取 $d^{(1)} = r^{(1)} + \beta_0 d^{(0)}$, 使得 $d^{(1)}$ 与 $d^{(0)}$ 是关于矩阵 A 的共轭向量, 即要求 $d^{(1)}$ 满足 $(d^{(1)}, Ad^{(0)}) = 0$, 由此可求得参数

$$\beta_0 = -\frac{r^{(1)T} A d^{(0)}}{d^{(0)T} A d^{(0)}}.$$

然后, 从 $x^{(1)}$ 出发, 沿方向 $d^{(1)}$ 进行搜索得

$$x^{(2)} = x^{(1)} + \alpha_1 d^{(1)},$$

其中 α_1 由式 (3.4.4) 确定.

一般地, 设已经求出 $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, 计算 $r^{(k+1)} = b - Ax^{(k+1)}$. 令 $d^{(k+1)} = r^{(k+1)} + \beta_k d^{(k)}$, 选取 β_k , 使得 $d^{(k+1)}$ 与 $d^{(k)}$ 是关于 A 的共轭向量, 即要求 $d^{(k+1)}$ 满足 $(d^{(k+1)}, Ad^{(k)}) = 0$. 由此可得

$$\beta_k = -\frac{r^{(k+1)T} A d^{(k)}}{d^{(k)T} A d^{(k)}}, \quad (3.4.5)$$

这就确定了搜索方向 $d^{(k+1)}$.

综上所述, 共轭梯度法的计算公式为

$$\begin{cases} d^{(0)} = r^{(0)} = b - Ax^{(0)}, \\ \alpha_k = \frac{r^{(k)T} d^{(k)}}{d^{(k)T} A d^{(k)}}, \\ x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}, \\ r^{(k+1)} = b - Ax^{(k+1)}, \\ \beta_k = -\frac{r^{(k+1)T} A d^{(k)}}{d^{(k)T} A d^{(k)}}, \\ d^{(k+1)} = r^{(k+1)} + \beta_k d^{(k)}. \end{cases} \quad (3.4.6)$$

下面讨论共轭梯度法中产生的非零残向量 $\{r^{(k)}\}$ 是正交向量组, 搜索方向 $\{d^{(k)}\}$ 是关于矩阵 A 的共轭向量组, 进而得出共轭梯度法的收敛性.

定理 3.4.3 设 $r^{(i)}, d^{(i)} (i = 0, 1, \dots, n-1)$ 分别是共轭梯度法中产生的非零残向量和搜索方向, 则

$$(1) (r^{(k)}, d^{(k-1)}) = 0;$$

$$(2) (r^{(k)}, d^{(k)}) = (r^{(k)}, r^{(k)});$$

(3) $r^{(0)}, r^{(1)}, \dots, r^{(k)} (k \leq n-1)$ 是正交向量组, $d^{(0)}, d^{(1)}, \dots, d^{(k)} (k \leq n-1)$ 是关于 A 的共轭向量组.

证明 (1) $r^{(k)} = b - Ax^{(k)} = b - A(x^{(k-1)} + \alpha_{k-1}d^{(k-1)}) = r^{(k-1)} - \alpha_{k-1}Ad^{(k-1)}$, 故

$$(r^{(k)}, d^{(k-1)}) = (r^{(k-1)}, d^{(k-1)}) - \alpha_{k-1}(Ad^{(k-1)}, d^{(k-1)}),$$

代入 α_{k-1} , 则

$$(r^{(k)}, d^{(k-1)}) = 0.$$

(2) $d^{(k)} = r^{(k)} + \beta_{k-1}d^{(k-1)}$, 并利用结论 (1) 可得

$$(r^{(k)}, d^{(k)}) = (r^{(k)}, r^{(k)}) + \beta_{k-1}(r^{(k)}, d^{(k-1)}) = (r^{(k)}, r^{(k)}).$$

(3) 用归纳法证明.

第 1 步 证当 $m = 1$ 时, $r^{(0)}, r^{(1)}$ 正交, $d^{(0)}, d^{(1)}$ 关于矩阵 A 共轭.

$$(r^{(1)}, r^{(0)}) = (r^{(0)} - \alpha_0 Ad^{(0)}, r^{(0)}) = (r^{(0)}, r^{(0)}) - \alpha_0(Ad^{(0)}, r^{(0)}),$$

注意到 $d^{(0)} = r^{(0)}$, 代入 α_0 则

$$(r^{(1)}, r^{(0)}) = 0,$$

$$(d^{(1)}, Ad^{(0)}) = (r^{(1)} + \beta_0 d^{(0)}, Ad^{(0)}) = (r^{(1)}, Ad^{(0)}) + \beta_0(d^{(0)}, Ad^{(0)}).$$

代入 β_0 , 则

$$(d^{(1)}, Ad^{(0)}) = 0.$$

第 2 步 假定当 $m = k$ 时结论成立, 即 $r^{(0)}, r^{(1)}, \dots, r^{(k)}$ 是正交向量组, $d^{(0)}, d^{(1)}, \dots, d^{(k)}$ 是关于矩阵 A 的共轭向量组.

第 3 步 证当 $m = k+1$ 时结论也成立, 即证 $r^{(0)}, r^{(1)}, \dots, r^{(k)}, r^{(k+1)}$ 是正交向量组, $d^{(0)}, d^{(1)}, \dots, d^{(k)}, d^{(k+1)}$ 是关于矩阵 A 的共轭向量组. 当 $i = 0, 1, \dots, k-1$ 时, 由 $d^{(i)} = r^{(i)} + \beta_{i-1}d^{(i-1)}$ 知 $r^{(i)} = d^{(i)} - \beta_{i-1}d^{(i-1)}$, 所以

$$\begin{aligned} (r^{(k+1)}, r^{(i)}) &= (r^{(k)} - \alpha_k Ad^{(k)}, r^{(i)}) = (r^{(k)}, r^{(i)}) - \alpha_k(Ad^{(k)}, r^{(i)}) \\ &= (r^{(k)}, r^{(i)}) - \alpha_k[(Ad^{(k)}, d^{(i)}) - \beta_{i-1}(Ad^{(k)}, d^{(i-1)})] \end{aligned}$$

$=0$ (根据第 2 步假定).

当 $i = k$ 时,

$$\begin{aligned}(r^{(k+1)}, r^{(k)}) &= (r^{(k)}, r^{(k)}) - \alpha_k [(Ad^{(k)}, d^{(k)}) - \beta_{k-1} (Ad^{(k)}, d^{(k-1)})] \\ &= (r^{(k)}, r^{(k)}) - \frac{(r^{(k)}, d^{(k)})}{(d^{(k)}, Ad^{(k)})} (Ad^{(k)}, d^{(k)}) \\ &= (r^{(k)}, r^{(k)}) - (r^{(k)}, d^{(k)}) \\ &= 0,\end{aligned}$$

即 $r^{(0)}, r^{(1)}, \dots, r^{(k)}, r^{(k+1)}$ 是正交向量组.

下面证

$$(d^{(k+1)}, Ad^{(i)}) = 0, \quad i = 0, 1, \dots, k.$$

由 $d^{(k+1)} = r^{(k+1)} + \beta_k d^{(k)}$ 知

$$(d^{(k+1)}, Ad^{(i)}) = (r^{(k+1)}, Ad^{(i)}) + \beta_k (d^{(k)}, Ad^{(i)}). \quad (3.4.7)$$

当 $i = 0, 1, \dots, k-1$ 时, 由 $r^{(i+1)} = r^{(i)} - \alpha_i Ad^{(i)}$ 得

$$Ad^{(i)} = (r^{(i)} - r^{(i+1)})/\alpha_i, \quad \alpha_i > 0.$$

代入式 (3.4.7), 并由 $\{r^{(i)}\}$ 的正交性及第 2 步假定得

$$(d^{(k+1)}, Ad^{(i)}) = \frac{(r^{(k+1)}, r^{(i)} - r^{(i+1)})}{\alpha_i} + \beta_k (d^{(k)}, Ad^{(i)}) = 0.$$

当 $i = k$ 时,

$$(d^{(k+1)}, Ad^{(k)}) = (r^{(k+1)}, Ad^{(k)}) + \beta_k (d^{(k)}, Ad^{(k)}) = 0,$$

所以有

$$(d^{(k+1)}, Ad^{(i)}) = 0, \quad i = 0, 1, \dots, k,$$

即 $d^{(0)}, d^{(1)}, \dots, d^{(k)}, d^{(k+1)}$ 是关于矩阵 A 的共轭向量组, 也即结论在 $m = k+1$ 时也成立. 根据归纳法, 结论得证. 证毕.

利用定理 3.4.3 的结论, 可以将 α_k 和 β_k 的表达式化简为

$$\alpha_k = \frac{r^{(k)T} r^{(k)}}{d^{(k)T} Ad^{(k)}}, \quad (3.4.4')$$

$$\beta_k = \frac{r^{(k+1)T} r^{(k+1)}}{r^{(k)T} r^{(k)}}. \quad (3.4.5')$$

事实上, 将定理 3.4.3 的结论 (2) $(r^{(k)}, d^{(k)}) = (r^{(k)}, r^{(k)})$ 代入 α_k 的表达式 (3.4.4) 即得 (3.4.4'). 由 $Ad^{(k)} = (r^{(k)} - r^{(k+1)})/\alpha_k$, 则

$$\beta_k = -\frac{r^{(k+1)T} Ad^{(k)}}{d^{(k)T} Ad^{(k)}} = -\frac{r^{(k+1)T} (r^{(k)} - r^{(k+1)})}{\alpha_k d^{(k)T} Ad^{(k)}} = \frac{r^{(k+1)T} r^{(k+1)}}{\alpha_k d^{(k)T} Ad^{(k)}}.$$

将 α_k 的表达式 (3.4.4') 代入上式, 即得式 (3.4.5').

另外, 注意到 $r^{(k)} = b - Ax^{(k)} = -\nabla f(x^{(k)})$, 若 $r^{(k)} \neq 0$, 由定理 3.4.3 的结论 (2) 有

$$(-\nabla f(x^{(k)}), d^{(k)}) = (r^{(k)}, d^{(k)}) = (r^{(k)}, r^{(k)}) > 0.$$

这说明 $d^{(k)}$ 是 $f(x)$ 在 $x^{(k)}$ 处的下降方向, 即有 $f(x^{(k+1)}) < f(x^{(k)})$, 即 $\{f(x^{(k)})\}$ 是单调下降数列. 据此, 共轭梯度法是一个下降算法.

定理 3.4.4 设 A 是 n 阶对称正定矩阵, 用共轭梯度法求解线性方程组 $Ax = b$, 或等价地, 求二次函数 $f(x) = \frac{1}{2}x^T Ax - b^T x$ 的极小点, 若计算过程中无舍入误差, 则最多迭代 n 次就可得到方程组 $Ax = b$ 的解或二次函数的最小点.

证明 在共轭梯度法的迭代过程中, 若某个 $r^{(k)} = b - Ax^{(k)} = 0$ ($0 \leq k \leq n-1$), 则 $x^{(k)}$ 是方程组的准确解; 若 $r^{(k)} \neq 0$ ($k = 0, 1, \dots, n-1$), 由定理 3.4.3 的结论 (3) 知, 残向量 $r^{(0)}, r^{(1)}, \dots, r^{(n-1)}, r^{(n)}$ 是正交向量组, 而在向量空间 R^n 中最多有 n 个互相正交的向量, 必然有 $r^{(n)} = b - Ax^{(n)} = 0$, 即 $x^{(n)}$ 是方程组的准确解. 证毕.

定理 3.4.5 设 $\{x^{(k)}\}$ 是用共轭梯度法求得的迭代序列, 则有误差估计式

$$\|x^{(k)} - x^*\|_A \leq 2 \left(\frac{\sqrt{K} - 1}{\sqrt{K} + 1} \right)^k \|x^{(0)} - x^*\|_A,$$

其中范数 $\|x\|_A = \sqrt{x^T Ax}$, $K = \text{Cond}_2(A)$.

证明 略.

从定理 3.4.5 知, 方程组系数矩阵的条件数越大, 收敛比越接近于 1, 收敛速度越慢; 反之, 收敛速度越快.

由定理 3.4.4 的证明过程知, 若 $r^{(k)} = 0$, 则 $x^{(k)}$ 就是线性方程组的解, 故在计算过程中将 $\|r^{(k+1)}\| \leq \varepsilon$ 作为迭代终止的条件.

共轭梯度的计算过程如下:

- (1) 给定初始近似向量 $x^{(0)}$ 及精度 $\varepsilon > 0$;
- (2) 计算 $r^{(0)} = b - Ax^{(0)}$, 取 $d^{(0)} = r^{(0)}$;
- (3) for $k = 0$ to $n - 1$ do

$$(i) \quad \alpha_k = \frac{r^{(k)T} r^{(k)}}{d^{(k)T} Ad^{(k)}};$$

- (ii) $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$;
 (iii) $r^{(k+1)} = b - Ax^{(k+1)}$;
 (iv) 若 $\|r^{(k+1)}\| \leq \varepsilon$ 或 $k+1 = n$, 则输出近似解 $x^{(k+1)}$, 停止;
 否则, 转 (v);
 (v) $\beta_k = \frac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2}$;
 (vi) $d^{(k+1)} = r^{(k+1)} + \beta_k d^{(k)}$;
 end do

定理 3.4.4 表明, 计算中若无舍入误差, 则至多迭代 n 次就求得了方程组的准确解. 但实际计算中存在舍入误差, 受舍入误差的影响, 残向量 $\{r^{(k)}\}$ 间不能精确满足正交关系, 搜索方向 $\{d^{(k)}\}$ 也不可能精确满足共轭关系, 因而迭代 n 次不一定就能求得方程组的准确解. 一般来说, 若 $r^{(n)} \neq 0$ (方程组的系数矩阵越病态, $r^{(n)}$ 就偏离零越远), 这时, 仍继续进行迭代, 函数 $f(x)$ 的值将继续减少, 总能求得一个更好的近似解. 在大多数情况下, 方程组的系数矩阵不是十分病态的, 往往不必迭代 n 次, $\|r^{(k)}\|$ 就充分小, 这时可以终止迭代, 取 $x^{(k)}$ 作为方程组的近似解.

计算经验表明, 对于不是十分病态的问题, 共轭梯度法收敛较快, 迭代次数远小于系数矩阵的阶数 n . 对于病态问题, 只要进行足够多次迭代 (迭代次数大约为矩阵阶数 n 的 3~5 倍) 后, 一般也能得到满意的结果, 因而共轭梯度法是求解高阶稀疏线性方程组的一个有效、常用的方法.

例 3.4.2 给定初始向量 $x^{(0)} = (0, 0, 0)^T$, 用共轭梯度法求解对称正定方程组

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix}.$$

解 取 $d^{(0)} = r^{(0)} = b - Ax^{(0)} = b = (3, 1, 3)^T$, 则

$$\|r^{(0)}\|_2^2 = 19, \quad d^{(0)T} A d^{(0)} = 55.$$

故

$$\alpha_0 = \frac{\|r^{(0)}\|_2^2}{d^{(0)T} A d^{(0)}} = \frac{19}{55}, \quad x^{(1)} = x^{(0)} + \alpha_0 d^{(0)} = \frac{19}{55} (3, 1, 3)^T.$$

又

$$r^{(1)} = b - Ax^{(1)} = \frac{6}{55} (-1, 6, -1)^T, \quad \|r^{(1)}\|_2^2 = 38 \times \frac{6^2}{55^2},$$

$$\beta_0 = \frac{\|r^{(1)}\|_2^2}{\|r^{(0)}\|_2^2} = \frac{72}{55^2}, \quad d^{(1)} = r^{(1)} + \beta_0 d^{(0)} = \frac{6 \times 19}{55^2} (-1, 18, -1)^T,$$

$$d^{(1)T}Ad^{(1)} = \frac{(6 \times 19)^2 \times 6}{55^3}, \quad \alpha_1 = \frac{\|r^{(1)}\|_2^2}{d^{(1)T}Ad^{(1)}} = \frac{55}{57},$$

$$x^{(2)} = x^{(1)} + \alpha_1 d^{(1)} = (1, 1, 1)^T, \quad r^{(2)} = b - Ax^{(2)} = (0, 0, 0)^T.$$

迭代两次求得了方程组的解 $x^* = x^{(2)} = (1, 1, 1)^T$. 解毕.

*3.5 基于伽辽金原理的迭代法

本节介绍一类基于伽辽金 (Galerkin) 原理建立的求解大型非对称线性方程组 $Ax = b$ 的迭代法. 该方法有些在理论上尚不十分完善, 但大量的实际计算表明, 这类算法还是十分有效的.

3.5.1 伽辽金原理和克雷洛夫子空间

设求解的线性方程组为

$$Ax = b, \quad |A| \neq 0, \quad (3.5.1)$$

其中 $A \in \mathbf{R}^{n \times n}$ 为大型非对称矩阵, $b \in \mathbf{R}^n$. 本节所用的范数 $\|\cdot\|$ 均为 2 范数.

任意给定 $x^{(0)} \in \mathbf{R}^n$, 令 $x = x^{(0)} + z$, 则方程组 (3.5.1) 等价于

$$Az = r^{(0)}, \quad (3.5.2)$$

其中 $r^{(0)} = b - Ax^{(0)}$. 若求得方程组 (3.5.2) 的近似解 $z^{(m)}$, 则原方程组 (3.5.1) 的近似解为 $x^{(m)} = x^{(0)} + z^{(m)}$. 以下讨论方程组 (3.5.2) 的求解问题.

设 $\mathbf{R}^n$ 中两个 m 维子空间 K_m, L_m 的基分别为 $\{v_i\}_{i=1}^m, \{w_i\}_{i=1}^m$, 则

$$K_m = \text{span}\{v_1, v_2, \dots, v_m\}, \quad L_m = \text{span}\{w_1, w_2, \dots, w_m\}.$$

下面给出求解方程组 (3.5.2) 的伽辽金原理.

伽辽金原理 在子空间 K_m 中找方程组 (3.5.2) 的近似解 $z^{(m)}$, 使得残向量 $r^{(0)} - Az^{(m)}$ 与 L_m 中的所有向量正交, 即求 $z^{(m)} \in K_m$, 使得

$$(r^{(0)} - Az^{(m)}, w_i) = 0, \quad i = 1, 2, \dots, m. \quad (3.5.3)$$

下面用矩阵表示式 (3.5.3). 记矩阵

$$V_m = (v_1, v_2, \dots, v_m) \in \mathbf{R}^{n \times m},$$

$$W_m = (w_1, w_2, \dots, w_m) \in \mathbf{R}^{n \times m}.$$

由 $z^{(m)} \in K_m$ 知

$$z^{(m)} = y_1 v_1 + y_2 v_2 + \dots + y_m v_m = V_m y^{(m)},$$

其中 $y^{(m)} = (y_1, y_2, \dots, y_m)^T$. 于是式 (3.5.3) 可以写为

$$w_i^T(r^{(0)} - AV_m y^{(m)}) = 0, \quad i = 1, 2, \dots, m,$$

其矩阵形式为 $W_m^T(r^{(0)} - AV_m y^{(m)}) = 0$, 即

$$(W_m^T AV_m) y^{(m)} = W_m^T r^{(0)}. \quad (3.5.4)$$

若 $W_m^T AV_m$ 可逆, 则 $y^{(m)} = (W_m^T AV_m)^{-1} W_m^T r^{(0)}$, 从而得到

$$z^{(m)} = V_m y^{(m)} = V_m (W_m^T AV_m)^{-1} W_m^T r^{(0)}. \quad (3.5.5)$$

选取不同的子空间 K_m 和 L_m 及它们的基 $\{v_i\}_{i=1}^m, \{w_i\}_{i=1}^m$, 便得到基于伽辽金原理求解方程组 (3.5.2) 的不同算法. $\text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{m-1}r^{(0)}\}$ 称为由矩阵 A 与向量 $r^{(0)}$ 生成的克雷洛夫子空间. 若取

$$K_m = \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{m-1}r^{(0)}\},$$

则基于伽辽金原理构成的算法称为克雷洛夫子空间法 (Krylov subspace method). 在克雷洛夫子空间法中, 存在以下两个问题:

- (1) 如何选取子空间 L_m , 使得矩阵 $W_m^T AV_m$ 可逆;
- (2) 当 $m = n$ 时, 由 (3.5.3) 知, $z^{(m)}$ 是方程组 (3.5.2) 的准确解 z^* ; 而当 $m < n$ 时, 如何估计误差 $\|z^{(m)} - z^*\|$.

就一般情况而言, 这两个问题的难度较大, 目前在理论上尚未解决. 本节介绍两个实用有效的、具有代表性的克雷洛夫子空间法——阿诺尔迪算法和广义极小残余算法.

3.5.2 阿诺尔迪过程

阿诺尔迪 (Arnoldi) 过程是利用格拉姆-施密特 (Gram-Schmidt) 正交化方法, 由向量组 $r^{(0)}, Ar^{(0)}, A^2r^{(0)}, \dots, A^{m-1}r^{(0)}$ 构造克雷洛夫子空间 K_m 的一组标准正交基的方法.

1. 阿诺尔迪过程

设 $r^{(0)}, Ar^{(0)}, A^2r^{(0)}, \dots, A^{m-1}r^{(0)}$ 线性无关, 由这组向量构造 K_m 的一组标准正交基 $\{v_i\}_{i=1}^m$, 其过程如下:

- (1) 令 $v_1 = r^{(0)} / \|r^{(0)}\|$.
- (2) 取 $\tilde{v}_2 = Av_1 - h_{11}v_1$, 要求 $\tilde{v}_2 \perp v_1$. 由 $(\tilde{v}_2, v_1) = 0$ 得 $h_{11} = (Av_1, v_1)$. 记 $h_{21} = \|\tilde{v}_2\|$, 令 $v_2 = \frac{\tilde{v}_2}{\|\tilde{v}_2\|} = \frac{\tilde{v}_2}{h_{21}}$.

(3) 取 $\tilde{v}_3 = Av_2 - h_{22}v_2 - h_{12}v_1$, 要求 $\tilde{v}_3 \perp v_1, \tilde{v}_3 \perp v_2$. 由 $(\tilde{v}_3, v_1) = 0, (\tilde{v}_3, v_2) = 0$ 得 $h_{12} = (Av_2, v_1), h_{22} = (Av_2, v_2)$. 记 $h_{32} = \|\tilde{v}_3\|$, 令 $v_3 = \frac{\tilde{v}_3}{\|\tilde{v}_3\|} = \frac{\tilde{v}_3}{h_{32}}$.

.....

(4) 一般地, 对于 $k = 1, 2, \dots, m-1$, 取 $\tilde{v}_{k+1} = Av_k - \sum_{i=1}^k h_{ik}v_i$, 要求 $\tilde{v}_{k+1} \perp v_i$ ($i = 1, 2, \dots, k$). 由 $(\tilde{v}_{k+1}, v_i) = 0$ 得 $h_{ik} = (Av_k, v_i)$ ($i = 1, 2, \dots, k$). 记 $h_{k+1,k} = \|\tilde{v}_{k+1}\|$, 令 $v_{k+1} = \frac{\tilde{v}_{k+1}}{\|\tilde{v}_{k+1}\|} = \frac{\tilde{v}_{k+1}}{h_{k+1,k}}$.

(5) 当 $k = m$ 时, 取 $\tilde{v}_{m+1} = Av_m - \sum_{i=1}^m h_{im}v_i$, 要求 $\tilde{v}_{m+1} \perp v_i$ ($i = 1, 2, \dots, m$). 由 $(\tilde{v}_{m+1}, v_i) = 0$ 得 $h_{im} = (Av_m, v_i)$ ($i = 1, 2, \dots, m$). 记 $h_{m+1,m} = \|\tilde{v}_{m+1}\|$, 当 $h_{m+1,m} \neq 0$ 时, 令 $v_{m+1} = \frac{\tilde{v}_{m+1}}{\|\tilde{v}_{m+1}\|} = \frac{\tilde{v}_{m+1}}{h_{m+1,m}}$. 当 $h_{m+1,m} = 0$ 时, 则令 $v_{m+1} = 0$.

为了便于编制计算机程序, 阿诺尔迪过程可以写为

(1) 取 $v_1 = r^{(0)} / \|r^{(0)}\|$;

(2) for $k = 1$ to m do

for $i = 1$ to k do

$h_{ik} := (Av_k, v_i)$

end do

$\tilde{v}_{k+1} := Av_k - \sum_{i=1}^k h_{ik}v_i$

$h_{k+1,k} := \|\tilde{v}_{k+1}\|$

if $k = m$ and $h_{m+1,m} = 0$, then $v_{m+1} := 0$

else $v_{k+1} := \tilde{v}_{k+1} / h_{k+1,k}$

end do

总结以上可得下述定理:

定理 3.5.1 设 $m < n$, 向量组 $r^{(0)}, Ar^{(0)}, A^2r^{(0)}, \dots, A^{m-1}r^{(0)}$ 线性无关, 则由阿诺尔迪过程产生的向量组 $\{v_i\}_{i=1}^m$ 是克雷洛夫子空间 K_m 的一组标准正交基, 并且有

$$v_{m+1} \perp K_m = \text{span}\{r^{(0)}, Ar^{(0)}, A^2r^{(0)}, \dots, A^{m-1}r^{(0)}\}.$$

2. 阿诺尔迪过程的矩阵表示

为了理论上讨论方便, 将阿诺尔迪过程用矩阵表示.

由 $\tilde{v}_{k+1} = Av_k - \sum_{i=1}^k h_{ik}v_i$ ($k = 1, 2, \dots, m$) 及 $v_{k+1} = \tilde{v}_{k+1} / h_{k+1,k}$ 得

$$Av_k = \sum_{i=1}^{k+1} h_{ik}v_i, \quad k = 1, 2, \dots, m,$$

即

$$\begin{aligned} Av_1 &= h_{11}v_1 + h_{21}v_2, \\ Av_2 &= h_{12}v_1 + h_{22}v_2 + h_{32}v_3, \\ &\dots\dots\dots \\ Av_m &= h_{1m}v_1 + h_{2m}v_2 + \dots + h_{mm}v_m + h_{m+1,m}v_{m+1}. \end{aligned}$$

记矩阵

$$H_m = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1,m-1} & h_{1m} \\ h_{21} & h_{22} & \cdots & h_{2,m-1} & h_{2m} \\ & h_{32} & \cdots & h_{3,m-1} & h_{3m} \\ & & \ddots & \vdots & \vdots \\ & & & h_{m,m-1} & h_{mm} \end{pmatrix},$$

则有

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T, \quad (3.5.6)$$

其中 $e_m = (0, 0, \dots, 1)^T \in \mathbf{R}^m$. 像 H 这样的下次对角线下方元素全是零的矩阵称为上海森伯格 (Hessenberg) 矩阵.

3.5.3 阿诺尔迪算法

阿诺尔迪算法是给定适当的初始向量 $x^{(0)}$ 和正整数 m , 使得向量组

$$r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \dots, A^{m-1} r^{(0)}$$

线性无关, 取子空间 $L_m = K_m = \text{span}\{r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \dots, A^{m-1} r^{(0)}\}$, 由阿诺尔迪过程构造子空间 $\text{span}\{r^{(0)}, Ar^{(0)}, A^2 r^{(0)}, \dots, A^{m-1} r^{(0)}\}$ 的一组标准正交基 $\{v_i\}_{i=1}^m$, 利用伽辽金原理建立起来的求方程组 (3.5.2) 的一种方法.

由 $L_m = K_m$ 知矩阵 $W_m = V_m$. 在式 (3.5.6) 两端左乘矩阵 V_m^T , 并注意到 $V_m^T V_m = I$ 及 v_{m+1} 与 V_m 的列向量正交, 则得

$$V_m^T A V_m = H_m. \quad (3.5.7)$$

又

$$V_m^T r^{(0)} = \begin{pmatrix} v_1^T \\ \vdots \\ v_m^T \end{pmatrix} (\|r^{(0)}\| v_1) = \|r^{(0)}\| e_1 = \beta e_1,$$

其中 $\beta = \|r^{(0)}\|$, $e_1 = (1, 0, \dots, 0)^T \in \mathbf{R}^m$. 于是在 $L_m = K_m$ 的标准正交基 $\{v_i\}_{i=1}^m$ 下, 方程组 (3.5.4) 可以写成

$$H_m y^{(m)} = \beta e_1. \quad (3.5.8)$$

如果 H_m 非奇异, 则可由方程组 (3.5.8) 解出唯一的 $y^{(m)}$, 于是便求得了方程组 (3.5.2) 的近似解 $z^{(m)} = V_m y^{(m)}$. 如果 H_m 奇异, 则称算法发生了恶性中断, 这正是阿诺尔迪算法的缺陷.

对于给定的整数 $m > 0$, 下面的定理给出了阿诺尔迪算法求出的近似解 $z^{(m)}$ 的残向量 $\|r^{(0)} - Az^{(m)}\|$ 的大小.

定理 3.5.2 对于给定的 $m > 0$, 设 $y^{(m)}$ 是方程组 (3.5.8) 的唯一解, $z^{(m)} = V_m y^{(m)}$, 则

$$\|r^{(0)} - Az^{(m)}\| = h_{m+1,m} |e_m^T y^{(m)}|. \quad (3.5.9)$$

证明 由式 (3.5.6) 可得

$$\begin{aligned} r^{(0)} - Az^{(m)} &= r^{(0)} - AV_m y^{(m)} = r^{(0)} - (V_m H_m + h_{m+1,m} v_{m+1} e_m^T) y^{(m)} \\ &= r^{(0)} - V_m H_m y^{(m)} - h_{m+1,m} e_m^T y^{(m)} v_{m+1} \\ &= r^{(0)} - V_m (\beta e_1) - h_{m+1,m} e_m^T y^{(m)} v_{m+1} \\ &= -h_{m+1,m} e_m^T y^{(m)} v_{m+1}. \end{aligned}$$

当 $\tilde{v}_{m+1} \neq 0$ 时, $\|v_{m+1}\| = 1$. 对上式两端取范数, 即得式 (3.5.9). 当 $\tilde{v}_{m+1} = 0$ 时, $h_{m+1,m} = \|\tilde{v}_{m+1}\| = 0$, 式 (3.5.9) 仍然成立. 证毕.

由定理 3.5.2 知, 如果对于某个 $m > 0$, H_m 非奇异, 而 $\tilde{v}_{m+1} = 0$, 这时 $r^{(0)} - Az^{(m)} = 0$, 即知 $z^{(m)} = V_m y^{(m)}$ 是方程 (3.5.2) 的准确解 z^* .

1. 原理型阿诺尔迪算法

原理型阿诺尔迪算法是给定误差限 $\varepsilon > 0$, 通过不断地增大 m , 确定 $z^{(m)}$, 使得 $\|r^{(0)} - Az^{(m)}\| < \varepsilon$. 其计算步骤如下:

(1) 给定误差限 $\varepsilon > 0$, 初始向量 $x^{(0)} \in \mathbb{R}^n$, 取 $m = 1$;

(2) 计算 $r^{(0)} = b - Ax^{(0)}$;

(3) 由阿诺尔迪过程求出 $\{v_i\}_{i=1}^m$ 及 v_{m+1} ;

若 H_m 为奇异矩阵, 则算法产生恶性中断, 更换 $x^{(0)}$, 转 (2);

若 H_m 为非奇异矩阵, 解方程 $H_m y = \beta e_1$, 得解 $y^{(m)}$, 计算 $z^{(m)} = V_m y^{(m)}$;

(4) 若 $v_{m+1} = 0$, 则原方程 (3.5.1) 的解为 $x^* = x^{(0)} + z^{(m)}$, 停止;

当 $v_{m+1} \neq 0$ 时, 若 $\|r^{(0)} - Az^{(m)}\| = h_{m+1,m} |e_m^T y^{(m)}| < \varepsilon$, 取 $x^* \approx x^{(0)} + z^{(m)}$, 停止; 若 $\|r^{(0)} - Az^{(m)}\| \geq \varepsilon$, $m := m + 1$, 转 (3).

上述算法用于求解大型方程组 ($n \gg 1$) 时, 一般需要相当大的 m . 从计算过程可以看到, m 越大, 不仅 $z^{(m)}$ 的计算量越大, 而且保存 V_m 所占用的存储量就越大.

这两点使得算法失去了实用价值, 为克服这一缺陷, 对阿诺尔迪算法改进的一个方法是循环型阿诺尔迪算法.

2. 循环型阿诺尔迪算法

循环型阿诺尔迪算法是固定 m , 周期性地重新开始的迭代方法. 计算步骤如下:

(1) 给定 $m \ll n$, 选取 $x^{(0)} \in \mathbf{R}^n$, 计算 $r^{(0)} = b - Ax^{(0)}$, $k := 0$;

(2) 用阿诺尔迪算法求出 $z^{(m)} = V_m y^{(m)}$, 令

$$x^{(k+1)} = x^{(k)} + z^{(m)};$$

(3) 若 $\|b - Ax^{(k+1)}\| < \varepsilon$, 则取 $x^* \approx x^{(k+1)}$, 停止; 否则, 令 $r^{(0)} = b - Ax^{(k+1)}$, $k := k + 1$, 转 (2).

在循环型阿诺尔迪算法中, 由于 $m \ll n$, $z^{(m)}$ 的计算量和 V_m 存储量相对较小. 但是, 当 $k \geq 1$ 时, 如果原理型阿诺尔迪算法中产生恶性中断, 则前面所有计算完全作废, 需另取 $x^{(0)}$ 重新计算. 阿诺尔迪算法存在两个问题: 其一, 事先无法知道什么时候会发生恶性中断; 其二, 对于一般矩阵而言, 它的收敛性目前尚未证明. 尽管如此, 大量的数值计算表明, 循环型阿诺尔迪算法确实有效, 它的收敛速度有时会超过共轭梯度法 [10].

3. 对称兰乔斯 (Lanczos) 算法

若用阿诺尔迪算法求解系数矩阵 A 是对称矩阵的线性方程组 (3.5.1), 则由式 (3.5.7) 知, H_m 是对称矩阵. 这时, 阿诺尔迪算法就简化为对称兰乔斯算法 (symmetric Lanczos algorithm). 再由 H_m 是上海森伯格矩阵知, H_m 是对称的三对角矩阵

$$H_m = \begin{pmatrix} h_{11} & h_{12} & & & \\ h_{21} & h_{22} & h_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & h_{m-1,m-2} & h_{m-1,m-1} & h_{m-1,m} \\ & & & h_{m,m-1} & h_{mm} \end{pmatrix}.$$

注意到 $h_{jk} = h_{kj}$, 记 $\alpha_k = h_{kk}$, $\beta_k = h_{k,k-1} = h_{k-1,k}$, 用 T_m 表示 H_m , 则有

$$T_m = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & \beta_m & \alpha_m \end{pmatrix}.$$

这时, 阿诺尔迪过程

$$Av_k = \sum_{i=1}^{k+1} h_{ik} v_i, \quad k = 1, 2, \dots, m.$$

可以简化为三项递推公式:

$$\begin{aligned} \beta_1 &= 0, \\ \beta_{k+1} v_{k+1} &= Av_k - \alpha_k v_k - \beta_k v_{k-1}, \quad k = 1, 2, \dots, m. \end{aligned}$$

阿诺尔迪算法简化为对称兰乔斯算法, 其计算步骤如下:

- (1) 给定误差限 $\varepsilon > 0$, 初始向量 $x^{(0)} \in \mathbf{R}^n$;
- (2) 计算 $r^{(0)} = b - Ax^{(0)}$, $\beta = \|r^{(0)}\|$, $v_1 = r^{(0)}/\beta$;
- (3) for $m = 1$ to n do
 - $\beta_1 := 0$
 - for $k = 1$ to m do
 - $\alpha_k := (Av_k, v_k)$
 - $\tilde{v}_{k+1} := Av_k - \alpha_k v_k - \beta_k v_{k-1}$
 - $\beta_{k+1} := \|\tilde{v}_{k+1}\|$, if $\beta_{k+1} = 0$, then $m := k$, 转 (4)
 - else $v_{k+1} := \tilde{v}_{k+1}/\beta_{k+1}$
 - end do
 - 如果 $T_m = \text{Tridiag}(\beta_i, \alpha_i, \beta_{i+1})$ 为奇异矩阵, 更换 $x^{(0)}$, 转 (2); 否则, 转 (4);
- (4) 用追赶法解三对角方程组 $T_m y = \beta e_1$, 得解 $y^{(m)}$, 计算 $z^{(m)} = V_m y^{(m)}$;
- (5) 如果 $\|r^{(0)} - Az^{(m)}\| < \varepsilon$, 取 $x^* = x^{(0)} + z^{(m)}$, 停止; 否则, end do.

3.5.4 广义极小残余算法

鉴于阿诺尔迪算法的恶性中断问题难以解决, 以及在理论上很难分析其收敛性, 这促使人们研究其他的伽辽金型算法. 下面介绍的广义极小残余算法 (generalized minimal residual algorithm, GMRES) 就是基于伽辽金原理建立的一个有效算法, 它是当前求解大型非对称 (稀疏) 线性方程组的主要方法.

设求解的方程组为 (3.5.2), 选取正整数 m , $x^{(0)} \in \mathbf{R}^n$, 使得 $r^{(0)}, Ar^{(0)}, \dots, A^{m-1}r^{(0)}$ 线性无关. 由于矩阵 A 非奇异, 故 $Ar^{(0)}, A^2r^{(0)}, \dots, A^m r^{(0)}$ 也线性无关. 取

$$K_m = \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{m-1}r^{(0)}\},$$

$$L_m = \text{span}\{Ar^{(0)}, A^2r^{(0)}, \dots, A^m r^{(0)}\},$$

简记 $L_m = AK_m$. 利用伽辽金原理求方程组 (3.5.2) 的近似解 $z^{(m)}$, 即取 $z^{(m)} \in K_m$, 使得 $(r^{(0)} - Az^{(m)}) \perp L_m$. 下述定理表明这个问题等价于一个最小二乘问题:

定理 3.5.3 对于给定的 $x^{(0)} \in \mathbf{R}^n$ 和固定的正整数 m , $z^{(m)} \in K_m$, 使得 $(r^{(0)} - Az^{(m)}) \perp L_m$ 的充分必要条件是

$$\|r^{(0)} - Az^{(m)}\| = \min_{z \in K_m} \|r^{(0)} - Az\|.$$

证明 必要性. 设 $z^{(m)} \in K_m$ 使得 $(r^{(0)} - Az^{(m)}) \perp L_m$, 则对任意的 $z \in K_m$ 有

$$\begin{aligned} \|r^{(0)} - Az\|^2 &= \|(r^{(0)} - Az^{(m)}) - A(z - z^{(m)})\|^2 \\ &= \|r^{(0)} - Az^{(m)}\|^2 - 2(r^{(0)} - Az^{(m)}, A(z - z^{(m)})) + \|A(z - z^{(m)})\|^2. \end{aligned}$$

由 $z^{(m)}, z \in K_m$ 知 $z - z^{(m)} \in K_m$, $A(z - z^{(m)}) \in AK_m = L_m$. 又由 $(r^{(0)} - Az^{(m)}) \perp L_m$ 知 $(r^{(0)} - Az^{(m)}, A(z - z^{(m)})) = 0$. 因此,

$$\|r^{(0)} - Az\|^2 = \|r^{(0)} - Az^{(m)}\|^2 + \|A(z - z^{(m)})\|^2 \geq \|r^{(0)} - Az^{(m)}\|^2.$$

充分性. 若 $\|r^{(0)} - Az^{(m)}\| = \min_{z \in K_m} \|r^{(0)} - Az\|$, 则对任意的 $v \in K_m, \alpha \in \mathbf{R}$ 都有

$$\|r^{(0)} - A(z^{(m)} + \alpha v)\|^2 \geq \|r^{(0)} - Az^{(m)}\|^2.$$

记

$$\begin{aligned} q(\alpha) &= \|r^{(0)} - A(z^{(m)} + \alpha v)\|^2 = \|(r^{(0)} - Az^{(m)}) - \alpha Av\|^2 \\ &= \|r^{(0)} - Az^{(m)}\|^2 - 2\alpha(r^{(0)} - Az^{(m)}, Av) + \alpha^2 \|Av\|^2. \end{aligned}$$

显然, 当 $\alpha = 0$ 时, $q(\alpha)$ 达到最小值 $\|r^{(0)} - Az^{(m)}\|^2$. 令 $q'(\alpha) = 0$ 得

$$\alpha = \frac{(r^{(0)} - Az^{(m)}, Av)}{\|Av\|^2}.$$

由 $\alpha = 0$ 知 $(r^{(0)} - Az^{(m)}, Av) = 0$. 再由 $v \in K_m$ 的任意性及 $Av \in L_m$, 故得 $(r^{(0)} - Az^{(m)}) \perp L_m$. 证毕.

定理 3.5.3 表明, 求 $z^{(m)}$ 等价于在 K_m 中极小化残向量 $r^{(0)} - Az$ 的 2 范数, 它给出了求 $z^{(m)}$ 一个的方法, 命名为广义极小残余算法, 下面予以介绍.

首先利用阿诺尔迪过程建立 K_m 中的一组标准正交基 $\{v_i\}_{i=1}^m$ 及 v_{m+1} , 记矩阵

$$\bar{H}_m = \begin{pmatrix} H_m \\ h_{m+1,m} e_m^T \end{pmatrix},$$

其中 $e_m = (0, \dots, 0, 1)^T \in \mathbf{R}^m$, 则式 (3.5.6) 可以写为

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T = (V_m, v_{m+1}) \begin{pmatrix} H_m \\ h_{m+1,m} e_m^T \end{pmatrix} = V_{m+1} \bar{H}_m,$$

其中矩阵 $V_{m+1} = (V_m, v_{m+1})$,

设 $z \in K_m$, 则存在唯一的 $y \in \mathbf{R}^m$, 使得 $z = V_m y$.

当 $v_{m+1} \neq 0$ 时, 注意到 $V_{m+1}^T V_{m+1} = I$. 因此,

$$\begin{aligned} \|r^{(0)} - Az\| &= \|r^{(0)} - AV_m y\| = \|r^{(0)} - V_{m+1} \bar{H}_m y\| \\ &= \|V_{m+1}(\beta \bar{e}_1 - \bar{H}_m y)\| = \|\beta \bar{e}_1 - \bar{H}_m y\|, \end{aligned}$$

其中 $\bar{e}_1 = (1, 0, \dots, 0)^T \in \mathbf{R}^{m+1}$.

当 $v_{m+1} = 0$ 时, $h_{m+1,m} = 0$, 式 (3.5.6) 为 $AV_m = V_m H_m$, 并且有

$$\begin{aligned} \|r^{(0)} - Az\| &= \|r^{(0)} - AV_m y\| = \|r^{(0)} - V_m H_m y\| \\ &= \|V_m(\beta e_1 - H_m y)\| = \|\beta e_1 - H_m y\| = \|\beta \bar{e}_1 - \bar{H}_m y\|, \end{aligned}$$

其中 $e_1 = (1, 0, \dots, 0)^T \in \mathbf{R}^m$. 由以上讨论可得

$$\|r^{(0)} - Az^{(m)}\| = \min_{z \in K_m} \|r^{(0)} - Az\| = \min_{y \in \mathbf{R}^m} \|\beta \bar{e}_1 - \bar{H}_m y\|.$$

于是将求解 $\min_{z \in K_m} \|r^{(0)} - Az\|$ 的问题转化为求解一个与之等价的最小二乘问题

$$\min_{y \in \mathbf{R}^m} \|\beta e_1 - \bar{H}_m y\|, \quad e_1 \in \mathbf{R}^{m+1}. \quad (3.5.10)$$

显然, 它的理论解为 $y^{(m)} = \beta(\bar{H}_m^T \bar{H}_m)^{-1} \bar{H}_m^T e_1$. 由此可知, GMRES 算法不会发生恶性中断现象.

由以上所述可构成原理型 GMRES 算法.

1. 原理型 GMRES 算法

原理型 GMRES 算法通过不断地增大 m 来确定 $z^{(m)}$, 使得 $\|r^{(0)} - Az^{(m)}\| < \varepsilon$. 其计算步骤如下:

- (1) 给定误差限 $\varepsilon > 0$, 初始向量 $x^{(0)} \in \mathbf{R}^n$, 取 $m = 1$;
- (2) 计算 $r^{(0)} = b - Ax^{(0)}$, $\beta = \|r^{(0)}\|$;
- (3) 由阿诺尔迪过程求出 $\{v_i\}_{i=1}^m$, v_{m+1} 及 $\bar{H}_m$;
- (4) 求解最小二乘问题 $\min_{y \in \mathbf{R}^m} \|\beta e_1 - \bar{H}_m y\|$ (其中 $e_1 \in \mathbf{R}^{m+1}$), 设其解为 $y^{(m)}$;
- (5) 计算 $z^{(m)} = V_m y^{(m)}$, 若 $\|r^{(0)} - Az^{(m)}\| < \varepsilon$, 取 $x^* \approx x^{(0)} + z^{(m)}$, 停止; 若 $\|r^{(0)} - Az^{(m)}\| \geq \varepsilon$, $m := m + 1$, 转 (3).

理论上, 当 $m = n$ 时, 由 $(r^{(0)} - Az^{(n)}) \perp L_n$ 可知 $r^{(0)} - Az^{(n)} = 0$, 即原理型 GMRES 算法求出的 $z^{(n)}$ 是方程组 (3.5.2) 的准确解 z^* . 由此可得方程组 (3.5.1) 的准确解 $x^* = x^{(0)} + z^* = x^{(0)} + z^{(n)}$. 然而, 对于 $n \gg 1$ 的大型方程组, 当 m 很大时, V_m 的存储量和 $z^{(m)}$ 的计算量都很大, 这使算法失去了实用价值. 为使 GMRES 算法实用, 对其改进, 得到下述循环型 GMRES 算法:

2. 循环型 GMRES 算法

循环型 GMRES 算法是固定 m , 周期性地重新开始的迭代方法. 用 GMRES(m) 表示给定 m 的循环型 GMRES 算法, 其计算步骤如下:

- (1) 给定误差限 $\varepsilon > 0$, 选取适当的 m ($m \ll n$) 及初始向量 $x^{(0)} \in \mathbf{R}^n$;
- (2) 计算 $r^{(0)} = b - Ax^{(0)}$, $\beta = \|r^{(0)}\|$;
- (3) 由阿诺尔迪过程求出 $\{v_i\}_{i=1}^m$, v_{m+1} 及 $\bar{H}_m$;
- (4) 求解 $\min_{y \in \mathbf{R}^m} \|\beta e_1 - \bar{H}_m y\|$ (其中 $e_1 \in \mathbf{R}^{m+1}$), 设其解为 $y^{(m)}$;
- (5) 计算 $z^{(m)} = V_m y^{(m)}$, 令 $x^{(m)} = x^{(0)} + z^{(m)}$;
- (6) 令 $r^{(m)} = b - Ax^{(m)}$;
- (7) 若 $\|r^{(m)}\| < \varepsilon$, 取 $x^* \approx x^{(m)}$, 停止; 否则, $x^{(0)} := x^{(m)}$, $r^{(0)} := r^{(m)}$, $\beta = \|r^{(0)}\|$, 转 (3).

在 GMRES(m) 算法中, 正整数 m 的选取是个关键性的问题, 可以举例说明循环型 GMRES 算法并非对于任意给定的整数 m 都收敛, 但可以证明当线性方程组 $Ax = b$ 的系数矩阵 A 满足一定的条件时 GMRES(m) 算法的收敛性. 下面不加证明地给出两个关于 GMRES(m) 算法的收敛性定理.

定理 3.5.4 设线性方程组 $Ax = b$ 的系数矩阵 A 是对称正定矩阵, 则对于任意的整数 $m \geq 1$ 和任意的初始向量 $x^{(0)}$, GMRES(m) 算法收敛.

定理 3.5.5 设线性方程组 $Ax = b$ 的系数矩阵 A 可对角化, 即存在非奇异矩阵 X , 使得 $A = X\Lambda X^{-1}$, 其中 $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, λ_i ($i = 1, 2, \dots, n$) 为矩阵 A 的特征值, 则当 m 充分大时, 对任意的初始向量 $x^{(0)}$, GMRES(m) 算法收敛.

证明 略, 参见文献 [10].

3. 最小二乘问题 $\min_{y \in \mathbf{R}^m} \|\beta e_1 - \bar{H}_m y\|$ 的求解

在 GMRES(m) 算法中, 每一次迭代都要解一个最小二乘问题 (3.5.10):

$$\min_{y \in \mathbf{R}^m} \|\beta e_1 - \bar{H}_m y\|, \quad e_1 \in \mathbf{R}^{m+1}.$$

下面讨论最小二乘问题 (3.5.10) 的求解.

由 $\text{rank}(V_m) = m$ 及 A 是可逆矩阵知 $\text{rank}(AV_m) = m$. 再由 $AV_m = V_{m+1}\bar{H}_m$ 知 $\text{rank}(\bar{H}_m) = m$. 针对 $\bar{H}_m$ 是 $(m+1) \times m$ 阶上海森伯格矩阵的特征, 对矩阵 $\bar{H}_m$

和向量 βe_1 同时作吉文斯变换, 则有

$$P\bar{H}_m = \begin{pmatrix} R \\ 0^T \end{pmatrix}, \quad P(\beta e_1) = (c_1, \dots, c_m, c_{m+1})^T,$$

其中 P 为 $m+1$ 阶正交矩阵, R 为 m 阶可逆的上三角矩阵. 因此,

$$\begin{aligned} \|\beta e_1 - \bar{H}_m y\|^2 &= \|P(\beta e_1 - \bar{H}_m y)\|^2 = \left\| \begin{pmatrix} c_1 \\ \vdots \\ c_m \\ c_{m+1} \end{pmatrix} - \begin{pmatrix} R \\ 0^T \end{pmatrix} y \right\|^2 \\ &= \left\| \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} - Ry \right\|^2 + c_{m+1}^2. \end{aligned}$$

由此可知, 最小二乘问题 $\min_{y \in \mathbf{R}^m} \|\beta e_1 - \bar{H}_m y\|$ 的解 $y^{(m)}$ 是上三角方程组

$$Ry = (c_1, c_2, \dots, c_m)^T \quad (3.5.11)$$

的唯一解, 并且 $\|\beta e_1 - \bar{H}_m y^{(m)}\| = |c_{m+1}|$.

由以上讨论, 对于 GMRES(m) 算法, 则有

$$\|b - Ax^{(m)}\| = \|r^{(0)} - Az^{(m)}\| = \min_{y \in \mathbf{R}^m} \|\beta e_1 - \bar{H}_m y\| = \|\beta e_1 - \bar{H}_m y^{(m)}\| = |c_{m+1}|.$$

因此, 可将其计算过程中的迭代终止准则 $\|b - Ax^{(m)}\| < \varepsilon$ 改为 $|c_{m+1}| < \varepsilon$, 使用更为简便.

小 结

本章介绍了求解线性方程组 $Ax = b$ 的一些迭代法. 迭代法利用了线性方程组系数矩阵具有大量零元素的特点, 它具有占用存储空间小、程序简单等优点, 是求解大型稀疏线性方程组的有效方法.

一般地, 当雅可比迭代法和高斯-赛德尔迭代法都收敛时, 高斯-赛德尔迭代法比雅可比迭代法收敛得快, 最为常用.

超松弛迭代法的松弛因子若选取适当, 则收敛得更快. 本章给出了求解对称正定三对角线性方程组最佳松弛因子的计算公式. 一般地, 选择最佳松弛因子比较困难. 通常, 根据计算实践经验, 给出不同的参数, 采用试算的办法, 选择出一个比较合适的松弛因子.

共轭梯度法计算只涉及矩阵与向量的乘法和向量的内积运算. 在理论上对于 n 阶线性方程组最多 n 次迭代就可求得方程组的准确解. 从这点上来看, 共轭梯度法是一种直接法. 对于良态问题, 共轭梯度法所需的迭代次数远小于方程组的阶数 n .

基于伽辽金原理建立的求解大型非对称线性方程组的克雷洛夫子空间法虽然在理论上尚不十分完善, 但大量的实际计算表明这类算法还是十分有效的.

阿诺尔迪算法和广义极小残余算法是两个实用有效的、具有代表性的克雷洛夫子空间法.

用阿诺尔迪算法求解大型方程组, 一般需要相当大的 m . 改进后的循环型阿诺尔迪算法则克服了阿诺尔迪算法 $z^{(m)}$ 计算量大、 V_m 存储量大的缺陷. 它固定 m , 是周期性重新开始的迭代方法.

当求解的线性方程组系数矩阵 A 是对称矩阵时, 阿诺尔迪算法就简化为对称兰乔斯算法.

在理论上, 广义极小残余算法 (GMRES) 当 $m = n$ 时, 由原理型 GMRES 算法求出的 $z^{(n)}$ 是方程组 $Az = r^{(0)}$ 的准确解, 由此得到线性方程组 $Ax = b$ 的准确解. 但对于大型方程组, 当 m 很大时, 为了减少 $z^{(m)}$ 的计算量和 V_m 的存储量, 对原理型 GMRES 算法进行改进, 得到循环型 GMRES 算法——GMRES(m). 在一定的条件下, GMRES(m) 算法是收敛的. 它是当前求解大型非对称 (稀疏) 线性方程组的主要方法.

在克雷洛夫子空间法中, 还存在一些理论上尚待解决的问题, 它是当前研究的热点之一, 相关的理论和方法仍在继续发展之中.

习 题

3.1 用雅可比迭代法与高斯-赛德尔迭代法求解方程组

$$\begin{cases} -8x_1 + x_2 + x_3 = 1, \\ x_1 - 5x_2 + x_3 = 16, \\ x_1 + x_2 - 4x_3 = 7, \end{cases}$$

取初始向量 $x^{(0)} = (0, 0, 0)^T$, 准确到 2 位小数.

3.2 设有方程组

$$\begin{cases} 10x_1 + 4x_2 + 4x_3 = 13, \\ 4x_1 + 10x_2 + 8x_3 = 11, \\ 4x_1 + 8x_2 + 10x_3 = 25, \end{cases}$$

写出雅可比迭代法、高斯-赛德尔迭代法与 $\omega = 1.2$ 的超松弛迭代法的迭代公式, 三种迭代法是否收敛? 为什么?

3.3 设有方程组

$$\begin{cases} 2x_1 - 3x_2 + 10x_3 = 3, \\ 5x_1 + 2x_2 + x_3 = -12, \\ -x_1 + 4x_2 + 2x_3 = 20, \end{cases}$$

写出三种收敛的迭代格式并说明收敛的理由.

3.4 欲用迭代法求解方程组

$$\begin{cases} x_1 - 4x_2 - 2x_4 = 11, \\ 4x_1 + 8x_3 + 3x_4 = 6, \\ 3x_2 + x_3 + 5x_4 = 2, \\ 5x_1 - x_2 + x_3 + 2x_4 = 5, \end{cases}$$

写出收敛的高斯-赛德尔迭代法与 $\omega = 0.8$ 的超松弛迭代法的迭代公式, 给定初始向量 $x^{(0)} = (0, 0, 0, 0)^T$, 用高斯-赛德尔迭代法计算 $x^{(1)}$.

3.5 设有方程组

$$\begin{cases} 2x_1 - x_2 + x_3 = 3, \\ x_1 + x_2 + x_3 = 2, \\ x_1 + x_2 - 2x_3 = -1, \end{cases}$$

证明雅可比迭代法对任意的初始向量都不收敛, 而高斯-赛德尔迭代法对任意的初始向量都收敛.

3.6 设有方程组

$$\begin{cases} x_1 + 2x_2 - 2x_3 = -3, \\ x_1 + x_2 + x_3 = 1, \\ 2x_1 + 2x_2 + x_3 = 1, \end{cases}$$

证明雅可比迭代法对任意的初始向量都收敛, 而高斯-赛德尔迭代法对任意的初始向量都不收敛.

3.7 用超松弛迭代法求解线性方程组

$$\begin{cases} 4x_1 + 3x_2 = 24, \\ 3x_1 + 4x_2 - x_3 = 30, \\ -x_2 + 4x_3 = -24, \end{cases}$$

选择使得收敛速度最快的松弛因子 ω , 给定初始向量 $x^{(0)} = (0, 0, 0)^T$, 计算 $x^{(2)}$.

3.8 (1) 求解线性方程组 $Ax = b$ (其中 A 为对称正定矩阵) 的迭代公式为

$$x^{(k+1)} = x^{(k)} - \omega(Ax^{(k)} - b), \quad k = 0, 1, 2, \dots,$$

证明当 $0 < \omega < \frac{2}{\beta}$ 时, 上述迭代法收敛 (其中 $0 < \alpha \leq \lambda(A) \leq \beta$);

(2) 若用 (1) 中的迭代法求解方程组

$$\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \end{pmatrix},$$

试确定使迭代法收敛的 ω 取值范围, 并求出使迭代法收敛速度最快的最优 ω 值 ω_{opt} .

3.9 取 $\omega = 0.9$, $x^{(0)} = (0, 0, 0)^T$, 用超松弛迭代法求解方程组

$$\begin{cases} 5x_1 + 2x_2 + x_3 = -12, \\ -x_1 + 4x_2 + 2x_3 = 20, \\ 2x_1 - 3x_2 + 10x_3 = 3 \end{cases}$$

的近似解 $x^{(k)}$, 使误差不超过 10^{-4} .

3.10 给定初始向量 $x^{(0)} = (0, 0, 0)^T$, 用共轭梯度法求解以下方程组:

$$(1) \begin{cases} 2x_1 - x_2 - x_3 = 2, \\ -x_1 + 2x_2 + x_3 = 0, \\ -x_1 + x_2 + 2x_3 = -2; \end{cases} \quad (2) \begin{cases} 4x_1 - x_2 + 2x_3 = 12, \\ -x_1 + 5x_2 + 3x_3 = 10, \\ 2x_1 + 3x_2 + 6x_3 = 18. \end{cases}$$

3.11 分别用循环阿诺尔迪算法和循环 GMRES 算法求解方程组 $Ax = b$, 其中

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

取 $m = 2$, 给定初始向量 $x^{(0)} = (0, 0, 0, 0)^T$, 算出 $x^{(2)}$.

计算实习

3.1 用雅可比迭代法和高斯-赛德尔迭代法求解方程组

$$\begin{pmatrix} 2.52 & 0.95 & 1.25 & -0.85 \\ 0.39 & 1.69 & -0.45 & 0.49 \\ 0.55 & -1.25 & 1.96 & -0.98 \\ 0.23 & -1.15 & -0.45 & 2.31 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1.38 \\ -0.34 \\ 0.67 \\ 1.52 \end{pmatrix},$$

使误差小于 10^{-3} , 并比较迭代次数.

3.2 用共轭梯度法求解线性方程组 $Ax = b$, 其中

$$A = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix}.$$

矩阵 A 的阶数 n 分别取为 100, 200, 400, 指出计算结果是否可靠.

3.3 分别取 $m = 10, 20$, 用循环型阿诺尔迪算法求解 $Ax = d$, 其中

$$A = \begin{pmatrix} B & -I & & \\ -I & B & -I & \\ & \ddots & \ddots & \ddots \\ & & -I & B & -I \\ & & & -I & B \end{pmatrix}_{200 \times 200}, \quad B = \begin{pmatrix} 4 & a & & \\ b & 4 & a & \\ & \ddots & \ddots & \ddots \\ & & b & 4 & a \\ & & & b & 4 \end{pmatrix}_{10 \times 10}$$

$$a = -1 + \delta, \quad b = -1 - \delta, \quad \delta = 0.01,$$

方程组的右端向量 $d = Af, f = (1, 1, \dots, 1)^T$. 显然, 准确解 $x = f$. 任意选取初始向量 $x^{(0)[7]}$.

第4章 插 值 法

在科学研究和工程技术中往往会遇到一些函数, 其表达式比较复杂, 或者不知其确切表达式, 只知道函数在区间 $[a, b]$ 上一些点 x_i ($i = 0, 1, \dots, n$) 处的函数值 $y_i = f(x_i)$ ($i = 0, 1, \dots, n$), 这些函数在实际中难以使用. 本章针对这类函数, 构造 $f(x)$ 的简单近似表达式 $p(x)$, 并估计其误差 $f(x) - p(x)$.

利用 $p(x)$ 可以近似计算 $f(x)$ 的函数值、导数值、极值、零点、积分以及微分方程、积分方程的解等.

4.1 多项式插值问题

由于多项式是最简单的函数, 容易计算, 故取多项式作为 $f(x)$ 的近似表达式.

多项式插值问题 在区间 $[a, b]$ 上给定 $n+1$ 个互不相同的点 x_i ($i = 0, 1, \dots, n$) ($a \leq x_0 < x_1 < \dots < x_n \leq b$), 及其在 x_i 处的函数值 $y_i = f(x_i)$ ($i = 0, 1, \dots, n$), 构造一个次数不超过 n 的多项式

$$p_n(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n,$$

使其满足

$$p_n(x_i) = y_i, \quad i = 0, 1, \dots, n. \quad (4.1.1)$$

取 $f(x) \approx p_n(x)$, 记 $R_n(x) = f(x) - p_n(x)$, 则 $p_n(x)$ 称为 $f(x)$ 的插值多项式, $f(x)$ 称为被插函数, 区间 $[a, b]$ 称为插值区间, 点 x_i ($i = 0, 1, \dots, n$) 称为插值节点 (插值节点 x_i 不一定按大小顺序排列), 式 (4.1.1) 称为插值条件, $R_n(x)$ 称为截断误差或插值余项, 求插值多项式的方法称为插值法.

由插值条件 (4.1.1) 知, 在插值节点上 $p_n(x_i) = f(x_i)$ ($i = 0, 1, \dots, n$). 为计算其他点 x ($x \neq x_i$) 处的函数值, 可用 $p_n(x)$ 的值作为 $f(x)$ 的近似值, 称之为插值, x 称为插值点. 若插值点位于插值区间内部, 则称为内插; 若插值点位于插值区间外部, 但接近插值区间端点, 则称为外插或外推.

从几何上来看, 插值法是求通过给定的 $n+1$ 个点 (x_i, y_i) ($i = 0, 1, \dots, n$) 的曲线 $y = p_n(x)$, 并用它近似已知曲线 $y = f(x)$, 其几何意义如图 4.1 所示.

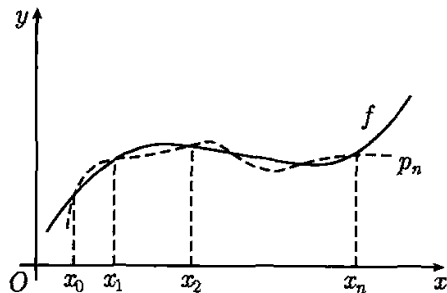


图 4.1

下面讨论插值多项式 $p_n(x)$ 的存在唯一性和误差估计.

定理 4.1.1 满足插值条件 (4.1.1) 的 n 次插值多项式 $p_n(x)$ 存在且唯一.

证明 设 $p_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$. 由插值条件 (4.1.1) 得

$$\begin{cases} a_0 + a_1x_0 + a_2x_0^2 + \cdots + a_nx_0^n = y_0, \\ a_0 + a_1x_1 + a_2x_1^2 + \cdots + a_nx_1^n = y_1, \\ \quad \quad \quad \cdots \quad \quad \quad \cdots \quad \quad \quad \cdots \\ a_0 + a_1x_n + a_2x_n^2 + \cdots + a_nx_n^n = y_n. \end{cases} \quad (4.1.2)$$

由于插值节点 x_i ($i = 0, 1, \cdots, n$) 互异, 它的系数矩阵行列式

$$\begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix} = \prod_{0 \leq j < i \leq n} (x_i - x_j) \neq 0.$$

故线性方程组 (4.1.2) 的解存在且唯一. 所以插值多项式 $p_n(x)$ 存在且唯一. 证毕.

定理 4.1.2 (截断误差估计式) 设 $f^{(n)}(x)$ 在区间 $[a, b]$ 上连续, $f^{(n+1)}(x)$ 在 (a, b) 内存在, 插值节点 $a \leq x_0 < x_1 < \cdots < x_n \leq b$, $p_n(x)$ 是满足插值条件 (4.1.1) 的 n 次插值多项式, 则对任意 $x \in [a, b]$, 存在 $\xi = \xi(x) \in (a, b)$, 使得

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n). \quad (4.1.3)$$

证明 由插值条件 (4.1.1) 知 $R_n(x_i) = 0$ ($i = 0, 1, \cdots, n$), 于是对于任意的 x ,

$$R_n(x) = k(x)(x - x_0)(x - x_1) \cdots (x - x_n), \quad (4.1.4)$$

其中 $k(x)$ 为与 x 有关的待定函数. 下面确定 $k(x)$.

任取区间 $[a, b]$ 上固定的一点 x ($x \neq x_i, i = 0, 1, \cdots, n$), 作辅助函数

$$\phi(t) = R_n(t) - k(x)(t - x_0)(t - x_1) \cdots (t - x_n).$$

显然, $\phi(x_i) = 0$ ($i = 0, 1, \cdots, n$), $\phi(x) = 0$, 即 $\phi(x)$ 在区间 $[a, b]$ 内有 $n+2$ 个零点 x, x_i ($i = 0, 1, \cdots, n$). 据定理条件可知, $\phi^{(n)}(t)$ 在区间 $[a, b]$ 连续, $\phi^{(n+1)}(t)$ 在区间 (a, b) 存在. 由罗尔定理知, $\phi'(t)$ 在两个零点之间至少有一个零点, 于是在 $\phi'(t)$ 在区间 $[a, b]$ 内至少有 $n+1$ 个零点, $\phi''(t)$ 在 $[a, b]$ 内至少有 n 个零点. 反复应用罗尔定理可知, $\phi^{(n+1)}(t)$ 在 (a, b) 内至少存在一个零点 $\xi \in (a, b)$, 使得 $\phi^{(n+1)}(\xi) = 0$, 即

$$\begin{aligned} \phi^{(n+1)}(\xi) &= R_n^{(n+1)}(\xi) - (n+1)! k(x) \\ &= f^{(n+1)}(\xi) - p_n^{(n+1)}(\xi) - (n+1)! k(x) \\ &= f^{(n+1)}(\xi) - (n+1)! k(x) = 0. \end{aligned}$$

由上式得

$$k(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

其中 $\xi \in (a, b)$ 且与 x 有关. 将 $k(x)$ 代入式 (4.1.4), 即得截断误差估计式 (4.1.3). 证毕.

为以后使用方便, 记 $\pi_{n+1}(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$, 则截断误差估计式 (4.1.3) 可以表示成

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \pi_{n+1}(x). \quad (4.1.3')$$

由式 (4.1.3') 可见, 如果 $f^{(n+1)}(x)$ 在区间 $[a, b]$ 上有界, 即存在常数 M_{n+1} , 使得 $|f^{(n+1)}(x)| \leq M_{n+1}$, 则有

$$|R_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |x - x_0| |x - x_1| \cdots |x - x_n|.$$

由此可知, 当插值点 x 靠近某插值节点 x_i 时, 插值多项式的误差 $|R_n(x)|$ 较小; 否则, 误差可能很大.

在截断误差的表达式中, ξ 是未知的, $f^{(n+1)}(\xi)$ 的大小很难估计, 在实际问题中, 误差估计式 (4.1.3) 难以应用. 下面介绍截断误差的一种实用估计法. 设 $p_n(x)$ 和 $\tilde{p}_n(x)$ 分别是以 $x_0, x_1, \cdots, x_n$ 和 $x_1, x_2, \cdots, x_{n+1}$ 构成的 n 次插值多项式, 则

$$\begin{aligned} f(x) - p_n(x) &= \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)(x - x_1) \cdots (x - x_n), \\ f(x) - \tilde{p}_n(x) &= \frac{f^{(n+1)}(\tilde{\xi})}{(n+1)!} (x - x_1)(x - x_2) \cdots (x - x_{n+1}). \end{aligned}$$

由于构造 $p_n(x)$ 和 $\tilde{p}_n(x)$ 只相差一个插值节点, 可以设想 $f^{(n+1)}(\xi) \approx f^{(n+1)}(\tilde{\xi})$. 以上两式相减得

$$\tilde{p}_n(x) - p_n(x) \approx \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_1)(x - x_2) \cdots (x - x_n)(x_{n+1} - x_0),$$

故 $\frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_1)(x - x_2) \cdots (x - x_n) \approx \frac{\tilde{p}_n(x) - p_n(x)}{x_{n+1} - x_0}$, 从而得到截断误差的实用估计式

$$\begin{cases} f(x) - p_n(x) \approx \frac{\tilde{p}_n(x) - p_n(x)}{x_{n+1} - x_0} (x - x_0), \\ f(x) - \tilde{p}_n(x) \approx \frac{\tilde{p}_n(x) - p_n(x)}{x_{n+1} - x_0} (x - x_{n+1}). \end{cases} \quad (4.1.5)$$

4.2 拉格朗日插值多项式

一般不采用通过解方程组 (4.1.2) 构造插值多项式的方法, 这是因为, 其一, 计算量大; 其二, 方程组 (4.1.2) 系数矩阵的条件数比较大, 是病态方程组. 本节和 4.3 节介绍简单实用的拉格朗日 (Lagrange) 插值法和牛顿 (Newton) 插值法, 将它们推出的插值多项式 $p_n(x)$ 分别记为 $L_n(x)$ 和 $N_n(x)$. 由插值多项式的存在唯一性定理 4.1.1 知 $L_n(x) \equiv N_n(x)$.

下面首先以构造二次拉格朗日插值多项式 $L_2(x)$ 为例来说明 n 次拉格朗日插值多项式 $L_n(x)$ 的构造方法.

给定数据

$$x_0 < x_1 < x_2, \quad y_i = f(x_i), \quad i = 0, 1, 2.$$

令二次拉格朗日插值多项式

$$L_2(x) = l_0(x)y_0 + l_1(x)y_1 + l_2(x)y_2, \quad (4.2.1)$$

其中 $l_i(x)$ ($i = 0, 1, 2$) 为二次多项式. 若令

$$\begin{aligned} l_0(x_0) &= 1, & l_1(x_0) &= 0, & l_2(x_0) &= 0, \\ l_0(x_1) &= 0, & l_1(x_1) &= 1, & l_2(x_1) &= 0, \\ l_0(x_2) &= 0, & l_1(x_2) &= 0, & l_2(x_2) &= 1, \end{aligned}$$

则有 $L_2(x_i) = y_i$ ($i = 0, 1, 2$), 即 $L_2(x)$ 满足插值条件 (4.1.1).

下面确定 $l_i(x)$ ($i = 0, 1, 2$).

由 $l_0(x_1) = l_0(x_2) = 0$ 知, x_1, x_2 是 $l_0(x)$ 的零点, 故 $l_0(x)$ 含有因式 $(x - x_1)(x - x_2)$. 由于 $l_0(x)$ 是二次多项式, 故

$$l_0(x) = c_0(x - x_1)(x - x_2), \quad (4.2.2)$$

其中 c_0 为待定常数. 再由 $l_0(x_0) = 1$ 可得数 $c_0 = \frac{1}{(x_0 - x_1)(x_0 - x_2)}$. 将 c_0 代入式 (4.2.2) 得

$$l_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}.$$

同理可得

$$l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}, \quad l_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

将 $l_i(x)$ 代入式 (4.2.1), 即得二次拉格朗日插值多项式

$$L_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}y_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}y_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}y_2.$$

由误差估计式 (4.1.3) 知, 其截断误差估计式为

$$R_3(x) = \frac{f^{(3)}(\xi)}{3!}(x-x_0)(x-x_1)(x-x_2).$$

下面用同样的方法构造 n 次拉格朗日插值多项式 $L_n(x)$.

给定 $n+1$ 个互不相同的插值节点 x_i ($i = 0, 1, \dots, n$) 和在 x_i 处的函数值 $y_i = f(x_i)$ ($i = 0, 1, \dots, n$), 设

$$L_n(x) = \sum_{i=0}^n l_i(x)y_i, \quad (4.2.3)$$

其中 $l_i(x)$ ($i = 1, 2, \dots, n$) 为 n 次多项式, 称为拉格朗日插值基函数. 显然, 若令

$$l_i(x_j) = \begin{cases} 1, & j = i, \\ 0, & j \neq i, \end{cases}$$

则 $L_n(x_j) = \sum_{i=0}^n l_i(x_j)y_i = l_j(x_j)y_j = y_j$ ($j = 0, 1, \dots, n$), 即 $L_n(x)$ 满足插值条件 (4.1.1).

现确定 $l_i(x)$ ($i = 0, 1, \dots, n$).

由 $l_i(x_j) = 0$ ($j = 0, 1, \dots, i-1, i+1, \dots, n$) 知, x_j 是 $l_i(x)$ 的零点, 故

$$l_i(x) = c_i(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n),$$

其中 c_i 为待定常数. 由 $l_i(x_i) = 1$ 得

$$c_i = \frac{1}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}.$$

故

$$l_i(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}.$$

将 $l_i(x)$ 代入式 (4.2.3) 得 n 次拉格朗日插值多项式

$$L_n(x) = \sum_{i=0}^n \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}y_i. \quad (4.2.4)$$

为书写简便起见, 改写 $L_n(x)$.

由 $\pi_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$ 知

$$\pi'_{n+1}(x_i) = (x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n),$$

则式 (4.2.4) 可写为

$$L_n(x) = \sum_{i=0}^n \frac{\pi_{n+1}(x)}{(x-x_i)\pi'_{n+1}(x_i)} y_i. \quad (4.2.4')$$

由插值多项式的唯一性及误差估计式 (4.1.3) 知, 拉格朗日插值多项式的误差为

$$R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-x_0)(x-x_1)\cdots(x-x_n).$$

拉格朗日插值多项式的缺点如下:

- (1) 乘除法计算量大;
- (2) 当增加一个插值节点时, 插值基函数 $l_i(x)$ 随之变化, 故构造新的插值多项式 $L_{n+1}(x)$ 不能利用原有的插值多项式 $L_n(x)$.

例 4.1.1 给定数据如下, 求其三次拉格朗日插值多项式及截断误差估计式:

x_i	-1	1	2	5
y_i	-7	7	-4	35

解 由式 (4.2.4) 知

$$\begin{aligned} L_3(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} y_0 + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} y_1 \\ &\quad + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} y_2 + \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} y_3 \\ &= \frac{(x-1)(x-2)(x-5)}{(-1-1)(-1-2)(-1-5)} \times (-7) + \frac{(x+1)(x-2)(x-5)}{(1+1)(1-2)(1-5)} \times 7 \\ &\quad + \frac{(x+1)(x-1)(x-5)}{(2+1)(2-1)(2-5)} \times (-4) + \frac{(x+1)(x-1)(x-2)}{(5+1)(5-1)(5-2)} \times 35 \\ &= 2x^3 - 10x^2 + 5x + 10, \end{aligned}$$

截断误差为

$$R_3(x) = f(x) - L_3(x) = \frac{f^{(4)}(\xi)}{4!} (x+1)(x-1)(x-2)(x-5).$$

解毕.

4.3 牛顿插值多项式

牛顿插值多项式克服了拉格朗日插值多项式的缺点, 具有计算量小、增加一个插值节点时能利用原有插值多项式的优点.

4.3.1 差商的定义

为建立牛顿插值多项式, 首先介绍差商的定义. 各阶差商定义如下:

零阶差商:

$$f[x_i] = y_i = f(x_i);$$

一阶差商:

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i};$$

二阶差商:

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i};$$

k 阶差商:

$$f[x_i, x_{i+1}, \dots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \dots, x_{i+k}] - f[x_i, x_{i+1}, \dots, x_{i+k-1}]}{x_{i+k} - x_i},$$

.....

n 阶差商:

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

计算差商可列成差商表. 以 $n = 4$ 为例, 列差商表如表 4.1 所示.

表 4.1 差商表

x_i	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	$f[x_0, x_1, x_2, x_3, x_4]$
x_0	$f[x_0]$				
x_1	$f[x_1]$	$f[x_0, x_1]$			
x_2	$f[x_2]$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
x_3	$f[x_3]$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$	
x_4	$f[x_4]$	$f[x_3, x_4]$	$f[x_2, x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$	$f[x_0, x_1, x_2, x_3, x_4]$

注: 差商表 4.1 中下方画线的差商是牛顿插值多项式中要用的数据.

4.3.2 牛顿插值多项式

过平面上两点 $(x_0, y_0), (x_1, y_1)$ 作线性插值, 则有

$$\begin{aligned} p_1(x) &= \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1 \\ &= y_0 + \frac{y_1 - y_0}{x_1 - x_0} (x - x_0) = f[x_0] + f[x_0, x_1](x - x_0). \end{aligned}$$

上式可以写为 $p_1(x) = c_0 + c_1(x - x_0)$. 由此设想 n 次插值多项式 $p_n(x)$ 可以写为

$$\begin{aligned} N_n(x) &= c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \dots + c_k(x - x_0)(x - x_1) \dots \\ &\quad \cdot (x - x_{k-1}) + \dots + c_n(x - x_0)(x - x_1) \dots (x - x_{n-1}). \end{aligned} \quad (4.3.1)$$

为使 $N_n(x)$ 是插值多项式, 选择系数 c_i , 使它满足插值条件 (4.1.1), 即

$$\begin{aligned} y_0 &= N_n(x_0) = c_0, \\ y_1 &= N_n(x_1) = c_0 + c_1(x_1 - x_0), \\ y_2 &= N_n(x_2) = c_0 + c_1(x_2 - x_0) + c_2(x_2 - x_0)(x_2 - x_1), \\ &\dots\dots\dots \\ y_n &= N_n(x_n) = c_0 + c_1(x_n - x_0) + c_2(x_n - x_0)(x_n - x_1) + \dots \\ &\quad + c_n(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}). \end{aligned}$$

显然,

$$\begin{aligned} c_0 &= y_0 = f(x_0) = f[x_0], \\ c_1 &= \frac{y_1 - y_0}{x_1 - x_0} = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = f[x_0, x_1], \\ c_2 &= \frac{y_2 - y_0 - c_1(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} = \frac{y_2 - y_1 + y_1 - y_0 - c_1(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{y_2 - y_1 + c_1(x_1 - x_0) - c_1(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} = \frac{y_2 - y_1 - c_1(x_2 - x_1)}{(x_2 - x_0)(x_2 - x_1)} \\ &= \frac{\frac{y_2 - y_1}{x_2 - x_1} - c_1}{x_2 - x_0} = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = f[x_0, x_1, x_2]. \end{aligned}$$

同理可得 $c_k = f[x_0, x_1, \dots, x_k]$ ($k = 0, 1, \dots, n$), 即差商表 4.1 中下方画线的各阶差商.

将 c_k 代入式 (4.3.1), 即得牛顿插值多项式

$$\begin{aligned} N_n(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \dots \\ &\quad + f[x_0, x_1, \dots, x_k](x - x_0)(x - x_1) \cdots (x - x_{k-1}) + \dots \\ &\quad + f[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned} \quad (4.3.2)$$

下面导出牛顿插值多项式的截断误差估计式.

牛顿插值多项式与拉格朗日插值多项式虽然在表达形式上不同, 但由插值多项式的唯一性知, 它们实质上是同一多项式. 因此, 误差项也相同, 均可用式 (4.1.3) 表示, 但用差商表示更为方便. 取 x_{n+1} 为区间 $[a, b]$ 上不等于 x_i ($i = 0, 1, \dots, n$) 的任意一点, 由式 (4.3.2) 知, 以 $x_0, x_1, \dots, x_n, x_{n+1}$ 为插值节点构造的 $n+1$ 次牛顿插值多项式

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

据插值条件知

$$\begin{aligned} f(x_{n+1}) &= N_{n+1}(x_{n+1}) \\ &= N_n(x_{n+1}) + f[x_0, x_1, \dots, x_{n+1}](x_{n+1} - x_0)(x_{n+1} - x_1) \cdots (x_{n+1} - x_n). \end{aligned}$$

由 x_{n+1} 的任意性, 将上式中的 x_{n+1} 改为 x , 则得

$$f(x) = N_n(x) + f[x_0, x_1, \dots, x_n, x](x - x_0)(x - x_1) \cdots (x - x_n).$$

由此可得 n 次牛顿插值多项式 $N_n(x)$ 的截断误差估计式

$$\begin{aligned} R_n(x) &= f(x) - N_n(x) \\ &= f[x_0, x_1, \dots, x_n, x](x - x_0)(x - x_1) \cdots (x - x_n). \end{aligned} \quad (4.3.3)$$

下面介绍牛顿插值多项式截断误差的实用估计法.

假定

$$f[x_0, x_1, \dots, x_n, x] \approx f[x_0, x_1, \dots, x_n, x_{n+1}],$$

由

$$N_{n+1}(x) = N_n(x) + f[x_0, x_1, \dots, x_n, x_{n+1}]\pi_{n+1}(x)$$

可得截断误差的实用估计式

$$\begin{aligned} R_n(x) &= f[x_0, x_1, \dots, x_n, x]\pi_{n+1}(x) \\ &\approx f[x_0, x_1, \dots, x_n, x_{n+1}]\pi_{n+1}(x) = N_{n+1}(x) - N_n(x). \end{aligned} \quad (4.3.4)$$

式 (4.3.4) 说明, n 次牛顿插值多项式 $N_n(x)$ 的截断误差大约等于 $n+1$ 次牛顿插值多项式的最后一项.

由 $N_{n+1}(x) = N_n(x) + f[x_0, x_1, \dots, x_n, x_{n+1}]\pi_{n+1}(x)$ 知, 当增加一个插值节点构造 $n+1$ 次牛顿插值多项式 $N_{n+1}(x)$ 时, 只需在 n 次牛顿插值多项式 $N_n(x)$ 后加一项 $f[x_0, x_1, \dots, x_n, x_{n+1}]\pi_{n+1}(x)$ 即可. 另外, 若用秦九韶算法计算多项式, 还可减少乘法运算次数, 这就克服了拉格朗日插值多项式的缺点.

构造牛顿插值多项式首先应列出差商表, 然后由差商表写出插值多项式.

例 4.3.1 给定数据点同例 4.1.1 求三次牛顿插值多项式及其截断误差估计式.

解 先列差商表 (表 4.2):

表 4.2

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_0, x_1, x_2, x_3]$
-1	<u>-7</u>			
1	7	<u>7</u>		
2	-4	-11	<u>-6</u>	
5	35	13	6	<u>2</u>

所以

$$\begin{aligned} N_3(x) &= -7 + 7(x+1) - 6(x+1)(x-1) + 2(x+1)(x-1)(x-2) \\ &= 2x^3 - 10x^2 + 5x + 10. \end{aligned}$$

截断误差为

$$R_3(x) = f[-1, 1, 2, 5, x](x+1)(x-1)(x-2)(x-5).$$

解毕.

4.3.3 差商的性质

由插值多项式的唯一性, 比较拉格朗日插值多项式 $L_n(x)$ 与牛顿插值多项式 $N_n(x)$ 最高次项 x^n 的系数得如下性质:

性质 4.3.1

$$f[x_0, x_1, \dots, x_n] = \sum_{i=0}^n \frac{f(x_i)}{\pi'_{n+1}(x_i)}.$$

由性质 4.3.1 可得如下性质:

性质 4.3.2 差商与节点的次序无关, 即对于任意的自然数 k ,

$$f[x_i, x_{i+1}, \dots, x_{i+k}] = f[x_{i_0}, x_{i_1}, \dots, x_{i_k}],$$

其中 $i_0, i_1, \dots, i_k$ 为 $i, i+1, \dots, i+k$ 的任意一种排列, 如

$$\begin{aligned} f[x_0, x_1] &= f[x_1, x_0], \\ f[x_0, x_1, x_2] &= f[x_0, x_2, x_1] = f[x_1, x_0, x_2] \\ &= f[x_1, x_2, x_0] = f[x_2, x_0, x_1] = f[x_2, x_1, x_0]. \end{aligned}$$

性质 4.3.3 (差商与导数的关系)

$$f[x_i, x_{i+1}, \dots, x_{i+k}] = \frac{f^{(k)}(\xi)}{k!}, \quad (4.3.5)$$

其中 ξ 介于 $x_i, x_{i+1}, \dots, x_{i+k}$ 之间.

证明 由插值多项式的唯一性知, 拉格朗日插值多项式与牛顿插值多项式的误差项相等, 故有

$$f[x_0, x_1, \dots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

由此知 $f[x_0, x_1, \dots, x_k] = \frac{f^{(k)}(\xi)}{k!}$, 其中 ξ 介于 $x_0, x_1, \dots, x_k$ 之间. 由此得式

(4.3.5). 证毕.

由式 (4.3.5) 知, k 阶差商对应 k 阶导数. 特别地, 当 $x_i = x_{i+1} = \cdots = x_{i+k}$ 时可得如下性质:

性质 4.3.4

$$f[\underbrace{x_i, x_i, \cdots, x_i}_{k+1 \text{ 个}}] = \frac{f^{(k)}(x_i)}{k!}. \quad (4.3.6)$$

在式 (4.3.6) 中若取 x_i 均为 x_0 , 则牛顿插值多项式变为泰勒多项式.

性质 4.3.5 (差商的导数)

$$\frac{df[x_0, x_1, \cdots, x_{k-1}, x]}{dx} = \frac{f^{(k+1)}(\xi)}{(k+1)!},$$

其中 ξ 介于 $x_0, x_1, \cdots, x_{k-1}, x$ 之间.

证明

$$\begin{aligned} \frac{df[x_0, x_1, \cdots, x_{k-1}, x]}{dx} &= \lim_{x_k \rightarrow x} \frac{f[x_0, x_1, \cdots, x_{k-1}, x_k] - f[x_0, x_1, \cdots, x_{k-1}, x]}{x_k - x} \\ &= \lim_{x_k \rightarrow x} f[x_0, x_1, \cdots, x_{k-1}, x_k, x] \\ &= \lim_{x_k \rightarrow x} \frac{f^{(k+1)}(\tilde{\xi})}{(k+1)!} = \frac{f^{(k+1)}(\xi)}{(k+1)!}. \end{aligned}$$

证毕.

由性质 4.3.3 和性质 4.3.5 可得

$$\frac{d\left(\frac{f^{(k)}(\tilde{\xi})}{k!}\right)}{dx} = \frac{f^{(k+1)}(\xi)}{(k+1)!}. \quad (4.3.7)$$

又由性质 4.3.3 可得如下性质:

性质 4.3.6 设 $f(x)$ 是 n 次多项式, 则一阶差商 $f[x_0, x]$ 是 x 的 $n-1$ 次多项式, 二阶差商 $f[x_0, x_1, x]$ 是 x 的 $n-2$ 次多项式.

一般地, 当 $k \leq n$ 时, $f(x)$ 的 k 阶差商 $f[x_0, x_1, \cdots, x_{k-1}, x]$ 是 x 的 $n-k$ 次多项式; 当 $k > n$ 时, k 阶差商 $f[x_0, x_1, \cdots, x_{k-1}, x] \equiv 0$.

4.4 埃尔米特插值多项式

前面介绍的拉格朗日插值多项式和牛顿插值多项式, 只要求在插值节点上插值多项式的值等于函数值, 而在一些问题中还要求在某些插值节点上, 插值多项式

的若干阶导数值等于函数的相应阶导数值. 这样的插值多项式称为埃尔米特 (Hermite) 插值多项式或带导数的插值多项式, 记为 $H_n(x)$. 它也是函数 $f(x)$ 的一种近似表达式, 下面介绍两种构造埃尔米特插值多项式的方法 —— 牛顿型方法和拉格朗日型方法.

方法 1 (牛顿型) 当插值节点 x_i 处的函数值、各阶导数值 $y_i, y'_i, y''_i, \dots, y_i^{(k)}$ 已知时, 把 x_i 视为 $k+1$ 重节点, 利用差商与导数的关系 (4.3.6), 由牛顿插值法构造之.

例 4.4.1 已知函数 $y = f(x)$ 的函数值、导数值如表 4.3 所示, 求其插值多项式及其误差项.

表 4.3

x_i	-1	0	1
$f(x_i)$	0	-4	-2
$f'(x_i)$		0	5
$f''(x_i)$		6	

解 由式 (4.3.6) 知

$$f[0, 0] = f'(0), \quad f[0, 0, 0] = \frac{f''(0)}{2!} = 3, \quad f[1, 1] = f'(1) = 5.$$

作差商表如表 4.4 所示.

表 4.4

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	...	...
-1	0					
0	-4	$\frac{-4}{1}$				
0	-4	0	$\frac{4}{2}$			
0	-4	0	3	$\frac{-1}{2}$		
1	-2	2	2	-1	$\frac{0}{3}$	
1	-2	5	3	1	2	$\frac{1}{4}$

所以

$$\begin{aligned} H_5(x) &= 0 - 4(x+1) + 4(x+1)x - (x+1)x^2 + 0(x+1)x^3 + (x+1)x^3(x-1) \\ &= x^5 - 2x^3 + 3x^2 - 4, \end{aligned}$$

$$R_3(x) = f[-1, 0, 0, 0, 1, 1, x](x+1)x^3(x-1)^2,$$

或者

$$R_3(x) = \frac{1}{6!} f^{(6)}(\xi)(x+1)x^3(x-1)^2.$$

解毕.

方法 2 (拉格朗日型) 下面只讨论含一阶导数的情况. 设已知函数 $f(x)$ 在 $n+1$ 个互异的插值节点 x_i ($i = 0, 1, \dots, n$) 处的函数值和导数值 $f(x_i), f'(x_i)$ ($i = 0, 1, \dots, n$), 要求构造一个不超过 $2n+1$ 次的多项式 $H_{2n+1}(x)$, 使其满足如下条件:

$$\begin{cases} H_{2n+1}(x_i) = f(x_i), \\ H'_{2n+1}(x_i) = f'(x_i), \end{cases} \quad i = 0, 1, \dots, n. \quad (4.4.1)$$

仿照拉格朗日插值多项式的构造方法, 令

$$H_{2n+1}(x) = \sum_{i=0}^n h_i(x)f(x_i) + \sum_{i=0}^n \bar{h}_i(x)f'(x_i), \quad (4.4.2)$$

其中 $h_i(x), \bar{h}_i(x)$ ($i = 0, 1, \dots, n$) 均为 $2n+1$ 次多项式, 称为插值基函数. 若它们分别满足如下条件:

$$h_i(x_j) = \begin{cases} 1, & j = i, \\ 0, & j \neq i, \end{cases} \quad i, j = 0, 1, \dots, n, \quad h'_i(x_j) = 0, \quad i, j = 0, 1, \dots, n, \quad (4.4.3)$$

$$\bar{h}_i(x_j) = 0, \quad i, j = 0, 1, \dots, n, \quad \bar{h}'_i(x_j) = \begin{cases} 1, & j = i, \\ 0, & j \neq i, \end{cases} \quad i, j = 0, 1, \dots, n, \quad (4.4.4)$$

则 $H_{2n+1}(x)$ 满足插值条件 (4.4.1), $H_{2n+1}(x)$ 就是所要求的埃尔米特插值多项式.

下面构造 $h_i(x)$ 和 $\bar{h}_i(x)$. 先构造 $h_i(x)$, 由条件 (4.4.3) 知, $x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ 是 $h_i(x)$ 的 2 重零点, 则 $h_i(x)$ 含有因子

$$(x-x_0)^2(x-x_1)^2 \cdots (x-x_{i-1})^2(x-x_{i+1})^2 \cdots (x-x_n)^2.$$

这是 $2n$ 次多项式, 因此, 可设

$$h_i(x) = (ax+b)l_i^2(x), \quad (4.4.5)$$

其中

$$l_i(x) = \frac{(x-x_0)(x-x_1) \cdots (x-x_{i-1})(x-x_{i+1}) \cdots (x-x_n)}{(x_i-x_0)(x_i-x_1) \cdots (x_i-x_{i-1})(x_i-x_{i+1}) \cdots (x_i-x_n)}.$$

由条件 $h_i(x_i) = 1, h'_i(x_i) = 0$ 得

$$\begin{aligned} (ax_i + b)l_i^2(x_i) &= 1, \\ al_i^2(x_i) + 2(ax_i + b)l_i(x_i)l'_i(x_i) &= 0, \end{aligned}$$

即

$$ax_i + b = 1,$$

$$a + 2(ax_i + b) \sum_{\substack{j=0 \\ j \neq i}}^n \frac{1}{x_i - x_j} = 0.$$

由此解得

$$a = -2 \sum_{\substack{j=0 \\ j \neq i}}^n \frac{1}{x_i - x_j}, \quad b = 1 + 2x_i \sum_{\substack{j=0 \\ j \neq i}}^n \frac{1}{x_i - x_j}.$$

将 a, b 代入式 (4.4.5) 得

$$h_i(x) = \left[1 - 2(x - x_i) \sum_{\substack{j=0 \\ j \neq i}}^n \frac{1}{x_i - x_j} \right] l_i^2(x). \quad (4.4.6)$$

现在构造 $\bar{h}_i(x)$. 由条件 (4.4.4) 知, x_i 是 $\bar{h}_i(x)$ 的 1 重零点, $x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ 是 $\bar{h}_i(x)$ 的 2 重零点, 故 $\bar{h}_i(x) = c(x - x_i)l_i^2(x)$, 而 $\bar{h}_i'(x) = c[l_i^2(x) + 2(x - x_i)l_i(x)l_i'(x)]$. 由 $\bar{h}_i'(x_i) = 1$ 可得常数 $c = 1$, 于是

$$\bar{h}_i(x) = (x - x_i)l_i^2(x). \quad (4.4.7)$$

将式 (4.4.6), (4.4.7) 代入式 (4.4.2) 得埃尔米特插值多项式

$$\begin{aligned} H_{2n+1}(x) &= \sum_{i=0}^n \left[1 - 2(x - x_i) \sum_{\substack{j=0 \\ j \neq i}}^n \frac{1}{x_i - x_j} \right] l_i^2(x) f(x_i) \\ &\quad + \sum_{i=0}^n (x - x_i) l_i^2(x) f'(x_i). \end{aligned} \quad (4.4.8)$$

按式 (4.1.3) 得其截断误差估计式为

$$\begin{aligned} R_{2n+1}(x) &= f(x) - H_{2n+1}(x) \\ &= \frac{f^{(2n+2)}(\xi)}{(2n+2)!} (x - x_0)^2 (x - x_1)^2 \cdots (x - x_n)^2. \end{aligned} \quad (4.4.9)$$

当 $n = 1$ 时, 有如下的三次埃尔米特插值多项式及其截断误差估计式:

$$\begin{aligned} H_3(x) &= \left(1 + 2 \frac{x_0 - x}{x_0 - x_1} \right) \left(\frac{x - x_1}{x_0 - x_1} \right)^2 f(x_0) + \left(1 + 2 \frac{x_1 - x}{x_1 - x_0} \right) \left(\frac{x - x_0}{x_1 - x_0} \right)^2 f(x_1) \\ &\quad + (x - x_0) \left(\frac{x - x_1}{x_0 - x_1} \right)^2 f'(x_0) + (x - x_1) \left(\frac{x - x_0}{x_1 - x_0} \right)^2 f'(x_1), \end{aligned} \quad (4.4.10)$$

$$R_3(x) = \frac{f^{(4)}(\xi)}{4!} (x - x_0)^2 (x - x_1)^2. \quad (4.4.11)$$

4.5 分段低次插值多项式

4.5.1 高次插值多项式的缺陷

人们为了获得较好的近似效果, 误以为在给定的插值区间 $[a, b]$ 上插值节点取得越多, 构造的插值多项式 $p_n(x)$ 的次数越高, $p_n(x)$ 逼近 $f(x)$ 的效果就越好, 但实际并非如此. 从插值多项式的截断误差估计式

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \pi_{n+1}(x)$$

可以看到, 截断误差与 $f^{(n+1)}(x)$ 及 $\pi_{n+1}(x)$ 有关. 随着 n 的增大, 不能保证插值多项式 $p_n(x)$ 充分逼近被插函数 $f(x)$. 20 世纪初, 龙格 (Runge) 就给出了一个这样的例子. 给定被插函数

$$f(x) = \frac{1}{1 + 25x^2}, \quad -1 \leq x \leq 1,$$

在区间 $[-1, 1]$ 上等距地取 $n+1$ 个节点 $x_i = -1 + i\frac{2}{n}$ ($i = 0, 1, \dots, n$), 构造拉格朗日插值多项式 $L_n(x)$. 例如, 取 $n = 10$, 插值多项式 $y = L_{10}(x)$ 及被插函数 $y = f(x)$ 的图形如图 4.2 所示.

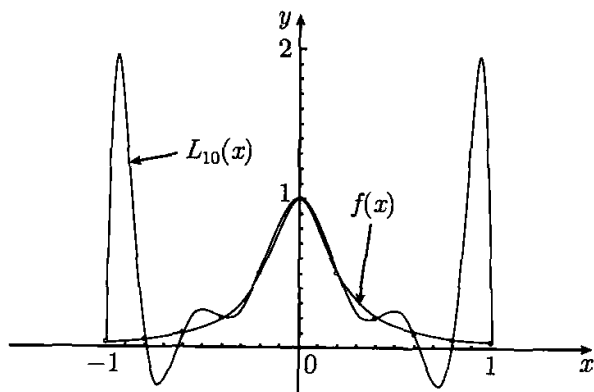


图 4.2

由图 4.2 可见, 在区间 $[-0.2, 0.2]$ 上 $L_{10}(x)$ 逼近 $f(x)$ 的效果较好, 但在靠近区间 $[-1, 1]$ 的两端, 二者相差甚远. 随着插值节点个数的增多, 构造的插值多项式的次数越高, 这种现象就越严重, 这种现象称为龙格现象. 另一方面, 由于存在数据误差, 当函数值 $f(x_i)$ 变为 $\tilde{f}(x_i)$ ($i = 0, 1, \dots, n$) 时, 设最大误差限为 ε , 则拉格朗日插值多项式的计算值 $\tilde{L}_n(x)$ 也有误差,

$$|L_n(x) - \tilde{L}_n(x)| = \left| \sum_{i=0}^n l_i(x) f(x_i) - \sum_{i=0}^n l_i(x) \tilde{f}(x_i) \right|$$

$$\leq \sum_{i=0}^n |l_i(x)| |f(x_i) - \tilde{f}(x_i)| \leq \varepsilon \sum_{i=0}^n |l_i(x)|,$$

而且很可能接近误差限 $\varepsilon \sum_{i=0}^n |l_i(x)|$. 当 n 很大时, $\sum_{i=0}^n |l_i(x)| \gg 1$, $\tilde{L}_n(x)$ 的误差就可能很大. 也就是说, 当 n 很大时, 插值多项式不稳定.

鉴于以上两个原因, 不宜采用高次插值多项式, 一般限定插值多项式的次数 $n < 7$.

4.5.2 分段低次插值法

为了避免高次插值多项式的缺陷, 得到 $f(x)$ 的较好近似式, 一般采用分段插值法, 即把插值区间 $[a, b]$ 分为若干个子区间, 在每个子区间上构造低次插值多项式, 如分段线性、分段二次、分段三次插值多项式等.

设 $a = x_0 < x_1 < \cdots < x_n = b$.

1. 分段线性插值多项式

在每个子区间 $[x_{i-1}, x_i]$ ($i = 1, 2, \cdots, n$) 上作线性插值, 即取

$$f(x) \approx N_{1,i}(x) = f(x_{i-1}) + f[x_{i-1}, x_i](x - x_{i-1}), \quad x_{i-1} \leq x \leq x_i$$

或

$$f(x) \approx L_{1,i}(x) = \frac{x - x_i}{x_{i-1} - x_i} f(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} f(x_i), \quad x_{i-1} \leq x \leq x_i.$$

从几何上来看, 分段线性插值就是用折线来近似曲线 $y = f(x)$. 截断误差

$$R_{1,i}(x) = \frac{f''(\xi_i)}{2!} (x - x_{i-1})(x - x_i),$$

其中 $\xi_i \in (x_{i-1}, x_i)$ 且与 x 有关.

记 $h_i = x_i - x_{i-1}$ ($i = 1, 2, \cdots, n$), $h = \max_{1 \leq i \leq n} h_i$, $M_2 = \max_{a \leq x \leq b} |f''(x)|$, 则

$$|R_{1,i}(x)| \leq \frac{h^2}{8} M_2. \quad (4.5.1)$$

2. 分段二次插值多项式

设 n 为偶数, 在每个子区间 $[x_{i-1}, x_{i+1}]$ ($i = 2k - 1, k = 1, 2, \cdots, n/2$) 上作二次插值多项式, 取

$$\begin{aligned} f(x) &\approx N_{2,i}(x) \\ &= f(x_{i-1}) + f[x_{i-1}, x_i](x - x_{i-1}) \\ &\quad + f[x_{i-1}, x_i, x_{i+1}](x - x_{i-1})(x - x_i), \quad x_{i-1} \leq x \leq x_{i+1} \end{aligned}$$

或

$$f(x) \approx L_{2,i}(x) = \frac{(x-x_i)(x-x_{i+1})}{(x_{i-1}-x_i)(x_{i-1}-x_{i+1})} f(x_{i-1}) + \frac{(x-x_{i-1})(x-x_{i+1})}{(x_i-x_{i-1})(x_i-x_{i+1})} f(x_i) \\ + \frac{(x-x_{i-1})(x-x_i)}{(x_{i+1}-x_{i-1})(x_{i+1}-x_i)} f(x_{i+1}), \quad x_{i-1} \leq x \leq x_{i+1},$$

截断误差

$$R_{2,i}(x) = \frac{f'''(\xi_i)}{3!} (x-x_{i-1})(x-x_i)(x-x_{i+1}), \quad \xi_i \in (x_{i-1}, x_{i+1}),$$

并且 ξ_i 与 x 有关,

$$|R_{2,i}(x)| \leq \frac{h^3}{12} M_3, \quad (4.5.2)$$

其中 $M_3 = \max_{a \leq x \leq b} |f'''(x)|$.

3. 分段三次埃尔米特插值多项式

若已知函数 $y = f(x)$ 在节点 x_i 上的函数值 $f(x_i)$ 与导数值 $f'(x_i)$ ($i = 0, 1, \dots, n$) 可在子区间 $[x_{i-1}, x_i]$ ($i = 1, 2, \dots, n$) 上构造三次插值多项式. 由式 (4.4.10), (4.4.11) 有

$$f(x) \approx H_{3,i}(x) \\ = \left(1 + 2\frac{x-x_{i-1}}{h_i}\right) \left(\frac{x-x_i}{h_i}\right)^2 f(x_{i-1}) + \left(1 + 2\frac{x_i-x}{h_i}\right) \left(\frac{x-x_{i-1}}{h_i}\right)^2 f(x_i) \\ + (x-x_{i-1}) \left(\frac{x-x_i}{h_i}\right)^2 f'(x_{i-1}) + (x-x_i) \left(\frac{x-x_{i-1}}{h_i}\right)^2 f'(x_i), \quad x_{i-1} \leq x \leq x_i, \\ R_{3,i}(x) = \frac{f^{(4)}(\xi_i)}{4!} (x-x_{i-1})^2 (x-x_i)^2, \quad \xi_i \in (x_{i-1}, x_i)$$

并且 ξ_i 与 x 有关, 则

$$|R_{3,i}(x)| \leq \frac{h^4}{384} M_4. \quad (4.5.3)$$

由式 (4.5.1)~(4.5.3) 可以看出, 若 $f(x)$ 的相应阶 (二、三、四阶) 导数连续, 则当 $h \rightarrow 0$ 时, 截断误差趋于零, 所以相应的插值多项式在插值区间 $[a, b]$ 上一致收敛于被插函数 $f(x)$. 若要提高逼近程度, 只需增加插值节点即可.

以上三种插值多项式具有公式简单、计算量小、稳定性好、一致收敛等优点, 但在插值子区间端点处, 分段线性插值多项式和分段二次插值多项式不能保证一阶导数的连续性, 分段三次插值多项式不能保证其二阶导数的连续性, 因而不能满足曲线充分光滑的要求.

4.6 三次样条插值函数

4.5 节介绍的分段低次插值多项式, 它的一阶或二阶导数不连续, 不满足许多实际问题的要求. 例如, 在飞机机翼外形设计、船体放样等许多实际问题中, 要求曲线二阶光滑, 即要求函数的二阶导数连续. 为了将一些已知点连成一条充分光滑的曲线, 早期的做法是绘图员用压铁把一条称为样条 (spline) 的富有弹性的细木条或细金属条固定在相邻的若干点上, 然后沿样条画下一条所需的光滑曲线, 这条曲线称为样条曲线. 实际上, 它是由分段三次曲线拼接而成的, 在连接点处二阶导数连续. 将样条曲线抽象为数学问题称为样条函数.

4.6.1 三次样条插值函数的定义

定义 4.6.1 设在区间 $[a, b]$ 上给定 $n+1$ 个节点 x_i ($a \leq x_0 < x_1 < \cdots < x_n \leq b$), 在节点 x_i 处的函数值为 $y_i = f(x_i)$ ($i = 0, 1, \cdots, n$). 若函数 $S(x)$ 满足以下三条:

- (1) 在每个子区间 $[x_{i-1}, x_i]$ ($i = 1, 2, \cdots, n$) 上, $S(x)$ 是三次多项式;
- (2) $S(x_i) = y_i$ ($i = 0, 1, \cdots, n$);
- (3) 在区间 $[a, b]$ 上, $S(x)$ 的二阶导函数 $S''(x)$ 连续,

则称 $S(x)$ 为函数 $y = f(x)$ 在区间 $[a, b]$ 上的三次样条插值函数.

由定义 4.6.1 知, $S(x)$ 是分段三次多项式, 故在每个子区间 $[x_{i-1}, x_i]$ ($i = 1, 2, \cdots, n$) 上, $S(x)$ 有 4 个待定参数, 共有 n 个子区间, 所以 $S(x)$ 共有 $4n$ 个待定参数. 根据定义 4.6.1 中的条件 (3) 知

$$\begin{cases} S_-(x_i) = S_+(x_i), \\ S'_-(x_i) = S'_+(x_i), \\ S''_-(x_i) = S''_+(x_i), \end{cases} \quad i = 1, 2, \cdots, n-1.$$

上式共有 $3n-3$ 个条件, 再加上定义 4.6.1 条件 (2) 中的 $n+1$ 个条件, 共有 $4n-2$ 个条件. 还需要增加两个条件才能确定 $4n$ 个待定参数, 即才能确定 $S(x)$. 所增加的条件称为边界条件或端点条件. 下面给出三种常用的边界条件:

- (1) 已知 $f(x)$ 在区间 $[a, b]$ 的两端点 a, b 处的二阶导数值 $f''(a), f''(b)$;
- (2) 已知 $f(x)$ 在区间 $[a, b]$ 的两端点 a, b 处的一阶导数值 $f'(a), f'(b)$;
- (3) 已知函数 $f(x)$ 是以 $T = b - a$ 为周期的周期函数.

4.6.2 三次样条插值函数的导出

1. 导出在子区间 $[x_{i-1}, x_i]$ ($i = 1, 2, \cdots, n$) 上 $S(x)$ 的表达式

由于 $S(x)$ 的二阶导数连续, 设 $S(x)$ 在节点 x_i 处的二阶导数值 $S''(x_i) =$

M_i ($i = 0, 1, \dots, n$), 其中 M_i 为未知的待定参数. 由 $S(x)$ 是分段三次多项式知, $S''(x)$ 是分段线性函数, $S''(x)$ 在子区间 $[x_{i-1}, x_i]$ 上可以表示为

$$\begin{aligned} S''(x) &= \frac{x - x_i}{x_{i-1} - x_i} M_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} M_i \\ &= \frac{x_i - x}{h_i} M_{i-1} + \frac{x - x_{i-1}}{h_i} M_i, \quad x_{i-1} \leq x \leq x_i, \end{aligned} \quad (4.6.1)$$

其中 $h_i = x_i - x_{i-1}$. 对式 (4.6.1) 两端积分两次得

$$\begin{aligned} S(x) &= \frac{(x_i - x)^3}{6h_i} M_{i-1} + \frac{(x - x_{i-1})^3}{6h_i} M_i \\ &\quad + b_i(x_i - x) + c_i(x - x_{i-1}), \quad x_{i-1} \leq x \leq x_i, \end{aligned} \quad (4.6.2)$$

其中 b_i, c_i 为积分常数. 现在确定 b_i, c_i . 由插值条件 $S(x_{i-1}) = y_{i-1}$, $S(x_i) = y_i$ 得

$$\frac{h_i^2}{6} M_{i-1} + b_i h_i = y_{i-1}, \quad \frac{h_i^2}{6} M_i + c_i h_i = y_i.$$

由此解得

$$b_i = \left(y_{i-1} - \frac{h_i^2}{6} M_{i-1} \right) / h_i, \quad c_i = \left(y_i - \frac{h_i^2}{6} M_i \right) / h_i.$$

将 b_i, c_i 代入式 (4.6.2), 则得 $S(x)$ 在子区间 $[x_{i-1}, x_i]$ ($i = 1, 2, \dots, n$) 上的表达式为

$$\begin{aligned} S(x) &= \frac{(x_i - x)^3}{6h_i} M_{i-1} + \frac{(x - x_{i-1})^3}{6h_i} M_i + \left(y_{i-1} - \frac{h_i^2}{6} M_{i-1} \right) \frac{x_i - x}{h_i} \\ &\quad + \left(y_i - \frac{h_i^2}{6} M_i \right) \frac{x - x_{i-1}}{h_i}, \quad x_{i-1} \leq x \leq x_i. \end{aligned} \quad (4.6.3)$$

2. 建立关于参数 M_i 的方程组

对式 (4.6.3) 求导得

$$\begin{aligned} S'(x) &= -\frac{(x_i - x)^2}{2h_i} M_{i-1} + \frac{(x - x_{i-1})^2}{2h_i} M_i \\ &\quad + \frac{y_i - y_{i-1}}{h_i} - \frac{M_i - M_{i-1}}{6} h_i, \quad x_{i-1} \leq x \leq x_i. \end{aligned} \quad (4.6.4)$$

在式 (4.6.4) 中, 令 $x = x_i$ 得 $S(x)$ 在 x_i 处的左导数

$$\begin{aligned} S'_-(x_i) &= \frac{h_i}{2} M_i + \frac{y_i - y_{i-1}}{h_i} - \frac{M_i - M_{i-1}}{6} h_i \\ &= \frac{h_i}{6} M_{i-1} + \frac{h_i}{3} M_i + \frac{y_i - y_{i-1}}{h_i}. \end{aligned} \quad (4.6.5)$$

在式 (4.6.4) 中, 令 $x = x_{i-1}$ 得 $S(x)$ 在 x_{i-1} 处的右导数

$$\begin{aligned} S'_+(x_{i-1}) &= -\frac{h_i}{2}M_{i-1} + \frac{y_i - y_{i-1}}{h_i} - \frac{M_i - M_{i-1}}{6}h_i \\ &= -\frac{h_i}{3}M_{i-1} - \frac{h_i}{6}M_i + \frac{y_i - y_{i-1}}{h_i}, \end{aligned}$$

从而

$$S'_+(x_i) = -\frac{h_{i+1}}{3}M_i - \frac{h_{i+1}}{6}M_{i+1} + \frac{y_{i+1} - y_i}{h_{i+1}}. \quad (4.6.6)$$

由 $S(x)$ 在节点 x_i 处一阶导数的连续性知

$$S'_-(x_i) = S'_+(x_i), \quad i = 1, 2, \dots, n-1,$$

即

$$\frac{h_i}{6}M_{i-1} + \frac{h_i + h_{i+1}}{3}M_i + \frac{h_{i+1}}{6}M_{i+1} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}, \quad i = 1, 2, \dots, n-1.$$

上式两端同乘 $\frac{6}{h_i + h_{i+1}}$ 得

$$\frac{h_i}{h_i + h_{i+1}}M_{i-1} + 2M_i + \frac{h_{i+1}}{h_i + h_{i+1}}M_{i+1} = \frac{6}{h_i + h_{i+1}} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right). \quad (4.6.7)$$

记

$$\mu_i = \frac{h_i}{h_i + h_{i+1}}, \quad \lambda_i = \frac{h_{i+1}}{h_i + h_{i+1}} = 1 - \mu_i,$$

$$d_i = \frac{6}{h_i + h_{i+1}} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right) = 6f[x_{i-1}, x_i, x_{i+1}], \quad i = 1, 2, \dots, n-1,$$

则关于 M_i 的方程组 (4.6.7) 可写为

$$\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = d_i, \quad i = 1, 2, \dots, n-1. \quad (4.6.8)$$

方程组 (4.6.8) 中的未知量 M_i 在力学上解释为细梁在 x_i 截面处的弯矩, 故称为三弯矩方程组. 三弯矩方程组中只有 $n-1$ 个方程, 不能完全确定 $n+1$ 个未知量 M_i ($i = 0, 1, \dots, n$), 欲解三弯矩方程组或者减少两个未知量, 或者再增加两个方程, 这由边界条件来确定.

3. 三种边界条件的三弯矩方程

(1) 第一种边界条件: 已知 $f''(a), f''(b)$. 取 $M_0 = f''(a), M_n = f''(b)$, 这时三弯矩方程组减少了两个未知量, 变成只含 $n-1$ 个未知量 M_i ($i = 1, 2, \dots, n-1$) 的

$n-1$ 个方程的方程组

$$\begin{cases} 2M_1 + \lambda_1 M_2 = d_1 - \mu_1 M_0, \\ \mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = d_i, & i = 2, 3, \dots, n-2. \\ \mu_{n-1} M_{n-2} + 2M_{n-1} = d_{n-1} - \lambda_{n-1} M_n, \end{cases} \quad (4.6.9)$$

用矩阵表示为

$$\begin{pmatrix} 2 & \lambda_1 & & & \\ \mu_2 & 2 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-2} & 2 & \lambda_{n-2} \\ & & & \mu_{n-1} & 2 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-2} \\ M_{n-1} \end{pmatrix} = \begin{pmatrix} d_1 - \mu_1 M_0 \\ d_2 \\ \vdots \\ d_{n-2} \\ d_{n-1} - \lambda_{n-1} M_n \end{pmatrix}. \quad (4.6.9')$$

注意到 $\mu_i + \lambda_i = 1 < 2$, 以上方程组的系数矩阵是严格三对角占优矩阵, 可用追赶法求解. 解出 M_i ($i = 1, 2, \dots, n-1$) 后, 代入式 (4.6.3), 即得三次样条插值函数的表达式.

若给定 $M_0 = M_n = 0$, 则相应的 $S(x)$ 称为自然三次样条插值函数.

(2) 第二种边界条件: 已知 $f'(a)$, $f'(b)$. 记 $y'_0 = f'(a)$, $y'_n = f'(b)$, 则有 $S'_+(x_0) = y'_0$, $S'_-(x_n) = y'_n$. 由式 (4.6.5), (4.6.6) 得

$$-\frac{h_1}{3}M_0 - \frac{h_1}{6}M_1 + \frac{y_1 - y_0}{h_1} = y'_0, \quad \frac{h_n}{6}M_{n-1} + \frac{h_n}{3}M_n + \frac{y_n - y_{n-1}}{h_n} = y'_n,$$

也即

$$2M_0 + M_1 = d_0, \quad M_{n-1} + 2M_n = d_n,$$

其中

$$\begin{aligned} d_0 &= \frac{6}{h_1} \left(\frac{y_1 - y_0}{h_1} - y'_0 \right) = 6f[x_0, x_0, x_1], \\ d_n &= \frac{6}{h_n} \left(y'_n - \frac{y_n - y_{n-1}}{h_n} \right) = 6f[x_{n-1}, x_n, x_n]. \end{aligned}$$

将以上两个方程添加到方程组 (4.6.8), 得到关于第二种边界条件的三弯矩方程组

$$\begin{cases} 2M_0 + M_1 = d_0, \\ \mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = d_i, & i = 1, 2, \dots, n-1, \\ M_{n-1} + 2M_n = d_n. \end{cases} \quad (4.6.10)$$

用矩阵表示为

$$\begin{pmatrix} 2 & 1 & & & \\ \mu_1 & 2 & \lambda_1 & & \\ & \mu_2 & 2 & \lambda_2 & \\ & & \ddots & \ddots & \ddots \\ & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & & 1 & 2 \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}. \quad (4.6.10')$$

同样, 该方程组的系数矩阵是严格三对角占优矩阵, 可用追赶法求解.

(3) 第三种边界条件: 周期型边界条件. 已知 $y = f(x)$ 是以 $T = b - a = x_n - x_0$ 为周期的周期函数, 则由周期性知, $y_n = y_0, y_{n+1} = y_1, M_n = M_0, M_{n+1} = M_1, h_{n+1} = h_1$. 这时, 将点 x_n 看成内节点, 则方程组 (4.6.8) 对 $i = n$ 也成立, 即有

$$\mu_n M_{n-1} + 2M_n + \lambda_n M_{n+1} = d_n,$$

也即

$$\lambda_n M_1 + \mu_n M_{n-1} + 2M_n = d_n,$$

其中

$$d_n = \frac{6}{h_n + h_{n+1}} \left(\frac{y_{n+1} - y_n}{h_{n+1}} - \frac{y_n - y_{n-1}}{h_n} \right) = \frac{6}{h_n + h_1} \left(\frac{y_1 - y_n}{h_1} - \frac{y_n - y_{n-1}}{h_n} \right).$$

方程组 (4.6.8) 的第 1 个 ($i = 1$) 方程为

$$\mu_1 M_0 + 2M_1 + \lambda_1 M_2 = d_1,$$

即

$$2M_1 + \lambda_1 M_2 + \mu_1 M_n = d_1.$$

于是可得关于第三种边界条件的三弯矩方程组

$$\begin{cases} 2M_1 + \lambda_1 M_2 + \mu_1 M_n = d_1, \\ \mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = d_i, \quad i = 2, 3, \dots, n-1, \\ \lambda_n M_1 + \mu_n M_{n-1} + 2M_n = d_n. \end{cases} \quad (4.6.11)$$

用矩阵表示为

$$\begin{pmatrix} 2 & \lambda_1 & & & \mu_1 \\ \mu_2 & 2 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 2 & \lambda_{n-1} \\ \lambda_n & & & \mu_n & 2 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}. \quad (4.6.11')$$

显然, 该方程组的系数矩阵也是严格对角占优矩阵, 可用习题 2.4 的方法求解.

根据不同的边界条件, 求解相应的三弯矩方程组, 将求得的 M_i 代入式 (4.6.3), 便得三次样条插值函数 $S(x)$.

例 4.6.1 已知函数 $y = f(x)$ 的数值如表 4.5 所示, 求它的自然三次样条插值函数 $S(x)$.

表 4.5

x_i	-3	-1	0	3	4
y_i	7	11	26	56	29

解 由 $h_i = x_i - x_{i-1}$ 得

$$h_1 = 2, \quad h_2 = 1, \quad h_3 = 3, \quad h_4 = 1.$$

由 $\mu_i = \frac{h_i}{h_i + h_{i+1}}, \lambda_i = 1 - \mu_i$ ($i = 1, 2, 3$) 得

$$\begin{aligned} \mu_1 &= \frac{2}{3}, & \mu_2 &= \frac{1}{4}, & \mu_3 &= \frac{3}{4}, \\ \lambda_1 &= \frac{1}{3}, & \lambda_2 &= \frac{3}{4}, & \lambda_3 &= \frac{1}{4}. \end{aligned}$$

由 $d_i = 6f[x_{i-1}, x_i, x_{i+1}]$ ($i = 1, 2, 3$) 得

$$d_1 = 26, \quad d_2 = -\frac{15}{2}, \quad d_3 = -\frac{111}{2}.$$

对于自然三次样条函数, $M_0 = M_4 = 0$, 所以三弯矩方程组为

$$\begin{cases} 2M_1 + \frac{1}{3}M_2 = 26, \\ \frac{1}{4}M_1 + 2M_2 + \frac{3}{4}M_3 = -\frac{15}{2}, \\ \frac{3}{4}M_2 + 2M_3 = -\frac{111}{2}. \end{cases}$$

从以上方程组解得 $M_1 = 12, M_2 = 6, M_3 = -30$. 将 M_i 代入 $S(x)$ 的表达式 (4.6.3) 得

$$S(x) = \begin{cases} x^3 + 9x^2 + 25x + 28, & -3 \leq x \leq -1, \\ -x^3 + 3x^2 + 19x + 26, & -1 \leq x \leq 0, \\ -2x^3 + 3x^2 + 19x + 26, & 0 \leq x \leq 3, \\ 5x^3 - 60x^2 + 208x - 163, & 3 \leq x \leq 4. \end{cases}$$

解毕.

4.6.3 三次样条插值函数的收敛性与误差估计

三次样条插值函数的收敛性与误差估计比较复杂,下面仅给出满足第一、二种边界条件的三次样条插值函数的结论.

定理 4.6.1 设 $f(x) \in C^4[a, b]$, $S(x)$ 是满足第一种或第二种边界条件的三次样条插值函数, 则

$$\max_{a \leq x \leq b} |f^{(k)}(x) - S^{(k)}(x)| \leq c_k \max_{a \leq x \leq b} |f^{(4)}(x)| h^{4-k} \leq c_k M_4 h^{4-k}, \quad k = 0, 1, 2, 3,$$

其中

$$c_0 = \frac{5}{384}, \quad c_1 = \frac{1}{24}, \quad c_2 = \frac{3}{8}, \quad c_3 = \frac{\beta + \beta^{-1}}{2},$$

$$\beta = \frac{\max h_i}{\min h_i}, \quad h = \max h_i, \quad M_4 = \max_{a \leq x \leq b} |f^{(4)}(x)|.$$

证明 略.

三次样条插值函数的优点如下:

- (1) 关于 M_i 的三弯矩方程组的系数矩阵是严格对角占优矩阵, 解存在且唯一, 求解稳定性好;
- (2) 定理 4.6.1 表明
 - (i) 若要提高精度, 只需增加插值节点, 不需要提高插值多项式的次数;
 - (ii) 随节点的加密, 即当 $h \rightarrow 0$ 时, $S(x), S'(x), S''(x)$ 分别一致地收敛于 $f(x), f'(x), f''(x)$; 当 β 有界时, $S'''(x)$ 一致地收敛于 $f'''(x)$.

小 结

本章主要介绍了 4 种插值多项式, 它们各有特点, 读者根据所解决问题的不同要求, 合理选择使用.

一般地, 在理论研究中, 插值节点及其对应的函数值用的是字符而非数字, 这时无法构造差商表, 多用拉格朗日插值多项式. 在工程技术问题中, 插值节点及其对应的函数值是通过试验或观测得到的数据, 这时最好选用计算简便且计算量较小的牛顿插值多项式. 若给定的数据点较多又无可导 (或光滑性) 要求, 为避免龙格现象, 宜选用分段低次插值多项式. 若要求构造的插值多项式除了满足插值条件外, 还要求与被插函数在某个或某些插值节点处若干阶导数相等, 这时选用埃尔米特插值多项式. 若要求构造的插值多项式在整个插值区间上二阶导函数连续 (二阶光滑), 则选用三次样条插值函数. 这时, 根据不同的端点条件, 首先要解一个三弯矩线性方程组, 这无疑增加了一些计算工作量.

总之, 由于高次插值多项式的龙格现象, 故在实际中很少使用它 (通常, 插值多项式的次数 $n < 7$), 而更多地使用分段低次插值多项式, 特别是三次样条插值函数,

由于它具有良好的收敛性、计算稳定性和二阶光滑性, 因此, 在理论研究和实际应用中具有重要的意义.

习 题

4.1 已知函数 $y = f(x)$ 在若干点处的函数值如下表所示, 求其拉格朗日插值多项式和牛顿插值多项式, 并写出截断误差表达式.

(1)	x_i	0	1	2	5
	y_i	2	3	12	147

;

(2)	x_i	-2	-1	0	1
	y_i	15	4	5	24

.

4.2 函数 $y = \sqrt{x}$ 在三点 100, 121, 144 处的值分别为 10, 11, 12, 由此建立二次插值多项式, 计算 $\sqrt{115}$ 的近似值并估计误差.

4.3 在 $x = -1, 1, 2$ 处, $f(x) = -3, 0, 4$, 求 $f(x)$ 的二次插值多项式.

(1) $p_2(x) = a_0 + a_1x + a_2x^2$;

(2) 拉格朗日插值多项式;

(3) 牛顿插值多项式,

证明上述三种方法求得的插值多项式是相同的.

4.4 已知二次式 $f(x)$ 在点 $x = 0, 1, 2$ 处的值分别为 $-0.81, 0.19, 3.19$, 求 $f(x)$ 的零点、极值点、在 $x = 1$ 处的导数和积分 $\int_0^2 f(x) dx$.

4.5 设函数 $f(x) = e^x (-4 \leq x \leq 4)$, 节点等距分布. 若用二次插值求 e^x 的近似值, 要使截断误差不超过 10^{-6} , 问步长 h 应取多大?

4.6 证明 k 阶差商的如下性质:

(1) 设 $p(x) = cf(x)$, 则

$$p[x_0, x_1, \dots, x_k] = cf[x_0, x_1, \dots, x_k];$$

(2) 设 $p(x) = f(x) + g(x)$, 则

$$p[x_0, x_1, \dots, x_k] = f[x_0, x_1, \dots, x_k] + g[x_0, x_1, \dots, x_k].$$

4.7 $f(x) = x^7 + x^4 + 3x + 1$, 求 $f[2^0, 2^1, \dots, 2^7]$ 和 $f[2^0, 2^1, \dots, 2^8]$.

4.8 设 $f(x)$ 在区间 $[a, b]$ 上三阶导数连续, $x_0, x_1 \in [a, b]$, 构造满足如下条件:

$$p(x_0) = f(x_0), \quad p'(x_0) = f'(x_0), \quad p(x_1) = f(x_1)$$

的二次插值多项式 $p(x)$, 并写出截断误差表达式.

4.9 已知函数 $y = f(x)$ 在若干点处的函数值、导数值如下表所示, 求埃尔米特插值多项式和截断误差表达式:

$$(1) \begin{array}{ccc|c} x_i & -1 & 0 & 1 \\ y_i & -1 & 0 & 1 \\ y'_i & 0 & 0 & 0 \end{array}; \quad (2) \begin{array}{ccc|c} x_i & 0 & 1 & 2 \\ y_i & 1 & -1 & 0 \\ y'_i & 0 & 0 & \\ y''_i & 2 & & \end{array}.$$

4.10 设 x_i ($i = 0, 1, \dots, n$) 是互不相同的插值节点, $l_i(x)$ ($i = 0, 1, \dots, n$) 是拉格朗日插值基函数. 证明:

$$(1) \sum_{i=0}^n l_i(x) = 1;$$

$$(2) \sum_{i=0}^n l_i(x) x_i^k = x^k \quad (k = 1, 2, \dots, n);$$

$$(3) \sum_{i=0}^n l_i(x) (x_i - x)^k = 0 \quad (k = 1, 2, \dots, n);$$

$$(4) \sum_{i=0}^n l_i(0) x_i^k = \begin{cases} 1, & k = 0, \\ 0, & k = 1, 2, \dots, n, \\ (-1)^n x_0 x_1 \cdots x_n, & k = n + 1. \end{cases}$$

4.11 设 $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1} + a_n x^n$, 方程 $f(x) = 0$ 有 n 个不同的实根 $x_1, x_2, \dots, x_n$, 证明

$$(1) \sum_{i=1}^n \frac{x_i^k}{f'(x_i)} = \begin{cases} 0, & 0 \leq k \leq n-2, \\ 1/a_n, & k = n-1. \end{cases}$$

(2) 设 $x_i \neq 0, -1$ ($i = 1, 2, \dots, n$), 则

$$\sum_{i=1}^n \frac{x_i^n f(x_i^{-1})}{f'(x_i)(1+x_i)} = (-1)^n (x_1 x_2 \cdots x_n - 1);$$

$$(3) \sum_{i=1}^n \frac{i^k}{(i-1) \cdots (i-i+1)(i-i-1) \cdots (i-n)} = 0, \quad k = 0, 1, \dots, n-2.$$

4.12 设 $f(x) = e^x$, 已知

$$f(-1) = 0.367879, \quad f(0) = 1, \quad f(1) = 2.718282.$$

利用以下方法计算 $e^{0.8}$ 并估计误差:

(1) 利用分段线性插值法;

(2) 利用分段三次插值法;

(3) 已知 $S'(-1) = f'(-1)$, $S'(1) = f'(1)$, 利用三次样条函数插值法.

4.13 已知函数 $y = f(x)$ 在若干点处的函数值如下表所示, 求满足以下端点条件的三次样条插值函数:

$$(1) S'(0) = 0, S'(4) = 48;$$

$$(2) S''(0) = 0, S''(4) = 24.$$

x_i	0	1	2	3	4
y_i	-8	-7	0	19	56

4.14 已知函数 $y = f(x)$ 在若干点处的函数值如下表所示, 求满足以下端点条件的三次样条插值函数:

$$(1) S'(0.25) = 1.0000, S'(0.53) = 0.6868;$$

$$(2) S''(0.25) = -2, S''(0.53) = -0.6479.$$

x_i	0.25	0.30	0.39	0.45	0.53
y_i	0.5000	0.5477	0.6245	0.6708	0.7280

计算实习

观察高次插值多项式的龙格现象.

给定函数 $f(x) = \frac{1}{1+25x^2}$ ($-1 \leq x \leq 1$). 取等距节点, 构造牛顿插值多项式 $N_5(x)$ 和 $N_{10}(x)$ 及三次样条插值函数 $S_{10}(x)$. 分别将三种插值多项式与 $f(x)$ 的曲线画在同一个坐标系上进行比较.

第5章 函数最优逼近

对于定义在区间 $[a, b]$ 上表达式比较复杂的连续函数 $f(x)$, 或者 $f(x)$ 无表达式, 只知道 $f(x)$ 在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的函数值的列表函数, 构造一个简单函数作为 $f(x)$ 的近似表达式, 这就是函数逼近问题. 由于多项式便于计算, 容易求其积分、微分. 因此, 近似式常取为多项式. 本章介绍的最优平方逼近和最优一致逼近多项式 $p(x)$, 使得在整个区间 $[a, b]$ 上或在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上 $p(x)$ 与 $f(x)$ 的误差在两种不同度量意义下达到最小.

5.1 函数的内积、范数和正交多项式

在讨论两种最优逼近之前, 先介绍函数的内积、范数和正交多项式的定义.

5.1.1 函数的内积和范数

将区间 $[a, b]$ 上所有连续函数构成的集合记为 $C[a, b]$, $C[a, b]$ 称为连续函数空间, 它是一个线性空间. 与欧氏空间 $\mathbf{R}^n$ 中向量的线性相关与线性无关的定义类似, 下面给出函数空间中函数线性相关与线性无关的定义.

定义 5.1.1 设 V 是 $C[a, b]$ 的一个线性子空间, 函数 $\phi_0(x), \phi_1(x), \dots, \phi_n(x) \in V \subset C[a, b]$, $c_i \in \mathbf{R}$ ($i = 0, 1, \dots, n$), 如果关系式

$$c_0\phi_0(x) + c_1\phi_1(x) + \dots + c_n\phi_n(x) = 0, \quad \forall x \in \mathbf{R}$$

当且仅当 $c_0 = c_1 = \dots = c_n = 0$ 时成立, 则称 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ 线性无关; 否则, 称 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ 线性相关.

若线性子空间 V 是由 $n+1$ 个线性无关的函数 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ 所构成的, 即 $\forall p(x) \in V$ 都有

$$p(x) = c_0\phi_0(x) + c_1\phi_1(x) + \dots + c_n\phi_n(x), \quad (5.1.1)$$

则称 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ 为线性子空间 V 的一组基函数, 称 V 为由 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ 生成的线性子空间, 记为

$$V = \text{span}\{\phi_0(x), \phi_1(x), \dots, \phi_n(x)\},$$

并称 V 为 $n+1$ 维空间, 系数 $c_0, c_1, \dots, c_n$ 称为 $p(x)$ 在基 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ 下的坐标. 由式 (5.1.1) 表示的 $p(x)$ 称为广义多项式.

特别地, 若取 $\phi_i(x) = x^i$ ($i = 0, 1, \dots, n$), 则显然, $\phi_i(x) \in C[a, b]$, $\phi_i(x)$ ($i = 0, 1, \dots, n$) 线性无关. 记 H_n 是由不超过 n 次的多项式

$$p_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

构成的集合, 则 $\phi_i(x) = x^i$ ($i = 0, 1, \dots, n$) 是它的一组基. $H_n = \text{span}\{1, x, x^2, \dots, x^n\}$ 是 $C[a, b]$ 的 $n+1$ 维子空间.

下面给出定义在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的列表函数和定义在区间 $[a, b]$ 上的连续函数内积的定义.

定义 5.1.2 已知函数 $f(x), g(x)$ 在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的函数值 $f(x_i), g(x_i)$ 和权系数 ω_i ($i = 1, 2, \dots, m$), 则

$$(f, g) = \sum_{i=1}^m \omega_i f(x_i) g(x_i) \quad (5.1.2)$$

称为函数 f 与 g 在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上关于权系数 ω_i 的内积.

定义 5.1.3 给定 $f, g \in C[a, b]$, $\omega(x)$ 是区间 $[a, b]$ 上的权函数, 则

$$(f, g) = \int_a^b \omega(x) f(x) g(x) dx \quad (5.1.3)$$

称为函数 f 与 g 在区间 $[a, b]$ 上关于权函数 $\omega(x)$ 的内积, 其中积分上、下限可取无穷大.

式 (5.1.2) 中的权系数 $\omega_i > 0$, 式 (5.1.3) 中的权函数 $\omega(x) \geq 0$ ($\forall x \in [a, b]$). 今后凡未给出 ω_i 或 $\omega(x)$ 时, 均默认为 $\omega_i \equiv 1$ 或者 $\omega(x) \equiv 1$.

容易验证上面定义的两内积具有以下性质:

- (1) $(f, g) = (g, f)$;
- (2) $\forall \alpha \in \mathbf{R}, (\alpha f, g) = \alpha(f, g)$;
- (3) $(f + h, g) = (f, g) + (h, g)$;
- (4) $(f, f) \geq 0, (f, f) = 0 \Leftrightarrow f \equiv 0$.

若 $f(x)$ 是列表函数, $f \neq 0$ 是指 $f(x)$ 在点 $x_1, x_2, \dots, x_m$ 上不全为零.

由内积可以定义函数 $f(x)$ 的 2 范数 $\|f(x)\|_2 = (f, f)^{1/2}$, 其表达式如下:

(1) 若 $f(x)$ 是定义在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的列表函数, 则

$$\|f(x)\|_2 = \left(\sum_{i=1}^m \omega_i f^2(x_i) \right)^{1/2}; \quad (5.1.4)$$

(2) 若 $f(x)$ 是定义在区间 $[a, b]$ 上的连续函数, 则

$$\|f(x)\|_2 = \left(\int_a^b \omega(x) f^2(x) dx \right)^{1/2}. \quad (5.1.5)$$

对于定义在区间 $[a, b]$ 上的连续函数 $f(x)$, 除了 2 范数以外, 还有以下两种常用范数:

$$\|f\|_\infty = \max_{a \leq x \leq b} |f(x)| \quad (\infty \text{ 范数}), \quad (5.1.6)$$

$$\|f\|_1 = \int_a^b |f(x)| dx \quad (1 \text{ 范数}). \quad (5.1.7)$$

可证以上定义的三种范数满足范数定义三条性质:

- (1) $\|f\| \geq 0$, $\|f\| = 0 \Leftrightarrow f \equiv 0$;
- (2) $\forall \alpha \in \mathbf{R}$, $\|\alpha f\| = |\alpha| \|f\|$;
- (3) $\|f + g\| \leq \|f\| + \|g\|$.

5.1.2 正交多项式

下面介绍在函数最优逼近和数值积分中有着重要应用的正交多项式.

1. 正交多项式的定义

定义 5.1.4 (1) 设 $f(x), g(x)$ 是定义在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的列表函数, 若内积

$$(f, g) = \sum_{i=1}^m \omega_i f(x_i) g(x_i) = 0,$$

则称 f 与 g 在点集 X 上关于权系数 ω_i 正交;

(2) 设 $f(x), g(x)$ 是定义在区间 $[a, b]$ 上的连续函数, 若内积

$$(f, g) = \int_a^b \omega(x) f(x) g(x) dx = 0,$$

则称 f 与 g 在区间 $[a, b]$ 上关于权函数 $\omega(x)$ 正交, $[a, b]$ 称为正交区间.

进而, 若函数族 $g_0(x), g_1(x), \dots, g_k(x), \dots$ 满足

$$(g_i, g_j) = \int_a^b \omega(x) g_i(x) g_j(x) dx = \begin{cases} 0, & j \neq i, \\ \gamma_i > 0, & j = i, \end{cases}$$

则称 $\{g_k(x)\}$ 为区间 $[a, b]$ 上关于权函数 $\omega(x)$ 正交的正交函数族. 若 $\gamma_i \equiv 1$, 则称为标准正交函数族. 特别地, 当 $\{g_k(x)\}$ 是 $(k \text{ 次})$ 多项式时, 简称 $g_0(x), g_1(x), \dots, g_k(x), \dots$ 为正交多项式.

2. 正交多项式的性质

性质 5.1.1 设 $g_0(x), g_1(x), \dots, g_n(x)$ (n 任意) 是正交多项式, 则它们线性无关.

证明 设存在常数 $c_0, c_1, \dots, c_n$, 使得

$$c_0 g_0(x) + c_1 g_1(x) + \dots + c_n g_n(x) = 0.$$

两边用 $g_k(x)$ 作内积得

$$0 = \left(\sum_{j=0}^n c_j g_j, g_k \right) = \sum_{j=0}^n c_j (g_j, g_k) = c_k (g_k, g_k).$$

由于 $(g_k, g_k) > 0$, 所以 $c_k = 0$ ($k = 0, 1, \dots, n$), 即 $g_0(x), g_1(x), \dots, g_n(x)$ 线性无关. 证毕.

推论 5.1.1 任意次数低于 n 的 k 次多项式 $p_k(x)$ ($k < n$) 与 n 次正交多项式 $g_n(x)$ 正交.

证明 由 $g_0(x), g_1(x), \dots, g_k(x)$ 线性无关知, $p_k(x)$ 可由 $g_0(x), g_1(x), \dots, g_k(x)$ 线性表示, 即

$$p_k(x) = c_0 g_0(x) + c_1 g_1(x) + \dots + c_k g_k(x),$$

从而

$$(p_k, g_n) = \left(\sum_{j=0}^k c_j g_j, g_n \right) = \sum_{j=0}^k c_j (g_j, g_n) = 0, \quad k < n.$$

证毕.

性质 5.1.2 在正交区间 $[a, b]$ 或 $[\min x_i, \max x_i]$ 上, n ($n \geq 1$) 次正交多项式 $g_n(x)$ 恰有 n 个不同的实零点.

证明 (1) 在正交区间上 $g_n(x)$ 必有奇重零点. 反证法. 假定方程 $g_n(x) = 0$ 无奇重根, 即 $g_n(x) = 0$ 全是偶重根, 则

$$g_n(x) = (x - \alpha_1)^{r_1} (x - \alpha_2)^{r_2} \dots (x - \alpha_j)^{r_j} q(x),$$

其中 $q(x) \neq 0$ 且不变号 (不妨设 $q(x) > 0$), $r_1, r_2, \dots, r_j$ 均是偶数且 $r_1 + r_2 + \dots + r_j \leq n$. 由 $q(x) > 0$ 知 $g_n(x) \geq 0$, 从而

$$0 = (g_0, g_n) = (1, g_n) = \sum_{i=1}^m \omega_i g_n(x_i) > 0$$

或者

$$0 = (g_0, g_n) = (1, g_n) = \int_a^b \omega(x) g_n(x) dx > 0,$$

得到矛盾. 这一矛盾说明 $g_n(x)$ 必有奇重零点.

(2) 证结论. 设奇重零点 $\alpha_1, \alpha_2, \dots, \alpha_k$ 的重数分别为 $r_1, r_2, \dots, r_k$ 且 $r_1 + r_2 + \dots + r_k \leq n$, 则

$$g_n(x) = (x - \alpha_1)^{r_1} (x - \alpha_2)^{r_2} \dots (x - \alpha_k)^{r_k} q(x),$$

其中 $q(x)$ 不变号. 设 $q(x) > 0$, 假定 $k < n$, 构造 k 次多项式

$$p_k(x) = (x - \alpha_1)(x - \alpha_2) \dots (x - \alpha_k).$$

由推论 5.1.1 知

$$0 = (p_k, g_n) = \sum_{i=1}^m \omega_i (x_i - \alpha_1)^{r_1+1} (x_i - \alpha_2)^{r_2+1} \dots (x_i - \alpha_k)^{r_k+1} q(x_i) > 0$$

或者

$$0 = (p_k, g_n) = \int_a^b \omega(x) (x - \alpha_1)^{r_1+1} (x - \alpha_2)^{r_2+1} \dots (x - \alpha_k)^{r_k+1} q(x) dx > 0,$$

得到矛盾, 这说明 $k \geq n$.

$g_n(x)$ 是 n 次多项式, 它最多有 n 个零点, 故 $k = n$. 由此可知 $r_1 = r_2 = \dots = r_n = 1$, 即 $g_n(x)$ 有 n 个不同的实零点. 证毕.

性质 5.1.3 设 $g_0(x), g_1(x), \dots, g_k(x), \dots$ 是最高次项系数为 1 的正交多项式, 则有以下三项递推关系:

$$g_{k+1}(x) = (x - b_k)g_k(x) - c_k g_{k-1}(x), \quad k = 1, 2, \dots,$$

其中

$$b_k = \frac{\beta_k}{\gamma_k}, \quad c_k = \frac{\gamma_k}{\gamma_{k-1}}, \quad \beta_k = (xg_k, g_k), \quad \gamma_k = (g_k, g_k).$$

证明 $xg_k(x)$ 是 $k+1$ 次多项式, 它可由正交多项式 $g_0(x), g_1(x), \dots, g_{k+1}(x)$ 线性表示, 即

$$xg_k(x) = b_0 g_0(x) + b_1 g_1(x) + \dots + b_k g_k(x) + b_{k+1} g_{k+1}(x).$$

比较上式两边 x^{k+1} 的系数得 $b_{k+1} = 1$. 对上式两边用 $g_j(x) (j = 0, 1, \dots, k)$ 作内积得 $(xg_k, g_j) = b_j (g_j, g_j)$, 所以

$$b_j = \frac{(xg_k, g_j)}{(g_j, g_j)}, \quad j = 0, 1, \dots, k.$$

当 $j = 0, 1, \dots, k-2$ 时, 由于 $xg_j(x)$ 是 $j+1$ ($1 \leq j+1 \leq k-1$) 次多项式, 于是

$$(xg_k, g_j) = (g_k, xg_j) = 0,$$

故 $b_j = 0$ ($j = 0, 1, \dots, k-2$).

当 $j = k-1$ 时, $xg_{k-1}(x)$ 是 k 次多项式, 于是

$$xg_{k-1}(x) = b_0^*g_0(x) + b_1^*g_1(x) + \dots + b_{k-1}^*g_{k-1}(x) + g_k(x),$$

则 $(xg_k, g_{k-1}) = (g_k, xg_{k-1}) = (g_k, g_k) = \gamma_k$, 故

$$b_{k-1} = \frac{(xg_k, g_{k-1})}{(g_{k-1}, g_{k-1})} = \frac{\gamma_k}{\gamma_{k-1}}.$$

当 $j = k$ 时,

$$b_k = \frac{(xg_k, g_k)}{(g_k, g_k)} = \frac{\beta_k}{\gamma_k}.$$

由以上可得

$$xg_k(x) = b_{k-1}g_{k-1}(x) + b_k g_k(x) + g_{k+1}(x) = \frac{\gamma_k}{\gamma_{k-1}}g_{k-1}(x) + \frac{\beta_k}{\gamma_k}g_k(x) + g_{k+1}(x),$$

从而

$$g_{k+1}(x) = \left(x - \frac{\beta_k}{\gamma_k}\right)g_k(x) - \frac{\gamma_k}{\gamma_{k-1}}g_{k-1}(x).$$

证毕.

下面求 $g_1(x)$. 由 $xg_0(x) = b_0g_0(x) + g_1(x)$ 有 $(g_0, xg_0) = b_0(g_0, g_0)$. 由此可得

$$b_0 = \frac{(g_0, xg_0)}{(g_0, g_0)} = \frac{\beta_0}{\gamma_0},$$

所以 $g_1(x) = (x - b_0)g_0(x) = x - \beta_0/\gamma_0$ (注意到 $g_0(x) \equiv 1$). 于是可由以下三项递推关系逐次构造正交多项式序列 $\{g_k(x)\}$:

$$\begin{cases} g_0(x) = 1, \\ g_1(x) = x - \beta_0/\gamma_0, \\ g_{k+1}(x) = (x - b_k)g_k(x) - c_k g_{k-1}(x), \quad k = 1, 2, \dots \end{cases}$$

3. 常用的几个正交多项式

(1) 三角函数系

$$1, \cos x, \sin x, \cos(2x), \sin(2x), \dots, \cos(kx), \sin(kx), \dots$$

是区间 $[-\pi, \pi]$ 上的正交函数系, 并且

$$(1, 1) = \int_{-\pi}^{\pi} dx = 2\pi,$$

$$(\sin(kx), \sin(kx)) = \int_{-\pi}^{\pi} \sin^2(kx) dx = \pi,$$

$$(\cos(kx), \cos(kx)) = \int_{-\pi}^{\pi} \cos^2(kx) dx = \pi.$$

(2) 勒让德 (Legendre) 多项式

$$p_k(x) = \frac{1}{2^k k!} \frac{d^k(x^2 - 1)^k}{dx^k}, \quad k = 0, 1, 2, \dots$$

是在区间 $[-1, 1]$ 上关于 $\omega(x) \equiv 1$ 正交的正交多项式.

前几个勒让德多项式为

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = \frac{1}{2}(3x^2 - 1), \quad p_3(x) = \frac{1}{2}(5x^3 - 3x),$$

$$p_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3), \quad p_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x), \quad \dots,$$

$$(p_k, p_k) = \int_{-1}^1 p_k^2(x) dx = \frac{2}{2k+1}.$$

三项递推关系为

$$(k+1)p_{k+1}(x) = (2k+1)x p_k(x) - k p_{k-1}(x), \quad k = 1, 2, \dots$$

$p_k(x)$ 的最高次项 x^k 的系数为

$$a_k = \frac{(2k)!}{2^k (k!)^2}.$$

(3) 拉盖尔 (Laguerre) 多项式

$$L_k(x) = e^x \frac{d^k(x^k e^{-x})}{dx^k}, \quad k = 0, 1, 2, \dots$$

在区间 $[0, +\infty)$ 上关于权函数 $\omega(x) = e^{-x}$ 正交.

前几个拉盖尔多项式为

$$L_0(x) = 1, \quad L_1(x) = -x + 1, \quad L_2(x) = x^2 - 4x + 2,$$

$$L_3(x) = -x^3 + 9x^2 - 18x + 6, \quad L_4(x) = x^4 - 16x^3 + 72x^2 - 96x + 24,$$

$$L_5(x) = -x^5 + 25x^4 - 200x^3 + 600x^2 - 600x + 120, \quad \dots,$$

$$(L_k, L_k) = \int_0^{+\infty} e^{-x} L_k^2(x) dx = (k!)^2.$$

三项递推关系为

$$L_{k+1}(x) = (1 + 2k - x)L_k(x) - k^2 L_{k-1}(x), \quad k = 1, 2, \dots$$

$L_k(x)$ 的最高次项 x^k 的系数为 $a_k = (-1)^k$.

(4) 埃尔米特 (Hermite) 多项式

$$H_k(x) = (-1)^k e^{x^2} \frac{d^k e^{-x^2}}{dx^k}, \quad k = 0, 1, 2, \dots$$

在区间 $(-\infty, +\infty)$ 上关于权函数 $\omega(x) = e^{-x^2}$ 正交.

前几个埃尔米特多项式为

$$\begin{aligned} H_0(x) &= 1, & H_1(x) &= 2x, & H_2(x) &= 4x^2 - 2, \\ H_3(x) &= 8x^3 - 12x, & H_4(x) &= 16x^4 - 48x^2 + 12, \\ H_5(x) &= 32x^5 - 160x^3 + 120x, & \dots, \\ (H_k, H_k) &= \int_{-\infty}^{\infty} e^{-x^2} H_k^2(x) dx = 2^k (k!) \sqrt{\pi}. \end{aligned}$$

三项递推关系为

$$H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x), \quad k = 1, 2, \dots$$

$H_k(x)$ 的最高次项 x^k 的系数为 $a_k = 2^k$.

(5) 切比雪夫 (Chebyshev) 多项式

$$T_k(x) = \cos(k \arccos x), \quad -1 \leq x \leq 1; \quad k = 0, 1, 2, \dots$$

显然, $T_0(x) = 1, T_1(x) = x$. 切比雪夫多项式具有以下性质:

性质 5.1.4 切比雪夫多项式在区间 $[-1, 1]$ 关于权函数 $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ 正交.

证明

$$(T_k, T_j) = \int_{-1}^1 \frac{T_k(x)T_j(x)}{\sqrt{1-x^2}} dx \stackrel{x=\cos\theta}{=} \int_0^\pi \cos(k\theta) \cos(j\theta) d\theta = \begin{cases} 0, & k \neq j, \\ \pi, & k = j = 0, \\ \frac{\pi}{2}, & k = j \neq 0. \end{cases}$$

证毕.

性质 5.1.5 三项递推关系

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \quad k = 1, 2, \dots$$

证明 $\theta \triangleq \arccos x$, 则 $T_k(x) = \cos(k\theta)$, 所以

$$T_{k+1}(x) + T_{k-1}(x) = \cos[(k+1)\theta] + \cos[(k-1)\theta] = 2\cos\theta \cos(k\theta) = 2xT_k(x).$$

证毕.

由三项递推关系可得前几个切比雪夫多项式为

$$\begin{aligned} T_0(x) &= 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x, \quad \dots \end{aligned}$$

由 $T_0(x) = 1, T_1(x) = x$ 及三项递推关系可得性质 5.1.6.

性质 5.1.6 $T_k(x)$ 是最高次项为 $2^{k-1}x^k$ 的 k 次多项式, 并且 $T_{2k}(x)$ 只含 x 的偶次幂, $T_{2k+1}(x)$ 只含 x 的奇次幂.

由切比雪夫多项式的定义可得下面两个性质.

性质 5.1.7 $T_k(x)$ 在区间 $[-1, 1]$ 上有 k 个零点

$$x_i = \cos\left(\frac{2i+1}{2k}\pi\right), \quad i = 0, 1, \dots, k-1.$$

性质 5.1.8 在区间 $[-1, 1]$ 上, $|T_k(x)| \leq 1$, 在 $k+1$ 个极值点 $x_i = \cos\left(\frac{i\pi}{k}\right)$ ($i = 0, 1, \dots, k$) 处, $T_k(x)$ 依次交替地取最大值 1 和最小值 -1 .

性质 5.1.9 (切比雪夫多项式的极性) 设 $p_n(x)$ 是最高次项系数为 1 的 n 次多项式, 则

$$\max_{-1 \leq x \leq 1} |p_n(x)| \geq \max_{-1 \leq x \leq 1} |2^{1-n}T_n(x)| = 2^{1-n}.$$

证明 反证法. 假定存在最高次项系数为 1 的 n 次多项式 $\bar{p}_n(x)$, 使得

$$\max_{-1 \leq x \leq 1} |\bar{p}_n(x)| < \max_{-1 \leq x \leq 1} |2^{1-n}T_n(x)| = 2^{1-n}.$$

令 $E(x) = 2^{1-n}T_n(x) - \bar{p}_n(x)$. 显然, $E(x)$ 是不超过 $n-1$ 次的多项式. 在 $T_n(x)$ 的 $n+1$ 个极值点 x_i ($i = 0, 1, \dots, n$) 处,

$$E(x_0) = 2^{1-n} - \bar{p}_n(x_0) > 0,$$

$$E(x_1) = -2^{1-n} - \bar{p}_n(x_1) < 0,$$

$$E(x_2) = 2^{1-n} - \bar{p}_n(x_2) > 0,$$

.....

$$E(x_n) = (-1)^n 2^{1-n} - \bar{p}_n(x_n) \begin{cases} > 0, & n \text{ 是偶数,} \\ < 0, & n \text{ 是奇数.} \end{cases}$$

据此, 由连续函数的零点存在定理知, 方程 $E(x) = 0$ 在区间 $[-1, 1]$ 上有 n 个根, 而 $E(x)$ 是不超过 $n-1$ 次的多项式, $E(x) = 0$ 最多有 $n-1$ 个根, 得到矛盾. 这说明假定不成立. 证毕.

5.2 最优平方逼近

本节讨论 $f(x)$ 是列表函数或定义在区间 $[a, b]$ 上连续函数的最优平方逼近问题.

5.2.1 最优平方逼近

设 $f(x)$ 是定义在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的列表函数或者 $f(x)$ 是定义在区间 $[a, b]$ 上表达式复杂的连续函数. 构造广义多项式

$$p(x) = c_0\phi_0(x) + c_1\phi_1(x) + \dots + c_n\phi_n(x), \quad (5.2.1)$$

使得

$$\|p - f\|_2^2 \quad (5.2.2)$$

达到最小, 其中 $c_0, c_1, \dots, c_n$ 为待定参数, $\phi_i(x)$ ($i = 0, 1, \dots, n$) 为已知的一组线性无关的函数 (称为基函数, 它的选取与具体问题有关). 取 $p(x)$ 作为 $f(x)$ 的近似表达式就是最优平方逼近问题.

下面针对函数 $f(x)$ 是离散和连续两种情况分别给出最优平方逼近问题的具体形式.

(1) 当 $f(x)$ 是定义在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的列表函数时, 设在点 x_i 处的函数值为 y_i , 式 (5.2.2) 即

$$\|p - f\|_2^2 = (p - f, p - f) = \sum_{i=1}^m \omega_i (p(x_i) - y_i)^2. \quad (5.2.3)$$

这时, 所求得的 $p(x)$ 称为**最小二乘拟合函数**. 若取 $\phi_i(x) = x^i$ ($i = 0, 1, \dots, n$), 则

$$p(x) = p_n(x) = c_0 + c_1x + c_2x^2 + \dots + c_nx^n$$

称为**最小二乘拟合多项式**. 取 $p(x)$ 作为函数 $f(x)$ 的近似表达式的误差为

$$\|p - f\|_2 = \sqrt{\sum_{i=1}^m \omega_i (p(x_i) - y_i)^2}. \quad (5.2.4)$$

(2) 当 $f(x)$ 是定义在区间 $[a, b]$ 上的连续函数时, 式 (5.2.2) 即

$$\|p - f\|_2^2 = (p - f, p - f) = \int_a^b \omega(x) (p(x) - f(x))^2 dx. \quad (5.2.5)$$

这时, 所求得的 $p(x)$ 称为 $f(x)$ 在区间 $[a, b]$ 上关于权函数 $\omega(x)$ 的**最优平方逼近函数**, 其误差为

$$\|p - f\|_2 = \left(\int_a^b \omega(x) (p(x) - f(x))^2 dx \right)^{1/2} \quad (5.2.6)$$

5.2.2 正规方程组

本小节导出最小二乘拟合函数或最优平方逼近函数 $p(x)$ 中的系数 $c_0, c_1, \dots, c_n$ 所满足的线性方程组.

注意到 $\|p - f\|_2^2$ 是参数 $c_0, c_1, \dots, c_n$ 的函数, 将其记为 $S(c_0, c_1, \dots, c_n)$. 求最小二乘拟合函数或最优平方逼近函数 $p(x)$ 的问题, 即求解以下无条件极值问题:

$$\min S(c_0, c_1, \dots, c_n) = \|p - f\|_2^2. \quad (5.2.7)$$

下面讨论问题 (5.2.7) 的求解. 因为

$$\begin{aligned} S &= \|p - f\|_2^2 = (p - f, p - f) = (p, p) - 2(p, f) + (f, f) \\ &= \left(\sum_{i=0}^n c_i \phi_i, \sum_{j=0}^n c_j \phi_j \right) - 2 \left(\sum_{i=0}^n c_i \phi_i, f \right) + (f, f) \\ &= \sum_{i=0}^n c_i \sum_{j=0}^n c_j (\phi_i, \phi_j) - 2 \sum_{i=0}^n c_i (\phi_i, f) + (f, f). \end{aligned}$$

由多元函数求极值的条件, 令

$$\frac{\partial S}{\partial c_k} = 0, \quad k = 0, 1, \dots, n, \quad (5.2.8)$$

而

$$\begin{aligned} \frac{\partial S}{\partial c_k} &= \sum_{j=0}^n c_j (\phi_k, \phi_j) + \sum_{i=0}^n c_i (\phi_i, \phi_k) - 2(\phi_k, f) \\ &= 2 \sum_{j=0}^n c_j (\phi_k, \phi_j) - 2(\phi_k, f). \end{aligned}$$

由上式和式 (5.2.8) 得到关于参数 $c_0, c_1, \dots, c_n$ 的线性方程组

$$\sum_{j=0}^n (\phi_k, \phi_j) c_j = (\phi_k, f), \quad k = 0, 1, \dots, n. \quad (5.2.9)$$

方程组 (5.2.9) 称为最优平方逼近问题的正规方程组或法方程组.

从正规方程组求得 $c_0, c_1, \dots, c_n$ 后, 代入式 (5.2.1), 即得最小二乘拟合函数或最优平方逼近函数 $p(x)$.

正规方程组可以改写为

$$\left(\phi_k, \sum_{j=0}^n c_j \phi_j - f \right) = 0, \quad k = 0, 1, \dots, n,$$

也即

$$(\phi_k, p - f) = 0, \quad k = 0, 1, \dots, n.$$

这说明 $p(x) - f(x)$ 与所有基函数 $\phi_k(x)$ ($k = 0, 1, \dots, n$) 正交.

正规方程组 (5.2.9) 的矩阵形式为

$$\begin{pmatrix} (\phi_0, \phi_0) & (\phi_0, \phi_1) & \cdots & (\phi_0, \phi_n) \\ (\phi_1, \phi_0) & (\phi_1, \phi_1) & \cdots & (\phi_1, \phi_n) \\ \vdots & \vdots & & \vdots \\ (\phi_n, \phi_0) & (\phi_n, \phi_1) & \cdots & (\phi_n, \phi_n) \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} (\phi_0, f) \\ (\phi_1, f) \\ \vdots \\ (\phi_n, f) \end{pmatrix}. \quad (5.2.9')$$

方程组 (5.2.9') 的系数矩阵是对称正定矩阵.

事实上, 若将方程组 (5.2.9') 的系数矩阵记为 A , A 的元素记为 a_{ij} . 由 $a_{ij} = (\phi_i, \phi_j) = (\phi_j, \phi_i) = a_{ji}$ 知, A 是对称矩阵. 以下证明 A 的正定性. $\forall c \in \mathbf{R}^{n+1}$, $c = (c_0, c_1, \dots, c_n)^T \neq 0$. 由 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$ 的线性无关知

$$u(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) \neq 0, \quad \forall x \in \mathbf{R}.$$

故二次型

$$\begin{aligned} c^T A c &= \sum_{i=0}^n \sum_{j=0}^n a_{ij} c_i c_j = \sum_{i=0}^n \sum_{j=0}^n (\phi_i, \phi_j) c_i c_j \\ &= \left(\sum_{i=0}^n c_i \phi_i, \sum_{j=0}^n c_j \phi_j \right) = (u, u) = \|u\|_2^2 > 0. \end{aligned}$$

这说明 A 是对称正定矩阵, 所以正规方程组的解存在且唯一, 即最小二乘拟合函数或最优平方逼近函数 $p(x)$ 存在且唯一. 正规方程组可用改进平方根法或共轭梯度法求解.

通常 n 较大时, 正规方程组是病态方程组, 一般取 $n \leq 6$. 若取 $\phi_0, \phi_1, \dots, \phi_n$ 是一组正交函数 $g_0, g_1, \dots, g_n$, 则正规方程组 (5.2.9) 简化为

$$(g_k, g_k) c_k = (g_k, f), \quad k = 0, 1, \dots, n, \quad (5.2.10)$$

从而 $c_k = (g_k, f) / (g_k, g_k)$ ($k = 0, 1, \dots, n$). 将 c_k ($k = 0, 1, \dots, n$) 代入式 (5.2.1), 即得 $f(x)$ 的最小二乘拟合函数或最优平方逼近函数

$$p(x) = \sum_{k=0}^n \frac{(g_k, f)}{(g_k, g_k)} g_k(x). \quad (5.2.11)$$

下面给出当 $f(x)$ 是定义在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的列表函数 (设 $f(x)$ 在点 x_i 处的函数值为 y_i) 时正规方程组 (5.2.9') 的另一种表示方式. 这时, 内积为

$$(\phi_k, \phi_j) = \sum_{i=1}^m \omega_i \phi_k(x_i) \phi_j(x_i), \quad k, j = 0, 1, \dots, n,$$

$$(\phi_k, f) = \sum_{i=1}^m \omega_i \phi_k(x_i) f(x_i) = \sum_{i=1}^m \omega_i \phi_k(x_i) y_i, \quad k = 0, 1, \dots, n.$$

于是正规方程组 (5.2.9') 可以写成

$$G^T W G c = G^T W y, \quad (5.2.12)$$

其中

$$G = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_n(x_2) \\ \vdots & \vdots & & \vdots \\ \phi_0(x_m) & \phi_1(x_m) & \cdots & \phi_n(x_m) \end{pmatrix},$$

$$W = \text{diag}(\omega_1, \omega_2, \dots, \omega_m), \quad y = (y_1, y_2, \dots, y_m)^T.$$

特别地, 当权系数 $\omega_i \equiv 1$ 时, 方程组 (5.2.12) 为

$$G^T G c = G^T y. \quad (5.2.13)$$

这时, 若取 $\phi_i(x) = x^i$ ($i = 0, 1, \dots, n$), 则矩阵

$$G = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ 1 & x_3 & x_3^2 & \cdots & x_3^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{pmatrix}. \quad (5.2.14)$$

再由内积 $(\phi_k, \phi_j) = \sum_{i=1}^m \phi_k(x_i) \phi_j(x_i) = \sum_{i=1}^m x_i^{k+j}$ ($k, j = 0, 1, \dots, n$) 知, 正规方程组 (5.2.9') 又可以表示为

$$\begin{pmatrix} \sum_{i=1}^m 1 & \sum_{i=1}^m x_i & \cdots & \sum_{i=1}^m x_i^n \\ \sum_{i=1}^m x_i & \sum_{i=1}^m x_i^2 & \cdots & \sum_{i=1}^m x_i^{n+1} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^m x_i^n & \sum_{i=1}^m x_i^{n+1} & \cdots & \sum_{i=1}^m x_i^{2n} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m y_i \\ \sum_{i=1}^m x_i y_i \\ \vdots \\ \sum_{i=1}^m x_i^n y_i \end{pmatrix}. \quad (5.2.15)$$

下面讨论最小二乘问题与插值问题的关系. 取 $\phi_i(x) = x^i$ ($i = 0, 1, \dots, n$) 当 $m = n + 1$ 时, 由式 (5.2.14) 表示的矩阵 G 是范德蒙德矩阵, 故 G 非奇异. 由方程组 (5.2.13) 有

$$Gc = y.$$

将上式写成分量形式, 即

$$c_0 + c_1 x_i + c_2 x_i^2 + \dots + c_n x_i^n = y_i, \quad i = 1, 2, \dots, n + 1,$$

也即 $p_n(x_i) = y_i$ ($i = 1, 2, \dots, n + 1$). 由此可知, 所求得最小二乘拟合多项式就是插值多项式. 这时,

$$\|p - f\|_2^2 = 0.$$

在求最小二乘拟合函数时, 函数类 (或基函数) 的选取十分重要. 同组数据在不同的函数类中获得的拟合函数是不同的, 函数类的选择直接影响拟合的效果. 选取函数类通常有以下三个原则:

(1) 直观性原则. 将数据点 (x_i, y_i) ($i = 1, 2, \dots, m$) 描绘在坐标纸上, 根据点的分布情况与已知各类函数曲线进行比较, 从中选出一个或几个函数类.

(2) 比较性原则. 在几个不同的函数类中, 分别求出拟合函数的均方差

$$\|p - f\|_2 = \sqrt{\sum_{i=1}^m \omega_i (p(x_i) - y_i)^2},$$

比较其优劣, 决定取舍.

(3) 根据实际问题的背景选择函数类. 有时给定的数据具有指数特征或者幂函数特征, 则取指数函数

$$y = be^{ax} \quad (5.2.16)$$

或幂函数

$$y = bx^a \quad (5.2.17)$$

作为最小二乘拟合函数较为合适. 此时, 对参数进行变换, 对函数 (5.2.16), (5.2.17) 分别取对数得

$$\ln y = \ln b + ax \quad (5.2.18)$$

或

$$\ln y = \ln b + a \ln x. \quad (5.2.19)$$

例如, 对指数型 $y = be^{ax}$ 的拟合函数, 通过求

$$S(a, b) = \sum_{i=1}^m (\ln b + ax_i - \ln y_i)^2$$

的极小, 就可以确定 $a, \ln b$, 最终确定参数 a, b . 对幂函数情形方法类似.

一般地, 求函数 $f(x)$ 的最小二乘拟合函数或最优平方逼近函数有以下三种方法:

(1) 根据问题的特点选择一组线性无关的基函数 $\phi_0(x), \phi_1(x), \dots, \phi_n(x)$, 通过解方程组 (5.2.9) 或 (5.2.9'), 求得最小二乘拟合函数或最优平方逼近函数.

(2) 利用三项递推关系构造正交多项式 $g_0(x), g_1(x), \dots, g_k(x), \dots$. 由式 (5.2.11) 得到最小二乘拟合多项式或最优平方逼近函数.

(3) 作变量替换, 利用已知的正交多项式作为基函数构造最优平方逼近函数. 例如, 设逼近区间为 $[a, b]$, 取勒让德正交多项式作为基函数, 则先利用变换

$$x = \frac{a+b}{2} + \frac{b-a}{2}t, \quad -1 \leq t \leq 1.$$

将 $x \in [a, b]$ 变为 $t \in [-1, 1]$, 这时, $f(x)$ 变为

$$\bar{f}(t) = f(x) = f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right).$$

此时, 原问题转化为求 $\bar{f}(t)$ 在区间 $[-1, 1]$ 上的最优平方逼近函数 $\bar{p}(t)$ 的问题. 求得 $\bar{p}(t)$ 后, 再利用其逆变换

$$t = \frac{2x - a - b}{b - a}$$

将 $\bar{p}(t)$ 变为 $p(x)$, 即得到 $f(x)$ 在区间 $[a, b]$ 上的最优平方逼近函数.

例 5.2.1 给定数据如表 5.1 所示, 求最小二乘拟合二次多项式.

表 5.1

x_i	0	0.25	0.50	0.75	1.00
y_i	1.000 0	1.284 0	1.648 7	2.117 0	2.718 3

解法 1 设所求的最小二乘拟合多项式 $p_2(x) = c_0 + c_1x + c_2x^2$, 则根据法方程组 (5.2.15) 有

$$\begin{pmatrix} 5 & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.3828 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 8.7680 \\ 5.4514 \\ 4.4015 \end{pmatrix}.$$

从该方程组解得 $c_0 = 1.0051$, $c_1 = 0.8647$, $c_2 = 0.8432$. 所以最小二乘二次拟合多项式为

$$p_2(x) = 1.0051 + 0.8647x + 0.8432x^2.$$

解法 2 利用三项递推关系构造正交多项式. 取

$$g_0(x) = 1, \quad g_1(x) = x - \frac{\beta_0}{\gamma_0}, \quad g_2(x) = (x - b_1)g_1(x) - c_1g_0(x),$$

而

$$\gamma_0 = (g_0, g_0) = (1, 1) = \sum_{i=1}^5 1 = 5,$$

$$\beta_0 = (xg_0, g_0) = (x, 1) = \sum_{i=1}^5 x_i = 2.5, \quad \frac{\beta_0}{\gamma_0} = \frac{1}{2},$$

故 $g_1(x) = x - \frac{1}{2}$. 又

$$\gamma_1 = (g_1, g_1) = \sum_{i=1}^5 \left(x_i - \frac{1}{2}\right)^2 = \frac{5}{8}, \quad \beta_1 = (xg_1, g_1) = \sum_{i=1}^5 x_i \left(x_i - \frac{1}{2}\right)^2 = \frac{5}{16},$$

$$b_1 = \frac{\beta_1}{\gamma_1} = \frac{1}{2}, \quad c_1 = \frac{\gamma_1}{\gamma_0} = \frac{1}{8},$$

所以

$$g_2(x) = \left(x - \frac{1}{2}\right)^2 - \frac{1}{8},$$

而

$$(g_2, g_2) = \sum_{i=1}^5 \left(\left(x_i - \frac{1}{2}\right)^2 - \frac{1}{8}\right)^2 = \frac{7}{128},$$

$$(g_0, f) = \sum_{i=1}^5 y_i = 8.7680,$$

$$(g_1, f) = \sum_{i=1}^5 \left(x_i - \frac{1}{2}\right) y_i = 1.0674,$$

$$(g_2, f) = \sum_{i=1}^5 \left(\left(x_i - \frac{1}{2}\right)^2 - \frac{1}{8}\right) y_i = 0.0461375.$$

据式 (5.2.11),

$$\begin{aligned} p_2(x) &= c_0g_0(x) + c_1g_1(x) + c_2g_2(x) \\ &= \frac{(g_0, f)}{(g_0, g_0)}g_0(x) + \frac{(g_1, f)}{(g_1, g_1)}g_1(x) + \frac{(g_2, f)}{(g_2, g_2)}g_2(x) \\ &= 1.7536 + 1.70784 \left(x - \frac{1}{2}\right) + 0.843657142 \left(\left(x - \frac{1}{2}\right)^2 - \frac{1}{8}\right) \\ &\approx 1.005137143 + 0.864182858x + 0.843657142x^2. \end{aligned}$$

解毕.

例 5.2.2 给定数据如表 5.2 所示, 求形如 $p(x) = c_0 + c_1 e^x + c_2 e^{-x}$ 的最小二乘拟合函数.

表 5.2

x_i	0.0	0.1	0.2	0.3	0.4	0.5	0.6
y_i	2.000 00	2.202 54	2.407 15	2.615 92	2.830 96	3.054 48	3.288 76

解 取 $\phi_0(x) = 1$, $\phi_1(x) = e^x$, $\phi_2(x) = e^{-x}$, 则正规方程组 (5.2.9') 为

$$\begin{pmatrix} 7 & 9.639\ 10 & 5.290\ 05 \\ 9.639\ 10 & 13.799\ 29 & 7 \\ 5.290\ 05 & 7 & 4.156\ 27 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 18.399\ 81 \\ 26.157\ 20 \\ 13.456\ 86 \end{pmatrix}.$$

解之得

$$c_0 = 1.986\ 14, \quad c_1 = 1.017\ 57, \quad c_2 = -1.002\ 06,$$

所求的最小二乘拟合函数为

$$p(x) = 1.986\ 14 + 1.017\ 57e^x - 1.002\ 06e^{-x}.$$

解毕.

例 5.2.3 某化学反应中, 测得生成物的浓度 y 与时间 t (min) 的数据如表 5.3 所示, 求 y 与 t 的近似函数关系.

表 5.3

t_i	1	2	3	4	5	6	7	8
y_i	4.00	6.40	8.00	8.80	9.22	9.50	9.70	9.86
t_i	9	10	11	12	13	14	15	16
y_i	10.00	10.20	10.32	10.42	10.50	10.55	10.58	10.60

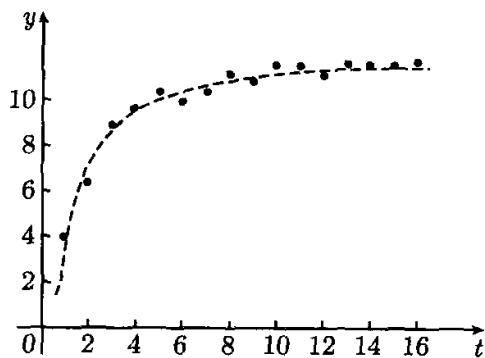


图 5.1

解法 1 将观测数据 (t_i, y_i) ($i = 1, 2, \dots, 16$) 描在坐标纸上, 如图 5.1 所示.

由图 5.1 所见, 反应初期浓度增加很快, 随后趋于平稳. 鉴于有理函数具有这种特征, 故取

$$y \approx \frac{at + b}{ct + d}.$$

由实际问题知, 反应开始时, $t = 0$, 浓度 $y = 0$, 可知 $b = 0$. 将分子分母同除以 a 后得 $y \approx \frac{t}{c_0 t + c_1}$, 或者取 $z = \frac{1}{y} \approx c_0 + c_1 t^{-1}$.

取 $\phi_0(t) = 1, \phi_1(t) = t^{-1}$, 则正规方程组 (5.2.9') 为

$$\begin{pmatrix} 16 & \sum_{i=1}^{16} t_i^{-1} \\ \sum_{i=1}^{16} t_i^{-1} & \sum_{i=1}^{16} t_i^{-2} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{16} y_i^{-1} \\ \sum_{i=1}^{16} t_i^{-1} y_i^{-1} \end{pmatrix},$$

即

$$\begin{cases} 16c_0 + 3.380\,729c_1 = 1.832\,912, \\ 3.380\,729c_0 + 1.584\,347c_1 = 0.528\,857. \end{cases}$$

解之得 $c_0 = 0.080\,174, c_1 = 0.162\,723$, 所以

$$y \approx \frac{t}{0.080\,174t + 0.162\,723}.$$

解法 2 由于曲线 $y = ae^{b/t}$ 也有图 5.1 的特征, 故取 $y \approx ae^{b/t}$ 或

$$z = \ln y \approx \ln a + bt^{-1} = c_0 + c_1 t^{-1},$$

其中 $c_0 = \ln a, c_1 = b$. 取 $\phi_0(t) = 1, \phi_1(t) = t^{-1}$, 于是可得正规方程组

$$\begin{pmatrix} 16 & \sum_{i=1}^{16} t_i^{-1} \\ \sum_{i=1}^{16} t_i^{-1} & \sum_{i=1}^{16} t_i^{-2} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{16} \ln y_i \\ \sum_{i=1}^{16} t_i^{-1} \ln y_i \end{pmatrix},$$

即

$$\begin{cases} 16c_0 + 3.380\,729c_1 = 35.260\,169, \\ 3.380\,729c_0 + 1.584\,347c_1 = 6.530\,988. \end{cases}$$

解之得 $c_0 = 2.427\,033, c_1 = -1.056\,684$, 故

$$a = e^{c_0} \approx 11.325\,232, \quad b = c_1,$$

从而

$$y \approx 11.325\,232e^{-1.056\,684/t}.$$

解毕.

下面比较两种解法得到的拟和函数的均方差.

对于解法 1,

$$\|p - f\|_2 = \sqrt{\sum_{i=1}^{16} (p(x_i) - y_i)^2} \approx 1.249\,000 \times 10^{-3},$$

而对于解法 2,

$$\|p - f\|_2 = \sqrt{\sum_{i=1}^{16} (p(x_i) - y_i)^2} \approx 0.340588 \times 10^{-3}.$$

由此可见, 解法 2 所得拟合函数近似程度较高.

例 5.2.4 求函数 $f(x) = e^x$ 在区间 $[0, 1]$ 上的最优逼近二次多项式.

解法 1 设 $p_2(x) = c_0 + c_1x + c_2x^2$, 则 $\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x^2$.

$$(\phi_k, \phi_j) = \int_0^1 x^{k+j} dx = \frac{1}{k+j+1},$$

$$(\phi_0, f) = \int_0^1 e^x dx = e - 1, \quad (\phi_1, f) = \int_0^1 xe^x dx = 1, \quad (\phi_2, f) = \int_0^1 x^2 e^x dx = e - 2.$$

由正规方程组 (5.2.9') 有

$$\begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} e - 1 \\ 1 \\ e - 2 \end{pmatrix}.$$

解之得

$$c_0 \approx 1.01299, \quad c_1 \approx 0.85114, \quad c_2 \approx 0.83917,$$

从而

$$p_2(x) = 1.01299 + 0.85114x + 0.83917x^2.$$

解法 2 利用三项递推关系构造正交多项式. 取

$$g_0(x) = 1, \quad g_1(x) = x - \frac{\beta_0}{\gamma_0}, \quad g_2(x) = (x - b_1)g_1(x) - c_1g_0(x),$$

而

$$\beta_0 = (xg_0, g_0) = (x, 1) = \int_0^1 x dx = \frac{1}{2},$$

$$\gamma_0 = (g_0, g_0) = (1, 1) = \int_0^1 dx = 1, \quad \frac{\beta_0}{\gamma_0} = \frac{1}{2},$$

故 $g_1(x) = x - \frac{1}{2}$. 又

$$\beta_1 = (xg_1, g_1) = \int_0^1 x \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{24},$$

$$\gamma_1 = (g_1, g_1) = \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{12},$$

$$b_1 = \frac{\beta_1}{\gamma_1} = \frac{1}{2}, \quad c_1 = \frac{\gamma_1}{\gamma_0} = \frac{1}{12},$$

所以 $g_2(x) = \left(x - \frac{1}{2}\right)^2 - \frac{1}{12}$. 而

$$\begin{aligned}(g_2, g_2) &= \int_0^1 \left(\left(x - \frac{1}{2}\right)^2 - \frac{1}{12} \right)^2 dx = \frac{1}{180}, \\(g_0, f) &= \int_0^1 e^x dx = e - 1, \\(g_1, f) &= \int_0^1 \left(x - \frac{1}{2}\right) e^x dx = \frac{1}{2}(3 - e), \\(g_2, f) &= \int_0^1 \left(\left(x - \frac{1}{2}\right)^2 - \frac{1}{12} \right) e^x dx = \frac{1}{6}(7e - 19),\end{aligned}$$

从而得

$$\frac{(g_0, f)}{(g_0, g_0)} = e - 1, \quad \frac{(g_1, f)}{(g_1, g_1)} = 18 - 6e, \quad \frac{(g_2, f)}{(g_2, g_2)} = 30(7e - 19).$$

由式 (5.2.11) 得最优平方逼近函数为

$$\begin{aligned}p(t) &= \sum_{k=0}^2 \frac{(g_k, f)}{(g_k, g_k)} g_k(x) \\&= e - 1 + (18 - 6e) \left(x - \frac{1}{2}\right) + 30(7e - 19) \left[\left(x - \frac{1}{2}\right)^2 - \frac{1}{12} \right] \\&\approx 1.012\,991\,31 + 0.851\,125\,07x + 0.839\,183\,97x^2.\end{aligned}$$

解毕.

解法 3 取勒让德正交多项式作为基函数, 先作变换 $x = \frac{t+1}{2} (-1 \leq t \leq 1)$, 则 $f(t) = e^{\frac{t+1}{2}}$. 前三个勒让德正交多项式为

$$p_0(t) = 1, \quad p_1(t) = t, \quad p_2(t) = \frac{1}{2}(3t^2 - 1),$$

则 $f(t)$ 的最优平方逼近函数为

$$p(x) = c_0 p_0(t) + c_1 p_1(t) + c_2 p_2(t).$$

对于勒让德正交多项式 $(p_k, p_k) = 2/(2k+1)$ ($k = 0, 1, 2$), 而

$$\begin{aligned}(p_0, f) &= \int_{-1}^1 e^{\frac{t+1}{2}} dt = 2(e - 1), \quad (p_1, f) = \int_{-1}^1 t e^{\frac{t+1}{2}} dt = 6 - 2e, \\(p_2, f) &= \frac{1}{2} \int_{-1}^1 (3t^2 - 1) e^{\frac{t+1}{2}} dt = 14e - 38,\end{aligned}$$

从而得

$$c_0 = \frac{(p_0, f)}{(p_0, p_0)} \approx 1.718\,282, \quad c_1 = \frac{(p_1, f)}{(p_1, p_1)} \approx 0.845\,155, \quad c_2 = \frac{(p_2, f)}{(p_2, p_2)} \approx 0.139\,864,$$

故

$$p(t) = 1.718\,282 + 0.845\,155t + 0.069\,932(3t^2 - 1).$$

再将 $t = 2x - 1$ 代入上式, 则得 $f(x)$ 的最优平方逼近函数

$$p(x) = 1.012\,991 + 0.851\,126x + 0.839\,184x^2.$$

5.3 最优一致逼近

设 $f(x)$ 是定义在区间 $[a, b]$ 上的连续函数, 构造 n 次多项式 $p_n(x)$ 作为 $f(x)$ 的近似表达式, 要求 $p_n(x)$ 在区间 $[a, b]$ 上的每一点算出的近似值都比较准确, 也即要求最大误差的绝对值达到最小, 构造这种 n 次多项式 $p_n(x)$ 的问题称为最优一致逼近问题.

5.3.1 最优一致逼近多项式

设 $f(x)$ 是定义在区间 $[a, b]$ 上的连续函数, 求 n 次多项式

$$p_n(x) = c_0 + c_1x + c_2x^2 + \cdots + c_nx^n,$$

使得

$$\|f - p_n\|_\infty = \max_{a \leq x \leq b} |f(x) - p_n(x)|$$

达到最小, 其中 $c_0, c_1, \dots, c_n$ 为待定参数, 则 $p_n(x)$ 称为 $f(x)$ 在区间 $[a, b]$ 上的 n 次最优一致逼近多项式.

换句话说, 若 n 次多项式 $p_n^*(x) = c_0^* + c_1^*x + c_2^*x^2 + \cdots + c_n^*x^n$ 为 $f(x)$ 在区间 $[a, b]$ 上的最优一致逼近多项式, 则

$$\max_{a \leq x \leq b} |f(x) - p_n^*(x)| = \min_{p_n \in H_n} \max_{a \leq x \leq b} |f(x) - p_n(x)|,$$

其中 H_n 为不超过 n 次的多项式集合.

定义 5.3.1 设函数 $f(x)$ 在区间 $[a, b]$ 上连续, $p_n(x) \in H_n$, 则

$$E = \|f - p_n\|_\infty = \max_{a \leq x \leq b} |f(x) - p_n(x)|$$

称为 $p_n(x)$ 与 $f(x)$ 的偏差. 若存在点 $\tilde{x} \in [a, b]$, 使得

$$|f(\tilde{x}) - p_n(\tilde{x})| = E,$$

则称点 $\tilde{x}$ 为 $p_n(x)$ 关于 $f(x)$ 的偏差点. 若 $p_n(\tilde{x}) - f(\tilde{x}) = E$, 则称 $\tilde{x}$ 为正偏差点; 若 $p_n(\tilde{x}) - f(\tilde{x}) = -E$, 则称 $\tilde{x}$ 为负偏差点.

下面不加证明地给出有关最优一致逼近多项式的几个定理.

定理 5.3.1 (切比雪夫定理) 设函数 $f(x)$ 在区间 $[a, b]$ 上连续, $p_n(x) \in H_n$, 则 $p_n(x)$ 是 $f(x)$ 的最优一致逼近多项式的充分必要条件是在区间 $[a, b]$ 上至少有 $n+2$ 个依次为正负的偏差点, 即至少有 $n+2$ 个点

$$a \leq \tilde{x}_0 < \tilde{x}_1 < \cdots < \tilde{x}_{n+1} \leq b,$$

使得

$$f(\tilde{x}_i) - p_n(\tilde{x}_i) = (-1)^i \sigma E = (-1)^i \sigma \|f - p_n\|_\infty, \quad \sigma = \pm 1; i = 0, 1, \cdots, n+1,$$

偏差点 $\tilde{x}_i$ ($i = 0, 1, \cdots, n+1$) 也称为 $f(x) - p_n(x)$ 的切比雪夫交错点组.

定理 5.3.2 设函数 $f(x)$ 在区间 $[a, b]$ 上连续, 则 $f(x)$ 在 H_n 中存在唯一的最优一致逼近多项式 $p_n(x)$.

定理 5.3.3 设函数 $f(x)$ 在区间 $[a, b]$ 上 $n+1$ 阶可导, 并且 $f^{(n+1)}(x)$ 在区间 (a, b) 中保持定号 (即 $f^{(n+1)}(x) > 0$ 或 $f^{(n+1)}(x) < 0$), 则区间 $[a, b]$ 的端点 a, b 属于 $f(x) - p_n(x)$ 的交错点组.

切比雪夫定理给出了一个构造最优一致逼近多项式的方法. 设最优一致逼近多项式为

$$p_n(x) = c_0 + c_1x + c_2x^2 + \cdots + c_nx^n,$$

偏差点为 $\tilde{x}_i$ ($i = 0, 1, \cdots, n+1$), 记 $\mu = \sigma E$, 则由切比雪夫定理有

$$f(\tilde{x}_i) - p_n(\tilde{x}_i) = (-1)^i \mu, \quad i = 0, 1, \cdots, n+1. \quad (5.3.1)$$

这是关于参数 c_i ($i = 0, 1, \cdots, n$), 偏差点 $\tilde{x}_i$ ($i = 0, 1, \cdots, n+1$) 和带有符号的偏差 μ 的 $2n+4$ 个未知量的 $n+2$ 个方程的非线性方程组.

通常, 求解方程组 (5.3.1) 有下面两种方法.

方法 1 列梅兹 (Remes) 方法 列梅兹方法的核心是由区间 $[a, b]$ 上给定的 $n+2$ 个点逐次求得交错点组, 最终求得最优一致逼近多项式的方法. 粗略地讲, 求解过程是将区间 $[a, b]$ 上给定的 $n+2$ 个点 $\tilde{x}_i$ ($a \leq \tilde{x}_0 < \tilde{x}_1 < \cdots < \tilde{x}_{n+1} \leq b$, 若 $f^{(n+1)}(x) > 0$ 或 $f^{(n+1)}(x) < 0$, 则取 $\tilde{x}_0 = a, \tilde{x}_{n+1} = b$) 代入方程组 (5.3.1), 求得 $p_n(x)$ 的一组系数 $c_0, c_1, \cdots, c_n$ 及 μ , 得到初始逼近函数 $p_n(x)$. 这时, $f(x) - p_n(x)$ 在点 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_{n+1}$ 上交错变号, 然后用 $f(x) - p_n(x)$ 的极值点 $\tilde{x}$ 替换 $\{\tilde{x}_i\}$ 中的

一点, 得到一个新的点集, 并使 $f(x) - p_n(x)$ 在新点集上正负相间取值, 将新点集代入方程组 (5.3.1) 再求解. 重复这个过程, 直至满足精度为止 (求解过程复杂, 在此不作详细讨论).

方法 2 由偏差点的定义知, 区间 (a, b) 内的偏差点必然是 $f(x) - p_n(x)$ 的极值点, 故偏差点 $\tilde{x}_i$ 满足方程

$$f'(\tilde{x}_i) - p'_n(\tilde{x}_i) = 0, \quad i = 1, 2, \dots, n.$$

又若 $f^{(n+1)}(x) > 0$ 或 $f^{(n+1)}(x) < 0$, 根据定理 5.3.3, 取偏差点 $\tilde{x}_0 = a, \tilde{x}_{n+1} = b$, 结合方程组 (5.3.1), 则得关于参数 c_i ($i = 0, 1, \dots, n$), 偏差点 $\tilde{x}$ ($i = 1, 2, \dots, n$) 和带有符号的偏差 μ 的 $2n + 2$ 个未知量的 $2n + 2$ 个方程的方程组

$$\begin{cases} f(\tilde{x}_i) - p_n(\tilde{x}_i) = (-1)^i \mu, & i = 0, 1, \dots, n+1, \\ f'(\tilde{x}_i) - p'_n(\tilde{x}_i) = 0, & i = 1, 2, \dots, n. \end{cases} \quad (5.3.2)$$

一般地, 方程组 (5.3.2) 是非线性方程组, 当 $n \geq 2$ 时求解比较复杂, 最终也只能得到近似解, 通常求最优一致逼近多项式采用其他近似方法. 下面给出一个由方法 2 求一次最优一致逼近多项式的例子.

例 5.3.1 求函数 $y = \arctan x$ 在区间 $[0, 1]$ 上的最优一致逼近一次多项式.

解 设 $p_1(x) = c_0 + c_1x$, 由于 $f''(x) = -2x(1+x^2)^{-2} < 0$ ($0 < x \leq 1$), 故取偏差点 $\tilde{x}_0 = 0, \tilde{x}_2 = 1$, 由方程组 (5.3.2) 有

$$\begin{cases} -c_0 = \mu, & \arctan \tilde{x}_1 - c_0 - c_1 \tilde{x}_1 = -\mu, \\ \frac{\pi}{4} - c_0 - c_1 = \mu, & \frac{1}{1 + \tilde{x}_1^2} - c_1 = 0. \end{cases}$$

这是关于 4 个未知量 $c_0, c_1, \tilde{x}_1, \mu$ 的方程组. 解之, 得

$$c_1 = \frac{\pi}{4} \approx 0.7854, \quad \tilde{x}_1 = \sqrt{\frac{1}{c_1} - 1} \approx 0.5227,$$

$$c_0 = \frac{\arctan \tilde{x}_1 - c_1 \tilde{x}_1}{2} \approx 0.0356, \quad \mu \approx -0.0356.$$

于是得到函数 $y = \arctan x$ 在区间 $[0, 1]$ 上的最优一致逼近一次多项式

$$\arctan x \approx p_1(x) = 0.0356 + 0.7854x,$$

最大误差 $E = -\mu = 0.0356$. 解毕.

5.3.2 近似最优一致逼近多项式

1. 切比雪夫插值多项式

由切比雪夫定理知, 若 $p_n(x)$ 是 $f(x)$ 的最优一致逼近多项式, 则在区间 $[a, b]$ 上至少有 $n+2$ 偏差点 $\tilde{x}_i$ ($i=0, 1, \dots, n+1$), 使得

$$f(\tilde{x}_i) - p_n(\tilde{x}_i) = (-1)^i \mu, \quad i = 0, 1, \dots, n+1.$$

这说明 $f(x) - p_n(x)$ 在区间 $[a, b]$ 上至少变号 $n+1$ 次, 从而至少有 $n+1$ 个不同的点 $x_i \in [a, b]$ ($i=0, 1, \dots, n$), 使得

$$f(x_i) - p_n(x_i) = 0, \quad i = 0, 1, \dots, n,$$

即

$$f(x_i) = p_n(x_i), \quad i = 0, 1, \dots, n.$$

据此可知, $p_n(x)$ 是以 x_i ($i=0, 1, \dots, n$) 为插值节点构成的 $f(x)$ 的 n 次插值多项式. 问题是如何选择插值节点 x_i ($i=0, 1, \dots, n$), 使得构造出的插值多项式的最大误差达到最小, 即

$$\max_{a \leq x \leq b} |f(x) - p_n(x)| = \max_{a \leq x \leq b} |R_n(x)| = \min.$$

使最大误差达到最小的插值多项式就是近似最优一致逼近多项式.

由插值多项式的误差估计式有

$$\begin{aligned} \max_{a \leq x \leq b} |f(x) - p_n(x)| &= \max_{a \leq x \leq b} |R_n(x)| \\ &= \max_{a \leq x \leq b} \left| \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-x_0)(x-x_1)\cdots(x-x_n) \right| \\ &\leq \frac{M_{n+1}}{(n+1)!} \max_{a \leq x \leq b} |(x-x_0)(x-x_1)\cdots(x-x_n)|, \end{aligned}$$

其中 $M_{n+1} = \max_{a \leq x \leq b} |f^{(n+1)}(x)|$.

若选取 x_i ($i=0, 1, \dots, n$) 使得 $\max_{a \leq x \leq b} |(x-x_0)(x-x_1)\cdots(x-x_n)|$ 达到最小, 则 $\max_{a \leq x \leq b} |R_n(x)|$ 达到最小.

(1) 所讨论的区间是 $[-1, 1]$. 由切比雪夫多项式的性质 5.1.9 知

$$\min \max_{-1 \leq x \leq 1} |(x-x_0)(x-x_1)\cdots(x-x_n)| = \max_{-1 \leq x \leq 1} \left| \frac{1}{2^n} T_{n+1}(x) \right| = \frac{1}{2^n},$$

即当选取插值节点 x_i ($i=0, 1, \dots, n$) 是 $n+1$ 次切比雪夫多项式 $T_{n+1}(x)$ 的零点

$$x_i = \cos \left[\frac{2i+1}{2(n+1)}\pi \right], \quad i = 0, 1, \dots, n$$

时, $\max_{-1 \leq x \leq 1} |(x-x_0)(x-x_1)\cdots(x-x_n)|$ 取到最小值 $\frac{1}{2^n}$. 这时, 所构造的插值多项式 $p_n(x)$ 的最大误差达到最小, 则 $p_n(x)$ 是 $f(x)$ 在区间 $[-1, 1]$ 上的近似最优一致逼近多项式, 其最大误差满足

$$\max_{-1 \leq x \leq 1} |R_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \frac{1}{2^n}. \quad (5.3.3)$$

(2) 所讨论的区间是有限区间 $[a, b]$.

(i) 通过变量替换 $x = \frac{a+b}{2} + \frac{b-a}{2}t$, 将 $x \in [a, b]$ 变为 $t \in [-1, 1]$. 在区间 $[-1, 1]$ 上构造多项式 $\tilde{p}_n(t)$, 然后将

$$t = \frac{2x - a - b}{b - a}$$

代入 $\tilde{p}_n(t)$, 即得到区间 $[a, b]$ 上的近似最优一致逼近多项式 $p_n(x)$.

(ii) 直接取插值节点

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}t_i, \quad i = 0, 1, \dots, n,$$

其中 t_i 为 $n+1$ 次切比雪夫多项式 $T_{n+1}(x)$ 的零点, 即取

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} \cos \left[\frac{2i+1}{2(n+1)}\pi \right], \quad i = 0, 1, \dots, n$$

作为插值节点, 所构造的插值多项式 $p_n(x)$ 就是 $f(x)$ 在区间 $[a, b]$ 上的近似最优一致逼近多项式. 此时,

$$\max_{a \leq x \leq b} |\pi_{n+1}(x)| = \max_{a \leq x \leq b} |(x-x_0)(x-x_1)\cdots(x-x_n)|$$

达到最小, 注意到

$$x - x_i = \frac{b-a}{2}(t - t_i),$$

于是有

$$\max_{a \leq x \leq b} |\pi_{n+1}(x)| \leq \left(\frac{b-a}{2} \right)^{n+1} \frac{1}{2^n}.$$

$p_n(x)$ 的最大误差满足

$$\max_{-1 \leq x \leq 1} |R_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \left(\frac{b-a}{2} \right)^{n+1} \frac{1}{2^n} = \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{2n+1}}. \quad (5.3.4)$$

例 5.3.2 利用切比雪夫插值多项式求 $y = \arctan x$ 在区间 $[0, 1]$ 上的近似最优一致逼近一次多项式.

解法 1 作变量替换, 令 $x = (t+1)/2$, 将 $x \in [a, b]$ 变为 $t \in [-1, 1]$, 先关于 t 作插值多项式, 然后将 t 换为 x , 即得近似最优一致逼近一次多项式.

$T_2(t)$ 的零点 $t_i = \cos\left(\frac{2i+1}{4}\pi\right)$ ($i = 0, 1$), 即

$$t_0 = \cos \frac{\pi}{4} = \frac{\sqrt{2}}{2}, \quad t_1 = \cos \frac{3\pi}{4} = -\frac{\sqrt{2}}{2}.$$

利用牛顿插值多项式 $N_1(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$, 则

$$\begin{aligned} y = \arctan x = \arctan \frac{t+1}{2} &\approx N_1(t) = \arctan t_0 + \frac{\arctan t_1 - \arctan t_0}{t_1 - t_0}(t - t_0) \\ &= \arctan \frac{1 + \frac{\sqrt{2}}{2}}{2} + \frac{\arctan \frac{1 - \frac{\sqrt{2}}{2}}{2} - \arctan \frac{1 + \frac{\sqrt{2}}{2}}{2}}{-\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}} \left(t - \frac{\sqrt{2}}{2}\right). \end{aligned}$$

将 $t = 2x - 1$ 代入上式得

$$\begin{aligned} y = \arctan x &\approx \arctan \frac{2 + \sqrt{2}}{4} + \left(\arctan \frac{2 + \sqrt{2}}{4} - \arctan \frac{2 - \sqrt{2}}{4}\right) \frac{1}{\sqrt{2}} \left(2x - 1 - \frac{\sqrt{2}}{2}\right) \\ &= \arctan \frac{2 + \sqrt{2}}{4} + \sqrt{2} \left(\arctan \frac{2 + \sqrt{2}}{4} - \arctan \frac{2 - \sqrt{2}}{4}\right) \left(x - \frac{2 + \sqrt{2}}{4}\right). \end{aligned}$$

解法 2 由解法 1 知 $t_0 = \frac{\sqrt{2}}{2}, t_1 = -\frac{\sqrt{2}}{2}$, 这时,

$$x_0 = \frac{t_0 + 1}{2} = \frac{2 + \sqrt{2}}{4}, \quad x_1 = \frac{t_1 + 1}{2} = \frac{2 - \sqrt{2}}{4},$$

则

$$\begin{aligned} y = \arctan x &\approx N_1(x) \\ &= \arctan \frac{2 + \sqrt{2}}{4} + \frac{\arctan \frac{2 - \sqrt{2}}{4} - \arctan \frac{2 + \sqrt{2}}{4}}{\frac{2 - \sqrt{2}}{4} - \frac{2 + \sqrt{2}}{4}} \left(x - \frac{2 + \sqrt{2}}{4}\right) \\ &= \arctan \frac{2 + \sqrt{2}}{4} + \sqrt{2} \left(\arctan \frac{2 + \sqrt{2}}{4} - \arctan \frac{2 - \sqrt{2}}{4}\right) \left(x - \frac{2 + \sqrt{2}}{4}\right) \\ &\approx 0.029197 + 0.793572x. \end{aligned}$$

下面给出误差估计. 据式 (5.3.4) 得

$$\max_{0 \leq x \leq 1} |R_1(x)| \leq \frac{M_2}{2!} \frac{1}{2^3} = \frac{3\sqrt{3}}{128} \approx 0.04059494,$$

其中 $M_2 = \max_{0 \leq x \leq 1} |y''| = \frac{3\sqrt{3}}{8}$. 解毕.

2. 截断切比雪夫级数法

设函数 $f(x)$ 在区间 $[-1, 1]$ 上连续, 当正交函数系取为切比雪夫多项式时, 广义傅里叶级数

$$\sum_{k=0}^{\infty} c_k T_k(x)$$

称为函数 $f(x)$ 的切比雪夫级数, 其中系数

$$c_k = \frac{(T_k, f)}{(T_k, T_k)}, \quad k = 0, 1, 2, \dots,$$

而

$$(T_0, T_0) = \pi, \quad (T_k, T_k) = \frac{\pi}{2}, \quad k = 1, 2, \dots,$$

于是

$$\begin{aligned} c_0 &= \frac{(T_0, f)}{\pi} = \frac{(1, f)}{\pi} = \frac{1}{\pi} \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \stackrel{x=\cos \theta}{=} \frac{1}{\pi} \int_0^\pi f(\cos \theta) d\theta, \\ c_k &= \frac{2(T_k, f)}{\pi} = \frac{2}{\pi} \int_{-1}^1 \frac{\cos(k \arccos x) f(x)}{\sqrt{1-x^2}} dx \stackrel{x=\cos \theta}{=} \frac{2}{\pi} \int_0^\pi f(\cos \theta) \cos(k\theta) d\theta, \\ &\quad k = 1, 2, \dots \end{aligned}$$

注意到 $f(x)$ 的傅里叶余弦级数为

$$f(x) \approx a_0 + \sum_{k=1}^{\infty} a_k \cos(kx),$$

其中系数

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx, \quad a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx) dx, \quad k = 1, 2, \dots$$

将 c_0, c_k 改写为

$$\begin{aligned} c_0 &= \frac{1}{\pi} \int_0^\pi f(\cos \theta) d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\cos \theta) d\theta, \\ c_k &= \frac{2}{\pi} \int_0^\pi f(\cos \theta) \cos(k\theta) d\theta = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\cos \theta) \cos(k\theta) d\theta, \quad k = 1, 2, \dots, \end{aligned}$$

而

$$\sum_{k=0}^{\infty} c_k T_k(x) = \sum_{k=0}^{\infty} c_k \cos(k \arccos x) \stackrel{x=\cos \theta}{=} \sum_{k=0}^{\infty} c_k \cos(k\theta).$$

由此可见, $f(x)$ 的切比雪夫级数就是 $f(\cos \theta)$ 的傅里叶余弦级数. 若将函数 $f(x)$ 的切比雪夫级数的部分和记为

$$S_n(x) = \sum_{k=0}^n c_k T_k(x),$$

则有下列收敛性定理:

定理 5.3.4 设在区间 $[-1, 1]$ 上 $f(x)$ 的一阶导数存在且分段连续, 则当 $n \rightarrow \infty$ 时, $S_n(x)$ 在 $[-1, 1]$ 上一致收敛于 $f(x)$.

证明 略.

据此可知,

$$f(x) = \sum_{k=0}^{\infty} c_k T_k(x).$$

当切比雪夫级数收敛较快时,

$$f(x) - S_n(x) = \sum_{k=n+1}^{\infty} c_k T_k(x) \approx c_{n+1} T_{n+1}(x).$$

由于 $T_{n+1}(x)$ 有 $n+2$ 个极值点, 所以可以近似地认为 $f(x) - S_n(x)$ 有 $n+2$ 个偏差点. 由切比雪夫定理知, $S_n(x)$ 是 $f(x)$ 的近似最优一致逼近 n 次多项式.

截断切比雪夫级数法的缺点: 在一般情况下, 系数 c_k 中的积分难以计算.

例 5.3.3 利用截断切比雪夫级数法求 $y = \arctan x$ 在区间 $[0, 1]$ 上的近似最优一致逼近一次多项式.

解 令 $x = \frac{t+1}{2}$, 则

$$y = \arctan \frac{t+1}{2}, \quad -1 \leq t \leq 1.$$

按切比雪夫级数系数计算公式

$$c_0 = \frac{1}{\pi} \int_0^{\pi} \arctan \left(\frac{\cos \theta + 1}{2} \right) d\theta \approx 0.427\,078\,586\,4,$$

$$c_1 = \frac{2}{\pi} \int_0^{\pi} \arctan \left(\frac{\cos \theta + 1}{2} \right) \cos \theta d\theta \approx 0.394\,736\,453\,9,$$

所以

$$\begin{aligned} \arctan x &\approx c_0 T_0(t) + c_1 T_1(t) = c_0 + c_1 t = c_0 + c_1 (2x - 1) \\ &\approx 0.032\,342\,132\,5 + 0.789\,472\,907\,8x. \end{aligned}$$

解毕.

3. 缩短幂级数法

在许多情况下, 可以用高阶泰勒 (Taylor) 多项式逼近 $f(x)$, 即取

$$f(x) \approx p_n(x) = f(0) + f'(x)x + \frac{1}{2!}f''(0)x^2 + \cdots + \frac{1}{n!}f^{(n)}(0)x^n.$$

随着多项式 $p_n(x)$ 的次数 n 的增大, 偏差

$$\max_{-1 \leq x \leq 1} |f(x) - p_n(x)|$$

逐渐减少. 然而, n 越大, $p_n(x)$ 的次数越高, 计算量随之增大. 为了减少计算量, 将 n 次多项式 $p_n(x)$ 降低为 $n-1$ 次多项式 $p_{n-1}(x)$, 要求最大误差达到最小.

一般地, 不妨设

$$p_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n, \quad -1 \leq x \leq 1. \quad (5.3.5)$$

将 $p_n(x)$ 降低为 $n-1$ 次多项式 $p_{n-1}(x)$, 不是直接去掉 a_nx^n 项, 而是减去一个包含 a_nx^n 项的 n 次多项式 $\bar{p}_n(x)$, 即 $p_{n-1}(x) = p_n(x) - \bar{p}_n(x)$, 要求 $p_n(x) - p_{n-1}(x)$ 的最大误差达到最小, 即要求 $\bar{p}_n(x)$ 满足

$$\max_{-1 \leq x \leq 1} |\bar{p}_n(x)| = \max_{-1 \leq x \leq 1} |p_n(x) - p_{n-1}(x)| = \min.$$

由切比雪夫多项式的性质 5.1.9 知

$$\min \max_{-1 \leq x \leq 1} |\bar{p}_n(x)| = \min \max_{-1 \leq x \leq 1} |p_n(x) - p_{n-1}(x)| = \max_{-1 \leq x \leq 1} \left| \frac{a_n}{2^{n-1}} T_n(x) \right|.$$

据此可知, 若取

$$\bar{p}_n(x) = \frac{a_n}{2^{n-1}} T_n(x),$$

即满足要求, 所以有

$$p_{n-1}(x) = p_n(x) - \frac{a_n}{2^{n-1}} T_n(x). \quad (5.3.6)$$

$p_{n-1}(x)$ 称为 $p_n(x)$ 的缩短多项式. 缩短多项式又可以用同样的方法逐步缩短, 最终得到满足所给精度要求的近似最优一致逼近多项式.

下面介绍缩短多项式 $p_{n-1}(x)$ 的另一种求法.

由切比雪夫多项式 $T_k(x)$ ($k = 0, 1, 2, \cdots$) 是区间 $[-1, 1]$ 上的正交多项式知, $p_n(x)$ 可以由切比雪夫多项式唯一地表示为

$$p_n(x) = c_0 T_0(x) + c_1 T_1(x) + \cdots + c_n T_n(x). \quad (5.3.7)$$

注意到含 x^n 的项只能出现在 $T_n(x)$ 中, 比较式 (5.3.5) 与式 (5.3.7) 中 x^n 的系数得

$$a_n = c_n 2^{n-1},$$

故 $c_n = \frac{a_n}{2^{n-1}}$. 由式 (5.3.6) 知, $p_{n-1}(x) = p_n(x) - c_n T_n(x)$, 即

$$p_{n-1}(x) = c_0 T_0(x) + c_1 T_1(x) + \cdots + c_{n-1} T_{n-1}(x). \quad (5.3.8)$$

为了得到 $p_{n-1}(x)$ 关于 x^k ($k = 0, 1, \cdots, n-1$) 的表达式, 将 $p_n(x)$ 的表达式 (5.3.5) 中的 x^k ($k = 0, 1, \cdots, n$) 用切比雪夫多项式 $T_k(x)$ ($k = 0, 1, \cdots, n$) 表示 (x^k 与 $T_k(x)$ 的关系如表 5.4 所示), 关于 $T_k(x)$ 合并同类项后, 删除含有 $T_n(x)$ 的项, 即得式 (5.3.8). 然后将 $T_k(x)$ ($k = 0, 1, \cdots, n-1$) 代入式 (5.3.8) 中 (切比雪夫多项式的前几项如表 5.5 所示), 就得到 $p_{n-1}(x)$ 关于 x^k ($k = 0, 1, \cdots, n-1$) 的表达式.

表 5.4 用切比雪夫多项式表示单项式 x^k

$1 = T_0$
$x = T_1$
$x^2 = \frac{1}{2}(T_0 + T_2)$
$x^3 = \frac{1}{4}(3T_1 + T_3)$
$x^4 = \frac{1}{8}(3T_0 + 4T_2 + T_4)$
$x^5 = \frac{1}{16}(10T_1 + 5T_3 + T_5)$
$x^6 = \frac{1}{32}(10T_0 + 15T_2 + 6T_4 + T_6)$
$x^7 = \frac{1}{64}(35T_1 + 21T_3 + 7T_5 + T_7)$
$x^8 = \frac{1}{128}(35T_0 + 56T_2 + 28T_4 + 8T_6 + T_8)$
$x^9 = \frac{1}{256}(126T_1 + 84T_3 + 36T_5 + 9T_7 + T_9)$
$x^{10} = \frac{1}{512}(126T_0 + 210T_2 + 120T_4 + 45T_6 + 10T_8 + T_{10})$

表 5.5 切比雪夫多项式

$T_0(x) = 1$
$T_1(x) = x$
$T_2(x) = 2x^2 - 1$
$T_3(x) = 4x^3 - 3x$
$T_4(x) = 8x^4 - 8x^2 + 1$
$T_5(x) = 16x^5 - 20x^3 + 5x$
$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$
$T_7(x) = 64x^7 - 112x^5 + 56x^3 - 7x$
$T_8(x) = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1$
$T_9(x) = 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x$
$T_{10}(x) = 512x^{10} - 1280x^8 + 1120x^6 - 400x^4 + 50x^2 - 1$

例 5.3.4 利用缩短幂级数法求 $y = \arctan x$ 在区间 $[0, 1]$ 上的近似最优一致逼近一次多项式.

解法 1 令 $x = \frac{1}{2}(t+1)$, 则

$$y = \arctan \frac{t+1}{2}, \quad -1 \leq t \leq 1,$$

而

$$\arctan \frac{t+1}{2} \approx p_2(t) = \arctan \frac{1}{2} + \frac{2}{5}t - \frac{2}{25}t^2.$$

由式 (5.3.6) 知

$$p_1(t) = p_2(t) - \left(-\frac{2}{25}\right) \times \frac{1}{2}T_2(t) = p_2(t) + \frac{1}{25}(2t^2 - 1) = \arctan \frac{1}{2} + \frac{2}{5}t - \frac{1}{25},$$

所以

$$\arctan x \approx p_1(x) = \arctan \frac{1}{2} - \frac{1}{25} + \frac{2}{5}(2x - 1) \approx 0.023\,647\,6 + 0.8x.$$

解法 2 由解法 1 和表 5.4 知

$$\begin{aligned} \arctan \frac{t+1}{2} \approx p_2(t) &= \arctan \frac{1}{2} + \frac{2}{5}t - \frac{2}{25}t^2 \\ &= \arctan \frac{1}{2} \times T_0 + \frac{2}{5}T_1 - \frac{2}{25} \times \frac{T_0 + T_2}{2} \\ &= \left(\arctan \frac{1}{2} - \frac{1}{25}\right)T_0 + \frac{2}{5}T_1 - \frac{1}{25}T_2, \end{aligned}$$

由式 (5.3.8) 知

$$\begin{aligned} p_1(t) &= \left(\arctan \frac{1}{2} - \frac{1}{25}\right)T_0 + \frac{2}{5}T_1 \\ &= \left(\arctan \frac{1}{2} - \frac{1}{25}\right) + \frac{2}{5}t, \end{aligned}$$

所以

$$\arctan x \approx p_1(x) = \arctan \frac{1}{2} - \frac{1}{25} + \frac{2}{5}(2x - 1).$$

解毕.

例 5.3.5 利用缩短幂级数法求函数 $f(x) = e^x$ 在区间 $[-1, 1]$ 上的近似最优一致逼近多项式, 使得误差不超过 0.005.

解 e^x 在 $x = 0$ 处的泰勒展开式为

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + R_n(x),$$

其中误差项

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

ξ 介于 0 与 x 之间. 显然,

$$\max_{-1 \leq x \leq 1} |R_n(x)| \leq \frac{e}{(n+1)!}.$$

当 $n=4$ 时,

$$\max_{-1 \leq x \leq 1} |R_4(x)| \leq \frac{e}{5!} \approx 0.022\,652\,348 > 0.005;$$

当 $n=5$ 时,

$$\max_{-1 \leq x \leq 1} |R_5(x)| \leq \frac{e}{6!} \approx 0.003\,775\,391,$$

所以

$$\begin{aligned} p_5(x) &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} \\ &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} \\ &= T_0 + T_1 + \frac{1}{2} \frac{T_0 + T_2}{2} + \frac{1}{6} \frac{3T_1 + T_3}{4} + \frac{1}{24} \frac{3T_0 + 4T_2 + T_4}{8} \\ &\quad + \frac{1}{120} \frac{10T_1 + 5T_3 + T_5}{16} \\ &= \frac{81}{64} T_0 + \frac{217}{192} T_1 + \frac{13}{48} T_2 + \frac{17}{384} T_3 + \frac{1}{192} T_4 + \frac{1}{1920} T_5. \end{aligned}$$

由于

$$\frac{e}{6!} + \frac{T_5}{1920} < 0.003\,78 + 0.000\,53 = 0.004\,31 < 0.005.$$

所以取

$$\begin{aligned} e^x &\approx P_4(x) = \frac{81}{64} T_0 + \frac{217}{192} T_1 + \frac{13}{48} T_2 + \frac{17}{384} T_3 + \frac{1}{192} T_4 \\ &= \frac{81}{64} + \frac{217}{192} x + \frac{13}{48} (2x^2 - 1) + \frac{17}{384} (4x^3 - 3x) + \frac{1}{192} (8x^4 - 8x^2 + 1) \\ &= 1 + \frac{383}{384} x + \frac{1}{2} x^2 + \frac{17}{96} x^3 + \frac{1}{24} x^4. \end{aligned}$$

解毕.

小 结

本章介绍了两种求函数近似表达式的方法——最优平方逼近法与最优一致逼近法.

最优平方逼近法是针对 $f(x)$ 是定义在点集 $X = \{x_1, x_2, \dots, x_m\}$ 上的列表函数或 $f(x)$ 是定义在区间 $[a, b]$ 上的连续函数的两种情况, 构造其逼近函数 (广义多项式)

$$p(x) = c_0\phi_0(x) + c_1\phi_1(x) + \dots + c_n\phi_n(x),$$

使得 $\|p - y\|_2^2$ 达到最小.

记 $S(c_0, c_1, \dots, c_n) = \|p - y\|_2^2$, 令 $\frac{\partial S}{\partial c_k} = 0$ ($k = 0, 1, \dots, n$), 得到关于参数 c_k ($k = 0, 1, \dots, n$) 的正规方程组 (5.2.9). 从正规方程组求得参数 c_k ($k = 0, 1, \dots, n$), 则得最小二乘拟合函数或最优平方逼近函数 $p(x)$. 但当正规方程组的阶数较高时, 往往是病态方程组. 为了避免求解病态方程组, 选取基函数 $\{\phi_k\}_0^n$ 为一组正交函数 $\{g_k\}_0^n$. 这时,

$$p(x) = \sum_{k=0}^n \frac{(g_k, f)}{(g_k, g_k)} g_k(x).$$

正交函数系常取为正交多项式. 正交多项式可以利用三项递推关系构造, 或者取为勒让德多项式. 由于勒让德多项式在区间 $[-1, 1]$ 上正交, 所以使用时首先利用变换 $x = \frac{a+b}{2} + \frac{b-a}{2}t$ 将区间 $[a, b]$ 变到 $[-1, 1]$. 利用勒让德多项式构造出多项式 $p(t)$ 后, 再利用反变换 $t = \frac{2x-a-b}{b-a}$ 将区间 $[-1, 1]$ 变回 $[a, b]$, 则得 $f(x)$ 在区间 $[a, b]$ 上的最优平方逼近多项式 $p(x)$.

而最优一致逼近法则是针对 $f(x)$ 是定义在区间 $[a, b]$ 上连续函数的情况, 构造 n 次多项式

$$p(x) = c_0 + c_1x + c_2x^2 + \dots + c_nx^n,$$

使得 $\|p - f\|_\infty$ 达到最小.

由切比雪夫定理进而得到关于最优一致逼近多项式的系数 c_k ($k = 0, 1, \dots, n$) 所满足的非线性方程组 (5.3.2). 当 $n = 1$ 时, 通过求解该方程组得到最优一致逼近一次多项式; 当 $n \geq 2$ 时, 由于求解非线性方程组比较复杂, 并且难以求出其准确解; 所以通常用切比雪夫插值法、截断切比雪夫级数法、缩短幂级数法求其近似最优一致逼近多项式.

习 题

5.1 求下列给定区间上关于权函数 $\omega(x)$ 的正交多项式 $g_0(x), g_1(x), g_2(x)$:

- (1) $[0, 1]$, $\omega(x) = \sqrt{x}$; (2) $[-1, 1]$, $\omega(x) = 1 + x^2$;
- (3) $[0, 1]$, $\omega(x) = \sqrt{x(1-x)}$; (4) $[-1, 1]$, $\omega(x) = \sqrt{1-x^2}$.

5.2 给定数据如下表中所示, 求其最小二乘拟合函数 $p(x)$:

(1) $p(x) = c_0 + c_1x + c_2x^2$;

x_i	1	3	4	5	6	7	8	9	10
y_i	2	7	8	10	11	11	10	9	8

(2) $p(x) = c_0 + c_1x + c_2x^2$;

x_i	0.78	1.56	2.34	3.12	3.81
y_i	2.50	1.20	1.12	2.25	4.28

(3) $p(x) = ae^{bx}$;

x_i	1	2	3	4	5	6	7	8
y_i	15.3	20.5	27.4	36.6	49.1	65.6	87.8	117.6

(4) $p(x) = c_0 + c_1x^2$.

x_i	19	25	31	38	44
y_i	19.0	32.3	49.0	73.3	97.8

5.3 求下列函数在指定区间上的最优平方逼近一次多项式:

(1) $y = \sqrt{x}$, $[0, 1]$; (2) $y = e^x$, $[-1, 1]$.

5.4 利用正交多项式求下列函数的最优平方逼近二次多项式:

(1) 数据同习题 5.2 (1); (2) $y = \arcsin x$, $[0, 1]$.

5.5 取基函数为勒让德多项式, 求函数 $f(x) = \sin \frac{\pi}{2}x$ 在区间 $[-1, 1]$ 上的最优平方逼近三次多项式.

5.6 求下列函数在指定区间上的最优一致逼近一次多项式:

(1) $y = \sqrt{x}$, $\left[\frac{1}{4}, 1\right]$; (2) $y = \ln x$, $[1, 2]$;

(3) $y = e^x$, $[0, 1]$; (4) $y = \sqrt{1+x^2}$, $[0, 1]$.

5.7 设 $f(x)$ 是区间 $[a, b]$ 上的连续函数, 其最小值、最大值分别为 m 与 M , 求 $f(x)$ 的最优一致逼近零次多项式.

5.8 求 $f(x) = e^x$ 在区间 $[-1, 1]$ 上的切比雪夫插值二次多项式, 并用范数 $\|\cdot\|_\infty$ 估计误差.

5.9 求 $p_4(x) = 1 + 5x^2 + x^3 + 2x^4$ 在 $[-1, 1]$ 上的近似最优一致逼近三次多项式.

5.10 用两种方法求下列函数在指定区间上的近似最优一致逼近一次式, 并求其偏差:

(1) $y = e^{-x}$, $[-1, 1]$; (2) $y = \sin x$, $\left[0, \frac{\pi}{4}\right]$.

5.11 证明切比雪夫多项式满足以下关系:

$$T_{m+n}(x) + T_{m-n}(x) = 2T_m(x)T_n(x), \quad m \geq n.$$

5.12 定义 $T_k^*(x) = T_k(2x - 1)$.

- (1) 求 $T_k^*(x)$ ($k = 0, 1, \dots, 4$);
- (2) 证明 $T_k^*(x)$ 在区间 $[0, 1]$ 上关于权函数 $\omega(x) = 1/\sqrt{x-x^2}$ 正交;
- (3) 证明 $T_k^*(x^2) = T_{2k}(x)$.

计 算 实 习

5.1 给定数据如下表所示, 求其最小二乘拟合四次多项式及其误差:

x_i	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
y_i	5.123 4	5.305 7	5.568 7	5.937 5	6.437 0	7.097 8	7.949 3	9.025 3	10.362 7

5.2 给定函数

$$f(x) = \frac{1}{1 + 25x^2}, \quad -1 \leq x \leq 1.$$

- (1) 求函数 $f(x)$ 在区间 $[-1, 1]$ 上的最优平方逼近三次多项式;
- (2) 在区间 $[-1, 1]$ 上取 $x_i = -1 + 0.4i$ ($i = 0, 1, \dots, 5$), 求函数 $f(x)$ 的最小二乘拟合三次多项式;
- (3) 在区间 $[-1, 1]$ 上取 $x_i = -1 + 0.2i$ ($i = 0, 1, \dots, 10$), 求函数 $f(x)$ 的最小二乘拟合三次多项式;
- (4) 试画出以上三种拟合函数的曲线, 并与第 4 章的计算实习中的结果进行比较.

第6章 数值积分与数值微分

在科学研究和工程技术中常常需要计算函数 $f(x)$ 的积分或导数. 然而, 在实际问题中, 常会遇到被积函数 $f(x)$ 是列表函数, 或者它的表达式比较复杂, 难以求出其原函数, 这时牛顿-莱布尼茨公式不能使用, 必须用数值积分方法计算. 本章介绍几类常用的数值积分和数值微分计算方法.

6.1 牛顿-科茨求积公式

6.1.1 数值积分的基本思想

设 $p(x)$ 是被积函数 $f(x)$ 的比较简单、易于积分的近似表达式, $f(x) = p(x) + R(x)$, 则

$$\int_a^b f(x) dx = \int_a^b p(x) dx + \int_a^b R(x) dx,$$

取积分 $\int_a^b f(x) dx \approx \int_a^b p(x) dx$, 则截断误差

$$R[f] = \int_a^b f(x) dx - \int_a^b p(x) dx = \int_a^b R(x) dx.$$

若记 $I[f] = \int_a^b f(x) dx$, $Q[f] = \int_a^b p(x) dx$, 则

$$I[f] = Q[f] + R[f].$$

例如, 在积分区间 $[a, b]$ 中插入 $n-1$ 个点 x_i ($i = 1, 2, \dots, n-1$), 满足 $a = x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n = b$, 取被积函数 $f(x)$ 的近似表达式为三次样条插值函数 $S(x)$, 即 $f(x) \approx S(x)$, 则有

$$\begin{aligned} I[f] &= \int_a^b f(x) dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) dx \approx \sum_{i=1}^n \int_{x_{i-1}}^{x_i} S(x) dx \\ &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \left[\frac{(x_i - x)^3}{6h_i} M_{i-1} + \frac{(x - x_{i-1})^3}{6h_i} M_i \right. \\ &\quad \left. + \left(y_{i-1} - \frac{h_i^2}{6} M_{i-1} \right) \frac{x_i - x}{h_i} + \left(y_i - \frac{h_i^2}{6} M_i \right) \frac{x - x_{i-1}}{h_i} \right] dx \end{aligned}$$

$$= \frac{1}{2} \sum_{i=1}^n h_i (y_{i-1} + y_i) - \frac{1}{24} \sum_{i=1}^n (M_{i-1} + M_i) h_i^3 = Q[f].$$

像这样近似计算积分的公式统称为数值积分公式.

6.1.2 牛顿-科茨求积公式

常用的几个基本数值积分公式是由被积函数 $f(x)$ 的近似表达式取为拉格朗日插值多项式, 然后对其积分所建立的, 即取拉格朗日插值多项式的积分作为函数 $f(x)$ 积分的近似值. 事实上, 对下式两边积分:

$$\begin{aligned} f(x) &= L_n(x) + R_n(x) \\ &= \sum_{i=0}^n l_i(x) f(x_i) + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-x_0)(x-x_1) \cdots (x-x_n) \end{aligned}$$

得

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b L_n(x) dx + \int_a^b R_n(x) dx \\ &= \sum_{i=0}^n \left[\int_a^b l_i(x) dx \right] f(x_i) + \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-x_0)(x-x_1) \cdots (x-x_n) dx. \end{aligned}$$

取

$$\begin{aligned} I[f] &= \int_a^b f(x) dx \\ &\approx \sum_{i=0}^n \left[\int_a^b l_i(x) dx \right] f(x_i) = \sum_{i=0}^n A_i f(x_i) = Q[f], \end{aligned} \quad (6.1.1)$$

其中

$$\begin{aligned} A_i &= \int_a^b l_i(x) dx \\ &= \int_a^b \frac{(x-x_0)(x-x_1) \cdots (x-x_{i-1})(x-x_{i+1}) \cdots (x-x_n)}{(x_i-x_0)(x_i-x_1) \cdots (x_i-x_{i-1})(x_i-x_{i+1}) \cdots (x_i-x_n)} dx. \end{aligned} \quad (6.1.2)$$

$$\begin{aligned} R[f] &= I[f] - Q[f] = \int_a^b f(x) dx - \int_a^b L_n(x) dx \\ &= \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-x_0)(x-x_1) \cdots (x-x_n) dx. \end{aligned} \quad (6.1.3)$$

x_i 称为求积节点, A_i 称为求积系数, $R[f]$ 称为截断误差或积分余项. 求积公式 (6.1.1) 称为插值型求积公式. 当求积节点在区间 $[a, b]$ 上等距分布时, 求积公式 (6.1.1) 称为牛顿-科茨 (Newton-Cotes) 求积公式.

下面建立几个常用的基本数值求积公式.

令 $h = \frac{b-a}{n}$, 取 $x_i = a + ih$ ($i = 0, 1, \dots, n$). 显然, $x_0 = a, x_n = b$. 现在用换元积分法计算积分常数 A_i .

令 $x = x_0 + th$, 则

$$x - x_i = x_0 + th - (x_0 + ih) = (t - i)h,$$

$$x_i - x_j = x_0 + ih - (x_0 + jh) = (i - j)h,$$

从而

$$\begin{aligned} & (x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n) \\ &= t(t-1) \cdots (t-i+1)(t-i-1) \cdots (t-n)h^n, \\ & (x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n) \\ &= i(i-1) \cdots 2 \times 1 \times (-1)(-2) \cdots [-(n-i)]h^n \\ &= (-1)^{n-i} i! \times (n-i)! \times h^n. \end{aligned}$$

将以上两式代入式 (6.1.2) 得

$$\begin{aligned} A_i &= \int_a^b \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} dx \\ &\stackrel{x=x_0+th}{=} \frac{(-1)^{n-i} h}{i! \times (n-i)!} \int_0^n t(t-1)(t-i+1)(t-i-1) \cdots (t-n) dt. \end{aligned} \quad (6.1.4)$$

由式 (6.1.4) 可得三个最基本的数值积分公式.

(1) 梯形求积公式 ($n = 1$). 令

$$x_0 = a, \quad x_1 = b, \quad h = b - a,$$

则

$$A_0 = -h \int_0^1 (t-1) dt = \frac{h}{2} = \frac{b-a}{2}, \quad A_1 = h \int_0^1 t dt = \frac{h}{2} = \frac{b-a}{2}.$$

由式 (6.1.1) 得

$$\int_a^b f(x) dx \approx Q(f) = \frac{b-a}{2} [f(a) + f(b)]. \quad (6.1.5)$$

为导出梯形求积公式的截断误差, 需用高等数学中的广义积分中值定理.

定理 6.1.1 (广义积分中值定理) 设函数 $f(x), g(x)$ 在区间 $[a, b]$ 上连续, $g(x)$ 在区间 $[a, b]$ 上不变号, 则存在一点 $\eta \in [a, b]$, 使得下式成立:

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

由式 (6.1.3), 利用广义积分中值定理, 可得梯形求积公式 (6.1.5) 的截断误差为

$$\begin{aligned} R_1[f] &= \int_a^b \frac{f''(\xi)}{2!} (x-a)(x-b) dx = \frac{f''(\eta)}{2} \int_a^b (x-a)(x-b) dx \\ &= -\frac{(b-a)^3}{12} f''(\eta), \quad a \leq \eta \leq b. \end{aligned} \quad (6.1.6)$$

(2) 辛普森 (Simpson) 求积公式 ($n=2$). 令

$$h = \frac{b-a}{2}, \quad x_0 = a, \quad x_1 = a+h = \frac{a+b}{2}, \quad x_2 = b.$$

由式 (6.1.1), (6.1.4) 得

$$\begin{aligned} \int_a^b f(x) dx &\approx Q[f] = \frac{h}{3} [f(x_0) + 4f(x_1) + f(x_2)] \\ &= \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]. \end{aligned} \quad (6.1.7)$$

辛普森求积公式的截断误差为

$$R_2[f] = -\frac{1}{90} h^5 f^{(4)}(\eta) = -\frac{(b-a)^5}{2880} f^{(4)}(\eta), \quad a \leq \eta \leq b. \quad (6.1.8)$$

注意: 导出辛普森求积公式的截断误差时不能利用广义积分中值定理, 证明见例 6.6.

(3) 科茨 (Cotes) 求积公式 ($n=4$). 令

$$h = \frac{b-a}{4}, \quad x_i = a + ih, \quad i = 0, 1, \dots, 4, \quad x_0 = a, \quad x_4 = b.$$

同样由式 (6.1.1), (6.1.4) 得

$$\int_a^b f(x) dx \approx Q[f] = \frac{b-a}{90} [7f(a) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(b)]. \quad (6.1.9)$$

科茨求积公式的截断误差为

$$R_4[f] = -\frac{8}{945} h^7 f^{(6)}(\eta) = -\frac{(b-a)^7}{1935360} f^{(6)}(\eta), \quad a \leq \eta \leq b. \quad (6.1.10)$$

6.1.3 复化求积公式

由插值多项式的龙格现象知, 高次插值多项式会产生很大的误差. 这说明通过增加插值节点构造高次插值多项式对其积分的方法不可能提高数值积分的精度. 提高精度的一种方法是把积分区间分成若干个 (如 n 个) 长度相等的子区间, 即令

$$h = \frac{b-a}{n}, \quad x_i = a + ih, \quad i = 0, 1, \dots, n.$$

在每个子区间 $[x_{i-1}, x_i]$ 上分别利用前述的三个基本求积公式, 将所得结果相加, 得到的求积公式称为复化求积公式.

1. 复化梯形求积公式

由式 (6.1.5) 得

$$\begin{aligned} I[f] &= \int_a^b f(x) dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) dx \approx \sum_{i=1}^n \frac{x_i - x_{i-1}}{2} [f(x_{i-1}) + f(x_i)] \\ &= \frac{h}{2} \left[f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + f(b) \right] \triangleq T_n. \end{aligned} \quad (6.1.11)$$

由式 (6.1.6) 知, 复化梯形求积公式的截断误差为

$$R_{T_n}[f] = I[f] - T_n = -\frac{h^3}{12} [f''(\eta_1) + f''(\eta_2) + \cdots + f''(\eta_n)], \quad (6.1.12)$$

其中 $\eta_i \in [x_{i-1}, x_i]$ ($i = 1, 2, \cdots, n$). 下面化简式 (6.1.12). 设 $f''(x)$ 在区间 $[a, b]$ 上连续, 则存在常数 $m < M$, 使得 $m \leq f''(x) \leq M$. 由此可得

$$nm \leq f''(\eta_1) + f''(\eta_2) + \cdots + f''(\eta_n) \leq nM,$$

即

$$m \leq \frac{f''(\eta_1) + f''(\eta_2) + \cdots + f''(\eta_n)}{n} \leq M,$$

从而由连续函数的介值定理知, 存在一点 $\eta \in [a, b]$, 使得

$$f''(\eta) = \frac{f''(\eta_1) + f''(\eta_2) + \cdots + f''(\eta_n)}{n}.$$

将上式代入式 (6.1.12), 注意到 $nh = b - a$, 则得复化梯形求积公式 (6.1.11) 的截断误差为

$$R_{T_n}[f] = -\frac{h^3}{12} n f''(\eta) = -\frac{b-a}{12} h^2 f''(\eta), \quad \eta \in [a, b]. \quad (6.1.13)$$

2. 复化辛普森求积公式

由式 (6.1.7) 得

$$\begin{aligned} \int_a^b f(x) dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) dx \\ &\approx \sum_{i=1}^n \frac{x_i - x_{i-1}}{6} \left[f(x_{i-1}) + 4f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i) \right] \\ &= \frac{h}{6} \left[f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + 4 \sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right) + f(b) \right] \triangleq S_n. \end{aligned} \quad (6.1.14)$$

设 $f^{(4)}(x)$ 在区间 $[a, b]$ 上连续, 根据式 (6.1.8), 同理可得复化辛普森求积公式的截断误差为

$$R_{S_n}[f] = -\frac{h^5}{2880} n f^{(4)}(\eta) = -\frac{b-a}{2880} h^4 f^{(4)}(\eta), \quad a \leq \eta \leq b. \quad (6.1.15)$$

3. 复化科茨求积公式

由式 (6.1.9) 可得

$$\begin{aligned}
 \int_a^b f(x) dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) dx \\
 &\approx \sum_{i=1}^n \frac{x_i - x_{i-1}}{90} \left[7f(x_{i-1}) + 32f\left(x_{i-1} + \frac{h}{4}\right) \right. \\
 &\quad \left. + 12f\left(x_{i-1} + \frac{h}{2}\right) + 32f\left(x_{i-1} + \frac{3h}{4}\right) + 7f(x_i) \right] \\
 &= \frac{h}{90} \left[7f(a) + 14 \sum_{i=1}^{n-1} f(x_i) + 12 \sum_{i=0}^{n-1} f\left(x_i + \frac{h}{2}\right) \right. \\
 &\quad \left. + 32 \sum_{i=0}^{n-1} f\left(x_i + \frac{h}{4}\right) + 32 \sum_{i=0}^{n-1} f\left(x_i + \frac{3h}{4}\right) + 7f(b) \right] \\
 &\triangleq C_n.
 \end{aligned}$$

设 $f^{(6)}(x)$ 在区间 $[a, b]$ 上连续, 根据式 (6.1.10), 同理可得复化科茨求积公式的截断误差为

$$R_{C_n}[f] = -\frac{h^7}{1935360} n f^{(6)}(\eta) = -\frac{b-a}{1935360} h^6 f^{(6)}(\eta), \quad a \leq \eta \leq b. \quad (6.1.16)$$

例 6.1.1 利用复化梯形求积公式计算积分

$$I[f] = \int_0^1 \frac{\sin x}{x} dx$$

使误差不超过 $\frac{1}{2} \times 10^{-3}$. 取相同的步长 h , 用复化辛普森计算, 给出结果和截断误差限.

解 首先由所给误差限确定子区间的长度 h . 由 $f(x) = \frac{\sin x}{x} = \int_0^1 \cos(tx) dt$ 知

$$\begin{aligned}
 f^{(k)}(x) &= \int_0^1 \frac{d^k(\cos(tx))}{dx^k} dt = \int_0^1 t^k \cos\left(tx + \frac{k\pi}{2}\right) dt, \\
 \max_{0 \leq x \leq 1} |f^{(k)}(x)| &\leq \int_0^1 t^k \left| \cos\left(tx + \frac{k\pi}{2}\right) \right| dt \leq \int_0^1 t^k dt = \frac{1}{k+1}.
 \end{aligned}$$

由复化梯形求积公式的截断误差 (6.1.13) 知, 要求选取 h 满足

$$|R_{T_n}[f]| = \frac{1}{12} h^2 |f''(\eta)| \leq \frac{h^2}{12} \times \frac{1}{3} \leq \frac{1}{2} \times 10^{-3} \Rightarrow h^2 \leq 18 \times 10^{-3}.$$

取 $h = \frac{1}{8} = 0.125$ 即满足要求, 而 $n = \frac{b-a}{h} = \frac{1}{h} = 8$, $x_i = 0.125i$ ($i = 0, 1, 2, \dots, 8$).

由式 (6.1.11) 有

$$I[f] \approx T_8 = \frac{1}{16} \left[f(0) + 2 \sum_{i=1}^7 f(0.125i) + f(1) \right] = 0.945\,690\,9,$$

其中定义 $f(0) = 1$.

取同样步长 $h = 0.125$, 则 $\frac{x_{i-1} + x_i}{2} = \left(i - \frac{1}{2}\right)h = 0.125(i - 0.5)$, 由式 (6.1.14) 得

$$\begin{aligned} I[f] \approx S_8 &= \frac{1}{48} \left[f(0) + 2 \sum_{i=1}^7 f(0.125i) + 4 \sum_{i=1}^8 f(0.125i - 0.0625) + f(1) \right] \\ &= 0.946\,083\,3. \end{aligned}$$

由式 (6.1.15) 知, 截断误差限为

$$|R_{S_8}[f]| \leq \frac{1}{2880} \times \left(\frac{1}{8}\right)^4 \times \frac{1}{5} \approx 0.17 \times 10^{-7}.$$

解毕.

6.1.4 变步长积分法

采用复化求积公式计算时, 为使截断误差不超过 ε , 需要估计被积函数高阶导数的最大值, 从而确定把积分区间 $[a, b]$ 分成等长子区间的个数 n , 但高阶导数的最大值很难求得. 然而, 由误差表达式可见, 只要被积函数的高阶导数有界, 则当 $h \rightarrow 0$ 时, 误差趋于零. 这说明可以采用逐步缩小步长 h 的方法, 以使积分的近似值满足精度要求, 这就是变步长积分法. 在实际使用中, 它是通过将区间逐次分半来实现的, 下面以复化梯形求积公式为例来说明变步长积分法的思想.

对于将积分区间 $[a, b]$ 分成 n 个等长子区间的复化梯形求积公式有

$$I[f] - T_n = -\frac{b-a}{12} h^2 f''(\eta), \quad a \leq \eta \leq b.$$

将区间 $[a, b]$ 分为 $2n$ 等份, 则有

$$I[f] - T_{2n} = -\frac{b-a}{12} \left(\frac{h}{2}\right)^2 f''(\eta_1), \quad a \leq \eta_1 \leq b.$$

假定 $f''(x)$ 在区间 $[a, b]$ 上连续且变化不大, 可以认为 $f''(\eta) \approx f''(\eta_1)$. 将以上两式相除得

$$\frac{I[f] - T_n}{I[f] - T_{2n}} \approx 4.$$

由此可得

$$I[f] \approx T_{2n} + \frac{1}{3}(T_{2n} - T_n). \quad (6.1.17)$$

故有

$$|I[f] - T_{2n}| \approx \frac{1}{3}|T_{2n} - T_n|.$$

若 $\frac{1}{3}|T_{2n} - T_n| \leq \varepsilon$, 取 $I[f] \approx T_{2n}$, 则大致满足精度要求. 实际计算时, 常用

$$|T_{2n} - T_n| \leq \varepsilon$$

作为判别计算终止的条件. 若满足, 则取 $I[f] \approx T_{2n}$; 否则, 将区间再分半进行计算, 直至满足精度要求.

为减少计算量, 在计算 T_{2n} 时, 可以利用 T_n 的结果. 下面导出 T_{2n} 与 T_n 的关系. 令 $h = (b-a)/n$, $x_i = a + ih$ ($i = 0, 1, \dots, n$), 则

$$\begin{aligned} T_n &= \frac{h}{2} \left[f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + f(b) \right], \\ T_{2n} &= \frac{h}{4} \left[f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + 2 \sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right) + f(b) \right] \\ &= \frac{1}{2}T_n + \frac{h}{2} \sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right). \end{aligned} \quad (6.1.18)$$

在实际计算时, 取 $n = 2^k$, 则 $h = \frac{b-a}{2^k}$,

$$\frac{x_{i-1} + x_i}{2} = \frac{a + (i-1)h + a + ih}{2} = a + (2i-1)\frac{h}{2} = a + (2i-1)\frac{b-a}{2^{k+1}}.$$

式 (6.1.18) 可以写为

$$T_{2^{k+1}} = \frac{1}{2}T_{2^k} + \frac{b-a}{2^{k+1}} \sum_{i=1}^{2^k} f\left(a + (2i-1)\frac{b-a}{2^{k+1}}\right). \quad (6.1.19)$$

开始计算时取

$$T_1 = \frac{b-a}{2}[f(a) + f(b)],$$

然后按式 (6.1.19) 计算. 若

$$|T_{2^{k+1}} - T_{2^k}| \leq \varepsilon,$$

则取 $I[f] \approx T_{2^{k+1}}$; 否则, 继续计算, 直到满足 $|T_{2^{k+1}} - T_{2^k}| \leq \varepsilon$ 为止.

6.1.5 龙贝格积分法

对于复化梯形求积公式而言, 由式 (6.1.17) 知, 可取积分的近似值为

$$I[f] \approx T_{2n} + \frac{1}{4-1}(T_{2n} - T_n) = \bar{T}_{2n}. \quad (6.1.20)$$

对于复化辛普森求积公式有

$$\begin{aligned} I[f] - S_n &= -\frac{b-a}{2880} h^4 f^{(4)}(\eta), \quad a \leq \eta \leq b, \\ I[f] - S_{2n} &= -\frac{b-a}{2880} \left(\frac{h}{2}\right)^4 f^{(4)}(\eta_1), \quad a \leq \eta_1 \leq b. \end{aligned}$$

假定 $f^{(4)}(x)$ 在区间 $[a, b]$ 上连续且变化不大, 可以认为 $f^{(4)}(\eta) \approx f^{(4)}(\eta_1)$. 将以上两式相除得

$$\frac{I[f] - S_n}{I[f] - S_{2n}} \approx 16.$$

由此解得

$$I[f] \approx S_{2n} + \frac{1}{4^2 - 1} (S_{2n} - S_n) \triangleq \bar{S}_{2n}. \quad (6.1.21)$$

同理, 对复化科茨求积公式有

$$I[f] \approx C_{2n} + \frac{1}{4^3 - 1} (C_{2n} - C_n) \triangleq \bar{C}_{2n}. \quad (6.1.22)$$

下面对式 (6.1.20)~(6.1.22) 作进一步的分析, 看一看它们的实质和它们之间的关系.

$$\begin{aligned} \bar{T}_{2n} &= T_{2n} + \frac{1}{3} (T_{2n} - T_n) = \frac{1}{3} (4T_{2n} - T_n) \\ &= \frac{1}{3} \left\{ 4 \left[\frac{1}{2} T_n + \frac{h}{2} \sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right) \right] - T_n \right\} \\ &= \frac{1}{3} \left\{ T_n + 2h \sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right) \right\} \\ &= \frac{1}{3} \left\{ \frac{h}{2} \left[f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + f(b) \right] + 2h \sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right) \right\} \\ &= \frac{h}{6} \left[f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + 4 \sum_{i=1}^n f\left(\frac{x_{i-1} + x_i}{2}\right) + f(b) \right] = S_n, \end{aligned}$$

即

$$S_n = T_{2n} + \frac{1}{4 - 1} (T_{2n} - T_n). \quad (6.1.23)$$

同理可得

$$C_n = S_{2n} + \frac{1}{4^2 - 1} (S_{2n} - S_n). \quad (6.1.24)$$

令

$$R_n = C_{2n} + \frac{1}{4^3 - 1} (C_{2n} - C_n), \quad (6.1.25)$$

式 (6.1.25) 称为龙贝格 (Romberg) 积分公式. 显然, 式 (6.1.23)~(6.1.25) 的计算最终归结为复化梯形求积公式的计算. 注意到梯形求积公式、辛普森求积公式、科茨求积公式的截断误差分别是 $c_T h^2 f''(\eta)$, $c_S h^4 f^{(4)}(\eta)$, $c_C h^6 f^{(6)}(\eta)$. 由 (6.1.23), (6.1.24) 可知, 龙贝格积分公式 (6.1.25) 的截断误差为 $c_R h^8 f^{(8)}(\eta)$. 依此类推可知, 若用 $R_{2n} + \frac{1}{4^4 - 1}(R_{2n} - R_n)$ 作为积分的近似值, 其截断误差应是 $ch^{10} f^{(10)}(\eta)$. 读者自然会设想, 取

$$R_{2n} + \frac{1}{4^4 - 1}(R_{2n} - R_n)$$

作为积分的近似值可能更精确. 事实上并非完全如此, 这是因为被积函数 $f(x)$ 的高阶导数性态很难估计, 特别是高阶导数性态不好的函数, $f^{(10)}(x)$ 的数量级可能很大, 虽然 h^{10} 很小, 但截断误差项 $ch^{10} f^{(10)}(\eta)$ 可能很大. 如此计算的效果可能并不理想. 再者, 一般地, $R_{2n} - R_n$ 是两个相近的数相减, 可能产生较大的误差. 因此, 一般不再继续进行外推, 计算到龙贝格积分公式 (6.1.25) 就已足够准确. 如果积分区间 $[a, b]$ 较大, 则最好事先把 $[a, b]$ 分成若干个子区间, 然后在每个子区间上再进行数值积分.

龙贝格积分法是将积分区间 $[a, b]$ 逐次分半, 利用式 (6.1.19) 逐次递推计算, 再用式 (6.1.23)~(6.1.25) 逐次计算, 以得到较高精度的积分近似值. 龙贝格积分法的计算公式如下:

$$T_1 = \frac{b-a}{2}[f(a) + f(b)],$$

$$T_{2^{k+1}} = \frac{1}{2}T_{2^k} + \frac{b-a}{2^{k+1}} \sum_{i=1}^{2^k} f\left(a + (2i-1)\frac{b-a}{2^{k+1}}\right), \quad k = 0, 1, 2, \dots,$$

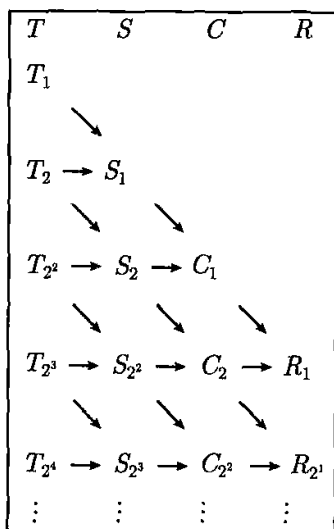


图 6.1

$$\begin{cases} S_{2^k} = T_{2^{k+1}} + \frac{1}{4-1}(T_{2^{k+1}} - T_{2^k}), \\ C_{2^k} = S_{2^{k+1}} + \frac{1}{4^2-1}(S_{2^{k+1}} - S_{2^k}), \\ R_{2^k} = C_{2^{k+1}} + \frac{1}{4^3-1}(C_{2^{k+1}} - C_{2^k}), \end{cases} \quad k = 0, 1, 2, \dots$$

计算过程按图 6.1 所示的顺序进行.

若 $|R_{2^{k+1}} - R_{2^k}| < \varepsilon$, 则取 $I[f] \approx R_{2^{k+1}}$; 否则, 继续计算, 直到满足精度要求为止.

例 6.1.2 用龙贝格积分法计算例 6.1.1 中的积分, 使误差不超过 10^{-12} .

解 利用龙贝格积分法计算, 结果如表 6.1 所示.

表 6.1

$T_{2k}(k=0,1,\cdots,5)$	S_{2k}	C_{2k}	R_{2k}
0.920 735 492 404			
0.939 793 284 806	0.946 145 882 273		
0.944 513 521 665	0.946 086 933 951	0.946 083 004 063	
0.945 690 863 583	0.946 083 310 889	0.946 083 069 352	0.946 083 070 388
0.945 985 029 934	0.946 083 085 384	0.946 083 070 350	0.946 083 070 366
0.946 058 560 963	0.946 083 071 306	0.946 083 070 367	0.946 083 070 367

由于

$$|0.946\,083\,070\,367 - 0.946\,083\,070\,366| = 10^{-12},$$

所以取 $I[f] \approx 0.946\,083\,070\,367$. 解毕.

6.2 待定系数法与高斯型求积公式

本节研究带权函数积分

$$I[f] = \int_a^b \omega(x)f(x) dx \quad (6.2.1)$$

的数值积分方法, 其中权函数 $\omega(x) \geq 0$, 它只在个别点处的函数值等于零. 当 $\omega(x) \equiv 1$ 时, 就是通常的积分, 以后若不特别指出, 就默认为 $\omega(x) \equiv 1$.

6.2.1 代数精度与待定系数法

用数值积分法计算积分, 有近似式 $I[f] \approx Q[f]$, 希望求积公式对尽可能多的被积函数 $f(x)$ 准确成立. 由插值型求积公式的截断误差估计式 (6.1.3) 知, 对于次数不超过 n 的多项式 $f(x)$, $f^{(n+1)}(x) = 0$, 则误差项 $R[f] = 0$, 即 $I[f] = Q[f]$. 下面给出一个评价近似式优劣的标准——代数精度的定义.

定义 6.2.1 (代数精度) 设某一近似式 (如数值积分公式 $I[f] \approx Q[f]$) 的截断误差为 $R[f]$, 如果对任意不超过 m 次的多项式 $f(x)$ 都有 $R[f] = 0$, 而当 $f(x)$ 是 $m+1$ 次多项式时 $R[f] \neq 0$, 则称该近似式的代数精度为 m .

对于数值积分公式, 由定义 6.2.1 知, 梯形求积公式的代数精度 $m=1$, 辛普森求积公式的代数精度 $m=3$, 科茨求积式的代数精度 $m=5$. 由式 (6.1.3) 知, 具有 $n+1$ 个插值节点的插值型求积公式 (6.1.1), 其代数精度 $m=n$. 显然, 代数精度越高的求积公式, 计算结果越准确.

由于多项式是由函数 $1, x^2, \cdots, x^m, \cdots$ 线性组合而成的, 所以近似式代数精度为 m 的充要条件是

$$R[x^k] = 0, \quad k = 0, 1, 2, \cdots, m, \quad R[x^{m+1}] \neq 0.$$

由此可知, 求积公式 (6.1.1) 中的求积系数 A_i 满足方程组

$$\sum_{i=0}^n A_i x_i^k = \int_a^b x^k dx = \frac{b^{k+1} - a^{k+1}}{k+1}, \quad k = 0, 1, 2, \dots, n. \quad (6.2.2)$$

在式 (6.2.2) 中取 $k = 0$, 则有

$$\sum_{i=0}^n A_i = b - a.$$

该式可以用来检验求积系数是否准确.

当节点 x_i 给定后, 通过解方程组

$$R[x^k] = 0, \quad k = 0, 1, \dots, m$$

(注意: 方程的个数应等于未知量 A_i 的个数) 求出求积系数 A_i , 由此得到求积公式的方法称为待定系数法.

例 6.2.1 确定求积公式

$$\int_{-2h}^{2h} f(x) dx \approx A_0 f(-h) + A_1 f(0) + A_2 f(h),$$

使其具有尽可能高的代数精度.

解 取 $f(x) = 1, x, x^2$, 令 $R[x^k] = 0$ ($k = 0, 1, 2$), 于是得方程组

$$\begin{cases} A_0 + A_1 + A_2 = 4h, \\ -A_0 h + A_2 h = 0, \\ A_0 h^2 + A_2 h^2 = \frac{16}{3} h^3. \end{cases}$$

解之得

$$A_0 = A_2 = \frac{8h}{3}, \quad A_1 = -\frac{4h}{3},$$

所以

$$\int_{-2h}^{2h} f(x) dx \approx \frac{4h}{3} [2f(-h) - f(0) + 2f(h)].$$

令 $f(x) = x^3$, 显然, $R[x^3] = 0$. 再令 $f(x) = x^4$, 则

$$I[x^4] = \int_{-2h}^{2h} x^4 dx = \frac{64h^5}{5}, \quad Q[x^4] = \frac{4h}{3} [2(-h)^4 + 2(h)^4] = \frac{16h^5}{3},$$

$$R[x^4] = I[x^4] - Q[x^4] = \frac{112}{15} h^5 \neq 0,$$

故该求积公式的代数精度 $m = 3$. 解毕.

例 6.2.2 确定求积公式

$$\int_a^b f(x) dx \approx A_0 f(a) + A_1 f(b) + A_2 f'(a),$$

使其具有尽可能高的代数精度.

注意: 当 $f(x)$ 的近似表达式取为埃尔米特插值多项式时, 求积公式中含有带导数的项.

解 为计算简单起见, 不妨假设 $a = 0, b = h$ (相当于作坐标平移). 取 $f(x) = 1, x, x^2$, 令 $R[x^k] = 0$ ($k = 0, 1, 2$), 得方程组

$$A_0 + A_1 = h, \quad A_1 h + A_2 = \frac{h^2}{2}, \quad A_1 h^2 = \frac{h^3}{3}.$$

解之得

$$A_0 = \frac{2h}{3}, \quad A_1 = \frac{h}{3}, \quad A_2 = \frac{h^2}{6},$$

所以

$$\int_a^b f(x) dx \approx \frac{h}{6} [4f(a) + 2f(b) + hf'(a)],$$

其中 $h = b - a$.

下面确定其代数精度. 令 $f(x) = x^3$, 则

$$I[x^3] = \int_0^h x^3 dx = \frac{h^4}{4}, \quad Q[x^3] = \frac{h}{6} [4f(0) + 2f(h) + hf'(0)] = \frac{h^4}{3},$$

$$R[x^3] = I[x^3] - Q[x^3] = -\frac{h^4}{12} \neq 0,$$

故该求积公式的代数精度 $m = 2$. 解毕.

6.2.2 广义佩亚诺定理

本小节介绍确定近似式截断误差的广义佩亚诺 (Peano) 定理.

一般地, 近似公式的截断误差 $R[f]$ 是函数 f 的线性泛函, 即

$$R[c_1 f_1 + c_2 f_2] = c_1 R[f_1] + c_2 R[f_2], \quad \forall c_1, c_2 \in \mathbf{R}.$$

可用下面的广义佩亚诺定理确定近似式的截断误差.

定理 6.2.1 (广义佩亚诺定理)^① 设近似式的截断误差 $R[f]$ 是区间 $[a, b]$ 上 $m+1$ 阶导数连续的函数 $f(x)$ 的线性泛函, 并且近似式的代数精度为 m , 则

$$R[f(x)] = R[e(x)],$$

^① 广义佩亚诺定理由西安交通大学邓建中教授于 1980 年提出, 1985 年载入《计算方法》, 1990 年在东北大学召开的专业会议上列入全国工科研究生“数值分析”教学大纲.

其中

$$e(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!} (x - \tilde{x}_0)(x - \tilde{x}_1) \cdots (x - \tilde{x}_m),$$

$\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 为区间 $[a, b]$ 上的任意点, ξ 与 $x, \tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 有关且位于 $x, \tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 之间.

证明 设 $p_m(x)$ 是以 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 为插值节点构成的插值多项式, 则

$$f(x) = p_m(x) + e(x),$$

所以

$$R[f] = R[p_m + e] = R[p_m] + R[e] = R[e].$$

证毕.

定理 6.2.1 中的插值节点 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 的取法比较灵活, 可以不同, 也可以相同. 当 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 中某些点相同时, $p_m(x)$ 是埃尔米特插值多项式. 点 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 的灵活选取是该定理的最大优点, 若选择适当, 则可以简化 $R[f]$ 的表达式.

由广义佩亚诺定理知, 数值积分公式的截断误差

$$R[f] = R[e] = I[e] - Q[e],$$

故选取插值节点 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 有以下两个原则:

(1) 为对 $I[e]$ 应用广义积分中值定理, 选取 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 应使

$$(x - \tilde{x}_0)(x - \tilde{x}_1) \cdots (x - \tilde{x}_m)$$

在积分区间 $[a, b]$ 上不变号;

(2) 选取 $\tilde{x}_0, \tilde{x}_1, \cdots, \tilde{x}_m$ 使计算 $Q[e]$ 尽可能简单, 最好使 $Q[e] = 0$.

例 6.2.3 利用广义佩亚诺定理确定例 6.2.1 和例 6.2.2 中数值积分公式的截断误差.

解 例 6.2.1 中数值积分公式的代数精度 $m = 3$, 取

$$e(x) = \frac{f^{(4)}(\xi)}{4!} (x - h)^2 (x + h)^2,$$

则

$$\begin{aligned} R[f] &= R[e] = I[e] - Q[e] \\ &= \int_{-2h}^{2h} \frac{f^{(4)}(\xi)}{4!} (x - h)^2 (x + h)^2 dx - \frac{4h}{3} [2e(-h) - e(0) + 2e(h)] \\ &= \frac{23h^5 f^{(4)}(\eta_1)}{90} + \frac{h^5 f^{(4)}(\xi)}{18} = \frac{14}{45} h^5 f^{(4)}(\eta), \quad -2h \leq \eta \leq 2h. \end{aligned}$$

例 6.2.2 中数值积分公式的代数精度 $m = 2$, 取 $a = 0, b = h$, 选取 $\bar{x}_0 = \bar{x}_1 = 0, \bar{x}_2 = h$, 则

$$e(x) = \frac{f'''(\xi)}{3!} x^2 (x - h).$$

这时

$$\begin{aligned} R[f] &= R[e] = I[e] - Q[e] \\ &= \int_0^h \frac{f'''(\xi)}{3!} x^2 (x - h) dx - \frac{h}{6} [4e(0) + 2e(h) + he'(0)] \\ &= \frac{1}{6} f'''(\eta) \int_0^h x^2 (x - h) dx \\ &= -\frac{h^4}{72} f'''(\eta) = -\frac{(b-a)^4}{72} f'''(\eta), \quad a \leq \eta \leq b. \end{aligned}$$

解毕.

例 6.2.4 利用广义佩亚诺定理确定辛普森求积公式的截断误差.

解 可以验证, 辛普森求积公式的代数精度 $m = 3$, 取

$$e(x) = \frac{f^{(4)}(\xi)}{4!} (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b),$$

则由广义佩亚诺定理,

$$\begin{aligned} R[f] &= R[e] = I[e] - Q[e] \\ &= \int_a^b \frac{f^{(4)}(\xi)}{4!} (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b) dx \\ &\quad - \frac{b-a}{6} \left[e(a) + 4e\left(\frac{a+b}{2}\right) + e(b) \right] \\ &= \frac{f^{(4)}(\eta)}{4!} \int_a^b (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b) dx \\ &= -\frac{(b-a)^5}{2880} f^{(4)}(\eta), \quad a \leq \eta \leq b. \end{aligned}$$

解毕.

6.2.3 高斯型求积公式

与上节类似, 将式 (6.2.1) 中的被积函数 $f(x)$ 用拉格朗日插值多项式 $L_n(x)$ 近似代替, 可得求积公式

$$I[f] = \int_a^b \omega(x) f(x) dx \approx \sum_{i=0}^n A_i f(x_i) = Q[f], \quad (6.2.3)$$

其中求积系数

$$A_i = \int_a^b \omega(x) l_i(x) dx, \quad i = 0, 1, \dots, n, \quad (6.2.4)$$

其截断误差

$$R[f] = I[f] - Q[f] = \int_a^b \omega(x) \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-x_0)(x-x_1)\cdots(x-x_n) dx. \quad (6.2.5)$$

由式 (6.2.5) 知, 数值积分公式 (6.2.3) 的代数精度 $m \geq n$. 若取被积函数

$$f(x) = (x-x_0)^2(x-x_1)^2\cdots(x-x_n)^2,$$

则

$$I[f] = \int_a^b \omega(x)(x-x_0)^2(x-x_1)^2\cdots(x-x_n)^2 dx > 0.$$

而 $Q[f] = \sum_{i=0}^n A_i f(x_i) = 0$. 此时 $I[f] \neq Q[f]$, 即 $R[f] \neq 0$, 故求积公式 (6.2.3) 的代数精度 $n \leq m \leq 2n+1$.

事实上, 若适当地选取插值节点 x_i ($i = 0, 1, \dots, n$), 则求积公式 (6.2.3) 的代数精度可以达到 $2n+1$. 将具有 $n+1$ 个节点的且其代数精度达到 $2n+1$ 的求积公式称为高斯 (Gauss) 型求积公式 (或最高代数精度求积公式), 节点 x_i 称为高斯点.

高斯型求积公式 (6.2.3) 中含有 $2n+2$ 个参数 x_i, A_i ($i = 0, 1, \dots, n$), 由其代数精度 $m = 2n+1$ 知, 这些参数必然满足 $R[x^k] = 0$ ($k = 0, 1, \dots, 2n+1$), 即得以下方程组:

$$\int_a^b \omega(x) x^k dx = \sum_{i=0}^n A_i x_i^k, \quad k = 0, 1, \dots, 2n+1. \quad (6.2.6)$$

由此可见, 可以通过解方程组 (6.2.6) 求出高斯型求积公式中的节点和求积系数 x_i, A_i ($i = 0, 1, \dots, n$), 从而得到高斯型求积公式.

例 6.2.5 确定以下高斯型求积公式及其误差项:

$$\int_{-1}^1 f(x) dx \approx A_0 f(x_0) + A_1 f(x_1).$$

解 这是 $n = 1$ 的高斯型求积公式, 其代数精度 $m = 2n+1 = 3$, 所以求积公式对 $f(x) = 1, x, x^2, x^3$ 准确成立, 由此得到关于未知量 x_0, x_1, A_0, A_1 的方程组

$$\begin{cases} A_0 + A_1 = 2, \\ x_0 A_0 + x_1 A_1 = 0, \\ x_0^2 A_0 + x_1^2 A_1 = \frac{2}{3}, \\ x_0^3 A_0 + x_1^3 A_1 = 0. \end{cases}$$

解之得

$$x_0 = -\frac{\sqrt{3}}{3}, \quad x_1 = \frac{\sqrt{3}}{3}, \quad A_0 = A_1 = 1,$$

所求高斯型求积公式为

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

下面利用广义佩亚诺定理确定误差项. 由于代数精度 $m = 3$, 取

$$e(x) = \frac{f^{(4)}(\xi)}{4!} \left(x + \frac{\sqrt{3}}{3}\right)^2 \left(x - \frac{\sqrt{3}}{3}\right)^2 = \frac{f^{(4)}(\xi)}{24} \left(x^2 - \frac{1}{3}\right)^2,$$

则

$$\begin{aligned} R[f] &= R[e] = I[e] - Q[e] = I[e] = \int_{-1}^1 \frac{f^{(4)}(\xi)}{24} \left(x^2 - \frac{1}{3}\right)^2 dx \\ &= \frac{f^{(4)}(\eta)}{24} \int_{-1}^1 \left(x^2 - \frac{1}{3}\right)^2 dx = \frac{1}{135} f^{(4)}(\eta), \quad -1 \leq \eta \leq 1. \end{aligned}$$

解毕.

方程组 (6.2.6) 是个非线性方程组, 不易求解. 因此, 通过解方程组 (6.2.6) 确定高斯型求积公式存在一定的困难. 下述定理则给出了一个构造高斯型求积公式的方法.

定理 6.2.2 求积公式的代数精度 $m = 2n + 1$ 的充分必要条件是节点 x_i ($i = 0, 1, \dots, n$) 为区间 $[a, b]$ 上关于权函数 $\omega(x)$ 正交的 $n + 1$ 次正交多项式 $g_{n+1}(x)$ 的零点, 并且求积系数由式 (6.2.4) 确定.

证明 由前面的讨论知, 积分公式 (6.2.3) 的代数精度 $n \leq m \leq 2n + 1$.

充分性. 对任意次数不超过 $2n + 1$ 的多项式 $p_{2n+1}(x)$, 设 $p_{2n+1}(x)$ 除以 $g_{n+1}(x)$ 的商为 $q_n(x)$, 余式为 $r_n(x)$, 则

$$p_{2n+1}(x) = g_{n+1}(x)q_n(x) + r_n(x),$$

$$\begin{aligned} R[p_{2n+1}] &= R[g_{n+1}q_n] + R[r_n] \quad (m \geq n, \text{ 则 } R[r_n] = 0) \\ &= I[g_{n+1}q_n] - Q[g_{n+1}q_n] \\ &= \int_a^b \omega(x)g_{n+1}(x)q_n(x) dx - \sum_{i=0}^n A_i g_{n+1}(x_i)q_n(x_i). \end{aligned}$$

由推论 5.1.1 知 $g_n(x)$ 必与 $g_{n+1}(x)$ 正交, 故上式第一项等于零. 又由于 x_i ($i = 0, 1, \dots, n$) 是 $g_{n+1}(x)$ 的零点, 故第二项等于零, 所以

$$R[p_{2n+1}] = 0,$$

即代数精度 $m \geq 2n + 1$. 结合前面的讨论结果 $m \leq 2n + 1$, 所以其代数精度 $m = 2n + 1$.

必要性. 设其代数精度为 $m = 2n + 1$, 则 $R[p_{2n+1}] = 0, R[r_n] = 0$, 于是

$$\begin{aligned} 0 &= R[p_{2n+1}] = R[g_{n+1}q_n] + R[r_n] = I[g_{n+1}q_n] - Q[g_{n+1}q_n] \\ &= \int_a^b \omega(x)g_{n+1}(x)q_n(x) dx - \sum_{i=0}^n A_i g_{n+1}(x_i)q_n(x_i) \\ &= - \sum_{i=0}^n A_i g_{n+1}(x_i)q_n(x_i), \end{aligned}$$

即 $\sum_{i=0}^n A_i g_{n+1}(x_i)q_n(x_i) = 0$. 由 $p_{2n+1}(x)$ 的任意性知, $q_n(x_i)$ 也具有任意性. 因此, $g_{n+1}(x_i) = 0$, 即 x_i 为 $g_{n+1}(x)$ 的零点. 证毕.

为了导出高斯型求积公式中求积系数 A_i 的计算公式, 首先给出以下引理.

引理 6.2.1 设 $\{g_k(x)\}$ 是区间 $[a, b]$ 上关于权函数 $\omega(x)$ 正交的、最高次项系数为 1 的 k 次正交多项式, x_i ($i = 0, 1, \dots, n$) 是 $g_{n+1}(x)$ 的零点, 则下式成立:

$$\frac{g_{n+1}(x)}{x - x_i} = \frac{\gamma_n}{g_n(x_i)} \sum_{k=0}^n \frac{g_k(x)g_k(x_i)}{\gamma_k}. \quad (6.2.7)$$

证明 由正交多项式的三项递推关系知

$$g_{k+1}(x) = (x - b_k)g_k(x) - \frac{\gamma_k}{\gamma_{k-1}}g_{k-1}(x),$$

$$g_{k+1}(x_i) = (x_i - b_k)g_k(x_i) - \frac{\gamma_k}{\gamma_{k-1}}g_{k-1}(x_i),$$

对以上两式分别乘以 $g_k(x_i)$ 和 $g_k(x)$, 然后相减得

$$\begin{aligned} &g_{k+1}(x)g_k(x_i) - g_{k+1}(x_i)g_k(x) \\ &= (x - x_i)g_k(x)g_k(x_i) - \frac{\gamma_k}{\gamma_{k-1}}[g_{k-1}(x)g_k(x_i) - g_{k-1}(x_i)g_k(x)], \end{aligned}$$

即

$$\begin{aligned} \frac{(x - x_i)g_k(x)g_k(x_i)}{\gamma_k} &= \frac{g_{k+1}(x)g_k(x_i) - g_{k+1}(x_i)g_k(x)}{\gamma_k} \\ &\quad - \frac{g_k(x)g_{k-1}(x_i) - g_k(x_i)g_{k-1}(x)}{\gamma_{k-1}}. \end{aligned}$$

对上式两边关于 k 求和得

$$(x - x_i) \sum_{k=1}^n \frac{g_k(x)g_k(x_i)}{\gamma_k}$$

$$\begin{aligned}
&= \sum_{k=1}^n \frac{g_{k+1}(x)g_k(x_i) - g_{k+1}(x_i)g_k(x)}{\gamma_k} - \sum_{k=1}^n \frac{g_k(x)g_{k-1}(x_i) - g_k(x_i)g_{k-1}(x)}{\gamma_{k-1}} \\
&= \frac{g_{n+1}(x)g_n(x_i) - g_{n+1}(x_i)g_n(x)}{\gamma_n} - \frac{g_1(x)g_0(x_i) - g_1(x_i)g_0(x)}{\gamma_0}.
\end{aligned}$$

注意到 $g_{n+1}(x_i) = 0$, $g_0(x) = 1$, $g_1(x) = x - \beta_0/\gamma_0$, 上式可写为

$$(x - x_i) \sum_{k=1}^n \frac{g_k(x)g_k(x_i)}{\gamma_k} = \frac{g_{n+1}(x)g_n(x_i)}{\gamma_n} - \frac{x - x_i}{\gamma_0}.$$

再注意到 $g_0(x)g_0(x_i) = 1$, 则上式为

$$(x - x_i) \sum_{k=0}^n \frac{g_k(x)g_k(x_i)}{\gamma_k} = \frac{g_{n+1}(x)g_n(x_i)}{\gamma_n},$$

即

$$\frac{g_{n+1}(x)}{(x - x_i)} = \frac{\gamma_n}{g_n(x_i)} \sum_{k=0}^n \frac{g_k(x)g_k(x_i)}{\gamma_k}.$$

证毕.

定理 6.2.3 在引理 6.2.1 的条件下, 高斯型求积公式的求积系数可以表示为

$$A_i = \frac{\gamma_n}{g'_{n+1}(x_i)g_n(x_i)}, \quad i = 0, 1, \dots, n. \quad (6.2.8)$$

证明 式 (6.2.4) 即

$$A_i = \int_a^b \omega(x)l_i(x) dx, \quad i = 0, 1, \dots, n,$$

其中

$$\begin{aligned}
l_i(x) &= \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} \\
&= \frac{\pi_{n+1}(x)}{(x - x_i)\pi'_{n+1}(x_i)}.
\end{aligned}$$

注意到 x_i ($i = 0, 1, \dots, n$) 是 $g_{n+1}(x)$ 的零点, 故 $l_i(x) = \frac{g_{n+1}(x)}{(x - x_i)g'_{n+1}(x_i)}$. 由式 (6.2.7) 得

$$\begin{aligned}
A_i &= \int_a^b \omega(x)l_i(x) dx = \frac{1}{g'_{n+1}(x_i)} \int_a^b \omega(x) \frac{g_{n+1}(x)}{(x - x_i)} dx \\
&= \frac{\gamma_n}{g'_{n+1}(x_i)g_n(x_i)} \sum_{k=0}^n \frac{g_k(x_i)}{\gamma_k} \int_a^b \omega(x)g_k(x)g_0(x) dx \\
&= \frac{\gamma_n}{g'_{n+1}(x_i)g_n(x_i)} \quad (\text{注意 } g_0(x) = 1).
\end{aligned}$$

证毕.

下面给出高斯型求积公式的截断误差.

定理 6.2.4 设被积函数 $f(x) \in C^{2n+2}$, 在引理 6.2.1 的条件下, 高斯型求积公式的截断误差为

$$R[f] = \frac{\gamma_{n+1}}{(2n+2)!} f^{(2n+2)}(\eta), \quad a \leq \eta \leq b. \quad (6.2.9)$$

证明 由于高斯型求积公式的代数精度 $m = 2n+1$, 根据广义佩亚诺定理, 取

$$e(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} (x-x_0)^2 (x-x_1)^2 \cdots (x-x_n)^2,$$

则

$$\begin{aligned} R[f] &= R[e] = I[e] - Q[e] = I[e] \\ &= \int_a^b \omega(x) \frac{f^{(2n+2)}(\xi)}{(2n+2)!} (x-x_0)^2 (x-x_1)^2 \cdots (x-x_n)^2 dx \\ &= \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \int_a^b \omega(x) g_{n+1}^2(x) dx \\ &= \frac{\gamma_{n+1}}{(2n+2)!} f^{(2n+2)}(\eta), \quad a \leq \eta \leq b. \end{aligned}$$

证毕.

对于最高次项系数为 a_k 的正交多项式 $\{g_k(x)\}$, 由式 (6.2.8) 和 (6.2.9) 可得高斯型求积公式的求积系数和截断误差分别为

$$A_i = \frac{a_{n+1}}{a_n} \frac{\gamma_n}{g'_{n+1}(x_i) g_n(x_i)}, \quad i = 0, 1, \cdots, n, \quad (6.2.10)$$

$$R[f] = \frac{\gamma_{n+1}}{a_{n+1}^2 (2n+2)!} f^{(2n+2)}(\eta), \quad a \leq \eta \leq b. \quad (6.2.11)$$

定理 6.2.5 高斯型求积公式的求积系数大于零.

证明 由于高斯型求积公式的代数精度 $m = 2n+1$, 它对不超过 $2n+1$ 次的多项式准确成立. 分别取 $f(x) = l_k^2(x)$ ($k = 0, 1, \cdots, n$), 其中 $l_k(x)$ 为 n 次拉格朗日插值基函数, 于是有

$$0 < \int_a^b \omega(x) l_k^2(x) dx = \sum_{i=0}^n A_i l_k^2(x_i) = A_k, \quad k = 0, 1, \cdots, n.$$

证毕.

6.2.4 常用的 4 种高斯型求积公式

综上所述, 高斯型求积公式 (6.2.3) 和求积系数 (6.2.10) 及截断误差 (6.2.11) 如下:

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

$$A_i = \frac{a_{n+1}}{a_n} \frac{\gamma_n}{g'_{n+1}(x_i) g_n(x_i)}, \quad i = 0, 1, \dots, n,$$

$$R[f] = \frac{\gamma_{n+1}}{a_{n+1}^2 (2n+2)!} f^{(2n+2)}(\eta), \quad a \leq \eta \leq b,$$

1. 高斯-勒让德 (Gauss-Legendre) 求积公式

勒让德多项式 $p_n(x)$ 是在区间 $[-1, 1]$ 上关于权函数 $\omega(x) \equiv 1$ 正交的正交多项式. $p_n(x)$ 最高次项 x^n 的系数 $a_n = \frac{(2n)!}{2^n (n!)^2}$, 内积 $\gamma_n = (p_n, p_n) = \frac{2}{2n+1}$. 取节点 x_i ($i = 0, 1, \dots, n$) 为勒让德多项式 $p_{n+1}(x)$ 的零点, 则由式 (6.2.3), (6.2.10), (6.2.11) 可得高斯-勒让德求积公式和求积系数及截断误差分别为

$$\int_{-1}^1 f(x) dx \approx \sum_{i=0}^n A_i f(x_i), \quad (6.2.12)$$

$$A_i = \frac{2}{(n+1)p'_{n+1}(x_i)p_n(x_i)}, \quad (6.2.13)$$

$$R[f] = \frac{2^{2n+3} [(n+1)!]^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\eta), \quad -1 \leq \eta \leq 1. \quad (6.2.14)$$

对于一般区间 $[a, b]$ 上的积分 $\int_a^b f(x) dx$, 通过变量代换

$$x = \frac{a+b}{2} + \frac{b-a}{2}t,$$

化为区间 $[-1, 1]$ 上的积分

$$\frac{b-a}{2} \int_{-1}^1 f\left(\frac{b+a}{2} + \frac{b-a}{2}t\right) dt,$$

便可使用高斯-勒让德求积公式进行计算.

前几个高斯-勒让德求积公式中的节点和求积系数如表 6.2 所示.

例 6.2.6 用 $n = 3$ 的高斯-勒让德求积公式计算积分

$$\int_0^1 \frac{\sin x}{x} dx.$$

表 6.2 高斯-勒让德求积公式中的节点 x_i 和求积系数 A_i

n	x_i	A_i
0	0	2
1	$\pm\sqrt{3}/3$	1
2	0 $\pm\sqrt{3/5}$	8/9 5/9
3	$\pm 0.339\,981\,043\,585$ $\pm 0.861\,136\,311\,594$	0.652 145 154 863 0.347 854 845 137
4	0 $\pm 0.538\,469\,310\,106$ $\pm 0.906\,179\,845\,939$	0.568 888 888 889 0.478 628 670 499 0.236 926 885 056
5	$\pm 0.238\,619\,186\,083$ $\pm 0.661\,209\,386\,466$ $\pm 0.932\,469\,514\,203$	0.467 913 934 573 0.360 761 573 048 0.171 324 492 379

解 令 $x = \frac{t+1}{2}$, 则

$$I[f] = \int_0^1 \frac{\sin x}{x} dx = \int_{-1}^1 \frac{\sin \frac{t+1}{2}}{t+1} dt,$$

记 $f(t) = \frac{\sin \frac{t+1}{2}}{t+1}$, $n=3$. 由表 6.1 知

$$t_0 = -0.861\,136\,311\,594, \quad t_1 = -0.339\,981\,043\,585,$$

$$t_2 = 0.339\,981\,043\,585, \quad t_3 = 0.861\,136\,311\,594.$$

$$A_0 = A_3 = 0.347\,854\,845\,137, \quad A_1 = A_2 = 0.652\,145\,154\,863,$$

$$I[f] \approx Q[f] = \sum_{i=0}^3 A_i f(t_i) = 0.946\,083\,070\,2.$$

解毕.

2. 高斯-拉盖尔 (Gauss-Laguerre) 求积公式

拉盖尔多项式 $\{L_n(x)\}$ 是在区间 $[0, +\infty)$ 上关于权函数 $\omega(x) = e^{-x}$ 正交的正交多项式. $L_n(x)$ 最高次项 x^n 的系数 $a_n = (-1)^n$, 内积 $\gamma_n = (n!)^2$, 取节点 x_i ($i = 0, 1, \dots, n$) 为拉盖尔多项式 $L_{n+1}(x)$ 的零点, 则由式 (6.2.3), (6.2.10), (6.2.11) 可得高斯-拉盖尔求积公式和求积系数及截断误差分别为

$$\int_0^\infty e^{-x} f(x) dx \approx \sum_{i=0}^n A_i f(x_i), \quad (6.2.15)$$

$$A_i = -\frac{(n!)^2}{L'_{n+1}(x_i)L_n(x_i)}, \quad (6.2.16)$$

$$R[f] = \frac{[(n+1)!]^2}{(2n+2)!} f^{(2n+2)}(\eta), \quad 0 \leq \eta < +\infty. \quad (6.2.17)$$

前几个高斯-拉盖尔求积公式中的节点 x_i 和求积系数 A_i 如表 6.3 所示.

表 6.3 高斯-拉盖尔求积公式中的节点 x_i 和求积系数 A_i

n	x_i	A_i
0	1	1
1	0.585 786 437 627	0.853 553 390 593
	3.414 213 562 373	0.146 446 609 407
2	0.415 774 556 783	0.711 093 009 929
	2.294 280 360 279	0.278 517 733 569
	6.289 945 082 937	$0.103 892 565 016 \times 10^{-1}$
3	0.322 547 689 619	0.603 154 104 342
	1.745 761 101 158	0.357 418 692 438
	4.536 620 296 921	$0.388 879 085 150 \times 10^{-1}$
	9.395 070 912 301	$0.539 294 705 561 \times 10^{-3}$
4	0.263 560 319 718	0.521 755 610 583
	1.413 403 059 107	0.398 666 811 083
	3.596 425 771 041	$0.759 424 496 817 \times 10^{-1}$
	7.085 810 005 859	$0.361 175 867 992 \times 10^{-2}$
	12.640 800 844 276	$0.233 699 723 858 \times 10^{-4}$
5	0.222 846 604 179	0.458 964 673 950
	1.188 932 101 673	0.417 000 830 772
	2.992 736 326 059	0.113 373 382 074
	5.775 143 569 105	$0.103 991 974 531 \times 10^{-1}$
	9.837 467 418 383	$0.261 017 202 815 \times 10^{-3}$
	15.982 873 980 602	$0.898 547 906 430 \times 10^{-6}$

3. 高斯-埃尔米特 (Gauss-Hermite) 求积公式

埃尔米特多项式 $\{H_n(x)\}$ 是在区间 $(-\infty, +\infty)$ 上关于权函数 $\omega(x) = e^{-x^2}$ 正交的正交多项式. $H_n(x)$ 最高次项 x^n 的系数 $a_n = 2^n$, 内积 $\gamma_n = 2^n(n!)\sqrt{\pi}$, 取节点 x_i ($i = 0, 1, \dots, n$) 为埃尔米特多项式 $H_{n+1}(x)$ 的零点, 则由式 (6.2.3), (6.2.10), (6.2.11) 可得高斯-埃尔米特求积公式和求积系数及截断误差分别为

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x) dx \approx \sum_{i=0}^n A_i f(x_i), \quad (6.2.18)$$

$$A_i = \frac{2^{n+1} n! \sqrt{\pi}}{H'_{n+1}(x_i) H_n(x_i)}, \quad (6.2.19)$$

$$R[f] = \frac{(n+1)! \sqrt{\pi}}{2^{n+1} (2n+2)!} f^{(2n+2)}(\eta), \quad -\infty < \eta < +\infty. \quad (6.2.20)$$

前几个高斯-埃尔米特求积公式中的节点 x_i 和求积系数 A_i 如表 6.4 所示.

表 6.4 高斯-埃尔米特求积公式中的节点 x_i 和求积系数 A_i

n	x_i	A_i
0	0	$\sqrt{\pi}$
1	$\pm 0.707\ 106\ 781\ 187$	0.886 226 925 453
2	0 $\pm 1.224\ 744\ 871\ 392$	1.181 635 900 604 0.295 408 975 151
3	$\pm 0.524\ 647\ 623\ 275$ $\pm 1.650\ 680\ 123\ 886$	0.804 914 090 006 $0.813\ 128\ 354\ 473 \times 10^{-1}$
4	0 $\pm 0.958\ 572\ 464\ 614$ $\pm 2.020\ 182\ 870\ 456$	0.945 308 720 483 0.393 619 323 152 $0.199\ 532\ 420\ 591 \times 10^{-1}$
5	$\pm 0.436\ 077\ 411\ 928$ $\pm 1.335\ 849\ 074\ 014$ $\pm 2.350\ 604\ 973\ 674$	0.724 629 595 224 0.157 067 320 323 $0.453\ 000\ 990\ 551 \times 10^{-2}$

4. 高斯-切比雪夫 (Gauss-Chebyshev) 求积公式

切比雪夫多项式 $\{T_n(x)\}$ 是在区间 $[-1, 1]$ 上关于权函数 $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ 正交的正交多项式. $T_n(x)$ 最高次项 x^n 的系数 $a_n = 2^{n-1}$, 内积 $\gamma_n = \frac{\pi}{2}$ ($n \geq 1$). 取节点 x_i ($i = 0, 1, \dots, n$) 为切比雪夫多项式 $T_{n+1}(x)$ 的零点

$$x_i = \cos \left[\frac{2i+1}{2(n+1)} \pi \right], \quad i = 0, 1, \dots, n.$$

下面计算求积系数 A_i .

由 $T_{n+1}(x) = \cos((n+1) \arccos x)$ 知

$$T'_{n+1}(x) = \frac{(n+1) \sin((n+1) \arccos x)}{\sqrt{1-x^2}},$$

故

$$T'_{n+1}(x_i) = \frac{(n+1) \sin(i\pi + \pi/2)}{\sin \theta_i} = \frac{(-1)^i (n+1)}{\sin \theta_i}, \quad \theta_i = \frac{2i+1}{2(n+1)} \pi.$$

$$\begin{aligned} T_n(x_i) &= \cos(n \arccos x_i) \stackrel{\theta_i = \arccos x_i}{=} \cos(n\theta_i) \\ &= \cos((n+1)\theta_i - \theta_i) = \cos(i\pi + \pi/2 - \theta_i) = (-1)^i \sin \theta_i, \end{aligned}$$

从而 $T'_{n+1}(x_i)T_n(x_i) = n+1$, 所以

$$A_i = \frac{a_{n+1}}{a_n} \frac{\gamma_n}{T'_{n+1}(x_i)T_n(x_i)} = \frac{\pi}{n+1}, \quad i = 0, 1, \dots, n. \quad (6.2.21)$$

由式 (6.2.3), (6.2.10), (6.2.11) 可得高斯-切比雪夫求积公式和截断误差分别为

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \approx \frac{\pi}{n+1} \sum_{i=0}^n f\left(\cos \frac{2i+1}{2(n+1)}\pi\right), \quad (6.2.22)$$

$$R[f] = \frac{\pi}{2^{2n+1}(2n+2)!} f^{(2n+2)}(\eta), \quad -1 \leq \eta \leq 1. \quad (6.2.23)$$

6.3 数值积分的稳定性

用数值积分法计算积分时, 除截断误差外, 计算函数值时还存在着舍入误差. 因此, 必须考虑舍入误差对计算结果的影响, 即数值积分的稳定性问题.

设函数值 $f(x_i)$ 的计算值为 $\tilde{f}(x_i)$, 舍入误差限为 ε , 即

$$|f(x_i) - \tilde{f}(x_i)| \leq \varepsilon_i \leq \varepsilon,$$

则数值积分的舍入误差

$$E = \left| \sum_{i=0}^n A_i f(x_i) - \sum_{i=0}^n A_i \tilde{f}(x_i) \right| \leq \sum_{i=0}^n |A_i| \varepsilon.$$

而由式 (6.2.6) 有

$$\sum_{i=0}^n A_i = \int_a^b \omega(x) dx = \gamma_0.$$

当求积系数 A_i 全为正时, $\sum_{i=0}^n |A_i| = \sum_{i=0}^n A_i = \gamma_0$, 这时

$$E \leq \gamma_0 \varepsilon.$$

这说明当求积系数 A_i 全为正时, 积分计算值的舍入误差限不会超过计算函数值舍入误差限的 γ_0 倍, 即舍入误差是可以控制的, 故数值积分方法是稳定的. 事实上, $n=1, 2, 4$ 的牛顿-科茨公式和高斯型求积公式的求积系数全大于零, 所以梯形、辛普森、科茨和高斯型求积公式都是稳定的数值求积方法. $n \geq 8$ 的牛顿-科茨求积公式中求积系数有正有负. 这时, 若每个 $|\varepsilon_i|$ 都达到 ε 且与相应的 A_i 正好都同号 (或异号), 则 $\left| \sum_{i=0}^n A_i \varepsilon_i \right| = \varepsilon \sum_{i=0}^n |A_i|$. 尽管 $\sum_{i=0}^n A_i = \gamma_0$, 但 $\sum_{i=0}^n |A_i|$ 有可能很大, 故不能保证其数值稳定性. 这从另一个角度说明了不宜采用高次插值多项式的原因.

6.4 数值微分

当 $f(x)$ 是列表函数时, 通常的求导法则无法求其导数. 本节介绍计算列表函

数近似导数的方法, 即数值微分法. 该法也常用于求其表达式比较复杂的函数的近似导数值和常微分方程、偏微分方程的数值求解中.

6.4.1 插值型数值微分公式

插值型数值微分公式是取 $f(x)$ 的插值多项式 (如 $L_n(x)$) 的 k 阶导数作为 $f(x)$ 的 k 阶导数的近似表达式, 即取 $f^{(k)}(x) \approx L_n^{(k)}(x)$, $f^{(k)}(x_i) \approx L_n^{(k)}(x_i)$. 由插值法知

$$f(x) = L_n(x) + R_n(x),$$

则 $f^{(k)}(x) = L_n^{(k)}(x) + R_n^{(k)}(x)$.

取

$$f^{(k)}(x) \approx L_n^{(k)}(x) = \sum_{i=0}^n l_i^{(k)}(x) f(x_i), \quad (6.4.1)$$

则截断误差

$$R[f] = f^{(k)}(x) - L_n^{(k)}(x) = R_n^{(k)}(x) = \frac{d^k}{dx^k} \left[\frac{f^{(n+1)}(\xi)}{(n+1)!} \pi_{n+1}(x) \right], \quad (6.4.2)$$

其中 ξ 与 $x, x_0, x_1, \dots, x_n$ 有关.

1. 两点数值微分公式 ($n=1, k=1$)

$$L_1(x) = \frac{x-x_1}{x_0-x_1} f(x_0) + \frac{x-x_0}{x_1-x_0} f(x_1),$$

$$R_1(x) = \frac{f''(x)}{2!} (x-x_0)(x-x_1),$$

则

$$L'_1(x) = \frac{f(x_1) - f(x_0)}{h}, \quad h = x_1 - x_0,$$

$$R'_1(x_0) = -\frac{h}{2} f''(\xi), \quad R'_1(x_1) = \frac{h}{2} f''(\xi),$$

所以由 $f'(x) = L'_1(x) + R'_1(x)$ 得

$$\begin{cases} f'(x_0) = \frac{f(x_1) - f(x_0)}{h} - \frac{h}{2} f''(\xi), \\ f'(x_1) = \frac{f(x_1) - f(x_0)}{h} + \frac{h}{2} f''(\xi). \end{cases} \quad (6.4.3)$$

2. 三点数值微分公式 ($n=2$)

设插值节点等距分布, 即 $x_i = x_0 + ih$ ($i=0, 1, 2$), 则二次拉格朗日插值多项式及误差项分别为

$$\begin{aligned}
 L_2(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1) \\
 &\quad + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2) \\
 &= \frac{(x-x_1)(x-x_2)}{2h^2}f(x_0) - \frac{(x-x_0)(x-x_2)}{h^2}f(x_1) + \frac{(x-x_0)(x-x_1)}{2h^2}f(x_2), \\
 R_2(x) &= \frac{f'''(\xi)}{3!}(x-x_0)(x-x_1)(x-x_2).
 \end{aligned}$$

(1) 求一阶导数 ($k=1$)

由 $f'(x_i) = L'_2(x_i) + R'_2(x_i)$ 可得一阶数值微分公式

$$\begin{cases} f'(x_0) = \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h} + \frac{h^2}{3}f'''(\xi), \\ f'(x_1) = \frac{f(x_2) - f(x_0)}{2h} - \frac{h^2}{6}f'''(\xi), \\ f'(x_2) = \frac{f(x_0) - 4f(x_1) + 3f(x_2)}{2h} + \frac{h^2}{3}f'''(\xi). \end{cases} \quad (6.4.4)$$

由式 (6.4.4) 第二式得到一个常用的数值微分公式

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6}f'''(\xi). \quad (6.4.5)$$

(2) 求二阶导数 ($k=2$)

$$\begin{aligned}
 L_2''(x) &= \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2}, \\
 R_2''(x) &= \frac{1}{3}(x-x_0+x-x_1+x-x_2)f'''(\xi) \\
 &\quad + 2[(x-x_1)(x-x_2) + (x-x_0)(x-x_2) + (x-x_0)(x-x_1)] \frac{d\left(\frac{f'''(\xi)}{3!}\right)}{dx} \\
 &\quad + \frac{1}{6}(x-x_0)(x-x_1)(x-x_2) \frac{d^2 f'''(\xi)}{dx^2}.
 \end{aligned}$$

由式 (4.3.7) 知 $\frac{d}{dx} \left(\frac{f'''(\xi)}{3!} \right) = \frac{f^{(4)}(\bar{\xi})}{4!}$, 故

$$\begin{aligned}
 R_2''(x) &= \frac{1}{3}(x-x_0+x-x_1+x-x_2)f'''(\xi) \\
 &\quad + \frac{2}{4!}[(x-x_1)(x-x_2) + (x-x_0)(x-x_2) + (x-x_0)(x-x_1)]f^{(4)}(\bar{\xi}) \\
 &\quad + \frac{1}{6}(x-x_0)(x-x_1)(x-x_2) \frac{d^2 f'''(\xi)}{dx^2},
 \end{aligned}$$

所以

$$\begin{cases} f''(x_0) = \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2} - hf'''(\xi) + \frac{h^2}{6}f^{(4)}(\bar{\xi}), \\ f''(x_1) = \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2} - \frac{h^2}{12}f^{(4)}(\bar{\xi}), \\ f''(x_2) = \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2} + hf'''(\xi) + \frac{h^2}{6}f^{(4)}(\bar{\xi}). \end{cases} \quad (6.4.6)$$

同样, 由式 (6.4.8) 第二式得到一个常用的二阶数值微分公式

$$f''(x) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi). \quad (6.4.7)$$

由数值微分公式可见, 其截断误差都随步长 h 的减小而减小, 但所有公式都是以 h 作除数, 因而随步长 h 的减小, 式中函数值的误差将给导数计算带来更大的误差, 所以 h 的选取要合适, 不宜太大, 也不宜太小, 原则上不能让舍入误差超过截断误差.

6.4.2 待定系数法

与数值积分法中的待定系数法类似, 也可用待定系数法导出数值微分公式, 以下例予以说明.

例 6.4.1 确定如下数值微分公式:

$$f''(x_0) \approx c_0 f(x_0) + c_1 f'(x_0) + c_2 f(x_1),$$

使其具有尽可能高的代数精度, 并给出截断误差表示式.

解 为了便于计算, 令 $x_0 = 0$, $x_1 = h$, 则截断误差

$$R[f] = f''(0) - c_0 f(0) - c_1 f'(0) - c_2 f(h).$$

分别取 $f = 1, x, x^2$, 令 $R[f] = 0$, 则得方程组

$$\begin{cases} -c_0 - c_2 = 0, \\ -c_1 - c_2 h = 0, \\ 2 - c_2 h^2 = 0. \end{cases}$$

解之得

$$c_0 = -\frac{2}{h^2}, \quad c_1 = -\frac{2}{h}, \quad c_2 = \frac{2}{h^2},$$

从而所求数值微分公式为

$$f''(x_0) \approx \frac{2}{h^2}[-f(x_0) - hf'(x_0) + f(x_1)],$$

其中 $h = x_1 - x_0$. 由于 $R[x^3] = -c_2 h^3 = -2h \neq 0$, 所以代数精度 $m = 2$. 根据广义佩亚诺定理, 取

$$e(x) = \frac{f'''(\xi)}{3!}(x - x_0)^2(x - x_1),$$

则

$$\begin{aligned} R[f] &= R[e] = e''(x_0) - \frac{2}{h^2}[-e(x_0) - he'(x_0) + e(x_1)] \\ &= e''(x_0) = -\frac{h}{3}f'''(\xi). \end{aligned}$$

解毕.

6.4.3 外推求导法

所谓外推法即利用几次计算结果的适当组合得到更精确的结果. 下面以计算一阶导数为例来说明外推法在数值微分中的应用. 根据泰勒公式

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \frac{h^4}{4!}f^{(4)}(x) + \frac{h^5}{5!}f^{(5)}(x) + \cdots, \\ f(x-h) &= f(x) - hf'(x) + \frac{h^2}{2!}f''(x) - \frac{h^3}{3!}f'''(x) + \frac{h^4}{4!}f^{(4)}(x) - \frac{h^5}{5!}f^{(5)}(x) + \cdots, \end{aligned}$$

以上两式相减得

$$f(x+h) - f(x-h) = 2hf'(x) + \frac{2h^3}{3!}f'''(x) + \frac{2h^5}{5!}f^{(5)}(x) + \frac{2h^7}{7}f^{(7)}(x) + \cdots.$$

由此可得

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{3!}f'''(x) - \frac{h^4}{5!}f^{(5)}(x) - \frac{h^6}{7!}f^{(7)}(x) - \cdots.$$

若记 $T(h) = \frac{f(x+h) - f(x-h)}{2h}$, 则

$$f'(x) = T(h) - \frac{h^2}{3!}f'''(x) - \frac{h^4}{5!}f^{(5)}(x) - \frac{h^6}{7!}f^{(7)}(x) - \cdots. \quad (6.4.8)$$

若取 $f(x) \approx T(h)$, 则截断误差

$$R[f] = f'(x) - T(h) = O(h^2).$$

将步长 h 缩小一半, 记

$$T(h/2) = \frac{f(x+h/2) - f(x-h/2)}{2 \times h/2},$$

则由式 (6.4.8) 有

$$f'(x) = T(h/2) - \frac{(h/2)^2}{3!} f'''(x) - \frac{(h/2)^4}{5!} f^{(5)}(x) - \frac{(h/2)^6}{7!} f^{(7)}(x) - \dots \quad (6.4.9)$$

式 (6.4.9) 两端乘以 4 减去式 (6.4.8) 得

$$\begin{aligned} f'(x) &= T(h/2) + \frac{1}{4-1} (T(h/2) - T(h)) + \frac{h^4}{4 \times 5!} f^{(5)}(x) + \frac{5h^6}{16 \times 7!} f^{(7)}(x) + \dots \\ &= T_1^0 + \frac{h^4}{4 \times 5!} f^{(5)}(x) + \frac{5h^6}{16 \times 7!} f^{(7)}(x) + \dots, \end{aligned} \quad (6.4.10)$$

其中

$$T_1^0 = T(h/2) + \frac{1}{4-1} (T(h/2) - T(h)).$$

若取 $f'(x) \approx T_1^0$, 则截断误差

$$R[f] = f'(x) - T_1^0 = O(h^4).$$

再将步长缩小一半, 由式 (6.4.10) 有

$$\begin{aligned} f'(x) &= T(h/4) + \frac{1}{4-1} (T(h/4) - T(h/2)) + \frac{(h/2)^4}{4 \times 5!} f^{(5)}(x) + \frac{5(h/2)^6}{16 \times 7!} f^{(7)}(x) + \dots \\ &= T_1^1 + \frac{(h/2)^4}{4 \times 5!} f^{(5)}(x) + \frac{5(h/2)^6}{16 \times 7!} f^{(7)}(x) + \dots, \end{aligned} \quad (6.4.11)$$

其中

$$T_1^1 = T(h/4) + \frac{1}{4-1} (T(h/4) - T(h/2)).$$

式 (6.4.11) 两端乘以 16 减去式 (6.4.10) 得

$$f'(x) = T_2^0 - \frac{h^6}{64 \times 7!} f^{(7)}(x) - \dots,$$

其中 $T_2^0 = T_1^1 + \frac{1}{4^2-1} (T_1^1 - T_1^0)$. 取 $f'(x) \approx T_2^0$, 则截断误差

$$R[f] = f'(x) - T_2^0 = O(h^6).$$

仿照以上过程可以继续计算下去, 直到相邻两次计算结果之差的绝对值小于预先给定的精度为止. 记

$$\begin{cases} h_i = \frac{h}{2^i}, \\ T_0^0 = T(h_0) = T(h) = \frac{f(x+h) - f(x-h)}{2h}, \\ T_0^i = T(h_i) = \frac{f(x+h_i) - f(x-h_i)}{2h_i}, \end{cases} \quad (6.4.12)$$

则

$$T_1^0 = T_0^1 + \frac{1}{4-1}(T_0^1 - T_0^0), \quad T_1^1 = T_0^2 + \frac{1}{4-1}(T_0^2 - T_0^1).$$

一般地,

$$T_k^i = T_{k-1}^{i+1} + \frac{1}{4^k - 1}(T_{k-1}^{i+1} - T_{k-1}^i). \quad (6.4.13)$$

若取 $f'(x) \approx T_k^0$, 则截断误差

$$R[f] = f'(x) - T_k^0 = O(h^{2(k+1)}).$$

单从上式来看, k 越大, 计算机算结果越准确, 但考虑到计算中存在舍入误差, 一般 k 不能取得太大.

外推求导法的计算过程如表 6.5 所示.

对于计算 $f(x)$ 的二阶导数也有类似的结果.

例 6.4.2 给定数据表如表 6.6 所示, 试用外推法计算 $y'(1)$.

表 6.5 外推求导法的计算过程

T_0^i	T_1^i	T_2^i	T_3^i	...
T_0^0				
T_0^1	T_1^0			
T_0^2	T_1^1	T_2^0		
T_0^3	T_1^2	T_2^1	T_3^0	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

表 6.6

x_i	0.2	0.6	0.8	1.2	1.4	1.8
y_i	1.221 403	1.822 119	2.225 541	3.320 117	4.055 200	6.049 647

解 按式 (6.4.12) 计算得

$$T_0^i = T(h_i), \quad h_0 = 0.8, \quad h_1 = 0.4, \quad h_2 = 0.2.$$

按式 (6.4.13) 计算 T_1^i, T_2^i , 由表 6.5 列出表 6.7.

表 6.7

T_0^i	T_1^i	T_2^i
3.017 653		
2.791 351	2.715 917	
2.736 440	2.718 136	2.718 284

取 $y'(1) \approx T_2^0 = 2.718 284$. 解毕.

实际上, 表中的数据由函数 $y = e^x$ 给出, 将外推法的计算结果与 $y'(1) = e \approx 2.718 281 828 \dots$ 比较, 外推法的计算结果准确到 5 位小数, 已相当准确.

6.4.4 利用三次样条插值函数求导法

前面介绍的数值微分公式的截断误差都包含 $f(x)$ 的高阶导数, 当高阶导数值较大时难以保证截断误差很小, 此时最好应用三次样条插值函数求导.

由第 4 章知

$$\max_{a \leq x \leq b} |f^{(k)}(x) - S^{(k)}(x)| \leq c_k M_4 h^{4-k}, \quad k = 0, 1, 2, 3.$$

显然, 当 $h \rightarrow 0$ 时, $S(x), S'(x), S''(x)$ 在区间 $[a, b]$ 上分别一致地收敛于 $f(x), f'(x), f''(x)$. 于是由式 (4.6.4), (4.6.1) 可得

$$\begin{aligned} f'(x) &\approx S'(x) \\ &= \frac{f(x_i) - f(x_{i-1})}{h_i} - \frac{h_i}{6}(M_i - M_{i-1}) \\ &\quad - \frac{M_{i-1}}{2h_i}(x_i - x)^2 + \frac{M_i}{2h_i}(x - x_{i-1})^2, \quad x_{i-1} \leq x \leq x_i, \end{aligned}$$

其中 $h_i = x_i - x_{i-1}$.

$$\begin{aligned} f''(x) &\approx S''(x) = \frac{x_i - x}{h_i} M_{i-1} + \frac{x - x_{i-1}}{h_i} M_i \\ &= M_{i-1} + \frac{M_i - M_{i-1}}{h_i}(x - x_{i-1}), \quad x_{i-1} \leq x \leq x_i. \\ f'''(x) &\approx S'''(x) = \frac{M_i - M_{i-1}}{h_i}, \quad x_{i-1} \leq x \leq x_i. \end{aligned}$$

小 结

本章介绍了三种不同类型的数值积分方法——基于插值多项式近似代替被积函数所建立的方法、待定系数法和高斯型求积公式.

首先利用等距节点的拉格朗日插值多项式近似代替被积函数 $f(x)$, 建立了三个最基本的数值积分公式——梯形求积公式、辛普森求积公式和科茨求积公式. 这几个公式单独使用, 精度不高. 为了提高精度, 将积分区间 $[a, b]$ 分为 n 个长度相等的子区间, 在每个子区间上分别应用以上三种求积公式, 所得结果相加得到相应的三种复化求积公式. 在给定精度的要求下, 若用复化求积公式计算, 当被积函数 $f(x)$ 比较复杂时, 截断误差中被积函数高阶导数的最大值往往难以估计, 因而无法确定将积分区间 $[a, b]$ 分为若干个等长子区间的个数 n (或子区间的长度 h), 这时复化求积公式难以应用. 变步长积分法采用逐步缩小步长的复化求积公式进行计算, 直到相邻两次计算结果之差满足给定的精度为止. 取小步长算出的结果作为积

分近似值. 而龙贝格积分法则是利用梯形变步长积分法将积分区间逐次分半计算, 用低精度公式逐次适当组合而获得较高计算精度的一种外推法. 它具有计算简单、数值稳定性好、算法易于实现等优点.

待定系数法是利用求积公式代数精度的定义, 通过解方程组得到求积系数, 然后利用广义佩亚诺定理确定截断误差而建立求积公式的一种方法.

高斯型求积公式则是以正交多项式的零点构造插值多项式近似代替被积函数所建立的方法. 它具有最高代数精度且是数值稳定的方法. 本章给出了 4 种常用的高斯型求积公式.

本章还介绍了 4 种类型的数值微分法, 即插值型数值微分公式、待定系数法、基于泰勒展开的外推求导法和利用三次样条插值函数的求导法. 数值微分公式中既含有相近数相减, 又含有小除数, 舍入误差随步长的减小而增大, 所以实际计算中, 步长不能取得太小, 原则上舍入误差不能超过截断误差.

习 题

6.1 设函数 $f(x)$ 在区间 $[a, b]$ 上二阶导数连续, 试导出以下求积公式:

$$(1) \int_a^b f(x) dx = (b-a)f(a) + \frac{(b-a)^2}{2} f'(\eta), \quad a < \eta < b;$$

$$(2) \int_a^b f(x) dx = (b-a)f(b) - \frac{(b-a)^2}{2} f'(\eta), \quad a < \eta < b;$$

$$(3) \int_a^b f(x) dx = (b-a)f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{24} f''(\eta), \quad a < \eta < b,$$

并说明求积公式的几何意义.

6.2 取 $n=8$, 用复化梯形、复化辛普森求积公式计算下列积分, 保留 7 位小数计算:

$$(1) \int_0^1 \frac{1}{1+x} dx; \quad (2) \int_0^1 \frac{\ln(1+x)}{1+x^2} dx;$$

$$(3) \int_0^1 \frac{\ln(1+x)}{x} dx; \quad (4) \int_0^{\pi/2} \frac{\sin x}{x} dx.$$

6.3 取 $n=10$, 用复化辛普森求积公式计算积分

$$\int_0^1 e^{-x^2} dx.$$

6.4 若用复化梯形求积公式计算积分 $I = \int_0^1 e^x dx$, 问区间 $[0, 1]$ 分成多少等份, 才能使截断误差不超过 $\frac{1}{2} \times 10^{-5}$? 若改用复化辛普森求积公式计算, 要达到同样的精度, 区间 $[0, 1]$ 应分成多少等份?

6.5 分别用复化梯形求积公式和复化辛普森求积公式计算积分

$$I = \int_0^1 \frac{1}{1+x^2} dx,$$

并与准确值 $I = \pi/4$ 比较 (取步长 $h = 0.1$, 计算结果取 5 位小数).

6.6 用龙贝格积分法求解习题 6.5, 使误差不超过 $\frac{1}{2} \times 10^{-5}$.

6.7 导出下列求积公式及其截断误差估计式:

$$(1) \int_0^h f(x) dx \approx A_0 f(0) + B_0 f'(0) + A_1 f(h) + B_1 f'(h);$$

$$(2) \int_0^{2h} f(x) dx \approx A_0 f(0) + A_1 f(h) + A_2 f(2h);$$

$$(3) \int_{-1}^1 x^2 f(x) dx \approx A_0 f(x_0) + A_1 f(x_1);$$

$$(4) \int_0^1 \sqrt{x} f(x) dx \approx A_0 f(x_0) + A_1 f(x_1);$$

$$(5) \int_0^{+\infty} e^{-x} f(x) dx \approx A_0 f(x_0) + A_1 f(x_1).$$

6.8 用高斯型求积公式计算下列积分:

$$(1) \int_1^3 e^x \sin x dx \text{ (用 3 点公式);}$$

$$(2) \int_0^{+\infty} e^{-x} \sqrt{x} dx \text{ (用 2 点公式);}$$

$$(3) \int_{-\infty}^{+\infty} e^{-x^2} \sqrt{1+x^2} dx \text{ (用 2 点公式);}$$

$$(4) \int_{-1}^1 \frac{x^2}{\sqrt{1-x^2}} dx \text{ (用 3 点公式).}$$

6.9 已知函数 $y = y(x)$ 的数值如下表所示, 用三点数值微分公式计算 $y'(1)$ 和 $y''(1)$.

x_i	0.6	0.8	0.9	1	1.1	1.2	1.4
y_i	0.736 0	0.836 5	0.909 5	1	1.110 5	1.244 6	1.601 7

6.10 确定下列数值微分公式的系数, 并导出截断误差表示式:

$$(1) f'(0) \approx af(-h) + bf(0) + cf(h);$$

$$(2) f'(h) \approx af'(0) + b[f(2h) - f(h)].$$

6.11 已知数据同习题 6.9, 用外推法计算 $y'(1)$.

计 算 实 习

6.1 用 $n = 4$ 的复化辛普森求积公式求积分方程

$$y(x) = \int_0^1 \left(\frac{1}{1+t} - x \right) y(t) dt + \frac{4x^3 + 5x^2 - 2x + 5}{8(x+1)^2}$$

的近似解, 然后用三次样条函数 $S(x)$ 近似 $y(x)$, 并比较 $S(x)$ 和真解 $y(x) = 1/(1+x)^2$ 在点 $x_i = ih$ ($i = 0, 1, \dots, 20, h = 1/20$) 处的函数值.

6.2 用龙贝格积分法计算习题 6.2 中的积分, 使计算结果尽可能准确.

第7章 非线性方程(组)的迭代解法

一般情况下, 非线性方程 $f(x) = 0$ 及方程组 $f_i(x_1, x_2, \dots, x_n) = 0$ ($i = 1, \dots, n$) 的解不能用解析式表示. 本章讨论求其近似解的迭代解法 (也称为逐次逼近法).

7.1 求解非线性方程的迭代法

本节讨论求解单变量非线性方程

$$f(x) = 0 \quad (7.1.1)$$

的迭代法. 若数 x^* 满足 $f(x^*) = 0$, 则称 x^* 为方程 (7.1.1) 的解或根, 或者称 x^* 为函数 $f(x)$ 的零点. 若 $f(x)$ 是 n 次多项式, 即

$$f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n,$$

其中 $a_n \neq 0$, 则称方程 (7.1.1) 为 n 次多项式方程或代数方程; 若 $f(x)$ 为超越函数, 则称方程 (7.1.1) 为超越方程. 对于代数方程, 当多项式的次数 $n \leq 4$ 时, 它的根可用求根公式表示, 但当 $n = 3, 4$ 时的求根公式比较复杂, 难以使用. 而当次数 $n \geq 5$ 时, 数学家已经证明它的根一般不能用公式表示. 这时, 必须用迭代解法进行数值求解. 将 $n > 1$ 的代数方程和超越方程统称为非线性方程. 本节介绍求非线性方程 (7.1.1) 实单根 x^* 的迭代法. 若要求多个根, 则可根据已求得的根 $x_1^*, \dots, x_m^*$, 由函数

$$f_m(x) = \frac{f(x)}{\prod_{i=1}^m (x - x_i^*)}$$

求方程 (7.1.1) 的第 $m+1$ 个根 x_{m+1}^* .

7.1.1 几种基本迭代法

本小节假定每个含根区间 $[a, b]$ 中只含有一个实根且是单根. 若在区间 $[a, b]$ 上存在多个实根, 则把 $[a, b]$ 分成若干个子区间, 使得每个子区间中只有一个实根且是单根 (称为根的隔离和重根的处理), 然后再求每个根的精确近似值.

1. 二分法

根据高等数学中的零点存在定理: “如果函数 $f(x)$ 在区间 $[a, b]$ 上连续, 并且 $f(a)f(b) < 0$, 则在区间 (a, b) 内必定存在一点 x^* , 使得 $f(x^*) = 0$ ” 建立了求 x^* 近似值的二分法, 其计算过程如下:

(1) 给定 $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, $a_0 := a$, $b_0 := b$, $k := 0$.

(2) 标号 1 $x_k := \frac{1}{2}(a_k + b_k)$, 做

若 $|f(x_k)| < \varepsilon_1$ 或 $|b_k - a_k| < \varepsilon_2$, 则取 $x^* \approx x_k$, 停止;

否则, 若 $f(a_k)f(x_k) < 0$, 则 $a_{k+1} := a_k$, $b_{k+1} := x_k$; 否则,

令 $a_{k+1} := x_k$, $b_{k+1} := b_k$.

$k := k + 1$, 转标号 1.

由以上计算过程可得到一系列的含根区间

$$[a, b] \supset [a_1, b_1] \supset [a_2, b_2] \supset \cdots \supset [a_k, b_k] \supset [a_{k+1}, b_{k+1}] \supset \cdots,$$

根 x^* 的近似值 x_k 的误差估计式如下:

$$|x^* - x_k| = \left| x^* - \frac{a_k + b_k}{2} \right| \leq \frac{b_k - a_k}{2} = \frac{1}{2} \frac{b_{k-1} - a_{k-1}}{2} = \cdots = \frac{b - a}{2^{k+1}}.$$

显然, 当 $k \rightarrow \infty$ 时, $x_k \rightarrow x^*$.

二分法计算简单、方法可靠且总是收敛, 但收敛速度较慢, 要求得根的比较精确的近似值则需要多次计算函数 $f(x)$ 的值, 运算量较大. 因此, 二分法常用于求根的大体范围 (即进行根的隔离) 或用于求其他快速收敛的迭代法 (如后面将要介绍的牛顿法) 所需的一个初始点.

2. 简单迭代法

迭代法的基本思想如下: 构造一个迭代格式 (递推关系式), 按迭代格式计算, 得到一个收敛于方程 (7.1.1) 的根的近似值序列.

为构造迭代格式, 将方程 $f(x) = 0$ 改写为同解方程 $x = \phi(x)$. 若 x^* 满足 $x^* = \phi(x^*)$, 则称 x^* 为 $\phi(x)$ 的不动点. 显然, 不动点 x^* 是方程 $f(x) = 0$ 的根. 于是将求方程 $f(x) = 0$ 的根转化为求 $x = \phi(x)$ 的不动点 x^* .

设 $\phi(x)$ 连续, 构造迭代格式

$$x_{k+1} = \phi(x_k), \quad k = 0, 1, 2, \cdots, \quad (7.1.2)$$

其中 $\phi(x)$ 称为迭代函数.

为求 x^* , 给定 x^* 的一个初始近似值 x_0 , 按迭代格式 (7.1.2) 逐次迭代, 得到一个迭代点列 $\{x_k\}$. 若 $\lim_{k \rightarrow \infty} x_k = x^*$, 对式 (7.1.2) 两端取极限得

$$x^* = \lim_{k \rightarrow \infty} x_{k+1} = \lim_{k \rightarrow \infty} \phi(x_k) = \phi\left(\lim_{k \rightarrow \infty} x_k\right) = \phi(x^*),$$

即 x^* 是 $\phi(x)$ 的不动点, 也即方程 (7.1.1) 的根.

在实际计算时, 若 $|f(x_{k+1})| < \varepsilon_1$ 或者 $|x_{k+1} - x_k| < \varepsilon_2$, 则取 x_{k+1} 作为 x^* 的近似值. 按式 (7.1.2) 迭代求方程 (7.1.1) 的根的方法称为简单迭代法.

为保证迭代序列逐次逼近方程 $f(x) = 0$ 的根, 关键是构造满足收敛性条件的迭代函数 $\phi(x)$ (收敛性条件详见 7.1.2 节).

例 7.1.1 用简单迭代法求区间 $(2, 3)$ 内方程 $x^3 - 2x - 5 = 0$ 的根.

解法 1 将原方程变形为 $x^3 = 2x + 5$, 开三次方得 $x = \sqrt[3]{2x + 5}$, 作迭代格式

$$x_{k+1} = \sqrt[3]{2x_k + 5}, \quad k = 0, 1, 2, \dots$$

取初始点 $x_0 = 2.5$, 按上式迭代得

$$\begin{aligned} x_1 &= 2.154\,434\,690, & x_2 &= 2.103\,612\,029, & x_3 &= 2.095\,927\,410, \\ x_4 &= 2.094\,760\,545, & x_5 &= 2.094\,583\,250, & x_6 &= 2.094\,556\,309, \\ x_7 &= 2.094\,552\,215, & x_8 &= 2.094\,551\,593, & x_9 &= 2.094\,551\,498, \\ x_{10} &= 2.094\,551\,484, & x_{11} &= 2.094\,551\,482 = x_{12}, \end{aligned}$$

取 $x^* \approx x_{11} = 2.094\,551\,482$.

解法 2 在方程两边同加 $2x^3 + 5$, 再同除 $3x^2 - 2$ 得同解方程 $x = (2x^3 + 5)/(3x^2 - 2)$, 作迭代格式

$$x_{k+1} = \frac{2x_k^3 + 5}{3x_k^2 - 2}, \quad k = 0, 1, 2, \dots$$

取初始点 $x_0 = 2.5$, 按上式迭代得

$$x_1 = 2.164\,179\,104, \quad x_2 = 2.097\,135\,356, \quad x_3 = 2.094\,555\,232, \quad x_4 = 2.094\,551\,482 = x_5,$$

取 $x^* \approx x_4$.

解法 3 在方程两边同加 $2x$, 再同除以 2 得同解方程 $x = \frac{(x^3 - 5)}{2}$, 作迭代格式

$$x_{k+1} = \frac{x_k^3 - 5}{2}, \quad k = 0, 1, 2, \dots$$

取初始点 $x_0 = 2.5$, 按上式迭代得

$$x_1 = 5.312\,5, \quad x_2 = 72.466\,430\,66, \quad x_3 = 190\,272.011\,8,$$

$$x_4 = 3.444\,250\,536 \times 10^{15}, \quad x_5 = 2.042\,933\,398 \times 10^{46},$$

计算 x_6 时溢出. 解毕.

从以上三种解法可见, 迭代序列是否收敛以及收敛的快慢同迭代函数 $\phi(x)$ 的选取有关.

3. 牛顿法

(1) 牛顿法

利用泰勒公式可以构造收敛更快的迭代格式. 设 $f(x)$ 在含根区间 $[a, b]$ 上二阶连续可微且 $f'(x) \neq 0$, x_k 是方程 $f(x) = 0$ 的根 x^* 的一个近似值, 则由泰勒公式得

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2!}f''(\xi_k)(x^* - x_k)^2.$$

略去二次项, 则得近似线性方程

$$f(x_k) + f'(x_k)(x^* - x_k) \approx 0.$$

由此解出

$$x^* \approx x_k - \frac{f(x_k)}{f'(x_k)}.$$

显然, $x_k - \frac{f(x_k)}{f'(x_k)}$ 是 x^* 的一个更好的近似值, 故令

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \dots \quad (7.1.3)$$

迭代格式 (7.1.3) 称为牛顿法.

在几何上, 线性化方程 $f(x_k) + f'(x_k)(x - x_k) = 0$ 的解 x_{k+1} 是曲线 $y = f(x)$ 在点 (x_k, y_k) 处的切线

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

与 x 轴的交点. 继续在点 $(x_{k+1}, f(x_{k+1}))$ 处作切线, 该切线与 x 轴的交点为 x_{k+2} , 它比 x_{k+1} 更靠近 x^* . 如此继续迭代下去. 若初始点 x_0 取得充分靠近 x^* , 则由牛顿法产生的迭代序列 $\{x_k\}$ 收敛于方程 (7.1.1) 的解 x^* . 因此, 牛顿法又称为切线法, 它的几何意义如图 7.1 所示. 例 7.1.1 的解法 2 实际上就是牛顿法的迭代过程.

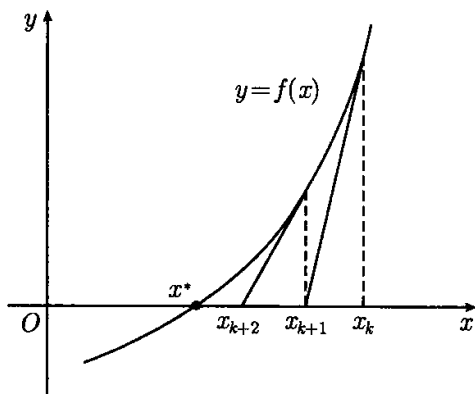


图 7.1 牛顿法几何意义

(2) 改进牛顿法 1

设 $f(x)$ 在含根区间 $[a, b]$ 上三阶连续可微且 $f'(x) \neq 0$, $f''(x^*) \neq 0$, x_k 是方程 $f(x) = 0$ 的根 x^* 的一个近似值, 则由泰勒公式有

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2!}f''(x_k)(x^* - x_k)^2 + \frac{1}{3!}f'''(\xi_k)(x^* - x_k)^3.$$

略去三次项, 则得近似二次方程

$$f''(x_k)(x^* - x_k)^2 + 2f'(x_k)(x^* - x_k) + 2f(x_k) \approx 0.$$

由此解出

$$x^* \approx x_k + \frac{-f'(x_k) \pm \sqrt{f'^2(x_k) - 2f(x_k)f''(x_k)}}{f''(x_k)}.$$

记

$$\begin{aligned}\tilde{x}_{k+1} &= x_k - \frac{f'(x_k) + \operatorname{sgn}(f'(x_k))\sqrt{f'^2(x_k) - 2f(x_k)f''(x_k)}}{f''(x_k)}, \\ \bar{x}_{k+1} &= x_k - \frac{2f(x_k)}{f'(x_k) + \operatorname{sgn}(f'(x_k))\sqrt{f'^2(x_k) - 2f(x_k)f''(x_k)}},\end{aligned}$$

取 $\tilde{x}_{k+1}, \bar{x}_{k+1}$ 中最靠近 x_k 者, 作为 x_{k+1} .

(3) 改进牛顿法 2

设函数 $y = f(x)$ 的反函数 $x = g(y)$ 三阶连续可微, 将 $g(y)$ 在 y_k 处三阶泰勒展开得

$$g(y) = g(y_k) + g'(y_k)(y - y_k) + \frac{1}{2}g''(y_k)(y - y_k)^2 + \frac{1}{3!}g'''(\eta_k)(y - y_k)^3.$$

由 $f(x^*) = 0$ 知 $x^* = g(0)$, 则

$$x^* = g(0) = g(y_k) - g'(y_k)y_k + \frac{1}{2}g''(y_k)y_k^2 - \frac{1}{3!}g'''(\eta_k)y_k^3.$$

取

$$x^* = g(0) \approx g(y_k) - g'(y_k)y_k + \frac{1}{2}g''(y_k)y_k^2,$$

由于 $y_k = f(x_k)$, $x_k = g(y_k)$, $g'(y_k) = \frac{1}{f'(x_k)}$ 及

$$g''(y_k) = \frac{d}{dy} \left(\frac{dx}{dy} \right) \Big|_{y=y_k} = \frac{d}{dx} \left(\frac{1}{f'(x)} \right) \frac{dx}{dy} \Big|_{y=y_k} = -\frac{f''(x_k)}{f'^3(x_k)},$$

所以

$$x^* = g(0) \approx x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f''(x_k)}{2f'^3(x_k)}f^2(x_k).$$

令

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f''(x_k)}{2f'^3(x_k)} f^2(x_k), \quad k = 0, 1, 2, \dots, \quad (7.1.4)$$

即得改进牛顿法 2 的迭代格式.

对例 7.1.1, 按改进牛顿法 2 的式 (7.1.4) 计算, 取 $x_0 = 2.5$ 得

$$x_1 = 2.113\,682\,534, \quad x_2 = 2.094\,555\,162, \quad x_3 = 2.094\,551\,482 = x_4.$$

(4) 简化牛顿法

在实际问题中, 若导数 $f'(x)$ 难以计算或计算量较大, 则此时可用 $f'(x_0)$ 或常数 c 代替牛顿法中的 $f'(x_k)$, 则得简化牛顿法

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, \quad k = 0, 1, \dots \quad \text{或} \quad x_{k+1} = x_k - \frac{f(x_k)}{c}, \quad k = 0, 1, \dots, \quad (7.1.5)$$

其几何意义是用斜率为 $f'(x_0)$ 或 c 的平行线与 x 轴的交点作为 x^* 的新近似点.

(5) 牛顿下山法

为了放宽牛顿法的初始点 x_0 的选择范围, 可将 (7.1.3) 修改为

$$x_{k+1} = x_k - \lambda \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \dots. \quad (7.1.6)$$

式 (7.1.6) 称为牛顿下山法, λ 称为下山因子. 选取 λ 应使

$$|f(x_{k+1})| < |f(x_k)|, \quad k = 0, 1, 2, \dots$$

成立. 实际计算时, 选取 λ 采用试算的办法, 即从 $\lambda = 1$ 开始, 逐次减半进行试算, 直到上式成立为止.

4. 弦割法

(1) 弦割法

在牛顿迭代法 (7.1.3) 中, 将曲线 $y = f(x)$ 在点 $(x_k, f(x_k))$ 处切线的斜率 $f'(x_k)$, 改用其上两点 $(x_{k-1}, f(x_{k-1}))$, $(x_k, f(x_k))$ 或 $(x_0, f(x_0))$, $(x_k, f(x_k))$ 连线 (弦) 的斜率

$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \quad \text{或} \quad \frac{f(x_k) - f(x_0)}{x_k - x_0}$$

来代替, 则得两点弦割法和单点弦割法的迭代格式

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} \\ &= x_k - \frac{f(x_k)/f(x_{k-1})}{f(x_k)/f(x_{k-1}) - 1} (x_k - x_{k-1}), \quad k = 1, 2, \dots \end{aligned} \quad (7.1.7)$$

(说明: 改写公式是为了避免两个相近的数相减),

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_0)}{f(x_k) - f(x_0)}, \quad k = 1, 2, \dots \quad (7.1.8)$$

从几何上来看, 弦割法是取过两点 $(x_{k-1}, f(x_{k-1}))$, $(x_k, f(x_k))$ 或 $(x_0, f(x_0))$, $(x_k, f(x_k))$ 弦线与 x 轴的交点作为 x^* 的新近似点 x_{k+1} , 如图 7.2 所示.

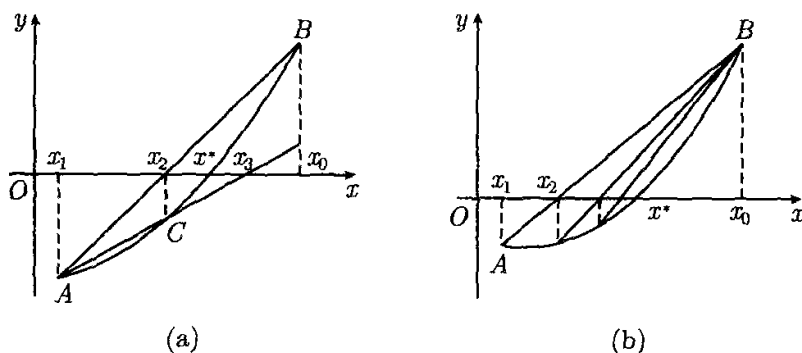


图 7.2 弦割法

对例 7.1.1, 取 $x_0 = 3$, $x_1 = 2$, 利用两点弦割法迭代格式可得

$$\begin{aligned} x_2 &= 2.058\,823\,529, & x_3 &= 2.096\,558\,637, & x_4 &= 2.094\,510\,554, \\ x_5 &= 2.094\,551\,435, & x_6 &= 2.094\,551\,482 = x_7. \end{aligned}$$

单点弦割法迭收敛较慢, 同样取 $x_0 = 3$, $x_1 = 2$, 一直迭代到 x_{23} , 才有 $x_{23} = 2.094\,551\,482 = x_{24}$.

(2) 改进弦割法

设函数 $y = f(x)$ 的反函数为 $x = g(y)$, 作 $g(y)$ 的二次牛顿插值多项式 $N_2(y)$, 则

$$\begin{aligned} x &= g(y) \approx N_2(y) \\ &= g(y_k) + g[y_k, y_{k-1}](y - y_k) + g[y_k, y_{k-1}, y_{k-2}](y - y_k)(y - y_{k-1}). \end{aligned}$$

取 $y = 0$, 则得

$$x^* = g(0) \approx g(y_k) - g[y_k, y_{k-1}]y_k + g[y_k, y_{k-1}, y_{k-2}]y_k y_{k-1}. \quad (7.1.9)$$

而差商

$$\begin{aligned} g[y_k, y_{k-1}] &= \frac{g(y_k) - g(y_{k-1})}{y_k - y_{k-1}} = \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}, \\ g[y_k, y_{k-1}, y_{k-2}] &= \frac{g[y_k, y_{k-1}] - g[y_{k-1}, y_{k-2}]}{y_k - y_{k-2}} \\ &= \left(\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} - \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})} \right) \frac{1}{f(x_k) - f(x_{k-2})}. \end{aligned}$$

将以上差商代入式 (7.1.9), 取近似号右端作为 x_{k+1} , 则得改进弦割法的迭代格式

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k) + \left(\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} - \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})} \right) \frac{f(x_k)f(x_{k-1})}{f(x_k) - f(x_{k-2})}. \quad (7.1.10)$$

对例 7.1.1, 取 $x_0 = 3$, $x_1 = 2$, $x_2 = 2.058\,823\,529$ (由两点弦割法求得), 按式 (7.1.10) 计算得

$$x_3 = 2.095\,658\,932, \quad x_4 = 2.094\,553\,689, \quad x_5 = 2.094\,551\,482 = x_6.$$

7.1.2 迭代法的收敛性

本小节讨论迭代法的收敛性.

定义 7.1.1 设 x^* 为方程 $f(x) = 0$ 的根, 若存在 x^* 的一个邻域 $N_\delta(x^*) = \{x | |x - x^*| \leq \delta, \delta > 0\}$, 对任意的初始点 $x_0 \in N_\delta(x^*)$, 由迭代法产生的迭代序列 $\{x_k\}$ 收敛于 x^* , 则称该迭代法局部收敛; 若对含根区间 $[a, b]$ 内任意的初始点 x_0 , 迭代序列 $\{x_k\}$ 都收敛于 x^* , 则称该迭代法全局收敛.

1. 简单迭代法的收敛性

定理 7.1.1 (迭代法收敛定理) 设迭代函数 $\phi(x)$ 一阶连续可微且满足如下条件:

(1) 当 $x \in [a, b]$ 时, $\phi(x) \in [a, b]$;

(2) 存在常数 L ($0 < L < 1$), 使得对任意 $x \in [a, b]$ 有 $|\phi'(x)| \leq L < 1$,

则对任意初点 $x_0 \in [a, b]$, 由迭代格式 $x_{k+1} = \phi(x_k)$ 产生的迭代序列 $\{x_k\}$ 收敛于 $\phi(x)$ 在区间 $[a, b]$ 上的唯一不动点 x^* , 并且有误差估计式

$$|x^* - x_k| \leq \frac{1}{1-L} |x_{k+1} - x_k| \leq \frac{L}{1-L} |x_k - x_{k-1}|, \quad (7.1.11)$$

$$|x^* - x_k| \leq \frac{L^k}{1-L} |x_1 - x_0|. \quad (7.1.12)$$

证明 (1) 证不动点的存在性. 定义函数 $g(x) = x - \phi(x)$, 则 $g(x)$ 在 $[a, b]$ 上连续, 由条件 (1) 知 $g(a) = a - \phi(a) \leq 0$, $g(b) = b - \phi(b) \geq 0$. 由零点存在定理知, 存在 $x^* \in [a, b]$, 使得 $g(x^*) = 0$, 即 $x^* = \phi(x^*)$.

(2) 证不动点的唯一性. 假定还有另一个不动点 $\bar{x}^* \neq x^*$, 则由微分中值定理及条件 (2) 得

$$0 < |\bar{x}^* - x^*| = |\phi(\bar{x}^*) - \phi(x^*)| = |\phi'(\xi)(\bar{x}^* - x^*)| \leq L|\bar{x}^* - x^*| < |\bar{x}^* - x^*|,$$

得到矛盾, 所以必有 $\bar{x}^* = x^*$.

(3) 证迭代序列的收敛性. 由条件 (2) 知

$$\begin{aligned}|x^* - x_k| &= |\phi(x^*) - \phi(x_{k-1})| = |\phi'(\xi_{k-1})(x^* - x_{k-1})| \\ &\leq L|x^* - x_{k-1}| \leq L^2|x^* - x_{k-2}| \leq \cdots \leq L^k|x^* - x_0|.\end{aligned}$$

因为 $0 < L < 1$, 所以当 $k \rightarrow \infty$ 时, $L^k \rightarrow 0$, 从而 $\lim_{k \rightarrow \infty} |x^* - x_k| = 0$, 即 $\lim_{k \rightarrow \infty} x_k = x^*$.

(4) 证误差估计式.

$$\begin{aligned}|x_{k+n} - x_k| &= |x_{k+n} - x_{k+n-1} + x_{k+n-1} - x_{k+n-2} + \cdots + x_{k+1} - x_k| \\ &\leq |x_{k+n} - x_{k+n-1}| + |x_{k+n-1} - x_{k+n-2}| + \cdots + |x_{k+1} - x_k| \\ &= \sum_{j=1}^n |x_{k+j} - x_{k+j-1}|.\end{aligned}\tag{7.1.13}$$

而

$$\begin{aligned}|x_{k+j} - x_{k+j-1}| &= |\phi(x_{k+j-1}) - \phi(x_{k+j-2})| = |\phi'(\xi)(x_{k+j-1} - x_{k+j-2})| \\ &\leq L|x_{k+j-1} - x_{k+j-2}| \leq L^2|x_{k+j-2} - x_{k+j-3}| \leq \cdots \\ &\leq L^{j-1}|x_{k+1} - x_k|.\end{aligned}\tag{7.1.14}$$

将式 (7.1.14) 代入式 (7.1.13) 得

$$|x_{k+n} - x_k| \leq \sum_{j=1}^n L^{j-1}|x_{k+1} - x_k| = \frac{1 - L^n}{1 - L}|x_{k+1} - x_k|.$$

对上式两端令 $k \rightarrow \infty$ 取极限, 注意到 $0 < L < 1$, 则得

$$|x^* - x_k| \leq \frac{1}{1 - L}|x_{k+1} - x_k| \leq \frac{L}{1 - L}|x_k - x_{k-1}|,$$

即得式 (7.1.11).

由式 (7.1.14) 知 $|x_{k+1} - x_k| \leq L^k|x_1 - x_0|$. 将它代入式 (7.1.11), 则得

$$|x^* - x_k| \leq \frac{L^k}{1 - L}|x_1 - x_0|,$$

即式 (7.1.12). 证毕.

由定理 7.1.1 的证明过程可见, L 越小, L^k 趋于零的速度越快, 迭代序列 $\{x_k\}$ 也就收敛得越快.

在 L 不很接近于 1 时, 由式 (7.1.11) 知, 若 $|x_k - x_{k-1}| < \varepsilon$, 则终止迭代, 取 $x^* \approx x_k$. 当给定允许误差限 ε 后, 可由式 (7.1.12) 估计迭代次数. 事实上, 由式

$|x^* - x_k| \leq \frac{L^k}{1-L} |x_1 - x_0| < \varepsilon$ 可得

$$L^k < \frac{\varepsilon(1-L)}{|x_1 - x_0|}.$$

对上式两端取对数得

$$k > \ln \left(\frac{\varepsilon(1-L)}{|x_1 - x_0|} \right) / \ln L.$$

上式给出了在给定精度要求下迭代次数 k 的一个大致估计.

迭代法的收敛性定理 7.1.1 中的条件 (1) 难以检验, 若将条件 (2) 改写如下: 存在常数 L ($0 < L < 1$), 在不动点的 x^* 某邻域 $|x - x^*| \leq \delta$ 内,

$$|\phi'(x)| \leq L < 1,$$

则条件 (1) 自然成立. 事实上,

$$|\phi(x) - x^*| = |\phi(x) - \phi(x^*)| = |\phi'(\xi)| |x - x^*| \leq L |x - x^*| < |x - x^*| \leq \delta,$$

即 $\forall x \in [x^* - \delta, x^* + \delta]$, 则 $\phi(x) \in [x^* - \delta, x^* + \delta]$, 即条件 (1) 成立. 据此, 由定理 7.1.1 可得以下局部收敛性定理:

定理 7.1.2 (迭代法的局部收敛性定理) 设迭代函数 $\phi(x)$ 在其不动点 x^* 的某邻域 $[x^* - \delta, x^* + \delta]$ 内连续可微, 并且存在常数 L ($0 < L < 1$), 使得

$$|\phi'(x)| \leq L < 1, \quad \forall x \in [x^* - \delta, x^* + \delta]$$

成立, 则对 $\forall x_0 \in [x^* - \delta, x^* + \delta]$, 由迭代公式 $x_{k+1} = \phi(x_k)$ 产生的迭代序列 $\{x_k\}$ 收敛于函数 $\phi(x)$ 的唯一不动点.

2. 牛顿法的收敛性

牛顿法相当于迭代函数为

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

的简单迭代法. 设 $f(x)$ 在其零点 x^* 附近二阶连续可微且 $f'(x^*) \neq 0$, 由

$$\phi'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$$

知, $\phi'(x^*) = 0$. 由此可见, 当 x 充分靠近 x^* 时 (即 x 属于 x^* 的某邻域), 存在常数 L ($0 < L < 1$), 使得 $|\phi'(x)| \leq L < 1$. 由此, 据定理 7.1.2 得到牛顿法的局部收敛性定理.

定理 7.1.3 (牛顿法的局部收敛性定理) 设 $f(x)$ 在其零点 x^* 附近二阶连续可微且 $f'(x^*) \neq 0$, 则当初始点 x_0 充分靠近 x^* 时, 由牛顿法产生的迭代序列 $\{x_k\}$ 收敛于 x^* .

牛顿法的全局收敛性定理如下:

定理 7.1.4 (牛顿法的全局收敛性定理) 设 $f(x)$ 区间 $[a, b]$ 上二阶连续可微, 并且满足

- (1) $f(a)f(b) < 0$;
- (2) $f'(x) \neq 0, \forall x \in [a, b]$;
- (3) $f''(x)$ 不变号, $\forall x \in [a, b]$;
- (4) 初始点 $x_0 \in [a, b]$ 且 $f(x_0)f''(x_0) > 0$,

则由牛顿法产生的迭代序列 $\{x_k\}$ 单调地收敛于方程 $f(x) = 0$ 在 $[a, b]$ 上的唯一根 x^* .

证明 定理中的条件 (1) 保证了根的存在性, 条件 (2) 表明函数 $f(x)$ 的严格单调性, 从而保证了根的唯一性, 条件 (3) 保证了 $f(x)$ 的曲线凹向不变, 条件 (4) 保证了切线总在 $f(x)$ 曲线的一侧, 即保证了序列 $\{x_k\}$ 单侧逼近 x^* , 也即保证了其收敛性. 下面证明收敛性.

为叙述方便起见, 不妨假定 $f'(x) > 0, f''(x) > 0$ (其他情况可类似证明). 由条件 (4) 有

$$0 < f(x_0) = f(x_0) - f(x^*) = f'(\xi)(x_0 - x^*),$$

可知 $x_0 - x^* > 0$, 即 $x_0 > x^*$, 从而

$$\begin{aligned} x_1 - x_0 &= -\frac{f(x_0)}{f'(x_0)} < 0 \Rightarrow x_1 < x_0, \\ x_1 - x^* &= x_0 - x^* - \frac{f(x_0) - f(x^*)}{f'(x_0)} \\ &= x_0 - x^* - \frac{f'(\xi)(x_0 - x^*)}{f'(x_0)} \quad (\xi \in (x^*, x_0)) \\ &= \frac{f'(x_0) - f'(\xi)}{f'(x_0)}(x_0 - x^*) \\ &= \frac{f''(\xi^*)(x_0 - \xi)(x_0 - x^*)}{f'(x_0)} > 0 \Rightarrow x^* < x_1, \end{aligned}$$

故 $x^* < x_1 < x_0$. 同理可证

$$x^* < x_k < x_{k-1} < \cdots < x_1 < x_0.$$

这表明序列 $\{x_k\}$ 单调减下有界, 必然收敛. 设收敛于 $\bar{x}^*$, 则由牛顿法 (7.1.3) 知, 当 $k \rightarrow \infty$ 时,

$$\bar{x}^* = \bar{x}^* - \frac{f(\bar{x}^*)}{f'(\bar{x}^*)},$$

故 $f(\bar{x}^*) = 0$, 而由根的唯一性知 $\bar{x}^* = x^*$, 于是 $\lim_{k \rightarrow \infty} x_k = x^*$. 证毕.

牛顿法的全局收敛性定理具有非常清楚的几何直观性, 详如图 7.3 所示.

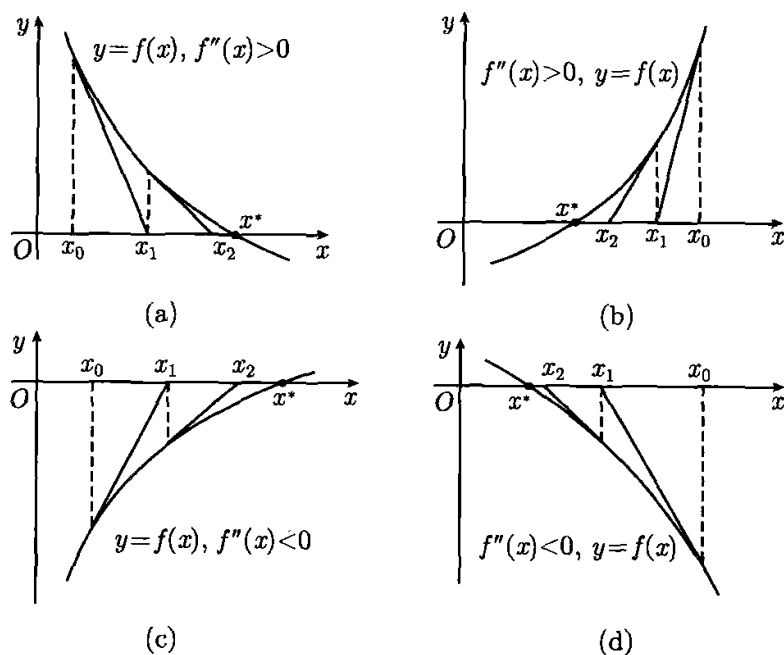


图 7.3

简化牛顿法的迭代函数为 $\phi(x) = x - \frac{f(x)}{c}$, 若 $|\phi'(x)| = \left|1 - \frac{f'(x)}{c}\right| < 1$, 即取 $0 < \frac{f'(x)}{c} < 2$ 在方程的根 x^* 附近成立, 则简化牛顿法局部收敛.

3. 弦割法的收敛性

下面给出弦割法的局部收敛性定理.

定理 7.1.5 (弦割法的局部收敛性定理) 设 $f(x)$ 在包含方程 $f(x) = 0$ 的根 x^* 的某个邻域上二阶连续可微, 并且 $f'(x^*) \neq 0$, 取初始点 x_0, x_1 充分接近 x^* , 则由弦割法迭代格式 (7.1.7) 产生的迭代序列 $\{x_k\}$ 收敛于 x^* .

证明 设函数 $y = f(x)$ 的反函数为 $x = g(y)$, 由牛顿插值公式有

$$x = g(y) = g(y_k) + g[y_k, y_{k-1}](y - y_k) + \frac{1}{2}g''(\eta_k)(y - y_k)(y - y_{k-1}).$$

取 $y = 0$, 则得

$$\begin{aligned}
 x^* &= g(0) = g(y_k) - g[y_k, y_{k-1}]y_k + \frac{1}{2}g''(\eta_k)y_k y_{k-1} \\
 &= x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}f(x_k) - \frac{1}{2}\frac{f''(\xi_k)}{f'^3(\xi_k)}f(x_k)f(x_{k-1}) \\
 &= x_{k+1} - \frac{1}{2}\frac{f''(\xi_k)}{f'^3(\xi_k)}(f(x_k) - f(x^*))(f(x_{k-1}) - f(x^*)) \\
 &= x_{k+1} - \frac{1}{2}\frac{f''(\xi_k)}{f'^3(\xi_k)}f'(\xi_k^*)(x_k - x^*)f'(\xi_{k-1}^*)(x_{k-1} - x^*),
 \end{aligned}$$

故

$$x_{k+1} - x^* = \frac{1}{2}\frac{f''(\xi_k)}{f'^3(\xi_k)}f'(\xi_k^*)(x_k - x^*)f'(\xi_{k-1}^*)(x_{k-1} - x^*). \quad (7.1.15)$$

由此可见, 当 x_{k-1}, x_k 充分靠近 x^* 时,

$$|x_{k+1} - x^*| < |x_k - x^*|,$$

由此可知 $\lim_{k \rightarrow \infty} x_k = x^*$. 证毕.

下面不加证明地给出弦割法的全局收敛性定理.

定理 7.1.6 (弦割法的全局收敛性定理) 在牛顿法的全局收敛性定理 7.1.4 的条件下, 取 $x_1 \in [a, b]$, 使得 $f(x_1)f''(x_1) < 0$, 则由弦割法迭代格式 (7.1.7) 产生的迭代序列 $\{x_k\}$ 收敛于方程 $f(x) = 0$ 在 $[a, b]$ 上的唯一根 x^* .

7.1.3 迭代法的收敛速度

为了度量迭代法收敛的快慢, 下面首先给出迭代序列收敛阶的定义.

定义 7.1.2 设迭代序列 $\{x_k\}$ 收敛于 x^* , 若存在常数 $p \geq 1$ 和 $c > 0$, 使得

$$\lim_{k \rightarrow \infty} \frac{|x^* - x_{k+1}|}{|x^* - x_k|^p} = c \neq 0, \quad (7.1.16)$$

则称该迭代序列为 p 阶收敛的, c 称为渐近误差常数. 特别地, 当 $p = 1$ ($0 < c < 1$) 时, 称为线性收敛; 当 $p > 1$ 时, 称为超线性收敛; 当 $p = 2$ 时, 称为平方收敛或二阶收敛.

如果迭代法产生迭代序列 $\{x_k\}$ 是 p 阶收敛的, 则称该迭代法具有 p 阶收敛速度.

定理 7.1.7 (收敛阶定理) 设迭代函数 $\phi(x)$ 在其不动点 x^* 的邻域内具有连续的 p ($p > 1$) 阶导数, 则由迭代格式 (7.1.2) 产生的迭代序列 $\{x_k\}$ p 阶收敛的充分必要条件是

$$\phi'(x^*) = \phi''(x^*) = \cdots = \phi^{(p-1)}(x^*) = 0, \quad \phi^{(p)}(x^*) \neq 0. \quad (7.1.17)$$

证明 由于 $p > 1, \phi'(x^*) = 0$, 由迭代法的局部收敛性定理知, 迭代序列 $\{x_k\}$ 局部收敛于 x^* . 下证序列 $\{x_k\}$ p 阶收敛.

充分性. 由泰勒公式及定理条件知

$$\begin{aligned} x_{k+1} &= \phi(x_k) = \phi(x^*) + \phi'(x^*)(x_k - x^*) + \cdots \\ &\quad + \frac{1}{(p-1)!} \phi^{(p-1)}(x^*)(x_k - x^*)^{p-1} + \frac{1}{p!} \phi^{(p)}(\xi_k)(x_k - x^*)^p \\ &= \phi(x^*) + \frac{1}{p!} \phi^{(p)}(\xi_k)(x_k - x^*)^p, \end{aligned}$$

其中 ξ_k 介于 x_k 与 x^* 之间. 由此可得

$$|x^* - x_{k+1}| = \frac{1}{p!} |\phi^{(p)}(\xi_k)(x_k - x^*)^p|,$$

所以

$$\lim_{k \rightarrow \infty} \frac{|x^* - x_{k+1}|}{|x^* - x_k|^p} = \lim_{k \rightarrow \infty} \frac{1}{p!} |\phi^{(p)}(\xi_k)| = \frac{1}{p!} |\phi^{(p)}(x^*)| \neq 0,$$

即迭代序列 $\{x_k\}$ p 阶收敛.

必要性. 反证法. 设迭代序列 $\{x_k\}$ p 阶收敛. 假定式 (7.1.17) 不成立, 则必有最小正整数 $p_0 < p$, 使得

$$\phi^{(j)}(x^*) = 0, \quad j = 1, 2, \cdots, p_0 - 1, \quad \phi^{(p_0)}(x^*) \neq 0.$$

由充分性的证明知, 序列 $\{x_k\}$ p_0 阶收敛, 得到矛盾, 故 (7.1.17) 成立. 证毕.

1. 简单迭代法线性收敛

简单迭代法 $x_{k+1} = \phi(x_k)$ 当 $0 < |\phi'(x^*)| < 1$ 时是线性收敛的. 事实上, 设 x^* 是 $x = \phi(x)$ 的不动点, 即 $x^* = \phi(x^*)$. 该式减去 $x_{k+1} = \phi(x_k)$, 则得 $x^* - x_{k+1} = \phi(x^*) - \phi(x_k) = \phi'(\xi_k)(x^* - x_k)$, 其中 ξ_k 介于 x^* 与 x_k 之间, 从而

$$\lim_{k \rightarrow \infty} \frac{|x^* - x_{k+1}|}{|x^* - x_k|} = \lim_{k \rightarrow \infty} |\phi'(\xi_k)| = |\phi'(x^*)|.$$

2. 牛顿法二阶收敛

设 $f(x)$ 在含根区间 $[a, b]$ 上二阶连续可微且 $f'(x) \neq 0$, x_k 是方程 $f(x) = 0$ 的根 x^* 的一个近似值, 则由泰勒公式有

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2!} f''(\xi_k)(x^* - x_k)^2.$$

上式两端同除以 $f'(x_k)$ 得

$$\begin{aligned} x^* - x_k + \frac{f(x_k)}{f'(x_k)} + \frac{1}{2} \frac{f''(\xi_k)}{f'(x_k)} (x^* - x_k)^2 \\ = x^* - x_{k+1} + \frac{1}{2} \frac{f''(\xi_k)}{f'(x_k)} (x^* - x_k)^2 = 0, \end{aligned}$$

所以

$$\lim_{k \rightarrow \infty} \frac{|x^* - x_{k+1}|}{|x^* - x_k|^2} = \frac{1}{2} \frac{|f''(\xi_k)|}{|f'(x_k)|} = \frac{1}{2} \frac{|f''(x^*)|}{|f'(x^*)|} \neq 0.$$

3. 弦割法超线性收敛

由 (7.1.15) 知

$$\begin{aligned} \frac{|x^* - x_{k+1}|}{|x^* - x_k|^p} &= \frac{1}{2} \left| \frac{f''(\xi_k) f'(\xi_k^*) f'(\xi_{k-1}^*)}{f'^3(\xi_k)} \right| \\ &\times \left(\frac{|x_k - x^*|}{|x^* - x_{k-1}|^p} \right)^{1-p} |x^* - x_{k-1}|^{1+p(1-p)}. \end{aligned}$$

在上式两端令 $k \rightarrow \infty$, 取极限得

$$c = \frac{1}{2} \left| \frac{f''(x^*)}{f'(x^*)} \right| c^{1-p} \lim_{k \rightarrow \infty} |x_{k-1} - x^*|^{1+p-p^2},$$

故知 $1 + p - p^2 = 0$, 由此得 $p = \frac{1 + \sqrt{5}}{2} \approx 1.618$, 即弦割法超线性收敛.

4. 单点弦割法线性收敛

单点弦割法实质上是简单迭代法, 其迭代函数

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

在点 x^* 处的导数 $\phi'(x^*) = 1 - \frac{f'(x^*)}{f'(\xi)}$, 其中 ξ 介于 x_0 与 x^* 之间. 由此可见, 单点弦割法一般线性收敛, 但当 $f'(x)$ 变化不大时, $\phi'(x^*) \approx 0$, 收敛可能很快.

弦割法的收敛阶虽然低于牛顿法, 但每次迭代只需计算一个函数值 $f(x_k)$, 不需计算导数值 $f'(x_k)$, 效率高, 实际问题中经常采用.

7.1.4 加速收敛技术

当迭代序列 $\{x_k\}$ 线性收敛时, 常用以下两种方法对 $\{x_k\}$ 进行加速. 即构造新的迭代序列 $\{\bar{x}_k\}$, 使得 $\{\bar{x}_k\}$ 比 $\{x_k\}$ 收敛的更快.

1. 松弛加速法

简单迭代法中的迭代函数 $\phi(x)$ 若满足条件

$$|\phi'(x)| \leq L < 1,$$

则当初始点 x_0 靠近 x^* 时, 可以保证其收敛性, 但当 L 接近于 1 时, 收敛得相当慢. 当此条件不满足时, 简单迭代法一般不收敛. 为使简单迭代法收敛得更快, 或者将

不收敛的迭代格式改造为收敛的迭代格式, 将原来的方程 $x = \phi(x)$ 作同解变形, 在方程两端减去 ωx ($\omega \neq 1$) (ω 称为松弛因子) 得

$$x - \omega x = \phi(x) - \omega x.$$

由此可得

$$x = \frac{\phi(x) - \omega x}{1 - \omega}.$$

记

$$\psi(x) = \frac{\phi(x) - \omega x}{1 - \omega},$$

则有 $x = \psi(x)$. 由此可得迭代格式

$$x_{k+1} = \psi(x_k) = \frac{\phi(x_k) - \omega x_k}{1 - \omega}. \quad (7.1.18)$$

为使该迭代格式产生的 $\{x_k\}$ 二阶收敛, 由定理 7.1.7 知, 应选取松弛因子 ω , 使得 $\psi'(x^*) = 0$. 由此得

$$\omega = \phi'(x^*).$$

因为 x^* 是所求的, 未知, 所以最佳松弛因子不能准确求出, 通常取 $\omega = \phi'(\bar{x})$, 其中 $\bar{x}$ 为 x^* 的一个好的近似值 (如 $\bar{x}$ 用二分法求得). 将 ω 代入 (7.1.18) 得迭代格式

$$x_{k+1} = \frac{\phi(x_k) - \phi'(\bar{x}) x_k}{1 - \phi'(\bar{x})}. \quad (7.1.19)$$

虽然该迭代格式的 $\psi'(x^*) \neq 0$, 但 $|\psi'(x^*)| \ll 1$. 这就大大地提高了收敛速度.

例如, 对例 7.1.1 解法 3, 可将不收敛的迭代格式改造为收敛的格式.

迭代函数 $\phi(x) = \frac{1}{2}(x^3 - 5)$ 的导数 $\phi'(x) = \frac{3}{2}x^2$. 取 $\bar{x} = 2.5$, 则 $\phi'(2.5) = 9.375$. 据式 (7.1.19), 构造迭代格式:

$$x_{k+1} = \frac{18.75x_k - x_k^3 + 5}{16.75}.$$

取 $x_0 = 2.5$, 按上式计算可得

$$\begin{aligned} x_1 &= 2.164\,179\,105, & x_2 &= 2.115\,943\,575, & x_3 &= 2.101\,516\,599, \\ x_4 &= 2.096\,857\,142, & x_5 &= 2.095\,318\,760, & x_6 &= 2.094\,807\,260, \\ x_7 &= 2.094\,636\,796, & x_8 &= 2.094\,579\,943, & x_9 &= 2.094\,560\,997\,7, \\ x_{10} &= 2.094\,554\,667, & x_{11} &= 2.094\,552\,545, & x_{12} &= 2.094\,551\,836, \\ &&&&& \dots\dots\dots \end{aligned}$$

由以上数据可见, 新迭代格式迭代 10 次后, 计算结果准确到 5 位小数. 若 $\bar{x}$ 取得越靠近 x^* , 则 $\{x_k\}$ 收敛得越快.

2. 艾特肯加速法

设迭代序列 $\{x_k\}$ 线性收敛于 x^* , 这时,

$$\frac{x^* - x_{k+1}}{x^* - x_k} \approx \frac{x^* - x_k}{x^* - x_{k-1}}.$$

由此可以解出

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

取上式的右端作为 x^* 的新近似值, 即令

$$\bar{x}_{k+1} = x_{k+1} - \frac{(x_{k+1} - x_k)^2}{x_{k+1} - 2x_k + x_{k-1}}, \quad (7.1.20)$$

则 $\bar{x}_{k+1}$ 比 x_{k+1} 更接近于 x^* . 式 (7.1.20) 称为艾特肯 (Aitken) 加速法, 它是利用迭代过程中相邻的三点 x_{k-1}, x_k, x_{k+1} 推出 x^* 的更好近似值的一种方法.

由艾特肯加速法可得斯特芬森 (Steffensen) 加速技术, 其计算过程如下:

给定 x_0, ε ,

for $k = 0$ to K do

$x_1 = \phi(x_0), \quad x_2 = \phi(x_1)$

$$\bar{x} = x_2 - \frac{(x_2 - x_1)^2}{x_2 - 2x_1 + x_0}$$

如果 $|\bar{x} - x_2| < \varepsilon$, 则取 $x^* \approx \bar{x}$, 停止;

否则, $x_0 := \bar{x}$

end do

可证经斯特芬森加速后的序列 $\{x_k\}$ 二阶收敛于 x^* , 留作作业, 请读者自证.

例如, 对例 7.1.1 解法 1, 取 $x_0 = 2.5$, 用斯特芬森加速技术计算, 当 $k = 3$ 时, 计算结果已准确到 9 位小数, 计算结果为

$$x^* \approx 2.094551482.$$

7.2 求解非线性代数方程组的迭代法

本节介绍求解如下非线性方程组的迭代解法:

$$\begin{cases} f_1(x_1, x_2, \dots, x_n) = 0, \\ f_2(x_1, x_2, \dots, x_n) = 0, \\ \dots\dots\dots \\ f_n(x_1, x_2, \dots, x_n) = 0, \end{cases} \quad (7.2.1)$$

其中 $f_i(x_1, x_2, \dots, x_n)$ 为定义在区域 $D \subset \mathbf{R}^n$ 上的实值函数, $f_i(x_1, x_2, \dots, x_n)$ 中至少有一个是非线性函数. 为了书写简便, 记 $x = (x_1, x_2, \dots, x_n)^T$, $f(x) = (f_1(x), f_2(x), \dots, f_n(x))^T$, 则非线性方程组 (7.2.1) 可以表示成

$$f(x) = 0.$$

7.2.1 简单迭代法

与 7.1 节单个方程求根的简单迭代法相似, 将方程组 (7.2.1) 改写为同解方程组

$$x_i = \phi_i(x_1, x_2, \dots, x_n), \quad i = 1, 2, \dots, n.$$

由此建立迭代格式

$$x_i^{(k+1)} = \phi_i(x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}), \quad i = 1, 2, \dots, n. \quad (7.2.2)$$

选取初始向量 $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})^T$, 令 $k = 1, 2, \dots$, 由式 (7.2.2) 可得向量迭代序列 $\{x^{(k)}\}$, 如果方程组 (7.2.1) 有唯一解 x^* , 并且 $\{x^{(k)}\}$ 收敛, 则 $\{x^{(k)}\}$ 收敛于 x^* . 按式 (7.2.2) 求解方程组 (7.2.1) 的方法称为简单迭代法. 记

$$\phi(x) = (\phi_1(x), \phi_2(x), \dots, \phi_n(x))^T, \quad x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})^T.$$

$\phi(x)$ 称为迭代函数, 则式 (7.2.2) 可写成向量形式

$$x^{(k+1)} = \phi(x^{(k)}), \quad k = 0, 1, \dots, \quad (7.2.3)$$

与单个方程不动点的定义相同, 向量函数的不动点定义如下: 若 x^* 满足 $x^* = \phi(x^*)$, 则称 x^* 为 $\phi(x)$ 的不动点. 不动点 x^* 满足方程 $f(x^*) = 0$, 即 x^* 是方程组 (7.2.1) 的解.

类似于求解线性方程组的高斯-赛德尔迭代法, 求解非线性方程组的高斯-赛德尔迭代格式为

$$x_i^{(k+1)} = \phi_i(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k)}, \dots, x_n^{(k)}), \quad i = 1, 2, \dots, n. \quad (7.2.4)$$

例 7.2.1 用简单迭代法解方程组

$$\begin{cases} 4x_1 - x_2 + 0.1e^{x_1} = 1, \\ -x_1 + 4x_2 + \frac{x_1^2}{8} = 0. \end{cases}$$

解 作迭代格式

$$\begin{cases} x_1^{(k+1)} = \frac{1}{4}(1 + x_2^{(k)} - 0.1e^{x_1^{(k)}}), \\ x_2^{(k+1)} = \frac{1}{4}\left(x_1^{(k)} - \frac{1}{8}(x_1^{(k)})^2\right), \end{cases} \quad k = 0, 1, 2, \dots$$

取 $x^{(0)} = (0, 0)^T$, 按以上迭代格式计算得

$$\begin{aligned} x^{(1)} &= (0.225, 0)^T, \\ x^{(2)} &= (0.218\ 691\ 932\ 1, 0.054\ 667\ 968\ 8)^T, \\ x^{(3)} &= (0.232\ 555\ 796\ 1, 0.053\ 178\ 415\ 5)^T, \\ x^{(4)} &= (0.231\ 749\ 082\ 6, 0.056\ 448\ 880\ 3)^T, \\ &\dots\dots\dots \\ x^{(18)} &= (0.232\ 567\ 005\ 1, 0.056\ 451\ 519\ 7)^T = x^{(19)}. \end{aligned}$$

取 $x^* \approx x^{(18)}$. 解毕.

例 7.2.2 用高斯-赛德尔迭代格式 (7.2.4) 解例 7.2.1 的方程组.

解 高斯-赛德尔迭代格式

$$\begin{cases} x_1^{(k+1)} = \frac{1}{4}(1 + x_2^{(k)} - 0.1e^{x_1^{(k)}}), \\ x_2^{(k+1)} = \frac{1}{4}\left(x_1^{(k+1)} - \frac{1}{8}(x_1^{(k)})^2\right), \end{cases} \quad k = 0, 1, 2, \dots$$

取 $x^{(0)} = (0, 0)^T$ 得

$$\begin{aligned} x^{(1)} &= (0.225, 0.054\ 667\ 968\ 8)^T, \\ x^{(2)} &= (0.232\ 358\ 924\ 3, 0.056\ 402\ 522\ 6)^T, \\ x^{(3)} &= (0.232\ 561\ 319\ 2, 0.056\ 450\ 180\ 8)^T, \\ &\dots\dots\dots \\ x^{(7)} &= (0.232\ 567\ 005\ 1, 0.056\ 451\ 519\ 7)^T = x^{(8)}. \end{aligned}$$

取 $x \approx x^{(7)}$. 解毕.

下面简要介绍简单迭代法的收敛性.

设向量函数 $\phi(x)$ 的所有分量函数 $\phi_i(x)$ 在点 x 处可微, 矩阵

$$J_\phi(x) = \begin{pmatrix} \frac{\partial \phi_1(x)}{\partial x_1} & \frac{\partial \phi_1(x)}{\partial x_2} & \dots & \frac{\partial \phi_1(x)}{\partial x_n} \\ \frac{\partial \phi_2(x)}{\partial x_1} & \frac{\partial \phi_2(x)}{\partial x_2} & \dots & \frac{\partial \phi_2(x)}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \phi_n(x)}{\partial x_1} & \frac{\partial \phi_n(x)}{\partial x_2} & \dots & \frac{\partial \phi_n(x)}{\partial x_n} \end{pmatrix}$$

称为向量函数 $\phi(x)$ 在 x 处的导数或在 x 处的雅可比矩阵.

与 7.1 节单个方程迭代序列收敛阶的定义类似, 可定义向量序列 $\{x^{(k)}\}$ 的收敛阶.

定义 7.2.1 设向量序列 $\{x^{(k)}\}$ 收敛于 x^* , 若存在常数 $p \geq 1$ 和 $c > 0$, 使得

$$\lim_{k \rightarrow \infty} \frac{\|x^* - x^{(k+1)}\|}{\|x^* - x^{(k)}\|^p} = c, \quad (7.2.5)$$

则称 $\{x^{(k)}\}$ p 阶收敛于 x^* .

简单迭代法 (7.2.3) 的全局收敛性定理如下:

定理 7.2.1 设映射 $\phi: D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$, 在闭区域 $D_0 \subset D$ 上满足

(i) $\phi(x) \in D_0$ ($\forall x \in D_0$);

(ii) 存在常数 L ($0 < L < 1$), 使得 $\|\phi(x) - \phi(y)\| \leq L\|x - y\|$ ($\forall x, y \in D_0$), 则以下结论成立:

(1) ϕ 在 D_0 上存在唯一的不动点 x^* ;

(2) 对任意的初始向量 $x^{(0)} \in D_0$, 由简单迭代法 (7.2.3) 产生的迭代序列 $\{x^{(k)}\} \subset D_0$ 线性收敛于 x^* , 并且有误差估计式

$$\|x^* - x^{(k)}\| \leq \frac{L^k}{1-L} \|x^{(1)} - x^{(0)}\|, \quad (7.2.6)$$

$$\|x^* - x^{(k)}\| \leq \frac{L}{1-L} \|x^{(k)} - x^{(k-1)}\|. \quad (7.2.7)$$

定理 7.2.1 条件 (1) 表明 ϕ 将区域 D_0 映入自身, 条件 (2) 称为 ϕ 的压缩条件. 若 ϕ 是压缩的, 则它是连续的. 该定理也称为压缩映射原理.

证明 证明过程完全类似于定理 7.1.1 的证明, 在此从略.

简单迭代法 (7.2.3) 的局部收敛性定理如下:

定理 7.2.2 设映射 ϕ 在其定义域内有不动点 x^* , ϕ 的分量有连续的偏导数且矩阵 $J_\phi(x^*)$ 的谱半径

$$\rho(J_\phi(x^*)) < 1, \quad (7.2.8)$$

则存在 x^* 的一个邻域 $D_\delta = \{x \mid \|x - x^*\| < \delta, \delta > 0\} \subset D_0$, 对任意的初始向量 $x^{(0)} \in D_\delta$, 由简单迭代法 (7.2.3) 产生的迭代序列 $\{x^{(k)}\} \subset D_\delta$ 收敛于 x^* .

7.2.2 牛顿法

1. 牛顿法

牛顿法是求解非线性方程组的一个十分重要的算法. 它的基本思想是, 将非线性函数逐次线性化, 从而形成一个迭代过程.

仿照单个方程求根的牛顿法, 将方程组 (7.2.1) 的每个方程在点 $x^{(k)} = (x_1^{(k)}, x_2^{(k)})$,

$\dots, x_n^{(k)})^T$ 处按多元函数泰勒公式展开有

$$\begin{aligned} f_i(x) &= f_i(x^{(k)}) + \sum_{j=1}^n \frac{\partial f_i(x^{(k)})}{\partial x_j} (x_j - x_j^{(k)}) + \dots \\ &= f_i(x^{(k)}) + \nabla f_i(x^{(k)})^T (x - x^{(k)}) + \dots, \quad i = 1, 2, \dots, n. \end{aligned}$$

取线性项, 由 $f_i(x) = 0$ 得近似方程组

$$f_i(x^{(k)}) + \nabla f_i(x^{(k)})^T (x - x^{(k)}) = 0, \quad i = 1, 2, \dots, n. \quad (7.2.9)$$

将式 (7.2.9) 表示成矩阵向量形式, 即

$$J_f(x^{(k)})(x - x^{(k)}) = -f(x^{(k)}). \quad (7.2.10)$$

将此方程的解记为 $x^{(k+1)}$, 则

$$x^{(k+1)} = x^{(k)} - J_f^{-1}(x^{(k)})f(x^{(k)}). \quad (7.2.11)$$

式 (7.2.11) 称为求解非线性方程组 (7.2.1) 的牛顿法.

令 $\Delta x^{(k)} = x - x^{(k)}$, 则方程组 (7.2.10) 为

$$J_f(x^{(k)})\Delta x^{(k)} = -f(x^{(k)}). \quad (7.2.12)$$

求解时先从方程组 (7.2.12) 解出 $\Delta x^{(k)}$, 再令

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)}.$$

即按下式计算:

$$\begin{cases} J_f(x^{(k)})\Delta x^{(k)} = -f(x^{(k)}), \\ x^{(k+1)} = x^{(k)} + \Delta x^{(k)}. \end{cases} \quad (7.2.13)$$

求解非线性方程组 (7.2.1) 的牛顿法的计算过程如下:

- (1) 在 x^* 的附近选取 $x^{(0)} \in D$, 给定允许误差限 $\varepsilon_1, \varepsilon_2$ 和最大迭代次数 K ;
- (2) for $k = 0$ to K do
 - (i) 计算 $f(x^{(k)}), J_f(x^{(k)})$;
 - (ii) 求解关于 $\Delta x^{(k)}$ 的线性方程组 $J_f(x^{(k)})\Delta x^{(k)} = -f(x^{(k)})$;
 - (iii) 令 $x^{(k+1)} = x^{(k)} + \Delta x^{(k)}$;
 - (iv) 若 $\|\Delta x^{(k)}\|/\|x^{(k)}\| < \varepsilon_1$, 或者 $\|f(x^{(k+1)})\| < \varepsilon_2$, 则取 $x^* \approx x^{(k+1)}$, 停止

计算; 如果 $k = K$, 输出 K 次迭代不满足精度要求的信息, 则停机;
end do

例 7.2.3 用牛顿法解例 7.2.1 的方程组.

解 按式 (7.2.13), 牛顿迭代格式为

$$\begin{pmatrix} 4 + 0.1e^{x_1^{(k)}} & -1 \\ -1 + 0.25x_1^{(k)} & 4 \end{pmatrix} \begin{pmatrix} \Delta x_1^{(k)} \\ \Delta x_2^{(k)} \end{pmatrix} = \begin{pmatrix} -4x_1^{(k)} - 0.1e^{x_1^{(k)}} + x_2^{(k)} + 1 \\ x_1^{(k)} - 0.125(x_1^{(k)})^2 - 4x_2^{(k)} \end{pmatrix},$$

$$\begin{cases} x_1^{(k+1)} = x_1^{(k)} + \Delta x_1^{(k)}, \\ x_2^{(k+1)} = x_2^{(k)} + \Delta x_2^{(k)}. \end{cases}$$

取 $x^{(0)} = (0, 0)^T$ 得

$$\begin{aligned} x^{(1)} &= (0.233\ 766\ 233\ 8, 0.058\ 441\ 558\ 4)^T, \\ x^{(2)} &= (0.232\ 567\ 040\ 0, 0.056\ 451\ 572\ 8)^T, \\ x^{(3)} &= (0.232\ 567\ 005\ 1, 0.056\ 451\ 519\ 7)^T = x^{(4)}, \end{aligned}$$

故 $x^* \approx x^{(3)}$.

牛顿法的局部收敛性定理如下:

定理 7.2.3 设 $x^* \in \text{int}(D)$ 是方程组 $f(x) = 0$ 的解, $f(x)$ 在包含 x^* 的某个开区域 $D_0 \subset D$ 内连续可微且 $J_f(x^*)$ 非奇异, 则存在闭球

$$\bar{B}(x^*; \delta) = \{x \mid \|x - x^*\| \leq \delta, \delta > 0\} \subset D_0,$$

使得对任意的 $x^{(0)} \in \bar{B}(x^*; \delta)$, 由牛顿法产生的迭代序列 $\{x^{(k)}\} \subset \bar{B}(x^*; \delta)$ 且超线性收敛于 x^* .

进而, 若 $f(x)$ 在 D_0 上二阶连续可微, 则序列 $\{x^{(k)}\}$ 至少二阶收敛.

证明 略.

牛顿法的优点是收敛速度快, 一般能达到二阶收敛速度. 它的不足之处是每次迭代都要计算一个由 n^2 个偏导数构成的雅可比矩阵 $J_f(x^{(k)})$, 并且还要解一个线性方程组. 当 n 较大时, 其计算量是巨大的.

2. 简化牛顿

为了减少牛顿法的计算量, 将牛顿迭代格式 (7.2.12) 中的雅可比矩阵 $J_f(x^{(k)})$ 用常数矩阵 $J_f(x^{(0)})$ 代替得

$$J_f(x^{(0)})\Delta x^{(k)} = -f(x^{(k)}). \quad (7.2.14)$$

由此得到简化牛顿法的迭代格式为

$$\begin{cases} J(x^{(0)})\Delta x^{(k)} = -f(x^{(k)}), \\ x^{(k+1)} = x^{(k)} + \Delta x^{(k)}. \end{cases} \quad (7.2.15)$$

在迭代开始前, 可先将线性方程组的系数矩阵 $J(x^{(0)})$ 作三角分解, 以后每次迭代只需解两个三角方程组, 从而可以大大地减少每次迭代的计算量. 简化牛顿法的收敛速度一般比牛顿法低.

7.2.3 弦割法

牛顿法中的方程组 (7.2.9) 即

$$f_i(x^{(k)}) + \sum_{j=1}^n \frac{\partial f_i(x^{(k)})}{\partial x_j} (x_j - x_j^{(k)}) = 0, \quad i = 1, 2, \dots, n.$$

将上式中在 $x^{(k)}$ 处的偏导数改用差商代替得

$$f_i(x^{(k)}) + \sum_{j=1}^n \frac{1}{h} [f_i(x^{(k)} + e_j h) - f_i(x^{(k)})] (x_j - x_j^{(k)}) = 0, \quad i = 1, \dots, n.$$

由此得迭代格式

$$\sum_{j=1}^n \frac{1}{h} [f_i(x^{(k)} + e_j h) - f_i(x^{(k)})] \Delta x_j^{(k)} = -f_i(x^{(k)}), \quad i = 1, \dots, n, \quad (7.2.16)$$

其中 $\Delta x_j^{(k)} = x_j - x_j^{(k)}$, $e_j = (0, \dots, 0, 1, 0, \dots, 0)^T \in \mathbf{R}^n$ 为第 j 个坐标轴上的单位向量. 这时, 迭代法称为弦割法. 它的收敛速度一般介于牛顿法和简化牛顿法之间.

下面把式 (7.2.16) 改写成便于计算的形式. 记 $f_{ij}^{(k)} = f_i(x^{(k)} + e_j h)$, $f_i^{(k)} = f_i(x^{(k)})$, 则

$$f_i^{(k)} + \sum_{j=1}^n \frac{1}{h} (f_{ij}^{(k)} - f_i^{(k)}) \Delta x_j^{(k)} = 0, \quad i = 1, 2, \dots, n,$$

即

$$\left(1 - \sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h} \right) f_i^{(k)} + \sum_{j=1}^n f_{ij}^{(k)} \frac{\Delta x_j^{(k)}}{h} = 0, \quad i = 1, 2, \dots, n,$$

也即

$$\sum_{j=1}^n f_{ij}^{(k)} \frac{\frac{\Delta x_j^{(k)}}{h}}{\sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h} - 1} = f_i^{(k)}, \quad i = 1, 2, \dots, n.$$

将上式写成

$$\sum_{j=1}^n f_{ij}^{(k)} z_j = f_i^{(k)}, \quad i = 1, 2, \dots, n, \quad (7.2.17)$$

其中

$$z_j = \frac{\frac{\Delta x_j^{(k)}}{h}}{\sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h} - 1}. \quad (7.2.18)$$

解方程组 (7.2.17) 得 $z_j (j = 1, 2, \dots, n)$.

下面导出 z_j 与 $\Delta x_j^{(k)}$ 之间的关系. 由式 (7.2.18) 得

$$\Delta x_j^{(k)} = \left(\sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h} - 1 \right) h z_j. \quad (7.2.19)$$

对式 (7.2.18) 关于 j 求和得

$$\sum_{j=1}^n z_j = \frac{\sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h}}{\sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h} - 1} = 1 + \frac{1}{\sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h} - 1}.$$

由此得到 $\sum_{j=1}^n \frac{\Delta x_j^{(k)}}{h} - 1 = \frac{1}{\sum_{j=1}^n z_j - 1}$. 将该式代入式 (7.2.19) 得

$$\Delta x_j^{(k)} = \frac{h z_j}{\sum_{j=1}^n z_j - 1}.$$

由上可知, 弦割法首先由方程组 (7.2.17) 解出 $z_j (j = 1, 2, \dots, n)$, 然后令

$$x_j^{(k+1)} = x_j^{(k)} + \Delta x_j^{(k)} = x_j^{(k)} + \frac{h z_j}{\sum_{j=1}^n z_j - 1}, \quad j = 1, 2, \dots, n. \quad (7.2.20)$$

综合以上, 弦割法的迭代格式为

$$\left\{ \begin{array}{l} \text{求解线性方程组 } \sum_{j=1}^n f_{ij}^{(k)} z_j = f_i^{(k)} \quad (i = 1, 2, \dots, n), \\ \text{令 } x_j^{(k+1)} = x_j^{(k)} + \Delta x_j^{(k)} = x_j^{(k)} + \frac{h z_j}{\sum_{j=1}^n z_j - 1} \quad (j = 1, 2, \dots, n). \end{array} \right.$$

若 $\|x^{(k+1)} - x^{(k)}\| < \varepsilon_1$, 或 $\|f(x^{(k+1)})\| < \varepsilon_2$ 则终止迭代, 取 $x^* \approx x^{(k+1)}$.

7.2.4 布洛依登法

为了避免牛顿法中计算 $f_i(x)$ ($i = 1, 2, \dots, n$) 的偏导数和 n^2 个偏导数函数值, 用常数矩阵 A_k 替代牛顿法

$$x^{(k+1)} = x^{(k)} - J_f^{-1}(x^{(k)})f(x^{(k)})$$

中的雅可比矩阵 $J_f(x^{(k)})$, 得到如下的布洛依登 (Broyden) 方法的迭代公式:

$$x^{(k+1)} = x^{(k)} - A_k^{-1}f(x^{(k)}). \quad (7.2.21)$$

1. 矩阵 A_k 应满足的条件

取

$$f(x) \approx f(x^{(k)}) + J_f(x^{(k)})(x - x^{(k)}),$$

则

$$f(x^{(k-1)}) \approx f(x^{(k)}) + J_f(x^{(k)})(x^{(k-1)} - x^{(k)}).$$

用 A_k 代替上式中的 $J_f(x^{(k)})$, 并令

$$f(x^{(k-1)}) = f(x^{(k)}) + A_k(x^{(k-1)} - x^{(k)}),$$

由此可得

$$A_k(x^{(k)} - x^{(k-1)}) = f(x^{(k)}) - f(x^{(k-1)}).$$

记

$$s^{(k)} = x^{(k)} - x^{(k-1)}, \quad y^{(k)} = f(x^{(k)}) - f(x^{(k-1)}),$$

则有

$$A_k s^{(k)} = y^{(k)}. \quad (7.2.22)$$

2. 构造矩阵 A_k

令

$$A_k = A_{k-1} + \Delta A_{k-1} = A_{k-1} + u s^{(k)T},$$

其中 $u \in \mathbf{R}^n$, $s^{(k)} \in \mathbf{R}^n$, ΔA_{k-1} 称为修正矩阵. 下面讨论向量 u 的取法.

由式 (7.2.22) 知

$$y^{(k)} = A_k s^{(k)} = (A_{k-1} + u s^{(k)T}) s^{(k)} = A_{k-1} s^{(k)} + s^{(k)T} s^{(k)} u.$$

由此可得

$$u = \frac{y^{(k)} - A_{k-1} s^{(k)}}{s^{(k)T} s^{(k)}};$$

所以

$$A_k = A_{k-1} + u s^{(k)T} = A_{k-1} + \frac{(y^{(k)} - A_{k-1} s^{(k)}) s^{(k)T}}{s^{(k)T} s^{(k)}}.$$

3. 求 A_k^{-1}

定理 7.2.4 (谢尔曼-莫里森 (Sherman-Morrison)) 设 A 是 n 阶可逆矩阵, $u, v \in \mathbf{R}^n$ 是任意向量, 若 $1 + v^T A^{-1} u \neq 0$, 则 $A + uv^T$ 是可逆矩阵, 其逆矩阵为

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

当利用谢尔曼-莫里森定理求矩阵 A_k 的逆矩阵时, u 的取法同前, 取 $v = s^{(k)}$, 由定理 7.2.4 可求得矩阵 A_k 的逆矩阵为

$$A_k^{-1} = A_{k-1}^{-1} + \frac{(s^{(k)} - A_{k-1}^{-1}y^{(k)})s^{(k)T}A_{k-1}^{-1}}{s^{(k)T}A_{k-1}^{-1}y^{(k)}}.$$

4. 布洛依登算法的计算步骤

(1) 给定初始近似向量 $x^{(0)} \in D$, 允许误差限 $\varepsilon_1, \varepsilon_2$ 和最大迭代次数 K ; 取初始矩阵 $A_0 = J_f(x^{(0)})$, 计算 $x^{(1)} = x^{(0)} - A_0^{-1}f(x^{(0)})$;

(2) for $k = 1$ to K do

$$s^{(k)} := x^{(k)} - x^{(k-1)};$$

$$y^{(k)} := f(x^{(k)}) - f(x^{(k-1)});$$

$$A_k^{-1} := A_{k-1}^{-1} + \frac{(s^{(k)} - A_{k-1}^{-1}y^{(k)})s^{(k)T}A_{k-1}^{-1}}{s^{(k)T}A_{k-1}^{-1}y^{(k)}}.$$

$$x^{(k+1)} := x^{(k)} - A_k^{-1}f(x^{(k)});$$

如果 $\|x^{(k+1)} - x^{(k)}\| < \varepsilon_1$ 或者 $\|f(x^{(k+1)})\| < \varepsilon_2$, 取 $x^* \approx x^{(k+1)}$, 停止;

如果 $k = K$, 输出 K 次迭代不满足精度要求的信息, 停机;

end do

布洛依登算法在一定的条件下具有局部超线性收敛速度, 其理论比较复杂, 在此不予以讨论.

小 结

本章介绍了求解非线性方程和非线性方程组的迭代解法.

非线性方程的解法有二分法、简单迭代法、牛顿法和割线法. 在有限区间上, 二分法一定收敛, 但收敛速度一般很慢, 常用于作根的隔离和求其他迭代法的初始点. 简单迭代法在一定的条件下全局收敛, 一般具有线性收敛速度, 这时可用松弛加速法或艾特肯加速法进行加速. 构造能够保证迭代点列收敛的迭代函数是该法的一个难点. 实际计算时, 最好先判断收敛性, 但也不妨先试算几步, 以检查迭代序列是否收敛. 牛顿法和弦割法是两个最常用的方法, 它们在一定的条件下全局收敛,

但这些条件往往难以检验, 一般用二分法求其初始点. 牛顿法的优点是收敛速度快, 二阶收敛. 它的主要缺陷是在方程的根处函数的导数接近于零时, 它的收敛速度很慢. 弦割法用差商代替了牛顿法中的微商, 在迭代中不需要计算函数的导数值, 但开始计算时需要两个初始点, 它具有超线性收敛速度.

由于非线性方程组的向量形式与单个方程的形式完全相同, 因此, 求解单个方程的许多算法的思想和理论都可以推广到方程组. 本章简要介绍了求解非线性方程组的简单迭代法、牛顿法、弦割法和布洛依登法. 如果导数计算量不大, 则通常用牛顿法, 它具有二阶收敛速度; 如果导数计算量大, 则宜用简化牛顿法、弦割法或简单迭代法, 简单迭代法线性收敛. 当未知量个数较多时, 宜用布洛依登法, 该法局部超线性收敛.

为保证迭代序列的收敛性, 选取初始向量应充分靠近方程组的解.

习 题

7.1 证明方程 $1 - x - \sin x = 0$ 在 $[0, 1]$ 上有且仅有一个根. 若用二分法求此根, 欲使误差不超过 0.5×10^{-4} , 问需计算左边函数值多少次.

7.2 方程 $x^3 - x^2 - 1 = 0$ 在 $x_0 = 1.5$ 附近有根, 将方程作三种同解变形, 可得三种迭代格式:

$$(1) x = 1 + \frac{1}{x^2}, x_{k+1} = 1 + \frac{1}{x_k^2};$$

$$(2) x^3 = 1 + x^2, x_{k+1} = \sqrt[3]{1 + x_k^2};$$

$$(3) x^2 = \frac{1}{x-1}, x_{k+1} = \frac{1}{\sqrt{x_k-1}}.$$

判断各迭代格式在 $x_0 = 1.5$ 附近的收敛性; 选一种收敛最快的迭代格式, 计算在 $x_0 = 1.5$ 附近的根, 准确到 4 位小数.

7.3 设在区间 $[a, b]$ 上 $0 < m \leq f'(x) \leq M$, 证明对任意 $\lambda \in (0, 2/M)$, 取 $x_0 \in [a, b]$ 时, 由迭代格式

$$x_{k+1} = x_k - \lambda f(x_k), \quad k = 0, 1, 2, \dots$$

产生的迭代序列 $\{x_k\}$ 收敛于方程 $f(x) = 0$ 的根 x^* .

7.4 用迭代法的思想证明

$$\lim_{k \rightarrow \infty} \underbrace{\sqrt{2 + \sqrt{2 + \cdots + \sqrt{2 + \sqrt{2}}}}}_{\text{开方 } k \text{ 次}} = 2.$$

7.5 设在区间 $[a, b]$ 上 $0 < \phi'(x) \leq L < 1$, $x = \phi(x)$ 有根 x^* , $x_0 \in [a, b]$, 但 $x_0 \neq x^*$, 证明由迭代格式 $x_{k+1} = \phi(x_k)$ 产生的迭代序列 $\{x_k\}$ 单调地收敛于 x^* .

7.6 已知在区间 $[a, b]$ 上 $|\phi'(x)| \geq k > 1$, $x = \phi(x)$ 只有一个根, 试写出收敛的迭代格式, 并求 $x = \tan x$ 在 $x = 4.5$ 附近的根, 准确到 4 位小数.

7.7 设 $f(x^*) = 0$, $f'(x^*) \neq 0$, $\phi(x) = x - m(x)f(x)$, 证明简单迭代格式 $x_{k+1} = \phi(x_k)$ 当 $m(x^*) \neq [f'(x^*)]^{-1}$ 时一阶收敛, 当 $m(x^*) = [f'(x^*)]^{-1}$ 时至少二阶收敛.

7.8 怎样选取函数 $h(x)$, 使得当

$$\phi(x) = x - \frac{f(x)}{f'(x)} + h(x) \left[\frac{f(x)}{f'(x)} \right]^2$$

时, 简单迭代格式 (7.1.2) 至少三阶收敛.

7.9 用松弛法和斯特芬森加速技术求解习题 7.2.

7.10 设简单迭代法 $x_{k+1} = \phi(x_k)$ 的迭代函数 $\phi(x)$ 在其不动点 x^* 的某领域内二阶连续可微, 且 $\phi'(x^*) = 0, 1$, 证明斯特芬森加速技术二阶收敛于 x^* .

7.11 用牛顿法、弦割法和牛顿下山法求解习题 7.2.

7.12 应用牛顿法于方程 $x^n - a = 0$ 和 $1 - a/x^n = 0$, 分别导出求 $\sqrt[n]{a}$ 的迭代公式. 已知 $\sqrt[4]{3} \approx x_0 = 1.3$, 问用这两种公式迭代二次后能得到根的准确到几位小数的近似值? (已知 $\sqrt[4]{3} = 1.316074013 \dots$).

7.13 证明 $f(x) = (x - x^*)^m g(x)$ 且 $g(x^*) \neq 0$, 当 $m > 1$ 时, 求解方程 $f(x) = 0$ 的牛顿法一阶收敛, 但令 $u = f(x)/f'(x)$ 时, 迭代法

$$x_{k+1} = x_k - m \frac{f(x_k)}{f'(x_k)} \quad \text{或} \quad x_{k+1} = x_k - \frac{u(x_k)}{u'(x_k)}$$

二阶收敛.

7.14 用简单迭代法和高斯-赛德尔迭代法求解方程组

$$2x - \cos y = 0, \quad 2y - \sin x = 0,$$

$|x| < 1, |y| < 1$, 准确到 2 位小数.

7.15 用牛顿法求解如下方程组:

$$x_1 + 2x_2 - 3 = 0, \quad 2x_1^2 + x_2^2 - 5 = 0$$

在 $x_1 = -1, x_2 = 2$ 附近的解, 准确到 2 位小数.

7.16 用牛顿法求解方程组

$$x^2 + xy^3 - 9 = 0, \quad 3x^2y - y^3 - 4 = 0,$$

初始点 (x_0, y_0) 分别取为 $(1.2, 2.5), (-2, 2.5), (2, -2.5)$, 观察迭代序列收敛于哪个解及收敛速度.

7.17 给定 $x^{(0)} = (0.6, 0.8)^T$ 用布洛依登法求解如下方程组的近似解, 准确到 2 位小数:

$$x_1^2 + x_2^2 - 1 = 0, \quad 2x_1 + x_2 - 1 = 0.$$

计算实习

7.1 在习题 7.1 和习题 7.2 中选一题, 利用简单迭代法、两种加速迭代法、牛顿法、弦割法求解.

7.2 利用简单迭代法、牛顿法、弦割法求解方程

$$f(x) = x^6 - 5x^5 + 3x^4 + x^3 - 7x^2 + 7x - 20 = 0$$

在区间 $[-1, 5]$ 内的全部实根 (先用二分法作根的隔离).

7.3 用牛顿法、弦割法、布洛依登法求以下方程组的解:

$$(1) \begin{cases} x_1^2 + x_2^2 + x_3^2 - 1.0 = 0, \\ 2x_1^2 + x_2^2 - 4x_3 = 0, \\ 3x_1^2 - 4x_2^2 + x_3^2 = 0, \end{cases} \quad \text{给定初始向量 } x^{(0)} = (1.0, 1.0, 1.0)^T;$$

$$(2) \begin{cases} \cos(x_1^2 + 0.4x_2) + x_1^2 + x_2^2 - 1.6 = 0, \\ 1.5x_1^2 - \frac{1}{0.36}x_2^2 - 1.0 = 0, \end{cases} \quad \text{给定初始向量 } x^{(0)} = (1.04, 0.47)^T.$$

第8章 矩阵特征值与特征向量的计算

特征值和特征向量是纯数学和应用数学中非常重要的基本概念, 常用来研究微分方程、离散和连续的动力系统, 在物理、化学和众多工程领域中有着广泛的应用. 本章介绍求解矩阵特征值、特征向量的常用数值方法.

8.1 基本性质

首先介绍特征值和特征向量的基本概念和性质.

定义 8.1.1 A 为 $n \times n$ 阶矩阵, x 为非零向量, 若存在数 λ , 使得 $Ax = \lambda x$ 成立, 则称 λ 为 A 的特征值, x 为对应于 λ 的特征向量.

定义 8.1.2 若 λ 满足 $P_A(\lambda) = \det(\lambda I - A) = 0$, 则称 $P_A(\lambda)$ 为矩阵 A 的特征多项式, $\det(\lambda I - A) = 0$ 称为特征方程.

定义 8.1.3 若 λ_i 为特征方程的 k 重根, 称 k 为特征值 λ_i 的代数重数; 齐次线性方程组 $(\lambda_i I - A)x = 0$ 的基础解系所含向量的个数 (即 λ_i 的特征子空间的维数) 称为特征值 λ_i 的几何重数.

性质 8.1.1 设 n 阶矩阵 $A = (a_{ij})_{n \times n}$ 的全部特征值为 $\lambda_1, \lambda_2, \dots, \lambda_n$, 则有

$$\lambda_1 + \lambda_2 + \dots + \lambda_n = a_{11} + a_{22} + \dots + a_{nn}, \quad (8.1.1)$$

$$\lambda_1 \lambda_2 \cdots \lambda_n = \det(A). \quad (8.1.2)$$

性质 8.1.2 设 λ 为矩阵 A 的一个特征值, 则对任何多项式 $P(x) = a_0 + a_1 x + \dots + a_m x^m$, $P(\lambda)$ 为矩阵 $P(A) = a_0 I + a_1 A + \dots + a_m A^m$ 的一个特征值.

性质 8.1.3 设 λ 为 n 阶可逆矩阵 A 的一个特征值, 则 $\lambda \neq 0$ 且 $\frac{1}{\lambda}$ 为逆矩阵 A^{-1} 的一个特征值, $\frac{\det(A)}{\lambda}$ 为 A 的伴随矩阵 A^* 的一个特征值.

性质 8.1.4 设 $\lambda_1, \lambda_2, \dots, \lambda_m$ 为矩阵 A 的互不相同的特征值, x_i 为 A 的属于特征值 λ_i ($i = 1, 2, \dots, m$) 的特征向量, 则 $x_1, x_2, \dots, x_m$ 线性无关.

性质 8.1.5 矩阵 A 的任意特征值的几何重数不大于其代数重数.

8.2 求一般矩阵特征值的计算方法

8.2.1 乘幂法及反幂法

1. 乘幂法

乘幂法用于求模 (绝对值) 最大的特征值. 设 A 的特征值按模大小排列为 $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$, 并且对应的特征向量 $\xi_1, \xi_2, \cdots, \xi_n$ 线性无关. 此时, 任一非零向量 z_0 均可用 $\xi_1, \xi_2, \cdots, \xi_n$ 线性表示为

$$z_0 = c_1 \xi_1 + c_2 \xi_2 + \cdots + c_n \xi_n, \quad (8.2.1)$$

并且 $c_1, c_2, \cdots, c_n$ 不全为零. 作向量序列 $z_k = A^k z_0$, 则

$$z_k = \lambda_1^k \left(c_1 \xi_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k \xi_2 + \cdots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^k \xi_n \right).$$

由此可见, 若 $\lambda_1 \neq 0$, 则由于当 $k \rightarrow \infty$ 时, $\left(\frac{\lambda_j}{\lambda_1} \right)^k \rightarrow 0$ ($j = 2, \cdots, n$), 所以当 k 充分大时, 必有 $z_k \approx \lambda_1^k c_1 \xi_1$, 即 z_k 可近似看成 λ_1 对应的特征向量, 并且 z_k 与 z_{k-1} 的非零分量之比趋近于 λ_1 .

然而, 由于当 $|\lambda_1| > 1$ 时, $\lambda_1^k \rightarrow \infty$; 当 $|\lambda_1| < 1$ 时, $\lambda_1^k \rightarrow 0$, 所以在计算机上计算 z_k 时会出现“上溢”或“下溢”. 为避免这种现象的出现, 通常在每步计算时将 z_k 单位化. 于是得到如下的求 A 的按模最大的特征值 λ_1 和对应的特征向量 ξ_1 的算法, 称为乘幂法.

算法 8.2.1 (乘幂法)

(1) 任取 z_0 , 如令 $z_0 := (1, 1, \cdots, 1)^T$, $\varepsilon_1 > 0, \varepsilon_2 > 0$;

(2) for $k = 1, 2, \cdots$, do

(i) 令 $y_k := Az_{k-1}$;

(ii) 求 y_k 绝对值最大的分量 $m_k := \max(y_k)$;

(iii) 令 $z_k := y_k / m_k$;

(iv) 若 $|m_k - m_{k-1}| \leq \varepsilon_1$ 或 $\|z_k - z_{k-1}\| \leq \varepsilon_2$, 则取 $\lambda_1 \approx m_k$, $\xi_1 \approx z_k$, 停止计算.

定理 8.2.1 乘幂法定义的序列 m_k 和向量序列 z_k 满足

$$\lim_{k \rightarrow \infty} m_k = \lambda_1, \quad \lim_{k \rightarrow \infty} z_k = \frac{\xi_1}{\max(\xi_1)}. \quad (8.2.2)$$

例 8.2.1 设 $A = \begin{pmatrix} -1 & 2 & 1 \\ 2 & -4 & 1 \\ 1 & 1 & 6 \end{pmatrix}$, 求 A 按模最大的特征值 λ_1 和特征向量 ξ_1 .

解 取 $z_0 = (1, 1, 1)^T$. 按乘幂法计算可得表 8.1.

表 8.1

k	m_k	$(z_k)_1$	$(z_k)_2$
1	-4	-0.5	0.25
2	-6.25	-0.32	0.16
8	-6.348 585 263	-0.107 044 958 5	-0.258 456 308 0
20	-6.418 337 792	-0.048 413 346 90	-0.370 584 043 6
40	-6.421 055 907	-0.046 154 377 98	-0.374 904 120 5
90	-6.421 066 614	-0.046 145 483 04	-0.374 921 131 3
91	-6.421 066 614	-0.046 145 483 03	-0.374 921 131 3

故得

$$\lambda_1 \approx -6.421\,066\,614, \quad \xi_1 = (-0.046\,145\,483\,03, -0.374\,921\,131\,3, 1)^T.$$

解毕.

2. 加速收敛技术

在前述的条件下, 乘幂法线性收敛. 证明如下: 设当 k 充分大时, $A^k z_0$ 绝对值最大的分量是其第 i 个分量, 则

$$\begin{aligned} m_k - \lambda_1 &= \max y_k - \lambda_1 = \frac{A^k z_0}{A^{k-1} z_0} - \lambda_1 \\ &= \frac{[c_1 \lambda_1^k \xi_1 + c_2 \lambda_2^k \xi_2 + \cdots + c_n \lambda_n^k \xi_n]_i}{[c_1 \lambda_1^{k-1} \xi_1 + c_2 \lambda_2^{k-1} \xi_2 + \cdots + c_n \lambda_n^{k-1} \xi_n]_i} - \lambda_1 \\ &= \frac{[c_2 \lambda_2^{k-1} (\lambda_2 - \lambda_1) \xi_2 + \cdots + c_n \lambda_n^{k-1} (\lambda_2 - \lambda_1) \xi_n]_i}{[c_1 \lambda_1^{k-1} \xi_1 + c_2 \lambda_2^{k-1} \xi_2 + \cdots + c_n \lambda_n^{k-1} \xi_n]_i} \\ &= \left(\frac{\lambda_2}{\lambda_1} \right)^{k-1} M_k, \quad M_k \rightarrow M \quad (M \text{ 为常数}), \end{aligned}$$

所以当 $k \rightarrow \infty$ 时,

$$\frac{|m_{k+1} - \lambda_1|}{|m_k - \lambda_1|} = \frac{|M_{k+1}(\lambda_2/\lambda_1)^k|}{|M_k(\lambda_2/\lambda_1)^{k-1}|} \rightarrow \left| \frac{\lambda_2}{\lambda_1} \right|.$$

这就证明了乘幂法线性收敛. 渐近常数 $\frac{\lambda_2}{\lambda_1}$ 称为乘幂法的收敛率, 其值越小, 收敛越快.

乘幂法线性收敛, 可用艾特肯方法进行加速收敛.

$$m_k^* = m_k - \frac{(m_k - m_{k-1})^2}{m_k - 2m_{k-1} + m_{k-2}}. \quad (8.2.3)$$

由例 8.2.1 所得的数据, 按式 (8.2.3) 可得表 8.2, 可见加速效果不错.

表 8.2

k	m_k^*	$(z_k^*)_1$	$(z_k^*)_2$
10	-6.423 206 251	-0.044 799 036 75	-0.377 496 088 5
20	-6.421 074 976	-0.046 140 213 23	-0.374 931 210 7
40	-6.421 066 612	-0.046 145 482 73	-0.374 921 132 1

既然乘幂法的收敛快慢取决于比值 $|\lambda_2/\lambda_1|$, 可选取常数 p , 用 $A - pI$ 代替 A 作乘幂法. 适当选取 p , 可使 $A - pI$ 的特征值 $\lambda_i - p$ 满足

$$|\lambda_1 - p| > |\lambda_2 - p| \geq |\lambda_i - p|, \quad i = 3, \dots, n$$

且

$$\left| \frac{\lambda_2 - p}{\lambda_1 - p} \right| < \left| \frac{\lambda_2}{\lambda_1} \right|.$$

这时的乘幂法收敛速度快, $m_k + p \rightarrow \lambda_1$, 而 z_k 仍收敛于 A 的特征向量 ξ_1 . 这种加速收敛的方法称为原点位移法.

例 8.2.2 取 $p = -2$, $z_0 = (1, 1, 1)^T$. 按原点位移法计算例 8.2.1 可得表 8.3.

表 8.3

k	$m_k - 2$	$(z_k)_1$	$(z_k)_2$	$(z_k)_3$
1	2	1	0.25	-0.5
2	1.25	-0.307 692 307 7	0.307 692 307 7	1
3	-5.384 615 385	-0.568 181 818 2	-0.295 454 545 5	1
10	-6.421 653 861	-0.050 103 023 5	-0.358 219 761 8	1
30	-6.421 064 781	-0.046 146 794 14	-0.374 918 606 5	1
58	-6.421 066 614	-0.046 145 483 03	-0.374 921 131 3	1
59	-6.421 066 614	-0.046 145 483 03	-0.374 921 131 3	1

原点位移法 p 的选取比较困难, 常用于反幂法.

3. 反幂法

设 A 可逆, 则对 A^{-1} 作的乘幂法称为反幂法. 由于 A^{-1} 的特征值是 A 的特征值的倒数, A^{-1} 的特征向量是 A 对应的特征向量, 所以当 A 的特征值满足 $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_{n-1}| > |\lambda_n|$ 时, 按反幂法必有

$$m_k \rightarrow \frac{1}{\lambda_n}, \quad z_k \rightarrow \frac{\xi_n}{\max(\xi_n)},$$

并且收敛率为 $|\lambda_n/\lambda_{n-1}|$.

反幂法主要用于已知矩阵近似特征值时, 求矩阵的特征向量并提高特征值的精度. 这是因为已知 A 的特征值 λ_m 的近似值 $\tilde{\lambda}_m$ 时, 一般有

$$|\lambda_i - \tilde{\lambda}_m| \gg |\lambda_m - \tilde{\lambda}_m| \approx 0, \quad i \neq m,$$

故对 $A - \tilde{\lambda}_m I$ 应用反幂法时,

$$m_k \rightarrow \frac{1}{\lambda_m - \tilde{\lambda}_m}, \quad z_k \rightarrow \xi_m^0 = \frac{\xi_m}{\max(\xi_m)}.$$

而收敛率为 $\frac{|\lambda_m - \tilde{\lambda}_m|}{|\lambda_j - \tilde{\lambda}_m|} \approx 0$, 所以迭代收敛很快, 往往只迭代两三步就可达到很高的精度. 因此, 反幂法是目前求特征向量最有效的一种方法.

已知 A 的特征值近似值 $\tilde{\lambda}_m$ 时, 求矩阵的特征向量并提高特征值 λ_m 的精度, 反幂法步骤如下:

算法 8.2.2 (反幂法)

(1) 三角分解 $A - \tilde{\lambda}_m I = LR$;

(2) for $k = 1, 2, \dots$, do

(i) 当 $k = 1$ 时, 令 $u := (1, 1, \dots, 1)^T$;

当 $k \neq 1$ 时, 解 $Lu = z_{k-1}$;

(ii) 解 $Ry_k = u$;

(iii) 求绝对值最大的分量 $m_k := \max(y_k)$;

(iv) 令 $z_k := y_k / m_k$;

(v) 若 $|m_k - m_{k-1}| \leq \varepsilon_1$ 或 $\|z_k - z_{k-1}\| \leq \varepsilon_2$, 则停止计算; 取 $\xi_m^0 \approx z_k$, $\lambda_m \approx \tilde{\lambda}_m + 1/m_k$.

例 8.2.3 用反幂法求例 8.2.1 中矩阵近似于 -6.42 的特征值和特征向量, 可得表 8.4.

表 8.4

k	m_k	$(z_k)_1$	$(z_k)_2$	$(z_k)_3$
1	-820.410 368 7	-0.046 028 298 46	-0.375 872 762 3	1
2	-937.834 480 8	-0.046 145 715 84	-0.374 920 570 4	1
3	-937.545 857 7	-0.046 145 482 84	-0.374 921 131 7	1

故得

$$\lambda_1 \approx -6.421\,066\,614, \quad \xi_1 = (-0.046\,145\,482\,84, -0.374\,921\,131\,7, 1)^T.$$

解毕.

8.2.2 求矩阵全部特征值与特征向量的 QR 方法

1. 基本 QR 方法

QR 方法是求一般实矩阵 A 的全部特征值非常有效的方法之一, 其基本步骤如下: 令 $A_0 = A$, 对 $k = 1, 2, \dots$, 分解 $A_{k-1} = Q_{k-1}R_{k-1}$, 令 $A_k = R_{k-1}Q_{k-1}$, 其

中 Q_{k-1} 为正交矩阵, R_{k-1} 为上三角矩阵. 这样所得的矩阵序列 $\{A_k\}$ 均相似于 A , 从而与 A 有相同的特征值.

性质 8.2.1 对一个给定矩阵 $A \in \mathbb{R}^{n \times n}$, 存在一个正交矩阵 $Q \in \mathbb{R}^{n \times n}$, 使得

$$Q^T A Q = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ & R_{22} & \cdots & R_{2m} \\ & & \ddots & \vdots \\ & & & R_{mm} \end{pmatrix}, \quad (8.2.4)$$

其中对角子块 R_{ii} 为 1×1 或 2×2 矩阵.

这种矩阵称为 Schur 矩阵. 当 R_{ii} 为 1×1 矩阵时, R_{ii} 就是 A 的特征值; 当 R_{ii} 为 2×2 矩阵时, 其特征值是一对共轭复数, 也是 A 的特征值. 所谓基本收敛是指对角子块下方元素趋于零, 上方元素不一定收敛. 这种收敛足以近似求出特征值, 对应的特征向量可用反幂法求得.

2. 一般矩阵的简化

QR 方法每一步分解 $A_{k-1} = Q_{k-1} R_{k-1}$, 然后计算 $R_{k-1} Q_{k-1}$, 运算量非常大. 为节省运算量, 通常先用豪斯霍尔德矩阵 $H = I - 2uu^T$ 将 A 化为拟上三角矩阵 (又称海森伯格矩阵), 即次对角线下方元素全为零的特殊矩阵, 其中 $\|u\|_2^2 = u^T u = 1$, H 为对称、正交矩阵. 因此, 线性变换 $y = Hx$ 是保持长度不变的, 称此线性变换为镜射变换.

已知向量 $a = (a_1, a_2, \dots, a_n)^T$, 欲使镜射变换 H 将 a 变为 $b = (a_1, \dots, a_r, c, 0, \dots, 0)^T$, 只需令

$$u = \frac{a - b}{\|a - b\|}, \quad H = I - 2uu^T.$$

此处, c 应满足

$$a_1^2 + \cdots + a_n^2 = \|a\|_2^2 = \|b\|_2^2 = a_1^2 + \cdots + a_r^2 + c^2,$$

故

$$c = \pm(a_{r+1}^2 + a_{r+2}^2 + \cdots + a_n^2)^{1/2}.$$

为避免计算 $a_{r+1} - c$ 时损失准确数字, c 的符号宜取为 $-\text{sgn}(a_{r+1})$, 此时,

$$c = -\text{sgn}(a_{r+1})(a_{r+1}^2 + a_{r+2}^2 + \cdots + a_n^2)^{1/2}, \quad \rho = \sqrt{2c(c - a_{r+1})}, \quad (8.2.5)$$

$$u = \begin{bmatrix} 0_r \\ u_{n-r} \end{bmatrix}, \quad u_{n-r} = \frac{1}{\rho}(a_{r+1} - c, a_{r+2}, \dots, a_n)^T, \quad (8.2.6)$$

$$H = I - 2uu^T = \begin{bmatrix} I_r & \\ & H_{n-r} \end{bmatrix}, \quad H_{n-r} = I_{n-r} - 2u_{n-r}u_{n-r}^T, \quad (8.2.7)$$

对任意向量 $x = (x_1, x_2, \dots, x_n)^T = \begin{bmatrix} x_r \\ x_{n-r}^* \end{bmatrix}$, 按式 (8.2.7) 确定的镜射变换将 x 变为

$$y = Hx = \begin{bmatrix} I_r & \\ & H_{n-r} \end{bmatrix} \begin{bmatrix} x_r \\ x_{n-r}^* \end{bmatrix} = \begin{bmatrix} x_r \\ y_{n-r}^* \end{bmatrix}, \quad (8.2.8)$$

其中

$$y_{n-r}^* = x_{n-r}^* - 2u_{n-r}(u_{n-r}^T x_{n-r}^*). \quad (8.2.9)$$

利用上述镜射矩阵可将任意矩阵 A 化为相似的拟上三角矩阵 (即下次对角下方全是零的矩阵). 事实上, 可作镜射矩阵 P_1 , 使 $P_1 A$ 各列的第一个元素不变, 但使第一列第二个元素后面各元素变为零. 再令 $A_1 = P_1 A P_1^T$, 则因为 P_1 为正交矩阵, 于是有

$$A_1 = P_1 A P_1^{-1}, \quad A_1^T = P_1 (P_1 A)^T.$$

再作镜射矩阵 P_2 , 使 $P_2 A_1$ 各列第一、二个元素不变, 但使第二列第三个元素后面各元素变为零. $P_1 A$, $P_2 A_1$ 有如下形式:

$$P_1 A = \begin{pmatrix} * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ & * & * & \cdots & * \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ & * & * & \cdots & * \end{pmatrix}, \quad P_2 A_1 = \begin{pmatrix} * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ & * & * & \cdots & * \\ & & * & \cdots & * \\ & \vdots & \ddots & \ddots & \vdots \\ & & * & \cdots & * \end{pmatrix}, \quad (8.2.10)$$

作矩阵 $A_2 = P_1 A P_2^T$, 则有

$$A_2 = P_2 A P_2^{-1}, \quad A_2^T = P_2 (P_2 A)^T.$$

如此继续下去, 最多左、右乘 $n-2$ 次, 就可将 A 化为相似的拟上三角矩阵 A_{n-2} .

例 8.2.4 用镜射变换将例 8.2.1 中的矩阵 A 化为相似的拟上三角矩阵.

解 为用镜射变换将 A 的第一列 $(-1, 2, 1)^T$ 化为 $(-1, c, 0)^T$. 由 (8.2.5)~(8.2.7), 令

$$c = -\operatorname{sgn}(2)(2^2 + 1^2)^{1/2} = -\sqrt{5}, \quad \rho = \sqrt{2(-\sqrt{5})(-\sqrt{5} - 2)} = \sqrt{10 + 4\sqrt{5}}.$$

$$u = (0, 2 + \sqrt{5}, 1)^T / \rho,$$

$$H = I - 2uu^T = \begin{pmatrix} 1 & & \\ & -2/\sqrt{5} & -1/\sqrt{5} \\ & -1/\sqrt{5} & -2/\sqrt{5} \end{pmatrix},$$

于是

$$HAH^T = \begin{pmatrix} -1 & -\sqrt{5} & 0 \\ -\sqrt{5} & -3.6 & 0.2 \\ 0 & 0.2 & -6.4 \end{pmatrix}.$$

解毕.

例 8.2.4 中 A 为对称矩阵. 一般对称矩阵的正交相似矩阵保持对称性不变, 而对称拟上三角矩阵只能是三对角矩阵, 故对称矩阵用镜射变换化成的相似拟上三角矩阵必为三对角矩阵.

3. 拟上三角矩阵的 QR 算法

一般矩阵化成相似拟上三角矩阵后, 应用 QR 分解产生的矩阵序列 $\{A_k\}$ 都是拟上三角矩阵. 分解 $A_{k-1} = Q_{k-1}R_{k-1}$ 常用旋转矩阵来完成. 选取旋转矩阵 $P_i = \text{Rotate}(i+1, i, \theta_i)$, 则最多经过 $n-1$ 步旋转变换, A_{k-1} 变为上三角矩阵 R_{k-1} , 即

$$P_{n-1} \cdots P_2 P_1 A_{k-1} = R_{k-1}.$$

在上述变换过程中, 用旋转矩阵 P_i 左乘 $A^{(i-1)}$ 变为 $A^{(i)}$, 只改变矩阵第 $i, i+1$ 行, 即

$$\begin{cases} a_{ij}^{(i)} = a_{ij}^{(i-1)} \cos \theta_i + a_{i+1,j}^{(i-1)} \sin \theta_i, \\ a_{i+1,j}^{(i)} = -a_{ij}^{(i-1)} \sin \theta_i + a_{i+1,j}^{(i-1)} \cos \theta_i, \end{cases} \quad j = i, \cdots, n.$$

为使 $a_{i+1,i}^{(i)} = 0$, 应选取

$$\theta_i = \arctan(a_{i+1,i}^{(i-1)} / a_{ii}^{(i-1)}). \quad (8.2.11)$$

分解过程算法描述如下:

算法 8.2.3

for $i = 1, 2, \cdots$, do

- 令 $\theta := \arctan(a_{i+1,i}/a_{ii})$, $c_i := \cos \theta$, $s_i := \sin \theta$
- for $j = i, \cdots, n$, do
 - ◇ $z := a_{ij}c_i + a_{i+1,j}s_i$
 - ◇ $a_{i+1,j} := -a_{ij}s_i + a_{i+1,j}c_i$
 - ◇ $a_{ij} := z$

• end do

end do

为求 $A_k = R_{k-1}Q_{k-1}$, 由于

$$A_k = R_{k-1}(P_{n-1} \cdots P_2 P_1)^{-1} = R_{k-1}P_1^T P_2^T \cdots P_{n-1}^T,$$

因此, A_k 的计算过程可写成如下:

算法 8.2.4

for $i = 1, 2, \dots$, do

• for $j = i, i+1$, do

$$\diamond z := a_{ji}c_i + a_{j,i+1}s_i$$

$$\diamond a_{j,i+1} := -a_{ji}s_i + a_{j,i+1}c_i$$

$$\diamond a_{ji} := z$$

• end do

end do

反复进行上述过程, 直到 A_k 变换为近似的 Schur 矩阵, 便可得出 A 的近似特征值.

例 8.2.5 用 QR 方法求例 8.2.4 化得的拟上三角矩阵的特征值.

解 由 (8.2.11), 令

$$\theta_1 = \arctan \frac{-\sqrt{5}}{-1} = 1.150\,261\,992, \quad c_1 = \cos \theta_1, \quad s_1 = \sin \theta_1,$$

则

$$\begin{aligned} A^{(1)} &= \begin{pmatrix} c_1 & s_1 & & \\ -s_1 & c_1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \begin{pmatrix} -1 & -\sqrt{5} & 0 \\ -\sqrt{5} & -3.6 & 0.2 \\ 0 & 0.2 & -6.4 \end{pmatrix} \\ &= \begin{pmatrix} 2.449\,489\,743\,192 & -4.199\,206\,274\,149 & 0.182\,574\,186 & \\ 0 & 0.571\,547\,607 & 0.081\,649\,658 & \\ 0 & 0.2 & -6.4 & \end{pmatrix}. \end{aligned}$$

令

$$\theta_2 = \arctan \frac{0.2}{0.571\,547\,607} = 0.336\,609\,879, \quad c_2 = \cos \theta_2, \quad s_2 = \sin \theta_2,$$

于是

$$\begin{aligned}
 A_1 &= R_0 \begin{pmatrix} c_1 & -s_1 & \\ s_1 & c_1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} 1 & & \\ & c_2 & -s_2 \\ & s_2 & c_2 \end{pmatrix} \\
 &= \begin{pmatrix} -4.838\,383\,838 & 0.552\,770\,798 & 0 \\ 0.552\,770\,798 & -0.439\,393\,939 & -2.004\,127\,972 \\ 0 & -2.004\,127\,972 & -5.727\,272\,727 \end{pmatrix}.
 \end{aligned}$$

同理,

$$\begin{aligned}
 A_2 &= \begin{pmatrix} -4.901\,408\,451 & 0.231\,646\,256 & 0 \\ 0.231\,646\,256 & -6.264\,950\,145 & 0.892\,076\,154 \\ 0 & 0.892\,076\,154 & 0.166\,358\,595 \end{pmatrix}, \\
 A_6 &= \begin{pmatrix} -5.136\,079\,688 & 0.588\,098\,275 & 0 \\ 0.588\,098\,275 & -6.151\,912\,454 & 0.000\,003\,999 \\ 0 & 0.000\,003\,999 & 0.287\,992\,139 \end{pmatrix}, \\
 A_{30} &= \begin{pmatrix} -6.421\,054\,217 & 0.004\,390\,120 & 0 \\ 0.004\,390\,120 & -4.866\,937\,928 & 0 \\ 0 & 0 & 0.287\,992\,137 \end{pmatrix}, \\
 A_{88} &= \begin{pmatrix} -6.421\,066\,621 & & \\ & -4.866\,925\,530 & \\ & & 0.287\,992\,137 \end{pmatrix},
 \end{aligned}$$

故

$$\lambda_1 = -6.421\,066\,621, \quad \lambda_2 = -4.866\,925\,530, \quad \lambda_3 = 0.287\,992\,137.$$

解毕.

4. 带位移的 QR 方法

为了加速 QR 算法的收敛, 仿照乘幂法, 可将 QR 算法改写如下:

算法 8.2.5 (带位移的 QR 方法)

(1) 令 $A_0 = A$ (A 为拟上三角矩阵);

(2) for $i = 1, 2, \dots, n$ do

 分解 $A_{k-1} - \mu_{k-1}I = Q_{k-1}R_{k-1}$,

 令 $A_k = R_{k-1}Q_{k-1} + \mu_{k-1}I$.

end do

这里, μ_{k-1} 称为位移量, 最好取为 A 的某一特征值. 实际计算时, 通常取为 A_{k-1} 的右下角元素 $a_{nn}^{(k-1)}$ 或右下角 2×2 矩阵特征值中最接近 $a_{nn}^{(k-1)}$ 者.

例 8.2.6 用带位移的 QR 方法求例 8.2.4 化得的拟上三角矩阵的特征值.

解 取位移量 $\mu_0 = -6.4$, 分解 $A_{k-1} - \mu_{k-1}I$ 得

$$\begin{pmatrix} 5.4 & -\sqrt{5} & 0 \\ -\sqrt{5} & 2.8 & 0.2 \\ 0 & 0.2 & 0 \end{pmatrix} = (P_2 P_1)^{-1} R_0,$$

其中

$$R_0 = \begin{pmatrix} 5.844\,655\,679 & -3.137\,183\,510 & -0.076\,516\,671 \\ & 1.743\,000\,878\,5 & 0.183\,563\,711 \\ & & -0.221\,202\,899 \end{pmatrix},$$

$$P_1 = \begin{pmatrix} \cos \theta_1 & \sin \theta_1 & \\ -\sin \theta_1 & \cos \theta_1 & \\ & & 1 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 1 & & \\ & \cos \theta_2 & \sin \theta_2 \\ & -\sin \theta_2 & \cos \theta_2 \end{pmatrix},$$

$$\theta_1 = -0.392\,590\,761, \quad \theta_2 = 0.114\,997\,409,$$

于是得

$$A_1 = R_0(P_1^T P_2^T) + 6.4I = \begin{pmatrix} 0.200\,234\,192 & -0.666\,846\,149 & 0 \\ -0.666\,846\,149 & -4.799\,171\,336 & -0.002\,432\,908 \\ 0 & -0.002\,432\,908 & -6.421\,062\,856 \end{pmatrix}.$$

同理得

$$A_2 = \begin{pmatrix} 0.283\,205\,888 & -0.157\,002\,612 & 0 \\ -0.157\,002\,612 & -4.862\,139\,274 & -0.000\,000\,006 \\ 0 & -0.000\,000\,006 & -6.421\,066\,615 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 0.287\,735\,078 & -0.036\,401\,350 & 0 \\ -0.036\,401\,350 & -4.866\,668\,465 & 0 \\ 0 & 0 & -6.421\,066\,615 \end{pmatrix},$$

故 $\lambda_3 = -6.421\,066\,615$. 求左上角 2×2 矩阵特征值得 $\lambda_1 = 0.287\,992\,138$, $\lambda_2 = -4.866\,925\,525$. 解毕.

8.2.3 阿诺尔迪方法

在实际应用中, 大规模稀疏矩阵是非常广泛的. 对稀疏矩阵, 经典的 QR 方法在进行一次 QR 迭代后, 所得到的矩阵通常不是稀疏的, 产生了计算机无法承受的

大量非零填充. 因此, 经典 QR 方法对稀疏矩阵是不适用的. 本小节求解稀疏矩阵特征值问题的阿诺尔迪方法.

阿诺尔迪方法的基本思想是保留和应用前面步骤的信息. 在进行第 k 步时, 应用保留的 $k+1$ 个向量 $z_0, Az_0, A^2z_0, \dots, A^kz_0$ 的信息. 在这些向量张成的 $k+1$ 维空间, 求特征向量的最佳逼近.

算法 8.2.6 (阿诺尔迪方法)

(1) 任取 z_0 , 将 z_0 单位化 $z_1 := z_0 / \|z_0\|_2$;

(2) for $k = 1, 2, \dots$, do

(i) 对 z_k 作格拉姆-施密特正交化 $y_{k+1} := Az_k - \sum_{j=1}^k (Az_k, z_j) z_j$;

(ii) 将 y_{k+1} 单位化 $z_{k+1} := y_{k+1} / \|y_{k+1}\|$;

(iii) 若 $(Az_{k+1}, z_k) = 0$, 则停止计算;

end do

对任意 j , 空间 $\text{span}\{z_0, Az_0, \dots, A^{j-1}z_0\}$ 称为 A 和 z_0 的第 j 个克雷洛夫子空间, 记为 $K_j(A, z_0)$.

定理 8.2.2 设 $z_0, Az_0, \dots, A^{j-1}z_0$ 线性无关, $z_1, z_2, \dots, z_m$ 是由阿诺尔迪算法计算出的正交向量, 则

(1) 对 $k = 1, 2, \dots, m$, $\text{span}\{z_1, z_2, \dots, z_k\} = K_k(A, z_0)$;

(2) 对 $k = 1, 2, \dots, m$, $(Az_{k+1}, z_k) > 0$;

(3) $(Az_{m+1}, z_m) = 0$ 的充要条件是 $z_0, Az_0, \dots, A^mz_0$ 线性相关, 克雷洛夫子空间 $K_m(A, z_0)$ 是 A 的不变子空间.

设 $z_1, z_2, \dots, z_m$ 是由阿诺尔迪算法计算出的正交向量, 记 $Q_m = [z_1, z_2, \dots, z_m]$, 则有

$$AQ_m = Q_{m+1}H_{m+1,m}, \quad (8.2.12)$$

其中

$$H_{m+1,m} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1,m-1} & h_{1m} \\ h_{21} & h_{22} & \cdots & h_{2,m-1} & h_{2m} \\ & h_{32} & \cdots & h_{3,m-1} & h_{3m} \\ & & \ddots & \vdots & \vdots \\ & & & h_{m,m-1} & h_{mm} \\ & & & & h_{m+1,m} \end{pmatrix}, \quad h_{ij} = (Az_i, z_j)$$

为上海森伯格矩阵.

定理 8.2.3 设 $Q_m, H_{m,m}$ 是由阿诺尔迪算法计算出的矩阵, λ 为 $H_{m,m}$ 的一个特征值, 所对应的特征向量为 z . 记 $\xi = Q_m z$, 则

$$\|Az - \lambda\xi\|_2 = |h_{m+1,m}| |z_m|, \quad (8.2.13)$$

其中 z_m 为 z 的最后一个分量.

定理 8.2.3 中的向量 ξ 称为 A 与克雷洛夫子空间 $K_m(A, z_0)$ 对应的 Ritz 向量, λ 称为 ξ 的 Ritz 值, (λ, ξ) 称为 Ritz 对.

如果 (λ, ξ) 是 A 的特征值和特征向量, 则 $Az - \lambda\xi$ 的残差为零. 当 $\|Az - \lambda\xi\|_2$ 很小时, (λ, ξ) 接近于 A 的特征值和特征向量.

数值试验表明, 在 m 充分大使得 $h_{m+1,m}$ 足够小之前, 某些 Ritz 值就是 A 的特征值的很好的逼近. Ritz 值一般首先逼近最大或最小特征值, 因此, 阿诺尔迪算法一般不适合计算谱内部的特征值. 当然, 将阿诺尔迪算法和前面介绍的原点位移法结合就可计算谱内部的特征值. 如要计算 A 的近似于 μ 的内部特征值时, 可对 $(A - \mu I)^{-1}$ 应用阿诺尔迪算法, 它的边缘特征值 $\tilde{\lambda}_1 = (\lambda_i - \mu)^{-1}$ 对应于 A 的近似于 μ 的内部特征值 $\lambda_i = \mu + 1/\tilde{\lambda}_1$.

8.3 求实对称矩阵特征值的计算方法

实际问题中出现的矩阵往往是对称的, 因而实对称矩阵在理论研究和工程应用中占用非常重要的地位. 本节介绍实对称矩阵特征值问题的计算方法.

8.3.1 雅可比方法

雅可比方法用于求实对称矩阵 A 的全部特征值和特征向量, 其基本思想是通过一系列正交相似变换, 把 A 化为对角矩阵 $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, 即令 $A_0 = A$, 对 $k = 1, 2, \dots$, 令 $A_k = R_k A_{k-1} R_k^T$, 其中 R_k 为正交矩阵, 使 A_k 逐渐变为对角矩阵 Λ . 由于 R_k 为正交矩阵, $P_k = R_1^T R_2^T \cdots R_k^T$ 也是正交矩阵, 并且 $A_k = P_k^T A P_k$, 所以一旦 A_m 化为对角矩阵 Λ , 必有

$$P_m^T A P_m = \Lambda, \quad A P_m = P_m \Lambda.$$

记 P_m 的列向量为 $\xi_1, \xi_2, \dots, \xi_n$, 则

$$A\xi_i = \lambda_i \xi_i, \quad i = 1, 2, \dots, n.$$

也就是说, A 一旦化为对角矩阵 Λ , 则 Λ 的对角元就是 A 的特征值, 而变换矩阵的乘积 P_m 的列向量就是 A 的特征向量.

雅可比方法的正交矩阵 R_k 取为如下旋转阵:

$$R(p, q, \theta) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \cos \theta & & \sin \theta & \\ & & & & \ddots & & \\ & & & & & 1 & \\ & & & -\sin \theta & & \cos \theta & \\ & & & & & & 1 & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} \quad \begin{matrix} (p) \\ (q) \end{matrix},$$

对于旋转阵 R , 正交相似变换 $A_k = RA_{k-1}R^T$ 保持了 A_{k-1} 的对称性及范数不变, 而且只有 A_{k-1} 的第 p, q 行和 p, q 列发生变化, 其余元素不变. 变化的元素计算公式如下:

$$\begin{cases} a_{pj}^{(k)} = a_{pj}^{(k-1)} \cos \theta + a_{qj}^{(k-1)} \sin \theta = a_{jp}^{(k)}, & j \neq p, q, \\ a_{qj}^{(k)} = -a_{pj}^{(k-1)} \sin \theta + a_{qj}^{(k-1)} \cos \theta = a_{jq}^{(k)}, & j \neq p, q, \\ a_{pp}^{(k)} = a_{pp}^{(k-1)} \cos^2 \theta + 2a_{pq}^{(k-1)} \cos \theta \sin \theta + a_{qq}^{(k-1)} \sin^2 \theta, \\ a_{qq}^{(k)} = a_{pp}^{(k-1)} \sin^2 \theta - 2a_{pq}^{(k-1)} \cos \theta \sin \theta + a_{qq}^{(k-1)} \cos^2 \theta, \\ a_{pq}^{(k)} = (a_{qq}^{(k-1)} - a_{pp}^{(k-1)}) \cos \theta \sin \theta + a_{pq}^{(k-1)} (\cos^2 \theta - \sin^2 \theta) = a_{qp}^{(k)}. \end{cases} \quad (8.3.1)$$

为保证 A_k 逐步化为对角矩阵, 雅可比方法选取 p, q, θ , 使得 A_{k-1} 绝对值最大的非对角元 $a_{pq}^{(k-1)}$ 化为零, 即选取 θ 使 $a_{pq}^{(k)} = 0$. 此时, θ 满足

$$\tan 2\theta = \frac{2a_{pq}^{(k-1)}}{a_{pp}^{(k-1)} - a_{qq}^{(k-1)}}, \quad (8.3.2)$$

其中 θ 可限制于 $|\theta| \leq \frac{\pi}{4}$. 若 $a_{pp}^{(k-1)} = a_{qq}^{(k-1)}$, 则当 $a_{pq}^{(k-1)} > 0$ 时, 取 $\theta = \frac{\pi}{4}$; 当 $a_{pq}^{(k-1)} < 0$ 时, 取 $\theta = -\frac{\pi}{4}$.

例 8.3.1 用雅可比法求例 8.2.1 中矩阵特征值和特征向量.

解 令 $A_0 = A$, $P_0 = I$. A_0 中绝对值最大的非对角元为 $a_{12}^{(0)} = 2$, 故取

$$\tan 2\theta = \frac{2a_{12}^{(0)}}{a_{11}^{(0)} - a_{22}^{(0)}} = \frac{4}{3},$$

$$c = \cos \theta = 0.894\,427\,191, \quad s = \sin \theta = 0.447\,213\,595\,5,$$

则

$$\begin{aligned} A_1 &= \begin{pmatrix} c & s & \\ -s & c & \\ & & 1 \end{pmatrix} A_0 \begin{pmatrix} c & -s & \\ s & c & \\ & & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & & 1.341\,640\,787 \\ & -5 & 0.447\,213\,599\,5 \\ 1.341\,640\,787 & 0.447\,213\,599\,5 & -6 \end{pmatrix}, \\ P_1 &= P_0 \begin{pmatrix} c & -s & \\ s & c & \\ & & 1 \end{pmatrix} = \begin{pmatrix} 0.894\,427\,191\,0 & -0.447\,213\,595\,5 & \\ 0.447\,213\,595\,5 & 0.894\,427\,191\,0 & \\ & & 1 \end{pmatrix}. \end{aligned}$$

A_1 中绝对值最大的非对角元为 $a_{13}^{(1)}$, 故取

$$\tan 2\theta = \frac{2a_{13}^{(1)}}{a_{11}^{(1)} - a_{33}^{(1)}} = 0.447\,213\,589,$$

$$c = \cos \theta = 0.977\,975\,186\,1, \quad s = \sin \theta = 0.208\,721\,190\,6,$$

则

$$\begin{aligned} A_2 &= \begin{pmatrix} c & s & \\ & 1 & \\ -s & & c \end{pmatrix} A_1 \begin{pmatrix} c & -s & \\ & 1 & \\ s & & c \end{pmatrix} \\ &= \begin{pmatrix} 0.286\,335\,345\,0 & 0.093\,342\,954\,12 & 0 \\ 0.093\,342\,954\,12 & -5 & -0.437\,363\,799\,3 \\ 0 & -0.437\,363\,799\,3 & -6.286\,355\,345 \end{pmatrix}, \\ P_2 &= P_1 \begin{pmatrix} c & -s & \\ & 1 & \\ s & & c \end{pmatrix} = \begin{pmatrix} 0.874\,727\,598\,6 & -0.447\,213\,595 & -0.186\,685\,908\,2 \\ 0.437\,363\,799\,3 & 0.894\,427\,191\,0 & -0.093\,342\,954\,1 \\ 0.208\,721\,190\,6 & 0 & 0.977\,975\,186\,1 \end{pmatrix}. \end{aligned}$$

如此继续下去, 将计算中绝对值小于 5×10^{-10} 的元素认为是零, 则最后得

$$A_7 = \begin{pmatrix} 0.287\,992\,139\,0 & & \\ & -4.866\,925\,525 & \\ & & -6.421\,066\,615 \end{pmatrix},$$

$$P_7 = \begin{pmatrix} 0.866\,432\,153\,3 & -0.497\,425\,200\,5 & -0.043\,168\,204\,2 \\ 0.453\,057\,727\,1 & 0.819\,589\,012\,2 & -0.350\,731\,446\,1 \\ 0.209\,842\,845\,9 & 0.284\,327\,313\,5 & 0.935\,480\,603\,6 \end{pmatrix},$$

故得 A 的特征值

$$\lambda_1 = 0.287\,992\,139, \quad \lambda_2 = -4.866\,925\,525, \quad \lambda_3 = -6.421\,066\,615,$$

对应的特征向量为

$$\xi_1 = (0.866\,432\,153\,3, 0.453\,057\,727\,1, 0.209\,842\,845\,9)^T,$$

$$\xi_2 = (-0.497\,425\,200\,5, 0.819\,589\,012\,2, 0.284\,327\,313\,5)^T,$$

$$\xi_3 = (-0.043\,168\,204\,2, -0.350\,731\,446\,1, 0.935\,480\,603\,6)^T.$$

解毕.

注意: 在雅可比法计算过程中, 为节省计算时间, 可采用以下措施:

(1) 因为 A_k 对称, 所有运算只需在上三角部分进行;

(2) 因为 A_k, P_k 与 A_{k-1}, P_{k-1} 相比, 只有部分元素变化, 故只需对它们进行变化, P_k 仅 p, q 列变化,

$$\begin{cases} p_{ip}^{(k)} = p_{ip}^{(k-1)} \cos \theta + p_{iq}^{(k-1)} \sin \theta, \\ p_{iq}^{(k)} = -p_{ip}^{(k-1)} \sin \theta + p_{iq}^{(k-1)} \cos \theta, \end{cases} \quad i = 1, 2, \dots, n; \quad (8.3.3)$$

(3) 因为 $A_k \rightarrow \Lambda$ 是一个极限过程, 一般不可能经过有限步便变成对角矩阵. 因此, 实际计算时, 应在非对角线元充分小或对角线元变化不大时停止计算;

(4) 因为选取绝对值最大的非对角元 $a_{pq}^{(k-1)}$ 总要耗费时间, 所以为节省计算时间, 也可按行的顺序把非对角元依次化为零, 这时方法称为循环雅可比法; 也可规定阈值 α_1 , 按序比较非对角元, 将超过阈值者化为零. 当所有非对角元均不超过 α_1 时, 取 $\alpha_2 < \alpha_1$ 作新阈值, 重复上述过程. 这种方法称为限值雅可比方法或雅可比过关法. 通常取 $\alpha_i = s_0/n^{i+1}$, 其中 s_0 为 A 的非对角元平方和.

雅可比方法稳定、精度高、求得的特征向量正交性好. 当然, 对于稀疏矩阵不能保持矩阵的稀疏性.

8.3.2 吉文斯方法

本小节介绍实对称三对角矩阵特征值的计算. 考虑矩阵

$$A = \begin{pmatrix} d_1 & b_1 & & & \\ b_1 & d_2 & b_2 & & \\ & b_2 & d_3 & \ddots & \\ & & \ddots & \ddots & b_{n-2} \\ & & & b_{n-2} & d_{n-1} & b_{n-1} \\ & & & & b_{n-1} & d_n \end{pmatrix}. \quad (8.3.4)$$

不失一般性, 假定 $b_i \neq 0 (\forall i)$.

令 A_i 为矩阵 A 的 i 阶首主子矩阵, 定义如下的特征多项式 $p_i(x) = \det(A_i - xI_i)$ 序列:

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= d_1 - x, \\ p_i(x) &= (d_i - x)p_{i-1}(x) - b_{i-1}^2 p_{i-2}(x), \quad i = 2, \dots, n. \end{aligned} \quad (8.3.5)$$

可以证明, $p_n(x)$ 是矩阵 A 的特征多项式. 多项式序列 (8.3.5) 称为施图姆 (Strum) 序列, 它满足如下的性质:

性质 8.3.1 A_{i-1} 的特征值严格分离 A_i 的特征值, 即

$$\lambda_i(A_i) < \lambda_i(A_{i-1}) < \lambda_{i-1}(A_i) < \dots < \lambda_2(A_i) < \lambda_1(A_{i-1}) < \lambda_1(A_i),$$

而且对任意实数 μ , $S_\mu = \{p_0(\mu), p_1(\mu), \dots, p_n(\mu)\}$ 的符号变化数 $s(\mu)$ 恰恰是矩阵 A 小于 μ 的特征值的个数.

下面给出求解实对称三对角矩阵特征值的吉文斯算法.

算法 8.3.1 (吉文斯算法)

for $i = 1, 2, \dots, n$ do

(1) 令 $b_0 = b_n = 0$, 定义包含 A 的特征值的区间 $[a^{(0)}, b^{(0)}]$, 其中

$$a^{(0)} := \min_{1 \leq i \leq n} [d_i - (|b_{i-1}| + |b_i|)], \quad b^{(0)} := \max_{1 \leq i \leq n} [d_i + (|b_{i-1}| + |b_i|)];$$

(2) for $k = 0, 1, 2, \dots$, do

令 $c^{(k)} := (a^{(k)} + b^{(k)})/2$;

计算施图姆序列在 $c^{(k)}$ 的值得 $S_{c^{(k)}} = \{p_0(c^{(k)}), p_1(c^{(k)}), \dots, p_n(c^{(k)})\}$,

计算序列 $S_{c^{(k)}}$ 的符号变化数 $s(c^{(k)})$,

如果 $s(c^{(k)}) > (n - i)$, 令 $b^{(k+1)} = c^{(k)}$; 否则, 令 $a^{(k+1)} = c^{(k)}$;

如果满足精度要求, 给出 λ_i 的一个近似, $\lambda_i \approx (a^{(k+1)} + b^{(k+1)})/2$.

end do

end do

例 8.3.2 设三对角矩阵 $A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 2 \end{pmatrix}$, 求 A 的第二个特征值 λ_2 .

解 由 (8.3.5) 得施图姆序列, 再应用吉文斯方法计算可得表 8.5.

表 8.5

k	$a^{(k)}$	$b^{(k)}$	$c^{(k)}$	$S_{c^{(k)}}$	$s(c^{(k)})$
0	0	4.000	2.0000	(1, 0, -1, 0, 1)	2
1	2.0000	4.000	3.0000	(1, -1, 0, 1, -1)	3
2	2.0000	3.000	2.5000	(1, -0.5, -0.75, 0.875, 0.3125)	2
3	2.5000	3.000	2.7500	(1, -0.75, -0.4375, 1.0781, -0.3711)	3
4	2.5000	2.750	2.6250	(1, -0.625, -0.6094, 1.0059, -0.0193)	3
5	2.5000	2.625	2.5625	(1, -0.5625, -0.6836, 0.9470, 0.1509)	2
6	2.5625	2.625	2.5938	(1, -0.5938, -0.6474, 0.9782, 0.0665)	2
7	2.5938	2.625	2.6094	(1, -0.6094, -0.6286, 0.9925, 0.0238)	2
8	2.6094	2.625	2.6172	(1, -0.6172, -0.6191, 0.9993, 0.0023)	2
9	2.6172	2.625	2.6211	(1, -0.6211, -0.6142, 1.0026, -0.0085)	3
10	2.9172	2.621	2.6191	(1, -0.6191, -0.6167, 1.0009, -0.0029)	3
11	2.6172	2.619	2.6182	(1, -0.6182, -0.6178, 1.0001, -0.0004)	3
12	2.6172	2.618	2.6177	(1, -0.6177, -0.6184, 0.9997, 0.0009)	2

故得 $\lambda_2 \approx 2.6177$. 解毕.

8.3.3 兰乔斯方法

设 A 为对称稀疏矩阵, 其特征值按模大小排列为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} \geq \lambda_n.$$

兰乔斯算法用于求大规模对称稀疏矩阵的最大和最小特征值 λ_1, λ_n . 参见第 3 章, 兰乔斯算法产生一个三对角矩阵 H_m 的序列, 其中 H_m 的最大和最小特征值收敛于矩阵 A 的最大和最小特征值.

对非零向量 x , 引入 Rayleigh 商 $r(x) = (x^T A x) / (x^T x)$, 则由柯朗-费希尔 (Courant-Fisher) 定理^[4], 成立

$$\lambda_1(A) = \max_{x \in \mathbf{R}^n, x \neq 0} r(x), \quad \lambda_n(A) = \min_{x \in \mathbf{R}^n, x \neq 0} r(x).$$

应用到三对角矩阵 $H_m = V_m^T A V_m$ 有

$$\lambda_1(H_m) = \max_{x \in \mathbf{R}^n, x \neq 0} \frac{(V_m x)^T A (V_m x)}{x^T x} = \max_{\|x\|_2=1} r(H_m x) \leq \lambda_1(A),$$

$$\lambda_m(H_m) = \min_{x \in \mathbf{R}^n, x \neq 0} \frac{(V_m x)^T A (V_m x)}{x^T x} = \min_{\|x\|_2=1} r(H_m x) \geq \lambda_n(A).$$

兰乔斯算法的每一步, 上式给出了 A 的最大和最小特征值的一个上、下界, 下面性质则保证了算法的收敛性.

性质 8.3.2 设 A 为对称稀疏矩阵, 其特征值按模大小排列为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, $u_1, u_2, \cdots, u_n$ 为对应的正交的特征向量, 设 $\eta_1, \eta_2, \cdots, \eta_n$ 为 H_m 对应的特征值, 并且 $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_n$, 则

$$\lambda_1 \geq \eta_1 \geq \lambda_1 - \frac{(\lambda_1 - \lambda_n)(\tan \phi_1)^2}{(T_{m-1}(1 + 2\rho_1))^2},$$

其中 $\cos \phi_1 = |(q^{(1)})^T u_1|$, $\rho_1 = (\lambda_1 - \lambda_2)/(\lambda_2 - \lambda_n)$, $T_{m-1}(x)$ 为 $m-1$ 次切比雪夫多项式.

$$\lambda_n \leq \eta_m \leq \lambda_n + \frac{(\lambda_1 - \lambda_n)(\tan \phi_n)^2}{(T_{m-1}(1 + 2\rho_n))^2},$$

其中 $\cos \phi_n = |(q^{(n)})^T u_n|$, $\rho_n = (\lambda_{n-1} - \lambda_n)/(\lambda_1 - \lambda_{n-1})$.

兰乔斯算法可以很好地保持矩阵的稀疏特性, 并且收敛速度比乘幂法等快得多, 所以兰乔斯算法常用于求大规模对称稀疏矩阵的最大和最小特征值, 并可推广应用到广义特征值问题. 具体 Matlab 算法过程可参见文献 [6]–eiglanz.

8.4 奇异值 (SVD) 的计算

定义 8.4.1 设 $A \in \mathbf{C}^{m \times n}$, 若存在酉阵 $U \in \mathbf{C}^{m \times m}$, $V \in \mathbf{C}^{n \times n}$, 使得

$$U^H A V = \Sigma = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p) \in \mathbf{R}^{m \times n}, \quad p = \min\{m, n\}, \quad (8.4.1)$$

并且 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, 则式 (8.4.1) 称为 A 的奇异值分解 (SVD), σ_i 称为 A 的奇异值, V 的列 v_i 称为 A 对应于 σ_i 的右奇异向量, U 的列 u_i 称为 A 对应于 σ_i 的左奇异向量.

如果 A 为实矩阵, 则奇异值为

$$\sigma_i(A) = \sqrt{\lambda_i(A^T A)}, \quad i = 1, 2, \cdots, p.$$

也就是说, A 的 SVD 分解可通过计算 $A^T A$ 的特征值和特征向量得到, 计算 $A^T A$ 的特征值和特征向量得到 A 的 SVD 分解.

注意: 由线性代数知识知, $A^T A$ 和 AA^T 的非零特征值相同, 所以 A 的 SVD 分解可也通过计算 AA^T 的特征值和特征向量得到.

例 8.4.1 设 $A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix}$, 求矩阵 A 奇异值和奇异向量.

解 因为 $A^T A$ 是三阶矩阵, AA^T 是二阶矩阵, 所以由 AA^T 计算 A 的奇异值比较方便. 矩阵 $AA^T = \begin{pmatrix} 5 & 2 \\ 2 & 8 \end{pmatrix}$ 的特征值是 $\lambda_1 = 9, \lambda_2 = 4$, 因此, A 的奇异值为 $\sigma_1 = 3, \sigma_2 = 2$. 因为 A 的左奇异向量是 AA^T 的特征向量, 可以求得与 λ_1, λ_2 对应的单位特征向量, 即左奇异向量为

$$u_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad u_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

右奇异向量可由计算 $A^T A$ 的特征向量得到, 也可应用 $v_i = \sigma_i^{-1} A^T u_i$ 得到

$$v_1 = \frac{1}{3\sqrt{5}} \begin{pmatrix} 5 \\ 2 \\ 4 \end{pmatrix}, \quad v_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix},$$

再计算与 v_1 和 v_2 都正交的单位向量得 $v_3 = \frac{1}{3}(-2, 1, 2)^T$.

于是得到 A 的奇异值分解 $A = U\Sigma V^T$, 其中

$$U = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}, \quad V = \frac{1}{3\sqrt{5}} \begin{pmatrix} 5 & 0 & -2\sqrt{5} \\ 2 & 6 & \sqrt{5} \\ 4 & -3 & 2\sqrt{5} \end{pmatrix}.$$

解毕.

下面介绍另一种计算矩阵奇异值分解的近似算法. $A \in \mathbb{C}^{m \times n} (m \geq n)$, 此方法分为两步. 第 1 步将矩阵 A 进行一系列初等行变换, 将其约化为行阶梯形矩阵; 第 2 步对约化得到的方阵执行 QR 算法.

算法 8.4.1

(1) 将矩阵 A 进行一系列初等行变换, 将其约化为行阶梯形矩阵, 即寻找矩阵 U, V , 使得

$$U^T A V = \begin{pmatrix} B \\ 0 \end{pmatrix},$$

其中 B 为上两对角矩阵, $U = U_1 U_2 \cdots U_{m-1}$, $V = V_1 V_2 \cdots V_{n-2}$, U_i, V_j 为豪斯霍尔德矩阵;

(2) 对方阵 B 执行 QR 算法, 求解 B 的特征值, 即寻找矩阵 W, Z , 使得

$$W^T B Z = \Sigma = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n),$$

从而得到 A 的奇异值分解

$$U^T A V = \begin{pmatrix} \Sigma \\ 0 \end{pmatrix}, \quad U = U \text{diag}(W, I_{m-n}), \quad V = V Z.$$

8.5 广义特征值问题

定义 8.5.1 设 $A, B \in \mathbb{C}^{n \times n}$ 为两个给定的矩阵, 则广义特征值问题定义如下: 求 $\lambda \in \sigma(A, B)$ 和非零向量 $x \in \mathbb{C}^n$ 满足

$$Ax = \lambda Bx, \quad (8.5.1)$$

其中广义特征值 $\sigma(A, B)$ 定义为

$$\sigma(A, B) = \{\lambda \in \mathbb{C} \mid \det(A - \lambda B) = 0\}.$$

当 $B = I$ 时, 即为标准的矩阵特征值问题.

8.5.1 广义 Schur 分解

定义 8.5.2 称 $P(\lambda) = \det(A - \lambda B)$ 为特征多项式, $\det(A - \lambda B) = 0$ 称为特征方程. 如果 $\det(A - \lambda B) \neq 0$, 则称 (A, B) 是正则的; 否则, 称 (A, B) 奇异.

定义 8.5.2 显示 (A, B) 有 n 个特征值的充分必要条件是 B 非奇. 在这种情况下, 广义特征值问题转化为等价的特征值问题 $Cx = \lambda x$, 其中 C 满足 $BC = A$. 于是对矩阵 C 执行 QR 算法可求得特征值. 当然, 如果 B 是病态矩阵, 则这一求解过程并不能得到准确的结果.

一种更实用的求解广义特征值的方法基于如下的 Schur 分解:

性质 8.5.1 设 (A, B) 正则, 则存在酉阵 U, Z , 使得 $U^H A Z = T, U^H B Z = S$, 其中 T, S 为上三角矩阵, 并且 (A, B) 的特征值可表示为

$$\begin{aligned} \lambda_i &= \frac{t_{ii}}{s_{ii}}, \quad s_{ii} \neq 0, \\ \lambda_i &= \infty, \quad t_{ii} \neq 0, s_{ii} = 0. \end{aligned}$$

如果 A, B 为实矩阵, 则存在正交矩阵 U, Z , 使得 $T = U^T A Z$ 为上拟三角矩阵, $S = U^T B Z$ 为上三角矩阵, 此分解称为 (A, B) 的广义实 Schur 分解. 于是得到如下的求解实矩阵广义特征值的 QZ 迭代算法:

算法 8.5.1 (QZ 迭代)

(1) 分别对矩阵 A, B 进行矩阵变换, 将其约化为上海森伯格形式和上三角形, 即寻找正交矩阵 Q, Z , 使得 $A = Q^T A Z$ 为上海森伯格矩阵, $B = Q^T B Z$ 为上三角矩阵;

(2) 对矩阵 AB^{-1} 执行 QR 迭代, 求解 AB^{-1} 的特征值.

8.5.2 对称正定矩阵的广义 Schur 分解

工程问题中常见的情形是 A, B 为对称矩阵, 并且 B 为正定矩阵. 在这种情形下, 对称正定矩阵对 (A, B) 的广义特征值有如下结论:

定理 8.5.1 设 A, B 为对称矩阵, 并且 B 为正定矩阵, 则对称正定矩阵对 (A, B) 有实特征值和线性无关的特征向量. 并且矩阵 A, B 可同时对角化, 即存在非奇矩阵 $X \in \mathbb{R}^{n \times n}$, 使得

$$X^T A X = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n), \quad X^T B X = I_n.$$

于是得到如下求解对称正定矩阵广义特征值的 QR-楚列斯基算法:

算法 8.5.2 (QR-楚列斯基算法)

(1) 计算楚列斯基分解 $B = H^T H$;

(2) 计算 $C = H^{-T} A H^{-1}$;

(3) 对 $i = 1, 2, \dots, n$, 应用 QR 迭代计算对称矩阵 C 的特征值 λ_i 和特征向量 z_i , 应用格拉姆-施密特过程将 $\{z_i\}$ 正交化得到正交特征向量 $\{y_i\}$;

(4) 对 $i = 1, 2, \dots, n$, 求解 $H x_i = y_i$ 计算 (A, B) 的特征向量 x_i .

小 结

本章介绍了求解矩阵特征值和特征向量的常用方法. 乘幂法常用于求按模最大的特征值; 带位移的反幂法用于已知特征值的近似值时, 求其精确值及对应的特征向量; QR 方法用于求一般矩阵全部特征值和特征向量; 阿诺尔迪方法用于求稀疏矩阵特征值. 雅可比方法用于求实对称矩阵的全部特征值和特征向量; 吉文斯方法用于实对称三对角矩阵特征值的计算; 兰乔斯算法用于求大规模对称稀疏矩阵的最大最小特征值, 它具有很好的收敛性, 并可以很好地保持矩阵的稀疏特性; 广义 Schur 分解用于求解广义特征值问题.

习 题

8.1 已知 $A = \begin{pmatrix} 6 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}$, 用乘幂法求 A 按模最大的特征值及对应的特征向量, 要

求 $|m_k - m_{k-1}| < 0.01$.

8.2 已知 $A = \begin{pmatrix} -11 & 11 & 1 \\ 11 & 9 & -2 \\ 1 & -2 & 13 \end{pmatrix}$, 先用乘幂法执行适当步数, 然后用原点位移法求 A

按模最大的特征值 (取 $p = 15$), 最后用艾特肯加速法求其更精确的近似值.

8.3 已知 $A = \begin{pmatrix} 4 & 1 & 4 \\ 1 & 10 & 1 \\ 4 & 1 & 10 \end{pmatrix}$, 用反幂法求 A 接近于 12 的特征值及对应的特征向量

(精确到 4 位小数).

8.4 已知 $A = \begin{pmatrix} 1 & 4 & 3 \\ 4 & 5 & 5 \\ 3 & 5 & 10 \end{pmatrix}$, 利用镜射变换将 A 化为相似的三对角矩阵, 并用带位移

的 QR 方法求其特征值.

8.5 已知 $A = \begin{pmatrix} 1 & 1 & 1/2 \\ 1 & 1 & 1/4 \\ 1/2 & 1/4 & 2 \end{pmatrix}$, 用雅可比方法求 A 的全部特征值.

8.6 已知 $\alpha = (a_1, a_2, \dots, a_n)$, $A = \alpha^T \alpha$, 求 A 的全部特征值.

8.7 已知 $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}$, 对 A 进行奇异值分解.

8.8 已知 $A = \begin{pmatrix} 1 & 2 & 4 & 3 \\ 3 & -1 & 2 & -2 \\ 5 & -4 & 0 & -7 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 2 & -1 \\ 3 & 2 & 1 \\ -1 & -2 & -1 \\ 3 & 5 & 4 \end{pmatrix}$, 求 (A, B) 的广义特

征值.

计算实习

8.1 已知

$$A = \begin{pmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{pmatrix},$$

- (1) 用乘幂法求 A 按模最大的特征值及对应的特征向量, 要求 $|m_k - m_{k-1}| < 0.00001$;
- (2) 利用反幂法, 使用不同的 p 值, 求 A 的不同特征值和特征向量;
- (3) 并用 QR 方法求其全部特征值.

8.2 已知 $A = \begin{pmatrix} 2 & 1 & & & \\ 1 & 3 & & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 99 & 1 \\ & & & 1 & 100 \end{pmatrix}$, 用吉文斯算法求 A 的最大和最小特征值.

第9章 常微分方程数值解法

常微分方程课程中对一些典型的常微分方程给出了求其解析解的方法. 然而, 在科学研究和工程技术领域中的常微分方程往往都比较复杂, 难以求其解析解. 本章主要介绍求解一阶常微分方程初值问题和二阶边值问题的数值解法.

一阶常微分方程初值问题为

$$\begin{cases} y'(x) = f(x, y(x)), & a \leq x \leq b, \\ y(a) = y_0, \end{cases} \quad (9.0.1)$$

其中 y_0 为已知常数, $f(x, y(x))$ 为已知函数.

由常微分方程理论知, 初值问题 (9.0.1) 的解存在且唯一性定理如下:

定理 9.0.1 设 $f(x, y)$ 在区域 $D = \{(x, y) | a \leq x \leq b, -\infty < y < \infty\}$ 上连续, 并且关于 y 满足利普希茨 (Lipschitz) 条件, 即存在常数 L , 使得

$$|f(x, y) - f(x, \bar{y})| \leq \left| \frac{\partial f}{\partial y}(y - \bar{y}) \right| \leq L|y - \bar{y}|$$

对于任意的 $x \in [a, b]$ 及任意的 $y, \bar{y}$ 都成立, 则初值问题 (9.0.1) 在区间 $[a, b]$ 上存在唯一的连续可微解 $y = y(x)$.

常微分方程数值解法是求其准确解 $y(x)$ 在一系列点 x_i ($a = x_0 < x_1 < \cdots < x_n = b$) 处 $y(x_i)$ 的近似值 y_i ($i = 0, 1, \cdots, n$) 的方法, y_i ($i = 0, 1, \cdots, n$) 称为数值解. 记 $h_i = x_i - x_{i-1}$, h_i 称为 x_{i-1} 到 x_i 的步长, 通常步长取为常量 h , 即取 $h = (b - a)/n$, 这时 $x_i = a + ih$ ($i = 0, 1, \cdots, n$).

本章是在初值问题 (9.0.1) 的解存在且唯一的条件下求其数值解.

9.1 初值问题常用数值解法的建立与使用

9.1.1 基本数值解法的建立与隐式法的求解

下面介绍导出初值问题 (9.0.1) 数值解法的三种方法: 数值微分法、数值积分法和泰勒展开法.

1. 数值微分法

数值微分法是用数值微分公式代替常微分方程中的导数所建立的方法.

(1) 欧拉 (Euler) 法

在 $x = x_i$ 处, 初值问题 (9.0.1) 为

$$y'(x_i) = f(x_i, y(x_i)). \quad (9.1.1)$$

由数值微分公式 (6.4.3) 有

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_i)}{h} - \frac{h}{2}y''(\xi_i).$$

将其代入式 (9.1.1) 左边得

$$y(x_{i+1}) = y(x_i) + hf(x_i, y(x_i)) + \frac{h^2}{2}y''(\xi_i).$$

略去误差项 $\frac{h^2}{2}y''(\xi_i)$, 并用 y_i 代替 $y(x_i)$, 则得求解初值问题 (9.0.1) 数值解的欧拉法计算公式

$$y_{i+1} = y_i + hf(x_i, y_i). \quad (9.1.2)$$

定义 9.1.1 (局部截断误差) 设 $y(x_{i+1})$ 是初值问题 (9.0.1) 的准确解 $y(x)$ 在 x_{i+1} 处的值, 假定前面各步计算均无误差 (称为局部化假定), 即 $y_i = y(x_i)$, $y_{i-1} = y(x_{i-1})$, $\dots$. 而 y_{i+1} 是由某一数值解法求得的 $y(x_{i+1})$ 的近似值, 则 $R[y] = y(x_{i+1}) - y_{i+1}$ 称为该数值解法在 x_{i+1} 处的局部截断误差.

由定义 9.1.1 可知, 所谓局部截断误差即在局部化假定下当前一步计算所产生的截断误差. 可见, 欧拉法的局部截断误差

$$R[y] = y(x_{i+1}) - y_{i+1} = \frac{h^2}{2}y''(\xi_i).$$

(2) 后退欧拉法

在 $x = x_{i+1}$ 处, 初值问题 (9.0.1) 为

$$y'(x_{i+1}) = f(x_{i+1}, y(x_{i+1})). \quad (9.1.3)$$

由数值微分公式 (6.4.3) 有

$$y'(x_{i+1}) = \frac{y(x_{i+1}) - y(x_i)}{h} + \frac{h}{2}y''(\xi_i).$$

将其代入式 (9.1.3) 左边得

$$y(x_{i+1}) = y(x_i) + hf(x_{i+1}, y(x_{i+1})) - \frac{h^2}{2}y''(\xi_i).$$

同理可得后退欧拉法及其局部截断误差

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}), \quad (9.1.4)$$

$$R[y] = y(x_{i+1}) - y_{i+1} = -\frac{h^2}{2}y''(\xi_i).$$

(3) 中点法

由数值微分公式 (6.4.4) 有

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_{i-1}))}{2h} - \frac{h^2}{6}y'''(\xi_i).$$

将其代入式 (9.1.1) 左边得

$$y(x_{i+1}) = y(x_{i-1}) + 2hf(x_i, y(x_i)) + \frac{h^3}{3}y'''(\xi_i).$$

同理可得中点法及其局部截断误差

$$y_{i+1} = y_{i-1} + 2hf(x_i, y_i), \quad (9.1.5)$$

$$R[y] = y(x_{i+1}) - y_{i+1} = -\frac{h^3}{3}y'''(\xi_i).$$

例 9.1.1 取步长 $h = 0.02$, 分别用欧拉法、后退欧拉法以及中点法求解初值问题

$$\begin{cases} y' = -\frac{0.9y}{1+2x}, & 0 \leq x \leq 0.1, \\ y(0) = 1. \end{cases}$$

解 $x_i = ih = 0.02i, i = 0, 1, \dots, 5, y_0 = 1.$

(1) 欧拉法. 由 $y_{i+1} = y_i + hf(x_i, y_i)$ 有

$$y_{i+1} = y_i - 0.02 \frac{0.9y_i}{1+2x_i}, \quad i = 0, 1, 2, 3, 4.$$

由此算得

$$y_0 = 1.0000, \quad y_1 = 0.9820, \quad y_2 = 0.9650, \quad y_3 = 0.9489, \quad y_4 = 0.9337, \quad y_5 = 0.9192.$$

(2) 后退欧拉法. 由 $y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$ 有

$$y_{i+1} = y_i - 0.02 \frac{0.9y_{i+1}}{1+2x_{i+1}}, \quad i = 0, 1, \dots, 4.$$

解之得

$$y_{i+1} = \frac{1+2x_{i+1}}{1.018+2x_{i+1}} y_i, \quad i = 0, 1, \dots, 4.$$

由此算得

$$y_0=1.000\,0, \quad y_1=0.983\,0, \quad y_2=0.966\,9, \quad y_3=0.951\,6, \quad y_4=0.937\,0, \quad y_5=0.923\,2.$$

(3) 中点法. 由 $y_{i+1} = y_{i-1} + 2hf(x_i, y_i)$ 有

$$y_{i+1} = y_{i-1} - 0.04 \frac{0.9y_i}{1+2x_i}, \quad i = 0, 1, 2, 3, 4.$$

已知 $y_0 = 1$, 按后退欧拉法取 $y_1 = 0.983\,0$, 由此算得

$$y_2 = 0.966\,0, \quad y_3 = 0.950\,8, \quad y_4 = 0.935\,4, \quad y_5 = 0.921\,8.$$

读者可以将以上数值解与真解 $y(x) = (1+2x)^{-0.45}$ 在 x_i 处的值进行比较. 解毕.

2. 数值积分法

数值积分法是通过对方微分方程两端积分所建立的方法.

(1) 梯形法

在区间 $[x_i, x_{i+1}]$ 上对微分方程 $y' = f(x, y)$ 两端积分得

$$\int_{x_i}^{x_{i+1}} y'(x) dx = \int_{x_i}^{x_{i+1}} f(x, y(x)) dx.$$

对右端积分利用梯形求积公式计算得

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + \frac{h}{2} [f(x_i, y(x_i)) + f(x_{i+1}, y(x_{i+1}))] - \frac{h^3}{12} f''(\xi_i, y(\xi_i)) \\ &= y(x_i) + \frac{h}{2} [f(x_i, y(x_i)) + f(x_{i+1}, y(x_{i+1}))] - \frac{h^3}{12} y'''(\xi_i), \quad x_i \leq \xi_i \leq x_{i+1}. \end{aligned}$$

由此得梯形法及其局部截断误差

$$y_{i+1} = y_i + \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, y_{i+1})], \quad (9.1.6)$$

$$R[y] = y(x_{i+1}) - y_{i+1} = -\frac{h^3}{12} y'''(\xi_i).$$

(2) 辛普森法

在区间 $[x_{i-1}, x_{i+1}]$ 上对微分方程 $y' = f(x, y)$ 两端积分得

$$\int_{x_{i-1}}^{x_{i+1}} y'(x) dx = \int_{x_{i-1}}^{x_{i+1}} f(x, y(x)) dx.$$

用辛普森求积公式计算右端积分得

$$\begin{aligned} y(x_{i+1}) &= y(x_{i-1}) + \frac{x_{i+1} - x_{i-1}}{6} [f(x_{i-1}, y(x_{i-1})) + 4f(x_i, y(x_i)) + f(x_{i+1}, y(x_{i+1}))] \\ &\quad - \frac{h^5}{90} f^{(4)}(\xi_i, y(\xi_i)). \end{aligned}$$

由此得辛普森法及其局部截断误差

$$y_{i+1} = y_{i-1} + \frac{h}{3}[f(x_{i-1}, y_{i-1}) + 4f(x_i, y_i) + f(x_{i+1}, y_{i+1})], \quad (9.1.7)$$

$$R[y] = y(x_{i+1}) - y_{i+1} = -\frac{h^5}{90}y^{(5)}(\xi_i).$$

同理, 用数值积分法可得欧拉法、后退欧拉法、中点法的计算公式, 此处不作推导, 留作习题.

(3) 亚当斯 (Adams) 显式法

给定 $k+1$ 个数据点 $(x_{i-j}, f_{i-j}) (j = 0, 1, \dots, k)$, 其中节点 x_{i-j} 等距分布, $f_{i-j} = f(x_{i-j}, y_{i-j})$. 作 k 次拉格朗日插值多项式 $L_k(x)$, 则 $f(x, y) = L_k(x) + R_k(x)$. 在区间 $[x_i, x_{i+1}]$ 上对微分方程 $y' = f(x, y)$ 两端积分得

$$\int_{x_i}^{x_{i+1}} y'(x) dx = \int_{x_i}^{x_{i+1}} f(x, y(x)) dx = \int_{x_i}^{x_{i+1}} L_k(x) dx + \int_{x_i}^{x_{i+1}} R_k(x) dx.$$

取 $\int_{x_i}^{x_{i+1}} y'(x) dx \approx \int_{x_i}^{x_{i+1}} L_k(x) dx$, 可得求解初值问题 (9.0.1) 的亚当斯显式公式 (也称为 Adams-Bashforth 公式) 及其局部截断误差

$$y_{i+1} = y_i + \frac{h}{A}(b_0 f_i + b_1 f_{i-1} + \dots + b_k f_{i-k}), \quad i \geq k, \quad (9.1.8)$$

$$R[y] = B_k h^{k+2} y^{(k+2)}(\xi_i).$$

式 (9.1.8) 中的系数如表 9.1 所示.

表 9.1 亚当斯显式公式系数表

k	A	b_0	b_1	b_2	b_3	b_4	b_5	B_k
0	1	1						1/2
1	2	3	-1					5/12
2	12	23	-16	5				3/8
3	24	55	-59	37	-9			251/720
4	720	1901	-2774	2616	-1274	251		95/288
5	1440	4277	-7923	9982	-7298	2877	-475	10987/60480

例如, 当 $k=3$ 时的亚当斯显式公式及其局部截断误差为

$$y_{i+1} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}),$$

$$R[y] = \frac{251}{720}h^5 y^{(5)}(\xi_i).$$

(4) 亚当斯 (Adams) 隐式法

给定 $k+1$ 个数据点 $(x_{i-j}, f_{i-j})(j = -1, 0, 1, \dots, k-1)$, 与亚当斯显式法同理, 可得亚当斯隐式公式(也称为 Adams-Moulton 公式)及其局部截断误差

$$y_{i+1} = y_i + \frac{h}{A^*}(b_0^* f_{i+1} + b_1^* f_i + \dots + b_k^* f_{i-k+1}), \quad i \geq k-1, \quad (9.1.9)$$

$$R[y] = B_k^* h^{k+2} y^{(k+2)}(\xi_i).$$

式 (9.1.9) 中的系数如表 9.2 所示.

表 9.2 亚当斯隐式公式系数表

k	A^*	b_0^*	b_1^*	b_2^*	b_3^*	b_4^*	b_5^*	B_k^*
0	1	1						$-1/2$
1	2	1	1					$-1/12$
2	12	5	8	-1				$-1/24$
3	24	9	19	-5	1			$-19/720$
4	720	251	646	-264	106	-19		$-3/160$
5	1 440	475	1 427	-798	482	-173	27	$-863/60 480$

例如, 当 $k=3$ 时的亚当斯隐式公式及其局部截断误差为

$$y_{i+1} = y_i + \frac{h}{24}(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}),$$

$$R[y] = -\frac{19}{720}h^5 y^{(5)}(\xi_i).$$

比较表 9.1 与表 9.2 中局部截断误差的系数可见 $|B_k^*| < |B_k|$ ($k \geq 1$), 这说明隐式法的局部截断误差小于显式法的局部截断误差.

3. 泰勒级数法

泰勒级数法是基于泰勒展开所建立的方法. 设函数 f 充分可导, 将 $y(x_{i+1})$ 在 x_i 处泰勒展开有

$$y(x_{i+1}) = y(x_i + h) = y(x_i) + y'(x_i)h + \frac{1}{2}y''(x_i)h^2 + \frac{1}{3!}y'''(x_i)h^3 + \dots + \frac{1}{p!}y^{(p)}(x_i)h^p$$

$$+ \frac{1}{(p+1)!}y^{(p+1)}(\xi_i)h^{p+1}, \quad x_i < \xi_i < x_{i+1}.$$

略去上式中的误差项, 用 $y_i^{(k)}$ 代替 $y^{(k)}(x_i)$, 则得 p 阶泰勒级数法

$$y_{i+1} = y_i + y_i' h + \frac{1}{2}y_i'' h^2 + \frac{1}{3!}y_i''' h^3 + \dots + \frac{1}{p!}y_i^{(p)} h^p, \quad (9.1.10)$$

其中 $y_i^{(k)}$ 的计算公式为

$$\begin{cases} y_i' = f(x_i, y_i), \\ y_i'' = (f'_x + f'_y f)|_{(x_i, y_i)}, \\ y_i''' = (f''_{xx} + 2f''_{xy} f + f''_{yy} f^2 + f'_x f'_y + f_y'^2 f)|_{(x_i, y_i)}, \\ \dots\dots \end{cases} \quad (9.1.11)$$

显然, 式 (9.1.10) 的局部截断误差

$$R[y] = y(x_{i+1}) - y_{i+1} = \frac{h^{p+1}}{(p+1)!} y^{(p+1)}(\xi_i).$$

在式 (9.1.11) 中取 $p = 1$, 即欧拉法; 取 $p = 2$ 得

$$y_{i+1} = y_i + hf(x_i, y_i) + \frac{h^2}{2} [f'_x(x_i, y_i) + f'_y(x_i, y_i)f(x_i, y_i)], \quad (9.1.12)$$

其局部截断误差

$$R[y] = y(x_{i+1}) - y_{i+1} = \frac{h^3}{6} y'''(\xi_i).$$

当 $p > 2$ 时需要计算式 (9.1.10) 中的高阶导数, 当 $f(x, y)$ 的表达式比较复杂时, 计算其高阶导数往往相当困难, 因而泰勒级数法适用于求解比较简单的常微分方程.

4. 名词介绍

下面介绍几个与数值解有关的名词.

(1) **p 阶方法**. 若某一数值解法的局部截断误差 $R[y] = O(h^{p+1})$, 则称该法为 p 阶方法或称该法具有 p 阶精度. 例如, 欧拉法、后退欧拉法是一阶方法, 中点法、梯形法是二阶方法, 辛普森法是四阶方法.

(2) **显式法**. 由数值求解公式可以直接计算 y_{i+1} 的方法, 即求解公式中的 $f(x, y)$ 不含 y_{i+1} 的方法. 例如, 欧拉法、中点法.

(3) **隐式法**. 由数值求解公式不能直接计算 y_{i+1} 的方法, 即求解公式中的 $f(x, y)$ 含有 y_{i+1} . 例如, 后退欧拉法、梯形法、辛普森法. 一般地, $f(x, y)$ 是非线性函数, y_{i+1} 不能显化, 计算 y_{i+1} 需要解方程.

(4) **单步法**. 计算 y_{i+1} 只需要一个 y_i 值的方法. 例如, 欧拉法、后退欧拉法、梯形法.

(5) **多步法**. 计算 y_{i+1} 需要 $y_i, y_{i-1}, \dots$ 多个值的方法. 一般 k 步法要用到 $y_i, y_{i-1}, \dots, y_{i-k+1}$ 的值. 例如, 中点法、辛普森法、亚当斯显 (隐) 式法.

5. 线性多步法开始值的计算和隐式法的求解

(1) 以亚当斯显式法为例说明线性多步法开始值的计算. 亚当斯显式法 (9.1.8) 从 $i = k$ 开始计算, 此时计算 y_{k+1} 需要开始值 $y_0, y_1, y_2, \dots, y_k$. 除了 y_0 是由初始条件给出的外, 其余各值需用其他方法算出. 为了保证计算的精度, 计算开始值的方法应采用与多步法同阶的单步法, 其他多步法的开始值同理计算.

(2) 隐式法的求解一般采用非线性方程求根的方法. 下面以梯形法

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

为例来说明隐式法的求解.

(i) 简单迭代法. 记

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

则有 $y_{i+1} = \phi(y_{i+1})$. 给定 y_{i+1} 的初始值 $y_{i+1}^{(0)}$ (如由欧拉法 $y_{i+1}^{(0)} = y_i + hf(x_i, y_i)$ 给出 $y_{i+1}^{(0)}$) 后, 由迭代公式

$$y_{i+1}^{(k+1)} = \phi(y_{i+1}^{(k)}) = y_i + \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, y_{i+1}^{(k)})], \quad k = 0, 1, 2, \dots$$

计算 y_{i+1} . 当 h 充分小时, 由利普希茨条件知

$$|\phi'(y_{i+1})| = \left| \frac{h}{2} \frac{\partial f(x_{i+1}, y_{i+1})}{\partial y} \right| \leq \frac{h}{2} L < 1.$$

由求解非线性方程的简单迭代法的收敛性定理知, 迭代序列 $\{y_{i+1}^{(k)}\}$ 收敛. 当 $|y_{i+1}^{(k+1)} - y_{i+1}^{(k)}| \leq \varepsilon$ 时, 取 $y_{i+1} \approx y_{i+1}^{(k+1)}$.

(ii) 牛顿法. 记

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

梯形法即解方程 $F(y_{i+1}) = 0$. 由非线性方程求根的牛顿法有

$$y_{i+1}^{(k+1)} = y_{i+1}^{(k)} - \frac{F(y_{i+1}^{(k)})}{F'(y_{i+1}^{(k)})}, \quad k = 0, 1, 2, \dots,$$

即

$$y_{i+1}^{(k+1)} = y_{i+1}^{(k)} - \frac{y_{i+1}^{(k)} - y_i - \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, y_{i+1}^{(k)})]}{1 - \frac{h}{2} f'_y(x_{i+1}, y_{i+1}^{(k)})}, \quad k = 0, 1, 2, \dots$$

当 $|y_{i+1}^{(k+1)} - y_{i+1}^{(k)}| \leq \varepsilon$ 时, 取 $y_{i+1} \approx y_{i+1}^{(k+1)}$.

(iii) 改进欧拉法. 所谓改进欧拉法就是用简单迭代法求解 y_{i+1} , 并且只迭代一次即可. 只要步长 h 取得足够小, 欧拉法就可以算出 y_{i+1} 较好的迭代初始值 $y_{i+1}^{(0)}$, 再用梯形法计算一次就可以得到 y_{i+1} 更好的近似值, 即

$$\begin{cases} y_{i+1}^{(0)} = y_i + hf(x_i, y_i), \\ y_{i+1} = y_i + \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, y_{i+1}^{(0)})] \end{cases} \quad (9.1.13)$$

或

$$y_{i+1} = y_i + \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, y_i + hf(x_i, y_i))]. \quad (9.1.13')$$

式 (9.1.13), (9.1.13') 称为**改进欧拉法**. 式 (9.1.13) 中的第一式称为预测公式, 第二式称为校正公式. 一般地, 由显式公式作预测, 同阶的隐式公式作校正的两个公式组成的方法称为预测—校正法. 式 (9.1.13) 就是一种预测—校正法. 式 (9.1.13) 或 (9.1.3') 可以改写成

$$\begin{cases} y_{i+1} = y_i + \frac{1}{2}(K_1 + K_2), \\ K_1 = hf(x_i, y_i), \\ K_2 = hf(x_i + h, y_i + K_1). \end{cases} \quad (9.1.14)$$

式 (9.1.14) 是 9.1.2 小节将要介绍的一种龙格—库塔法.

9.1.2 龙格—库塔法

利用泰勒展开可以导出**龙格—库塔 (Runge-Kutta) 法**(简称 **R-K 方法**). m 级的龙格—库塔法的一般形式为

$$\begin{cases} y_{i+1} = y_i + \lambda_1 K_1 + \lambda_2 K_2 + \cdots + \lambda_m K_m, \\ K_1 = hf(x_i, y_i), \\ K_2 = hf(x_i + \alpha_2 h, y_i + \beta_{21} K_1), \\ K_3 = hf(x_i + \alpha_3 h, y_i + \beta_{31} K_1 + \beta_{32} K_2), \\ \quad \dots\dots\dots \\ K_m = hf(x_i + \alpha_m h, y_i + \beta_{m1} K_1 + \beta_{m2} K_2 + \cdots + \beta_{m,m-1} K_{m-1}), \end{cases}$$

其中 $\lambda_i, \alpha_i, \beta_{j,k}$ 均为常数, 由待定系数法确定. 确定的原则是将局部截断误差 $R[y] = y(x_{i+1}) - y_{i+1}$ 在 x_i 处泰勒展开, 适当选取 h 的系数, 使得局部截断误差 $R[y]$ 的阶尽可能高.

下面以 $m = 2$ 级 R-K 法为例来说明 R-K 法的导出思想. 2 级 R-K 法形式如下:

$$\begin{cases} y_{i+1} = y_i + \lambda_1 K_1 + \lambda_2 K_2, \\ K_1 = hf(x_i, y_i), \\ K_2 = hf(x_i + \alpha h, y_i + \beta K_1). \end{cases} \quad (9.1.15)$$

将 $y(x_{i+1})$ 在 x_i 处泰勒展开有

$$y(x_{i+1}) = y(x_i + h) = y(x_i) + y'(x_i)h + \frac{1}{2!}y''(x_i)h^2 + \frac{1}{3!}y'''(x_i)h^3 + O(h^4).$$

由 $y'(x) = f(x, y(x))$ 知

$$\begin{aligned} y''(x) &= f'_x + f'_y y' = f'_x + f'_y f, \\ y'''(x) &= f''_{xx} + 2f''_{xy} f + f''_{yy} f^2 + f'_x f'_y + f'^2_y f, \end{aligned}$$

所以

$$\begin{aligned} y(x_{i+1}) &= y(x_i) + fh + \frac{1}{2}(f'_x + f'_y f)h^2 \\ &\quad + \frac{1}{6}(f''_{xx} + 2f''_{xy} f + f''_{yy} f^2 + f'_x f'_y + f'^2_y f)h^3 + O(h^4), \end{aligned} \quad (9.1.16)$$

其中 f 为 $f(x_i, y(x_i))$ 的简写, 符号 $f'_x, f'_y, f''_{xx}, f''_{xy}, f''_{yy}$ 的含义亦然.

又由二元函数的泰勒展开式有

$$\begin{aligned} K_2 &= hf(x_i + \alpha h, y_i + \beta K_1) \\ &= h \left(f + \alpha f'_x h + \beta f'_y f h + \frac{1}{2} \alpha^2 f''_{xx} h^2 + \alpha \beta f''_{xy} f h^2 + \frac{1}{2} \beta^2 f''_{yy} f^2 h^2 + \dots \right), \end{aligned}$$

于是

$$\begin{aligned} y_{i+1} &= y_i + \lambda_1 K_1 + \lambda_2 K_2 \\ &= y_i + \lambda_1 fh + \lambda_2 h \left(f + \alpha f'_x h + \beta f'_y f h + \frac{1}{2} \alpha^2 f''_{xx} h^2 \right. \\ &\quad \left. + \alpha \beta f''_{xy} f h^2 + \frac{1}{2} \beta^2 f''_{yy} f^2 h^2 + \dots \right) \\ &= y_i + (\lambda_1 + \lambda_2) fh + \lambda_2 (\alpha f'_x + \beta f'_y f) h^2 \\ &\quad + \lambda_2 \left(\frac{1}{2} \alpha^2 f''_{xx} + \alpha \beta f''_{xy} f + \frac{1}{2} \beta^2 f''_{yy} f^2 \right) h^3 + O(h^4). \end{aligned} \quad (9.1.17)$$

由式 (9.1.16) 和 (9.1.17) 可得局部截断误差

$$\begin{aligned} R[y] &= y(x_{i+1}) - y_{i+1} \\ &= (1 - \lambda_1 - \lambda_2) fh + \left[\left(\frac{1}{2} - \alpha \lambda_2 \right) f'_x + \left(\frac{1}{2} - \beta \lambda_2 \right) f'_y f \right] h^2 + \left[\left(\frac{1}{6} - \frac{1}{2} \alpha^2 \lambda_2 \right) f''_{xx} \right. \\ &\quad \left. + \left(\frac{1}{3} - \alpha \beta \lambda_2 \right) f''_{xy} f + \left(\frac{1}{6} - \frac{1}{2} \beta^2 \lambda_2 \right) f''_{yy} f^2 + \frac{1}{6} (f'_x f'_y + f'^2_y f) \right] h^3 + O(h^4). \end{aligned}$$

为使 $R[y]$ 的阶尽可能得高, 由 f 的任意性, 应选取 $\alpha, \beta, \lambda_1, \lambda_2$, 使得 h, h^2 的系数为零, 即得方程组

$$\begin{cases} 1 - \lambda_1 - \lambda_2 = 0, \\ \frac{1}{2} - \alpha \lambda_2 = 0, \\ \frac{1}{2} - \beta \lambda_2 = 0. \end{cases} \quad (9.1.18)$$

由此解出 $\alpha, \beta, \lambda_1, \lambda_2$, 代入 (9.1.15) 便得到相应公式. 注意到 $\alpha, \beta, \lambda_1, \lambda_2$ 无论怎样选取, h^3 的系数中 $f'_x f'_y + f'^2_y f \neq 0$, 故 $R[y] = O(h^3)$, 所以通过解方程组 (9.1.18) 得到的方法均是二阶方法. 方程组 (9.1.18) 含有 3 个方程、4 个未知量, 其解有无穷多个.

(1) 取 $\lambda_1 = \lambda_2 = 1/2, \alpha = \beta = 1$, 代入式 (9.1.15), 则得

$$\begin{cases} y_{i+1} = y_i + \frac{1}{2}K_1 + \frac{1}{2}K_2, \\ K_1 = hf(x_i, y_i), \\ K_2 = hf(x_i + h, y_i + K_1), \end{cases}$$

即改进欧拉法 (9.1.14).

(2) 取 $\lambda_1 = 0$, 则 $\lambda_2 = 1, \alpha = \beta = 1/2$, 代入式 (9.1.15), 则得

$$\begin{cases} y_{i+1} = y_i + K_2, \\ K_1 = hf(x_i, y_i), \\ K_2 = hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}K_1\right). \end{cases} \quad (9.1.19)$$

式 (9.1.19) 也常写为

$$y_{i+1} = y_i + hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}hf(x_i, y_i)\right). \quad (9.1.19')$$

(9.1.19) 或 (9.1.19') 称为变形欧拉法.

类似于上述二阶方法的推导, 可得多种 4 级 4 阶 R-K 法. 应用最广泛的是如下标准 (经典) 4 级 4 阶 R-K 法:

$$\begin{cases} y_{i+1} = y_i + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4), \\ K_1 = hf(x_i, y_i), \\ K_2 = hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}K_1\right), \\ K_3 = hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}K_2\right), \\ K_4 = hf(x_i + h, y_i + K_3). \end{cases} \quad (9.1.20)$$

另外两个常用的 4 级 4 阶 R-K 法的计算公式为

$$\begin{cases} y_{i+1} = y_i + \frac{1}{6}(K_1 + 4K_3 + K_4), \\ K_1 = hf(x_i, y_i), \\ K_2 = hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}K_1\right), \\ K_3 = hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{4}K_1 + \frac{1}{4}K_2\right), \\ K_4 = hf(x_i + h, y_i - K_2 + 2K_3), \end{cases} \quad (9.1.21)$$

$$\begin{cases} y_{i+1} = y_i + \frac{1}{8}(K_1 + 3K_2 + 3K_3 + K_4), \\ K_1 = hf(x_i, y_i), \\ K_2 = hf\left(x_i + \frac{1}{3}h, y_i + \frac{1}{3}K_1\right), \\ K_3 = hf\left(x_i + \frac{2}{3}h, y_i - \frac{1}{3}K_1 + K_2\right), \\ K_4 = hf(x_i + h, y_i + K_1 - K_2 + K_3). \end{cases} \quad (9.1.22)$$

4 阶以上 R-K 法计算函数值的工作量较大, 对于大量的实际问题 4 阶 R-K 法已可满足精度要求. 若有更高的精度要求, 则可使用下面的 6 级 5 阶英格兰 (England) 公式:

$$\begin{cases} y_{i+1} = y_i + \frac{1}{24}K_1 + \frac{5}{48}K_4 + \frac{27}{56}K_5 + \frac{125}{336}K_6, \\ K_1 = hf(x_i, y_i), \\ K_2 = hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}K_1\right), \\ K_3 = hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{4}K_1 + \frac{1}{4}K_2\right), \\ K_4 = hf(x_i + h, y_i - K_2 + 2K_3), \\ K_5 = hf\left(x_i + \frac{2}{3}h, y_i + \frac{7}{27}K_1 + \frac{10}{27}K_2 + \frac{1}{27}K_4\right), \\ K_6 = hf\left(x_i + \frac{1}{5}h, y_i + \frac{1}{625}(28K_1 - 125K_2 + 546K_3 + 54K_4 - 378K_5)\right). \end{cases} \quad (9.1.23)$$

英格兰公式的另一用途是在实际计算中用于估计局部截断误差的大小. 由于 4 阶 R-K 法的局部截断误差的表达式过于复杂 (此处从略), 不便于应用. 注意到

(9.1.21), (9.1.23) 两组公式中的 K_1, K_2, K_3, K_4 完全相同, 若要估计局部截断误差的大小, 按式 (9.1.23) 加算 K_5, K_6 , 将两组公式计算出的 $y(x_{i+1})$ 的近似值分别记为 y_{i+1} 与 $\tilde{y}_{i+1}$, 则有

$$y(x_{i+1}) - y_{i+1} = ch^5, \quad y(x_{i+1}) - \tilde{y}_{i+1} = \tilde{c}h^6.$$

以上两式相减得 $\tilde{y}_{i+1} - y_{i+1} \approx ch^5$, 从而可得局部截断误差实用估计式

$$y(x_{i+1}) - y_{i+1} \approx \tilde{y}_{i+1} - y_{i+1}. \quad (9.1.24)$$

R-K 方法的优缺点如下:

优点: 精度高, 它是显式单步法, 计算 y_{i+1} 只需要 y_i 一个值, 每一步的步长可根据精度的要求单独考虑其大小.

缺点: 需要计算多个函数值, 当 $f(x, y)$ 比较复杂时, 计算量大.

R-K 方法也常用于线性多步法开始值的计算. 除了以上介绍的显式 R-K 法以外, 还有隐式 R-K 法. 由于隐式 R-K 法的计算比较复杂, 此处从略.

例 9.1.2 用标准 R-K 法求解例 9.1.1 的初值问题.

解
$$f(x, y) = -\frac{0.9y}{1+2x}, \quad y_0 = 1, \quad h = 0.02.$$

按式 (9.1.20), 当 $i = 0$ 时,

$$K_1 = hf(x_0, y_0) = -h \frac{0.9y_0}{1+2x_0} = -0.02 \times 0.9 = -0.018,$$

$$\begin{aligned} K_2 &= hf(x_0 + h/2, y_0 + K_1/2) = hf(h/2, 1 - 0.018/2) \\ &= 0.02f(0.01, 0.991) = 0.02 \left(-\frac{0.9 \times 0.991}{1+0.02} \right) = -0.017488235, \end{aligned}$$

$$\begin{aligned} K_3 &= hf(x_0 + h/2, y_0 + K_2/2) = hf(h/2, 1 - 0.017488235/2) \\ &= 0.02f(0.01, 0.991255882) = 0.02 \left(-\frac{0.9 \times 0.991255882}{1+0.02} \right) \\ &= -0.017492751, \end{aligned}$$

$$\begin{aligned} K_4 &= hf(x_0 + h, y_0 + K_3) = hf(0.02, 1 - 0.017492751) \\ &= 0.02f(0.02, 0.982507249) = 0.02 \left(-\frac{0.9 \times 0.982507249}{1+0.04} \right) \\ &= -0.017004933. \end{aligned}$$

$$y_1 = y_0 + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) = 0.982505515.$$

其余各步计算请读者自己完成. 解毕.

9.1.3 待定系数法、预测 — 校正公式

1. 用待定系数法建立线性多步法

线性多步法一般形式如下:

$$y_{i+1} = \sum_{j=0}^k \alpha_j y_{i-j} + h \sum_{j=-1}^k \beta_j f_{i-j}, \quad (9.1.25)$$

其中 α_j, β_j 为常数, $f_{i-j} = f(x_{i-j}, y_{i-j})$. 当 $\beta_{-1} = 0$ 时, (9.1.25) 是显式公式(法); 当 $\beta_{-1} \neq 0$ 时, 式 (9.1.25) 右边第二个和式中含有 f_{i+1} , 故是隐式公式(法). 式 (9.1.25) 的局部截断误差为

$$\begin{aligned} R[y] &= y(x_{i+1}) - y_{i+1} = y(x_{i+1}) - \sum_{j=0}^k \alpha_j y_{i-j} - h \sum_{j=-1}^k \beta_j f_{i-j} \\ &= y(x_{i+1}) - \sum_{j=0}^k \alpha_j y(x_{i-j}) - h \sum_{j=-1}^k \beta_j y'(x_{i-j}). \end{aligned}$$

待定系数法建立线性多步法的思想如下: 用待定系数法确定局部截断误差 $R[y]$ 中的系数 α_j, β_j , 应使它的阶尽可能高, 然后用广义佩亚诺定理确定其误差项. 以下例说明待定系数法的思想.

例 9.1.3 确定如下求解公式的系数和局部截断误差:

$$y_{i+1} = \alpha_0 y_i + \alpha_1 y_{i-1} + \alpha_2 y_{i-2} + h(\beta_{-1} f_{i+1} + \beta_0 f_i + \beta_1 f_{i-1}).$$

解

$$\begin{aligned} R[y] &= y(x_{i+1}) - y_{i+1} \\ &= y(x_i + h) - \alpha_0 y(x_i) - \alpha_1 y(x_i - h) - \alpha_2 y(x_i - 2h) \\ &\quad - h[\beta_{-1} y'(x_i + h) + \beta_0 y'(x_i) + \beta_1 y'(x_i - h)]. \end{aligned}$$

取 $x_i = 0$, 令 $R[x^k] = 0$ ($k = 0, 1, 2, 3, 4$) 得

$$\begin{cases} 1 - \alpha_0 - \alpha_1 - \alpha_2 = 0, \\ h[1 + \alpha_1 + 2\alpha_2 - (\beta_{-1} + \beta_0 + \beta_1)] = 0, \\ h^2[1 - \alpha_1 - 4\alpha_2 - 2(\beta_{-1} - \beta_1)] = 0, \\ h^3[1 + \alpha_1 + 8\alpha_2 - 3(\beta_{-1} + \beta_1)] = 0, \\ h^4[1 - \alpha_1 - 16\alpha_2 - 4(\beta_{-1} - \beta_1)] = 0. \end{cases} \quad (9.1.26)$$

方程组 (9.1.26) 含有 5 个方程、6 个未知量, 故有一个未知量可以任意选取. 设 α_1 任意, 则可解得

$$\begin{aligned}\alpha_0 &= \frac{9}{8}(1 - \alpha_1), & \alpha_2 &= -\frac{1}{8}(1 - \alpha_1), & \beta_{-1} &= \frac{1}{24}(9 - \alpha_1), \\ \beta_0 &= \frac{1}{12}(9 + 7\alpha_1), & \beta_2 &= \frac{1}{24}(-9 + 17\alpha_1).\end{aligned}$$

(1) 取 $\alpha_1 = 1$, 则

$$\alpha_0 = 0, \quad \alpha_2 = 0, \quad \beta_{-1} = \frac{1}{3}, \quad \beta_0 = \frac{4}{3}, \quad \beta_1 = \frac{1}{3},$$

从而得辛普森公式

$$y_{i+1} = y_{i-1} + \frac{h}{3}(f_{i+1} + 4f_i + f_{i-1}).$$

(2) 取 $\alpha_1 = 0$, 则

$$\alpha_0 = \frac{9}{8}, \quad \alpha_2 = -\frac{1}{8}, \quad \beta_{-1} = \frac{3}{8}, \quad \beta_0 = \frac{6}{8}, \quad \beta_1 = -\frac{3}{8}.$$

于是汉明 (Hamming) 公式

$$y_{i+1} = \frac{1}{8}[9y_i - y_{i-2} + 3h(f_{i+1} + 2f_i - f_{i-1})]. \quad (9.1.27)$$

下面估计汉明公式的局部截断误差.

$$\begin{aligned}R[y] &= y(x_{i+1}) - y_{i+1} \\ &= y(x_{i+1}) - \frac{1}{8}\{9y(x_i) - y(x_{i-2}) + 3h[y'(x_{i+1}) + 2y'(x_i) - y'(x_{i-1})]\} \\ &= y(x_i + h) - \frac{1}{8}\{9y(x_i) - y(x_i - 2h) + 3h[y'(x_i + h) + 2y'(x_i) - y'(x_i - h)]\}.\end{aligned}$$

取 $x_i = 0$, 令 $y = x^5$, 则

$$R[y] = -3h^5 \neq 0,$$

所以汉明公式的代数精度 $m = 4$. 取

$$e(x) = \frac{1}{5!}y^{(5)}(\xi)(x - x_{i-1})^2(x - x_i)(x - x_{i+1})^2,$$

注意到

$$e(x_{i-1}) = e'(x_{i-1}) = e(x_i) = e(x_{i+1}) = e'(x_{i+1}) = 0,$$

根据广义佩亚诺定理,

$$\begin{aligned}
 R[y] &= R[e(x)] \\
 &= e(x_{i+1}) - \frac{1}{8}\{9e(x_i) - e(x_{i-2}) + 3h[e'(x_{i+1}) + 2e'(x_i) - e'(x_{i-1}))]\} \\
 &= \frac{1}{8}e(x_{i-2}) - \frac{3}{4}he'(x_i) \\
 &= \frac{1}{8} \times \frac{1}{5!}y^{(5)}(\xi)h^2(-2h)(-3h)^2 - \frac{3}{4} \times \frac{1}{5!}y^{(5)}(\xi)h^5 \\
 &= -\frac{h^5}{40}y^{(5)}(\xi).
 \end{aligned}$$

同样, 用待定系数法可以导出米尔恩 (Milne) 公式及其局部截断误差

$$y_{i+1} = y_{i-3} + \frac{4}{3}h(2f_i - f_{i-1} + 2f_{i-2}), \quad (9.1.28)$$

$$R[y] = \frac{14}{45}h^5y^{(5)}(\xi).$$

2. 预测—校正公式

在实际应用中, 常将显式公式与同阶的隐式公式联合使用, 用显式公式计算的结果作出 y_{i+1} 的估计 (称为预测), 将其代入隐式法的等式右端算出 y_{i+1} 的值 (称为校正), 这一对公式构成了预测—校正公式, 如式 (9.1.13). 通常, 将亚当斯显式公式与同阶的亚当斯隐式公式联合使用, 构成预测—校正公式, 例如, 取 $k=1$, 则得预测—校正公式

$$\begin{cases} y_{i+1}^{(0)} = y_i + \frac{h}{2}(3f_i - f_{i-1}), \\ y_{i+1} = y_i + \frac{h}{2}[f(x_{i+1}, y_{i+1}^{(0)}) + f_i]; \end{cases} \quad (9.1.29)$$

取 $k=2$, 则得预测—校正公式

$$\begin{cases} y_{i+1}^{(0)} = y_i + \frac{h}{12}(23f_i - 16f_{i-1} + 5f_{i-2}), \\ y_{i+1} = y_i + \frac{h}{12}[5f(x_{i+1}, y_{i+1}^{(0)}) + 8f_i - f_{i-1}]; \end{cases} \quad (9.1.30)$$

取 $k=3$, 则得预测—校正公式

$$\begin{cases} y_{i+1}^{(0)} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}), \\ y_{i+1} = y_i + \frac{h}{24}[9f(x_{i+1}, y_{i+1}^{(0)}) + 19f_i - 5f_{i-1} + f_{i-2}]. \end{cases} \quad (9.1.31)$$

又如, 用米尔恩公式 (9.1.28) 作预测, 用汉明公式 (9.1.27) 作校正, 构成最常用的一个预测—校正公式

$$\begin{cases} y_{i+1}^{(0)} = y_{i-3} + \frac{4}{3}h(2f_i - f_{i-1} + 2f_{i-2}), \\ y_{i+1} = \frac{1}{8}(9y_i - y_{i-2}) + \frac{3}{8}h[f(x_{i+1}, y_{i+1}^{(0)}) + 2f_i - f_{i-1}]. \end{cases} \quad (9.1.32)$$

3. 预测—修正—校正—修正公式

将式 (9.1.32) 改写为

$$\begin{cases} p_{i+1} = y_{i-3} + \frac{4}{3}h(2f_i - f_{i-1} + 2f_{i-2}), \\ y_{i+1} = \frac{1}{8}(9y_i - y_{i-2}) + \frac{3}{8}h[f(x_{i+1}, p_{i+1}) + 2f_i - f_{i-1}], \end{cases} \quad (9.1.33)$$

则有

$$y(x_{i+1}) - p_{i+1} = \frac{14}{45}h^5y^{(5)}(\xi), \quad (9.1.34)$$

$$y(x_{i+1}) - y_{i+1} = -\frac{1}{40}h^5y^{(5)}(\xi_1). \quad (9.1.35)$$

假定 $y^{(5)}(\xi) \approx y^{(5)}(\xi_1)$, 以上两式相减得

$$y_{i+1} - p_{i+1} \approx \frac{121}{360}h^5y^{(5)}(\xi).$$

由此可得

$$h^5y^{(5)}(\xi) \approx \frac{360}{121}(y_{i+1} - p_{i+1}), \quad h^5y^{(5)}(\xi_1) \approx \frac{360}{121}(y_{i+1} - p_{i+1}).$$

将以上两式分别代入式 (9.1.34), (9.1.35) 得

$$y(x_{i+1}) \approx p_{i+1} + \frac{112}{121}(y_{i+1} - p_{i+1}), \quad (9.1.36)$$

$$y(x_{i+1}) \approx y_{i+1} - \frac{9}{121}(y_{i+1} - p_{i+1}). \quad (9.1.37)$$

以上两式表明, 对预测值 p_{i+1} 和校正值 y_{i+1} 分别作以修正, 将会得到更好的近似值. 据此将式 (9.1.33) 作以修改, 则得修正汉明预测—校正公式

$$\begin{cases} p_{i+1} = y_{i-3} + \frac{4}{3}h(2f_i - f_{i-1} + 2f_{i-2}), \\ m_{i+1} = p_{i+1} + \frac{112}{121}(c_i - p_i), \\ c_{i+1} = \frac{1}{8}(9y_i - y_{i-2}) + \frac{3}{8}h[f(x_{i+1}, m_{i+1}) + 2f_i - f_{i-1}], \\ y_{i+1} = c_{i+1} - \frac{9}{121}(c_{i+1} - p_{i+1}), \end{cases} \quad (9.1.38)$$

其中 m_{i+1} 为 p_{i+1} 的修正值, c_{i+1} 为校正值, y_{i+1} 为 c_{i+1} 的修正值. 式 (9.1.38) 的第二式中本应是 $c_{i+1} - p_{i+1}$, 而 c_{i+1} 在下一式中计算, m_{i+1} 在 c_{i+1} 算出之前是无法计算的, 只好用 $c_i - p_i$ 代替 $c_{i+1} - p_{i+1}$. 因为式 (9.1.38) 中含有 y_{i-3} , 所以该式在 $i \geq 3$ 时才能使用. 当 $i = 3$ 时, 需要事先知道 y_0, y_1, y_2, y_3 . 由初始条件 y_0 已知, 其余三值可用同阶 (4 阶) 的单步法 (如标准 R-K 法) 求得, 算出相应的 f_1, f_2, f_3 , 留作后用. 当 $i = 3$ 时, $c_3 - p_3$ 尚未算出, 可令 $c_3 - p_3 = 0$.

同理, 也可用其他线性多步法 (如亚当斯法) 的显式公式与同阶隐式公式构造类似于修正汉明预测—校正公式的预测—修正—校正—修正公式. 这类方法的优点是计算函数值的次数少, 并且可以随时估计误差、调整步长. 如在式 (9.1.38) 中, 每步只需计算 f_i 和 $f(x_{i+1}, m_{i+1})$ 两个函数值, 而在同阶的标准 R-K 法中每步要计算 4 个函数值. 当区间 $[a, b]$ 较大, 计算步数很多时, 则更显示出这类方法的优越性.

由式 (9.1.38) 的最后一式 $y_{i+1} - c_{i+1} = -\frac{9}{121}(c_{i+1} - p_{i+1})$ 知, 可用 $c_{i+1} - p_{i+1}$ 来估计误差或调整步长. 若 $|c_{i+1} - p_{i+1}| \leq \varepsilon$, 则 $|y_{i+1} - c_{i+1}| < \varepsilon$, 即知 y_{i+1} 满足精度要求. 又若 $|c_{i+1} - p_{i+1}|$ 非常小, 则说明步长 h 还可以放大; 若 $|c_{i+1} - p_{i+1}|$ 比较大, 则说明步长 h 取得过大, 需要进一步缩小. 因此, 这类方法常用来求解函数值计算量大, 或者要求随时估计误差的问题.

9.2 数值解中误差的积累、数值方法的收敛性和绝对稳定性

常微分方程初值问题 (9.0.1) 的数值解 y_i ($i = 0, 1, 2, \dots, n$) 是通过逐步计算得到的, 在每一步计算中都存在着局部截断误差和舍入误差. 本节研究这两种误差的积累对数值解的影响以及方法的收敛性和稳定性. 为简单起见, 本节只对显式单步法进行讨论.

9.2.1 数值解中误差的积累和数值方法的收敛性

显式单步法可以写为如下的一般形式:

$$y_{i+1} = y_i + h\phi(x_i, y_i, h). \quad (9.2.1)$$

假设该法是 p 阶方法, 即局部截断误差 $R[y] = y(x_{i+1}) - y_{i+1} = O(h^{p+1})$, 则

$$y(x_{i+1}) = y(x_i) + h\phi(x_i, y(x_i), h) + R[y], \quad (9.2.2)$$

并假设 $\max_{a \leq x \leq b} |y^{(p+1)}(x)| \leq M_{p+1}$, 则

$$|R[y]| \leq cM_{p+1}h^{p+1}. \quad (9.2.3)$$

在实际计算中, 由于存在舍入误差, 按式 (9.2.1) 算出的 $\tilde{y}_{i+1}$ 不是理论上的数值解 y_{i+1} , 而是 y_{i+1} 的近似值, 它应满足

$$\tilde{y}_{i+1} = \tilde{y}_i + h\phi(x_i, \tilde{y}_i, h) + \eta_i, \quad (9.2.4)$$

其中 η_i 表示第 i 步计算所产生的舍入误差.

下面的定理给出单步法 (9.2.1) 的全程误差估计式.

定理 9.2.1 设求解常微分方程初值问题 (9.0.1) 的显式单步法如式 (9.2.1) 所示, 其中 $\phi(x, y, h)$ 在区域 $D = \{(x, y) | a \leq x \leq b, -\infty < y < \infty\}$ 上关于 y 满足利普希茨条件, 即存在常数 L , 使得

$$|\phi(x, y, h) - \phi(x, \bar{y}, h)| \leq \left| \frac{\partial \phi}{\partial y}(y - \bar{y}) \right| \leq L|y - \bar{y}| \quad (9.2.5)$$

对于任意的 $x \in [a, b]$ 及任意的 $y, \bar{y}$ 都成立, 则显式单步法 (9.2.1) 的全程误差 $e_i = y(x_i) - \tilde{y}_i$ ($i = 0, 1, 2, \dots, n$) 估计式如下:

$$|e_i| \leq \begin{cases} |e_0| + (x_i - a)(cM_{p+1}h^p + \eta/h), & L = 0, \\ |e_0|e^{L(x_i-a)} + (e^{L(x_i-a)} - 1) \left(\frac{cM_{p+1}h^p}{L} + \frac{\eta}{hL} \right), & L \neq 0, \end{cases} \quad (9.2.6)$$

其中 η 为 η_i 的上界.

证明 式 (9.2.2) 减去式 (9.2.4) 得

$$|y(x_{i+1}) - \tilde{y}_{i+1}| \leq |y(x_i) - \tilde{y}_i| + h|\phi(x_i, y(x_i), h) - \phi(x_i, \tilde{y}_i, h)| + |R[y]| + |\eta_i|. \quad (9.2.7)$$

注意到 $\phi(x, y, h)$ 在区域 D 上关于 y 满足利普希茨条件 (9.2.5), 并设 $|\eta_i| \leq \eta$. 由式 (9.2.5) 和 (9.2.7) 得

$$|e_{i+1}| \leq (1 + hL)|e_i| + |R[y]| + \eta = \alpha|e_i| + \beta,$$

其中 $\alpha = 1 + hL, \beta = |R[y]| + \eta$. 由上式可知

$$\begin{aligned} |e_{i+1}| &\leq \alpha|e_i| + \beta \leq \alpha(\alpha|e_{i-1}| + \beta) + \beta, \\ &= \alpha^2|e_{i-1}| + (1 + \alpha)\beta \\ &= \alpha^3|e_{i-2}| + (1 + \alpha + \alpha^2)\beta \leq \dots \\ &\leq \alpha^{i+1}|e_0| + (1 + \alpha + \alpha^2 + \dots + \alpha^i)\beta. \end{aligned}$$

于是有

$$|e_i| \leq \alpha^i|e_0| + (1 + \alpha + \alpha^2 + \dots + \alpha^{i-1})\beta. \quad (9.2.8)$$

(1) 当 $L = 0$ 时, $\alpha = 1$, 由式 (9.2.8), (9.2.3) 得

$$\begin{aligned} |e_i| &\leq |e_0| + i\beta = |e_0| + i(|R[y]| + \eta) \\ &\leq |e_0| + i(cM_{p+1}h^{p+1} + \eta) \\ &= |e_0| + ih(cM_{p+1}h^p + \eta/h) \\ &= |e_0| + (x_i - a)(cM_{p+1}h^p + \eta/h). \end{aligned} \quad (9.2.9)$$

(2) 当 $L \neq 0$ 时,

$$1 + \alpha + \alpha^2 + \cdots + \alpha^{i-1} = \frac{1 - \alpha^i}{1 - \alpha} = \frac{\alpha^i - 1}{hL},$$

而

$$\alpha^i = (1 + hL)^i = [(1 + hL)^{1/hL}]^{ihL} < e^{ihL},$$

由 (9.2.8) 得

$$\begin{aligned} |e_i| &< |e_0|e^{ihL} + \frac{e^{ihL} - 1}{hL}\beta \\ &\leq |e_0|e^{L(x_i-a)} + (e^{L(x_i-a)} - 1) \left(\frac{cM_{p+1}h^p}{L} + \frac{\eta}{hL} \right). \end{aligned} \quad (9.2.10)$$

由式 (9.2.9), (9.2.10) 即得定理结论. 证毕.

显然, 欧拉法 (9.1.2) 中的 $\phi(x, y, h) = f(x, y)$ 满足莱布尼茨条件, 它的局部截断误差 $R[y] = \frac{1}{2}y''(\xi_i)h^2$, 则 $|R[y]| \leq \frac{1}{2}M_2h^2$. 由定理 9.2.1 可知, 其全程误差估计式为

$$|e_i| \leq \begin{cases} |e_0| + (x_i - a) \left(\frac{1}{2}M_2h + \eta/h \right), & L = 0, \\ |e_0|e^{L(x_i-a)} + (e^{L(x_i-a)} - 1) \left(\frac{M_2h}{2L} + \frac{\eta}{hL} \right), & L \neq 0. \end{cases} \quad (9.2.11)$$

由定理 9.2.1 可得以下两个结论:

(1) 局部截断误差和舍入误差均对计算结果产生影响. 局部截断误差(含 $cM_{p+1}h^p$ 的项)随 h 的减小而减小, 而舍入误差(含 e_0 和 η 的项)随 h 的减小而增大. 因此, 为减少计算解 $\tilde{y}_i$ 的误差, h 应选择合适的, 既不能取得太大, 也不能取得太小.

(2) 全程误差与解法的阶数 p , 真解的 $p+1$ 阶导数 $y^{(p+1)}(x)$ 有关. 为保证计算解的精度, 一般应选取阶数 p 较大的解法, 但对于一些真解不够光滑, 即 $p+1$ 阶导数不存在或者虽然存在但上界很大的问题, 则应采用低阶解法.

下面讨论显式单步法的收敛性, 首先给出常微分方程数值解的收敛性定义.

定义 9.2.1 设初值问题 (9.0.1) 满足定理 9.0.1 的条件, 其准确解 (真解) 为 $y(x)$. 对于任一固定的 $x \in [a, b]$, 令 $h = \frac{x-a}{n}$, $x = a + ih$. 若由某一数值解法产生的近似解 y_i , 当 $h \rightarrow 0$ (即 $n \rightarrow \infty$) 时, 一致地有

$$\lim_{h \rightarrow 0} y_i = y(x),$$

则称该法为收敛的, 或者说, 数值解 y_i 一致地收敛于初值问题 (9.0.1) 的真解 $y(x)$ 在点 x_i 处的值 $y(x_i)$.

假定在计算中的舍入误差可以忽略不计, 即 $e_0 = 0, \eta = 0$, 则计算出的解等于理论数值解, 即 $\tilde{y}_i = y_i$. 此时, 由定理 9.2.1 可得显式单步法的收敛性定理.

定理 9.2.2 在定理 9.2.1 的条件下并假设计算中不存在舍入误差, 则显式单步法收敛.

证明 在定理 9.2.1 的结论中取 $e_0 = \eta = 0$, 并注意到 $x_i - a \leq b - a$, 由式 (9.2.6) 得

$$|y(x_i) - y_i| \leq \begin{cases} (b-a)cM_{p+1}h^p, & L = 0, \\ (e^{L(b-a)} - 1)\frac{cM_{p+1}h^p}{L}, & L \neq 0. \end{cases} \quad (9.2.12)$$

在式 (9.2.12) 中令 $h \rightarrow 0$, 即得数值解 y_i 一致地收敛于初值问题 (9.0.1) 的准确解 $y(x)$ 在点 x_i 处的值 $y(x_i)$. 证毕.

$y(x_i) - y_i$ 称为全程 (或整体) 截断误差. 由定理 9.2.2 可知, 若一个方法的局部截断误差为 $O(h^{p+1})$, 则它的全程截断误差为 $O(h^p)$. 由式 (9.2.12) 知, 欧拉法的全程截断误差为

$$|y(x_i) - y_i| \leq \begin{cases} \frac{1}{2}(b-a)M_2h, & L = 0, \\ (e^{L(b-a)} - 1)\frac{M_2h}{2L}, & L \neq 0. \end{cases}$$

显然, 它是收敛的.

下面利用定理 9.2.2 讨论标准 (经典) 的 4 级 4 阶 R-K 法的收敛性.

设函数 $f(x, y)$ 在区域 $D = \{(x, y) | a \leq x \leq b, -\infty < y < +\infty\}$ 上关于 y 满足利普希茨条件, 并且步长 h 满足 $0 \leq h \leq h_0$. 记

$$\begin{aligned} K_1(x, y, h) &= hf(x, y), \\ K_2(x, y, h) &= hf\left(x + \frac{h}{2}, y + \frac{1}{2}K_1(x, y, h)\right), \\ K_3(x, y, h) &= hf\left(x + \frac{h}{2}, y + \frac{1}{2}K_2(x, y, h)\right), \\ K_4(x, y, h) &= hf(x + h, y + K_3(x, y, h)). \end{aligned}$$

此时, 式 (9.1.20) 中的 $\phi(x, y, h)$ 可以写成

$$\phi(x, y, h) = \frac{1}{6h} [K_1(x, y, h) + 2K_2(x, y, h) + 2K_3(x, y, h) + K_4(x, y, h)].$$

显然, $K_1(x, y, h) = hf(x, y)$ 满足不等式

$$|K_1(x, y, h) - K_1(x, \tilde{y}, h)| \leq hL|y - \tilde{y}|.$$

利用上述不等式可得

$$\begin{aligned} |K_2(x, y, h) - K_2(x, \tilde{y}, h)| &\leq hL \left| y - \tilde{y} + \frac{1}{2}K_1(x, y, h) - \frac{1}{2}K_1(x, \tilde{y}, h) \right| \\ &\leq hL \left(1 + \frac{hL}{2} \right) |y - \tilde{y}|. \end{aligned}$$

同理, 容易得到下述不等式:

$$\begin{aligned} |K_3(x, y, h) - K_3(x, \tilde{y}, h)| &\leq hL \left[1 + \frac{hL}{2} + \frac{1}{4}(hL)^2 \right] |y - \tilde{y}|, \\ |K_4(x, y, h) - K_4(x, \tilde{y}, h)| &\leq hL \left[1 + hL + \frac{1}{2}(hL)^2 + \frac{1}{4}(hL)^3 \right] |y - \tilde{y}|. \end{aligned}$$

于是关于有

$$\begin{aligned} |\phi(x, y, h) - \phi(x, \tilde{y}, h)| &\leq L \left(1 + \frac{1}{2}hL + \frac{1}{6}(hL)^2 + \frac{1}{24}(hL)^3 \right) |y - \tilde{y}| \\ &\leq L \left(1 + \frac{1}{2}h_0L + \frac{1}{6}(h_0L)^2 + \frac{1}{24}(h_0L)^3 \right) |y - \tilde{y}|. \end{aligned}$$

因此, $\phi(x, y, h)$ 关于 y 满足利普希茨条件, 根据定理 9.2.2 知, 标准 (经典) 的 4 级 4 阶 R-K 法是收敛的.

9.2.2 绝对稳定性

前面介绍的算法的收敛性是在假定数值计算过程中的每一步计算都是准确的, 即在不舍入误差的前提下讨论的. 在实际计算中, 由于计算机的字长有限等原因, 每一步计算不可避免地存在舍入误差. 如果每一步计算中的舍入误差不能得到控制, 则必然会形成舍入误差的积累. 这种误差的积累甚至于“淹没”了真解, 致使计算出的 $y_1, y_2, \dots$ 不是初值问题的近似解. 因此, 必须研究舍入误差的控制问题. 数值稳定性是研究数值方法在计算中各步产生的舍入误差是否能得到控制的问题. 若某个算法在计算中的每一步产生的舍入误差在以后各步计算中都能逐步削弱, 则称这个算法为绝对稳定的方法. 下面给出算法绝对稳定性的定义.

定义 9.2.2 若某个数值方法在计算 y_i 时有误差 e_i , 如果在计算以后各步 y_j ($j > i$) 时, 由 e_i 引起的误差 e_j 都满足

$$|e_j| < |e_i|, \quad j > i,$$

则称该算法为绝对稳定的方法.

由于初值问题的计算过程与方程中的 $f(x, y)$ 有关, 这给研究方法的绝对稳定性带来困难. 数值稳定性是数值方法本身的特性, 为简单起见, 通常考察数值方法

用于求解下面满足利普希茨条件的典型常微分方程 (称为试验方程)

$$y' = \lambda y, \quad \operatorname{Re}(\lambda) < 0 \quad (9.2.13)$$

(其中 λ 为常数, 可以为复数) 时的绝对稳定性.

若用某种方法求解方程 (9.2.13), 取步长为 h , 如果计算开始时产生的误差在以后的计算中能逐步削弱, 则称这个解法相对于 $\bar{h} = \lambda h$ 为绝对稳定的, $\bar{h}$ 的全体称为绝对稳定区域. 它与实轴的交称为绝对稳定区间. 若一个算法绝对稳定区域越大, 则称这个算法的绝对稳定性越好.

下面首先考察欧拉法的稳定性.

用欧拉法 (9.1.2) 求解方程 (9.2.13), 数值解

$$y_{i+1} = y_i + hf(x_i, y_i) = y_i + \lambda h y_i.$$

当 y_i 因计算中的舍入误差变为 $\tilde{y}_i$ 时, y_{i+1} 变为 $\tilde{y}_{i+1}$, 则

$$\tilde{y}_{i+1} = \tilde{y}_i + \lambda h \tilde{y}_i.$$

以上两式相减, 令 $e_i = y_i - \tilde{y}_i$ 得

$$e_{i+1} = e_i + \lambda h e_i = (1 + \bar{h})e_i.$$

由此可见, 欲使欧拉法稳定, e_i 应逐步削弱, 即 $|e_{i+1}| < |e_i|$. 这时, $\bar{h}$ 应满足

$$|1 + \bar{h}| < 1. \quad (9.2.14)$$

由于 $\bar{h}$ 可能是复数, 所以在 $\bar{h}$ 的复平面上, 式 (9.2.14) 表示以实轴上点 -1 为圆心, 半径为 1 的单位圆的内部, 所以欧拉法的绝对稳定区域是圆域, 如图 9.1 所示, 其绝对稳定区间为 $-2 < \lambda h < 0$.

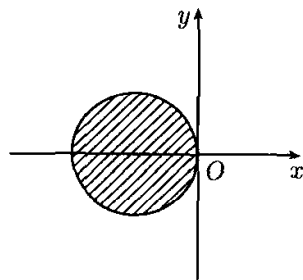


图 9.1 欧拉法的稳定区域

其次考察后退欧拉法, 用后退欧拉法 (9.1.4) 求解方程 (9.2.13), 则有

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}) = y_i + \lambda h y_{i+1}.$$

与欧拉法同理可得

$$e_{i+1} = e_i + \lambda h e_{i+1} = e_i + \bar{h} e_{i+1}.$$

由此可得 $e_{i+1} = \frac{e_i}{1 - \bar{h}}$, 其绝对稳定区域是 $\left| \frac{1}{1 - \bar{h}} \right| < 1$, 即

$$|1 - \bar{h}| > 1. \quad (9.2.15)$$

这是复平面上以实轴上点 1 为圆心, 半径为 1 的单位圆的外部, 如图 9.2 所示.

由图 9.1 和图 9.2 可见, 后退欧拉法的局部稳定区域比欧拉法的大得多. 一般地, 隐式法的绝对稳定区域 (性) 都比同阶的显式法大 (好).

最后讨论标准 4 级 4 阶 R-K 法 (9.1.20) 的绝对稳定区域. 对于方程 (9.2.13), 由式 (9.1.20) 可以推出

$$y_{i+1} = \left(1 + \bar{h} + \frac{\bar{h}^2}{2} + \frac{\bar{h}^3}{6} + \frac{\bar{h}^4}{24} \right) y_i.$$

若计算 y_0 时有舍入误差 $e_0 = y_0 - \tilde{y}_0$, 而在以后的计算中无误差存在, 则在 y_{i+1} 处引起的误差为

$$e_{i+1} = \left(1 + \bar{h} + \frac{\bar{h}^2}{2} + \frac{\bar{h}^3}{6} + \frac{\bar{h}^4}{24} \right) e_i.$$

因此, 在复平面上, 绝对稳定区域由不等式

$$\left| 1 + \bar{h} + \frac{\bar{h}^2}{2} + \frac{\bar{h}^3}{6} + \frac{\bar{h}^4}{24} \right| < 1 \quad (9.2.16)$$

所确定, 其绝对稳定区域如图 9.3 所示, 其绝对稳定区间为 $-2.78 < \lambda h < 0$, 即 $0 < h < -2.78/\lambda$.

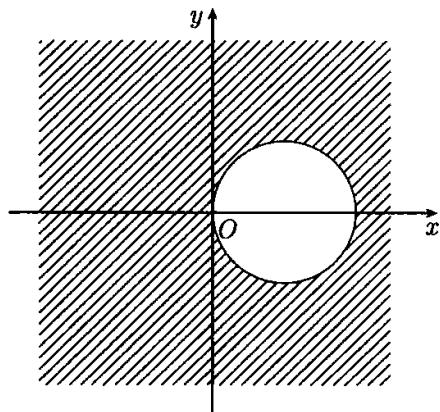


图 9.2 后退欧拉法的稳定区域

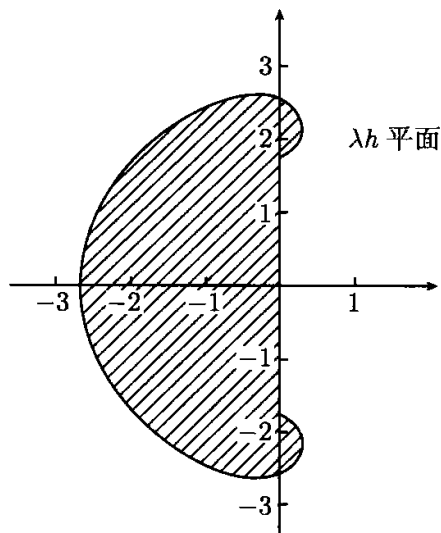


图 9.3 经典 R-K 法的稳定区域

从以上讨论可以看到, 算法的稳定性与步长 h 密切相关, 某个算法在一种步长下是稳定的计算方法, 取大一点的步长就可能不稳定. 在计算时应选取步长 h , 使得 $\bar{h} = \lambda h$ 在绝对稳定区域以内, 只有既收敛又稳定的方法才能在实际计算中使用.

由于多步法的绝对稳定域比较复杂, 在此不予以讨论, 但给出两个不稳定的方法的例子. 通过较为复杂的推导可知, 中点法和辛普森法的绝对稳定区域不存在, 即它们都不是绝对稳定的方法, 故不能使用.

9.3 一阶微分方程组与高阶方程的数值解法

9.3.1 一阶微分方程组

在科学研究和工程技术领域常常会遇到求解如下的一阶常微分方程组的初值问题:

$$\begin{cases} y_1'(x) = f_1(x, y_1(x), y_2(x), \dots, y_n(x)), \\ y_2'(x) = f_2(x, y_1(x), y_2(x), \dots, y_n(x)), \\ \dots\dots\dots \\ y_n'(x) = f_n(x, y_1(x), y_2(x), \dots, y_n(x)), \\ y_1(x_0) = y_{10}, y_2(x_0) = y_{20}, \dots, y_n(x_0) = y_{n0}, \end{cases} \quad (9.3.1)$$

其中 $a \leq x \leq b$.

若记

$$y(x) = (y_1(x), y_2(x), \dots, y_n(x))^T, \quad y_0 = (y_{10}, y_{20}, \dots, y_{n0})^T, \\ f(x, y) = (f_1(x, y), f_2(x, y), \dots, f_n(x, y))^T,$$

则微分方程组 (9.3.1) 可以写成向量形式

$$\begin{cases} y'(x) = f(x, y(x)), & a \leq x \leq b, \\ y(a) = y_0. \end{cases} \quad (9.3.1')$$

微分方程组初值问题 (9.3.1) 的解存在唯一性定理如下:

定理 9.3.1 设 $f(x, y)$ 在 $n+1$ 维区域 $D = \{(x, y) | a \leq x \leq b, -\infty < y_i < +\infty, i = 1, 2, \dots, n\}$ 上连续, 关于 y 满足利普希茨条件, 即存在常数 L , 使得

$$\|f(x, y) - f(x, \bar{y})\| \leq L\|y - \bar{y}\|$$

对于任意的 $x \in [a, b]$ 及任意的 $y, \bar{y}$ 都成立, 则微分方程组 (9.3.1') 存在唯一的连续解 $y = y(x)$.

在以后的讨论中, 总假定微分方程组 (9.3.1) 中的 $f(x, y)$ 满足定理 9.3.1 的条件.

微分方程组初值问题 (9.3.1') 在形式上与单个微分方程初值问题 (9.0.1) 完全相同, 只是数量函数在此变成了向量函数. 因此, 9.1 节建立的求解单个一阶微分方程初值问题的数值方法, 可以完全平移到求解一阶微分方程组的初值问题中, 只不

过是将单个方程中的函数换为向量函数即可. 例如, 标准 4 级 4 阶 R-K 法 (9.1.20) 的向量形式如下:

$$\begin{cases} y_{i+1} = y_i + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4), \\ K_1 = hf(x_i, y_i), \\ K_2 = hf\left(x_i + \frac{h}{2}, y_i + \frac{1}{2}K_1\right), \\ K_3 = hf\left(x_i + \frac{h}{2}, y_i + \frac{1}{2}K_2\right), \\ K_4 = hf(x_i + h, y_i + K_3). \end{cases} \quad (9.3.2)$$

其分量形式为

$$\begin{cases} y_{j,i+1} = y_{i,j} + \frac{1}{6}(K_{j1} + 2K_{j2} + 2K_{j3} + K_{j4}), \\ K_{j1} = hf_j(x_i; y_{1i}, y_{2i}, \dots, y_{ni}), \\ K_{j2} = hf_j\left(x_i + \frac{h}{2}; y_{1i} + \frac{K_{11}}{2}, y_{2i} + \frac{K_{21}}{2}, \dots, y_{ni} + \frac{K_{n1}}{2}\right), \quad j = 1, 2, \dots, n. \\ K_{j3} = hf_j\left(x_i + \frac{h}{2}; y_{1i} + \frac{K_{12}}{2}, y_{2i} + \frac{K_{22}}{2}, \dots, y_{ni} + \frac{K_{n2}}{2}\right), \\ K_{j4} = hf_j(x_i + h; y_{1i} + K_{13}, y_{2i} + K_{23}, \dots, y_{ni} + K_{n3}). \end{cases}$$

汉明修正预测—校正公式 (9.1.38) 的向量形式为

$$\begin{cases} p_{i+1} = y_{i-3} + \frac{4}{3}h(2f_i - f_{i-1} + 2f_{i-2}), \\ m_{i+1} = p_{i+1} + \frac{112}{121}(c_i - p_i), \\ c_{i+1} = \frac{1}{8}(9y_i - y_{i-2}) + \frac{3}{8}h[f(x_{i+1}, m_{i+1}) + 2f_i - f_{i-1}], \\ y_{i+1} = c_{i+1} - \frac{9}{121}(c_{i+1} - p_{i+1}). \end{cases} \quad (9.3.3)$$

读者不难写出式 (9.3.3) 的分量形式.

在化学反应、自动控制、电力系统等领域常会遇到一类病态方程组, 通常称之为刚性方程组或 stiff 方程组. 一般地, 设线性微分方程组初值问题为

$$\begin{cases} y' = Ay + b(x), \\ y(a) = y_0, \end{cases} \quad (9.3.4)$$

其中 A 为 n 阶矩阵, A 的特征值为 $\lambda_i (i = 1, 2, \dots, n)$ 且 $\operatorname{Re}(\lambda_i) < 0$. 记

$$s = \max_{1 \leq i \leq n} |\operatorname{Re}(\lambda_i)| / \min_{1 \leq i \leq n} |\operatorname{Re}(\lambda_i)|,$$

则 s 称为微分方程组 (9.3.4) 的刚性比. 若 $s \gg 1$, A 称为病态矩阵, 故方程组 (9.3.4) 为刚性方程组. 对于一般的非线性方程组 (9.3.1'), 将向量函数 $f = (f_1, f_2, \dots, f_n)^T$ 关于变量 $y = (y_1, y_2, \dots, y_n)^T$ 的雅可比矩阵为

$$J_f(y) = \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} & \cdots & \frac{\partial f_1}{\partial y_n} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \cdots & \frac{\partial f_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial y_1} & \frac{\partial f_n}{\partial y_2} & \cdots & \frac{\partial f_n}{\partial y_n} \end{pmatrix}.$$

设其特征值为 $\lambda_i(x) (i = 1, 2, \dots, n)$, 并且 $\operatorname{Re}(\lambda_i(x)) < 0$, 这时的刚性方程和刚性比的定义同前. 通常 $s \geq 10$ 就认为方程组是刚性的, s 越大, 方程组病态就越严重. 若用通常的方法求解刚性方程组, 则误差的积累往往会淹没真解, 故应采取绝对稳定性较好的方法. 这些方法都有现成的数学软件可供使用, 限于篇幅, 此处不作介绍 (刚性方程组的数值解法一般进行专题讨论). 有兴趣的读者可参见有关文献, 如文献 [9], [19], [27].

9.3.2 高阶常微分方程

m 阶常微分方程初值问题为

$$\begin{cases} y^{(m)} = f(x, y, y', y'', \dots, y^{(m-1)}), & a \leq x \leq b, \\ y(a) = y_0, y'(a) = y'_0, \dots, y^{(m-1)}(a) = y_0^{(m-1)}. \end{cases} \quad (9.3.5)$$

求解高阶微分方程初值问题是将其转化为一阶微分方程组来求解. 为此, 引进新的变量 $y_1 = y, y_2 = y', \dots, y_m = y^{(m-1)}$, 即可将 m 阶微分方程 (9.3.5) 转化为如下的一阶微分方程组:

$$\begin{cases} y'_1 = y_2, \\ y'_2 = y_3, \\ \dots\dots\dots \\ y'_{m-1} = y_m, \\ y'_m = f(x, y_1, y_2, \dots, y_m), \\ y_1(a) = y_0, y_2(a) = y'_0, \dots, y_m(a) = y_0^{(m-1)}. \end{cases} \quad (9.3.6)$$

例 9.3.1 取 $h = 0.1$, 用标准 4 级 4 阶 R-K 法求解二阶初值问题

$$\begin{cases} y'' - 2y' + 2y = e^{2x} \sin x, & 0 \leq x \leq 1, \\ y(0) = -0.4, & y'(0) = -0.6. \end{cases}$$

解 令 $y_1 = y$, $y_2 = y'$, 则得

$$\begin{cases} y'_1 = y_2, \\ y'_2 = e^{2x} \sin x - 2y_1 + 2y_2, \\ y_1(0) = -0.4, & y_2(0) = -0.6. \end{cases}$$

记

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad y(0) = \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} -0.4 \\ -0.6 \end{pmatrix},$$

$$f(x, y) = \begin{pmatrix} f_1(x, y) \\ f_2(x, y) \end{pmatrix} = \begin{pmatrix} y_2 \\ e^{2x} \sin x - 2y_1 + 2y_2 \end{pmatrix}.$$

则标准 4 级 4 阶 R-K 法计算公式为

$$y_{i+1} = y_i + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4),$$

其中

$$K_1 = hf(x_i, y_i) = \begin{pmatrix} K_{11} \\ K_{12} \end{pmatrix} = \begin{pmatrix} hy_{2i} \\ h(e^{2x_i} \sin x_i - 2y_{1i} + 2y_{2i}) \end{pmatrix},$$

$$\begin{aligned} K_2 &= hf\left(x_i + \frac{h}{2}, y_i + \frac{K_1}{2}\right) = \begin{pmatrix} K_{21} \\ K_{22} \end{pmatrix} \\ &= \begin{pmatrix} h(y_{2i} + K_{12}/2) \\ h(e^{2(x_i+h/2)} \sin(x_i + h/2) - 2(y_{1i} + K_{11}/2) + 2(y_{2i} + K_{12}/2)) \end{pmatrix}, \end{aligned}$$

$$\begin{aligned} K_3 &= hf\left(x_i + \frac{h}{2}, y_i + \frac{K_2}{2}\right) = \begin{pmatrix} K_{31} \\ K_{32} \end{pmatrix} \\ &= \begin{pmatrix} h(y_{2i} + K_{22}/2) \\ h(e^{2(x_i+h/2)} \sin(x_i + h/2) - 2(y_{1i} + K_{21}/2) + 2(y_{2i} + K_{22}/2)) \end{pmatrix}, \end{aligned}$$

$$\begin{aligned}
 K_4 &= hf(x_i + h, y_i + K_3) = \begin{pmatrix} K_{14} \\ K_{24} \end{pmatrix} \\
 &= \begin{pmatrix} h(y_{2i} + K_{23}) \\ h(e^{2(x_i+h)} \sin(x_i + h) - 2(y_{1i} + K_{13}) + 2(y_{2i} + K_{23})) \end{pmatrix}.
 \end{aligned}$$

读者不难按以上公式计算出方程在点 $x_i = ih$ ($i = 1, 2, \dots, 10$) 处的近似解 y_i , 并与方程的解析解 $y(x) = 0.2e^{2x}(\sin x - 2\cos x)$ 作以比较.

9.4 边值问题的数值解法

本节介绍两种求解二阶常微分方程

$$y'' = f(x, y, y'), \quad a < x < b \quad (9.4.1)$$

边值问题的数值解法——差分法和打靶法, 其边值条件可分为如下三类:

第一边值条件:

$$y(a) = \alpha, \quad y(b) = \beta;$$

第二边值条件:

$$y'(a) = \alpha, \quad y'(b) = \beta;$$

第三边值条件:

$$y'(a) - \alpha_0 y(a) = \alpha_1, \quad y'(b) + \beta_0 y(b) = \beta_1.$$

其中 $\alpha_0 \geq 0$, $\beta_0 \geq 0$, $\alpha_0 + \beta_0 > 0$.

微分方程 (9.4.1) 分别结合第一、第二、第三边值条件, 则称其为第一、第二、第三边值问题.

9.4.1 有限差分法

有限差分法是求解常微分方程边值问题的一种基本数值方法, 该法用数值微分公式代替微分方程和边界条件中的导数, 略去误差项, 把微分方程离散化成一个差分方程组. 求解这个差分方程组, 将差分方程组的解作为微分方程边值问题的近似解. 这种解称为差分解.

1. 差分方程的建立与求解

下面介绍求解一个典型的线性常微分方程

$$y''(x) - q(x)y(x) = f(x), \quad a < x < b$$

边值问题的差分法. 首先介绍求解第一边值问题

$$\begin{cases} y''(x) - q(x)y(x) = f(x), & a < x < b, \\ y(a) = \alpha, & y(b) = \beta, \end{cases} \quad (9.4.2)$$

其中 $q(x), f(x)$ 为已知函数且 $q(x) \geq 0, \alpha, \beta$ 为已知常数.

为建立差分方程组, 先将微分方程离散化. 为此, 将求解区间 $[a, b]$ 分成 n 等份, 即取步长 $h = (b - a)/n$, 令节点 $x_i = a + ih$ ($i = 0, 1, 2, \dots, n$). 在内节点 x_i ($i = 1, 2, \dots, n-1$) 处, 微分方程为

$$y''(x_i) - q(x_i)y(x_i) = f(x_i), \quad i = 1, 2, \dots, n-1. \quad (9.4.3)$$

代入数值微分公式

$$y''(x_i) = \frac{y(x_{i-1}) - 2y(x_i) + y(x_{i+1}))}{h^2} - \frac{h^2}{12}y^{(4)}(\xi_i) \quad (9.4.4)$$

得

$$\frac{y(x_{i-1}) - 2y(x_i) + y(x_{i+1}))}{h^2} - q(x_i)y(x_i) = f(x_i) + \frac{h^2}{12}y^{(4)}(\xi_i), \quad i = 1, 2, \dots, n-1. \quad (9.4.5)$$

略去误差项, 并记 $q_i = q(x_i), f_i = f(x_i)$, 令 $\{y_i\}$ 满足方程

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - q_i y_i = f_i, \quad i = 1, 2, \dots, n-1, \quad (9.4.6)$$

即

$$y_{i-1} - (2 + h^2 q_i)y_i + y_{i+1} = h^2 f_i, \quad i = 1, 2, \dots, n-1. \quad (9.4.6')$$

方程组 (9.4.6') 称为差分方程组, 略去的项

$$R[x_i] = \frac{h^2}{12}y^{(4)}(\xi_i)$$

称为差分方程逼近微分方程的截断误差.

式 (9.4.6') 是一个含有 $n+1$ 个未知量 y_i ($i = 0, 1, \dots, n$), 而只有 $n-1$ 个方程的线性方程组. 要使该方程组有唯一解, 需要根据边值条件, 或者减少两个未知量, 或者增补两个方程.

对于第一边值问题 (9.4.2), 由于边值条件 $y_0 = \alpha, y_n = \beta$ 已知, 则其差分方程组为

$$\begin{cases} -(2 + h^2 q_1)y_1 + y_2 = h^2 f_1 - \alpha, \\ y_{i-1} - (2 + h^2 q_i)y_i + y_{i+1} = h^2 f_i, & i = 2, 3, \dots, n-2, \\ y_{n-2} - (2 + h^2 q_{n-1})y_{n-1} = h^2 f_{n-1} - \beta. \end{cases} \quad (9.4.7)$$

注意到 $q_i \geq 0$ ($q(x) \geq 0$), 这是一个严格主对角占优的三对角方程组, 其解存在且唯一, 用追赶法解之.

对于第二边值问题, 为保证略去的截断误差为 $O(h^2)$, 在边界处的一阶导数用三点数值微分公式代替, 即

$$\begin{cases} y'(x_0) = \frac{-3y(x_0) + 4y(x_1) - y(x_2)}{2h} + \frac{h^2}{3}y'''(\xi_0), \\ y'(x_n) = \frac{y(x_{n-2}) - 4y(x_{n-1}) + 3y(x_n)}{2h} + \frac{h^2}{3}y'''(\xi_n). \end{cases} \quad (9.4.8)$$

略去上式中的误差并与方程组 (9.4.6') 联立, 可得第二边值问题的差分方程组

$$\begin{cases} -3y_0 + 4y_1 - y_2 = 2h\alpha, \\ y_{i-1} - (2 + h^2q_i)y_i + y_{i+1} = h^2f_i, \quad i = 1, 2, \dots, n-1, \\ y_{n-2} - 4y_{n-1} + 3y_n = 2h\beta. \end{cases}$$

消去以上方程中的 y_0 和 y_n 得

$$\begin{cases} -\left(\frac{2}{3} + h^2q_1\right)y_1 + \frac{2}{3}y_2 = h^2f_1 + \frac{2}{3}h\alpha, \\ y_{i-1} - (2 + h^2q_i)y_i + y_{i+1} = h^2f_i, \quad i = 2, \dots, n-2, \\ \frac{2}{3}y_{n-2} - \left(\frac{2}{3} + h^2q_{n-1}\right)y_{n-1} = h^2f_{n-1} - \frac{2}{3}h\beta. \end{cases} \quad (9.4.9)$$

显然, 这也是一个严格主对角占优的三对角方程组, 其解存在且唯一, 用追赶法解之.

对于第三边值问题, 用与第二边值问题同样的方法可以建立其差分方程组为

$$\begin{cases} -(3 + 2h\alpha_0)y_0 + 4y_1 - y_2 = 2h\alpha_1, \\ y_{i-1} - (2 + h^2q_i)y_i + y_{i+1} = h^2f_i, \quad i = 1, 2, \dots, n-1, \\ y_{n-2} - 4y_{n-1} + (3 + 2h\beta_0)y_n = 2h\beta_1. \end{cases} \quad (9.4.10)$$

可以证明, 这个差分方程组的解存在且唯一. 但用追赶法求解这个线性方程组时, 不能保证计算过程中的数值稳定性. 为了使离散化后得到的差分方程组求解是数值稳定的, 采用一种新的节点划分方法, 即令

$$h = \frac{b-a}{n}, \quad x_i = a + \left(i - \frac{1}{2}\right)h, \quad i = 0, 1, \dots, n+1.$$

在内点 x_i ($i = 1, 2, \dots, n$) 上的差分方程仍然为 (9.4.6'), 而对在边界点 a, b 处的 $y'(a)$, $y'(b)$ 和 $y(a)$, $y(b)$ 分别利用公式

$$\begin{cases} y'(x) = \frac{y(x+h) - y(x-h)}{2h} - \frac{h^2}{6} y'''(\xi), \\ y(x) = \frac{y(x-h) + y(x+h)}{2} - \frac{h^2}{2} y''(\xi). \end{cases}$$

近似计算. 于是由第三边值条件得两个方程

$$\frac{y_1 - y_0}{h} - \alpha_0 \frac{y_0 + y_1}{2} = \alpha_1, \quad \frac{y_{n+1} - y_n}{h} + \beta_0 \frac{y_n + y_{n+1}}{2} = \beta_1,$$

即

$$\begin{aligned} -\frac{1 + h\alpha_0/2}{1 - h\alpha_0/2} y_0 + y_1 &= \frac{h\alpha_1}{1 - h\alpha_0/2}, \\ y_n - \frac{1 + h\beta_0/2}{1 - h\beta_0/2} y_{n+1} &= -\frac{h\beta_1}{1 - h\beta_0/2}. \end{aligned}$$

将以上两个方程与方程 (9.4.6') 联立, 则得第三边值条件的另一种差分方程组

$$\begin{cases} -\frac{1 + h\alpha_0/2}{1 - h\alpha_0/2} y_0 + y_1 = \frac{h\alpha_1}{1 - h\alpha_0/2}, \\ y_{i-1} - (2 + h^2 q_i) y_i + y_{i+1} = h^2 f_i, \quad i = 1, 2, \dots, n, \\ y_n - \frac{1 + h\beta_0/2}{1 - h\beta_0/2} y_{n+1} = -\frac{h\beta_1}{1 - h\beta_0/2}. \end{cases} \quad (9.4.11)$$

这是一个严格对角占优的三对角方程组, 可用追赶法求解.

对于更一般的线性常微分方程边值问题

$$\begin{cases} y''(x) + p(x)y'(x) + q(x)y(x) = f(x), & a < x < b, \\ \alpha_0 y'(a) + \alpha_1 y(a) = \alpha_2, \\ \beta_0 y'(b) + \beta_1 y(b) = \beta_2 \end{cases}$$

(其中 $|\alpha_0| + |\alpha_1| \neq 0, |\beta_0| + |\beta_1| \neq 0$) 的离散化, 令 $h = \frac{b-a}{n}, x_i = a + ih$ ($i = 0, 1, 2, \dots, n$). 为保证略去的截断误差为 $O(h^2)$, 二阶导数用式 (9.4.4), 在边界处的一阶导数用 (9.4.8), 方程中的一阶导数用

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_{i-1}))}{2h} - \frac{h^2}{6} y'''(\xi_i)$$

代替, 略去误差项后得到差分方程组

$$\begin{cases} \frac{\alpha_0(-3y_0 + 4y_1 - y_2)}{2h} + \alpha_1 y_0 = \alpha_2, \\ \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + \frac{p_i(y_{i+1} - y_{i-1})}{2h} + q_i y_i = f_i, \quad i = 1, 2, \dots, n-1, \\ \frac{\beta_0(y_{n-2} - 4y_{n-1} + 3y_n)}{2h} + \beta_1 y_n = \beta_2, \end{cases}$$

解此方程组可得差分解 y_i ($i = 0, 1, \dots, n$).

例 9.4.1 取 $h = 0.25$, 用差分法解

$$\begin{cases} y'' - y = -x, & 0 < x < 1, \\ y(0) = 0, \quad y(1) = 0. \end{cases}$$

解

$$n = \frac{1}{h} = 4, \quad x_i = ih = 0.25i, \quad i = 0, 1, 2, 3, 4.$$

据差分方程 (9.4.7) 有

$$\begin{cases} -2.0625y_1 + y_2 = -0.015625, \\ y_1 - 2.0625y_2 + y_3 = -0.03125, \\ y_2 - 2.0625y_3 = -0.046875. \end{cases}$$

解之得

$$y_1 = 0.0348852, \quad y_2 = 0.0563258, \quad y_3 = 0.0500365.$$

解毕.

注意: 例 9.4.1 的真解是

$$y(x) = x - \frac{e^x - e^{-x}}{e - e^{-1}},$$

$y(x)$ 在节点 x_1, x_2, x_3 处的值分别为

$$0.0350476, \quad 0.0565908, \quad 0.0502758.$$

可见, 差分解准确到 3 位小数.

2. 差分解的误差估计与收敛性

为简单起见, 本段只讨论第一边值问题 (9.4.2) 的差分解的收敛性. 为此, 先介绍一个引理及其推论.

引理 9.4.1 (极值原理) 设 y_i ($i = 0, 1, \dots, n$) 是一组不全相等的数, 并且满足 $l(y_i) \triangleq \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - q_i y_i \geq 0$ ($i = 1, 2, \dots, n-1$), 其中 $q_i \geq 0$, 则非负 y_i 的最大值必为 y_0 或 y_n , 即 $\max_{y_i \geq 0} \{y_i\} = \max\{y_0, y_n\}$.

证明 用反证法. 假定存在 $k \in \{1, 2, \dots, n-1\}$, 使得 $y_k = \max_{y_i \geq 0} \{y_i\} = M > y_{k-1}$ 或 y_{k+1} , 则

$$l(y_k) = \frac{y_{k-1} - 2M + y_{k+1}}{h^2} - q_k M < \frac{M - 2M + M}{h^2} - q_k M \leq 0.$$

这与 $l(y_i) \geq 0$ 矛盾, 故结论成立. 证毕.

推论 9.4.1 设两组数 $\{y_i\}, \{z_i\}$ ($i = 0, 1, \dots, n$) 满足以下关系:

$$|l(y_i)| \leq -l(z_i), \quad i = 1, 2, \dots, n-1,$$

$$|y_0| \leq z_0, \quad |y_n| \leq z_n,$$

则 $|y_i| \leq z_i$ ($i = 0, 1, \dots, n$).

证明 已知条件可以改写为

$$l(z_i) \leq l(y_i) \leq -l(z_i), \quad i = 1, 2, \dots, n-1,$$

$$-z_0 \leq y_0 \leq z_0, \quad -z_n \leq y_n \leq z_n.$$

由此可知

$$\begin{cases} l(y_i - z_i) = l(y_i) - l(z_i) \geq 0, & i = 1, 2, \dots, n-1, \\ y_0 - z_0 \leq 0, \quad y_n - z_n \leq 0, \end{cases}$$

$$\begin{cases} l(-y_i - z_i) = -l(y_i) - l(z_i) \geq 0, & i = 1, 2, \dots, n-1, \\ -y_0 - z_0 \leq 0, \quad -y_n - z_n \leq 0. \end{cases}$$

根据极值原理知

$$\max_{y_i - z_i \geq 0} \{y_i - z_i\} = \max\{y_0 - z_0, y_n - z_n\} \leq 0,$$

$$\max_{-y_i - z_i \geq 0} \{-y_i - z_i\} = \max\{-y_0 - z_0, -y_n - z_n\} \leq 0,$$

故有

$$y_i \leq z_i, \quad -z_i \leq y_i, \quad i = 0, 1, \dots, n,$$

即

$$|y_i| \leq z_i, \quad i = 0, 1, \dots, n.$$

证毕.

定理 9.4.1 设 $y(x)$ 是第一边值问题 (9.4.2) 的解, 并且 $y(x)$ 在区间 $[a, b]$ 上 4 阶连续可微, y_i ($i = 0, 1, \dots, n$) 是差分方程 (9.4.7) 的解, 则有误差估计式

$$|y(x_i) - y_i| \leq \frac{h^2}{24} M_4 (x_i - a)(b - x_i) \leq \frac{h^2}{96} M_4 (b - a)^2,$$

其中 $M_4 = \max_{a \leq x \leq b} |y^{(4)}(x)|$.

证明 记 $\varepsilon_i = y(x_i) - y_i$, 则 $\varepsilon_0 = \varepsilon_n = 0$. 式 (9.4.5) 减式 (9.4.6) 得

$$\frac{\varepsilon_{i-1} - 2\varepsilon_i + \varepsilon_{i+1}}{h^2} - q_i \varepsilon_i = \frac{h^2}{12} y^{(4)}(\xi_i), \quad i = 1, 2, \dots, n-1,$$

即

$$l(\varepsilon_i) = \frac{h^2}{12} y^{(4)}(\xi_i), \quad i = 1, 2, \dots, n-1.$$

于是有

$$\begin{cases} l(\varepsilon_i) = \frac{h^2}{12} y^{(4)}(\xi_i), & i = 1, 2, \dots, n-1, \\ \varepsilon_0 = 0, \varepsilon_n = 0, \end{cases}$$

令数 ρ_i ($i = 0, 1, \dots, n$) 满足差分方程

$$\begin{cases} l(\rho_i) = -\frac{h^2}{12} M_4, & i = 1, 2, \dots, n-1, \\ \rho_0 = 0, \rho_n = 0, \end{cases}$$

则有

$$\begin{cases} |l(\varepsilon_i)| \leq -l(\rho_i), & i = 1, 2, \dots, n-1, \\ |\varepsilon_0| \leq \rho_0, |\varepsilon_n| \leq \rho_n. \end{cases}$$

根据推论 9.4.1 有

$$|\varepsilon_i| \leq \rho_i, \quad i = 0, 1, \dots, n.$$

若记

$$L(\rho_i) = \frac{\rho_{i-1} - 2\rho_i + \rho_{i+1}}{h^2},$$

则 ρ_i ($i = 0, 1, \dots, n$) 满足的差分方程可写为

$$\begin{cases} L(\rho_i) = -\frac{h^2}{12} M_4 + q_i \rho_i, & i = 1, 2, \dots, n-1, \\ \rho_0 = 0, \rho_n = 0. \end{cases}$$

令 $\eta_i (i = 0, 1, \dots, n)$ 满足

$$\begin{cases} L(\eta_i) = -\frac{h^2}{12}M_4, & i = 1, 2, \dots, n-1, \\ \eta_0 = 0, \eta_n = 0. \end{cases} \quad (9.4.12)$$

以上两式相减得

$$\begin{cases} L(\rho_i - \eta_i) = q_i \rho_i \geq 0, & i = 1, 2, \dots, n-1, \\ \rho_0 - \eta_0 = 0, \rho_n - \eta_n = 0. \end{cases}$$

根据极值原理知

$$\max_{\rho_i - \eta_i \geq 0} \{\rho_i - \eta_i\} = \max\{\rho_0 - \eta_0, \rho_n - \eta_n\} = 0.$$

由此即得 $\rho_i - \eta_i \leq 0$ ($i = 0, 1, \dots, n$), 从而有

$$|\varepsilon_i| \leq \rho_i \leq \eta_i, \quad i = 0, 1, \dots, n.$$

求解差分方程 (9.4.12) 得

$$\eta_i = \frac{h^4}{24}M_4i(n-i) = \frac{h^2}{24}M_4ih(nh-ih) = \frac{h^2}{24}M_4(x_i-a)(b-x_i), \quad i = 0, 1, \dots, n,$$

所以

$$\begin{aligned} |y(x_i) - y_i| &= |\varepsilon_i| \leq \eta_i = \frac{h^2}{24}M_4(x_i-a)(b-x_i) \\ &\leq \frac{h^2}{24}M_4 \max_{a \leq x \leq b} (x-a)(b-x) = \frac{h^2}{24}M_4 \left(\frac{a+b}{2} - a \right) \left(b - \frac{a+b}{2} \right) \\ &= \frac{h^2}{96}M_4(b-a)^2. \end{aligned}$$

证毕.

由误差估计式可见, 若 $|y^{(4)}(x)|$ 有界 (即 $q(x)$ 和 $f(x)$ 的二阶导数有界), 则当 $h \rightarrow 0$ 时, $|y(x_i) - y_i| \rightarrow 0$, 即当步长 $h \rightarrow 0$ 时, 差分解收敛于微分方程的真解.

9.4.2 打靶法

打靶法 (也称为试射法) 是把常微分方程边值问题转化为初值问题来求解. 下面以第一边值问题为例来说明打靶法的思想. 考虑二阶常微分方程第一边值问题

$$\begin{cases} y''(x) = f(x, y, y'), & a < x < b, \\ y(a) = \alpha, y(b) = \beta. \end{cases} \quad (9.4.13)$$

打靶法将边值问题 (9.4.13) 转化为求解以下初值问题:

$$\begin{cases} y''(x) = f(x, y, y'), & a < x < b, \\ y(a) = \alpha, \quad y'(a) = m. \end{cases} \quad (9.4.14)$$

其思想就是设法确定 $y'(a)$ 的值 m , 使得初值问题 (9.4.14) 的解也满足边值问题 (9.4.13) 的另一个边值条件 $y(b) = \beta$. 从几何上来看, 也就是从微分方程经过点 (a, α) 而具有不同斜率的积分曲线中, 寻找一条过点 (b, β) 的曲线, 如图 9.4 所示.

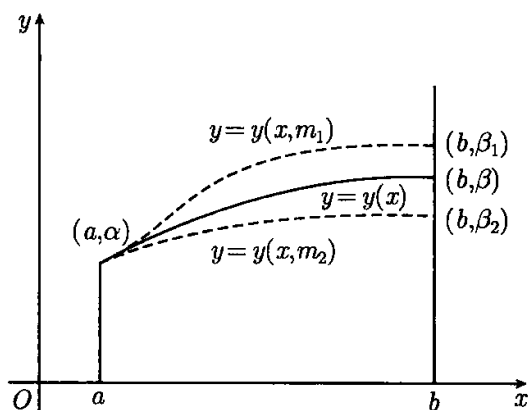


图 9.4 打靶法的几何意义

在实际计算时, 对于具体的问题或作定性分析, 或根据经验选取一个斜率 m_1 , 进行试算 (试射), 即求解

$$\begin{cases} y''(x) = f(x, y, y'), & a < x < b, \\ y(a) = \alpha, \quad y'(a) = m_1. \end{cases} \quad (9.4.15)$$

方程 (9.4.15) 的解与 m_1 有关, 将其记为 $y(x, m_1)$. 若 $y(b, m_1) = \beta$, 或者 $|y(b, m_1) - \beta| < \varepsilon$ (ε 为允许误差界), 则 $y(x, m_1)$ 即是所求的解; 否则, 根据 $\beta_1 = y(b, m_1)$ 与 β 的差距修正 m_1 , 设将 m_1 修正为 m_2 , 然后再求解

$$\begin{cases} y''(x) = f(x, y, y'), & a < x < b, \\ y(a) = \alpha, \quad y'(a) = m_2. \end{cases}$$

反复进行以上过程, 直到满足精度要求为止.

下面介绍修正 m 的方法. 设初值问题 (9.4.14) 的解为 $y(x, m)$. 理想的 m 应使 $y(x, m)$ 在边界点 b 处 $y(b, m) = \beta$ (中靶), 即 m 是方程

$$y(b, m) - \beta = 0$$

的根. 但对于数值解法, $y(b, m)$ 无明确的解析表达式, 这个方程难以求解.

修正 m 的一个简单常用的方法是取 $m_2 = \frac{\beta}{\beta_1} m_1$, 此后由非线性方程求根的割线法

$$m_k = m_{k-1} - \frac{y(b, m_{k-1}) - \beta}{y(b, m_{k-1}) - y(b, m_{k-2})} (m_{k-1} - m_{k-2}), \quad k = 3, 4, \dots$$

生成点列 $\{m_k\}$, 反复求解

$$\begin{cases} y''(x) = f(x, y, y'), & a < x < b, \\ y(a) = \alpha, \quad y'(a) = m_k, & k = 1, 2, \dots \end{cases}$$

设其解为 $y(x, m_k)$, 直到满足 $|y(b, m_k) - \beta| < \varepsilon$ 为止.

这时, 取 $y(x_1, m_k), y(x_2, m_k), \dots, y(x_{n-1}, m_k), y(x_n, m_k)$ 作为边值问题 (9.4.13) 的真解 $y(x)$ 在点 x_i ($i = 1, 2, \dots, n$) 处的近似值.

小 结

本章介绍了求解常微分方程初、边值问题的基本数值方法.

对于初值问题, 介绍了导出数值解法的 4 种方法, 即数值微分法、数值积分法、基于泰勒展开的 R-K 法、待定系数法及由这些方法建立的求解公式. 同时引进了若干基本概念, 并且讨论了显式单步法的收敛性和稳定性. 简要介绍了一阶常微分方程组和高阶方程的数值解法.

在实际应用中, 要针对求解问题的具体情况, 选择适当的求解方法.

4 级 4 阶 R-K 法精度高, 是显式单步法, 易于调节步长, 计算过程稳定, 但它计算 $f(x, y)$ 函数值的次数较多. 因此, 适用于求解函数 $f(x, y)$ 比较简单的问题.

线性多步法和由它构成的预测—校正公式, 每一步计算函数值的次数少, 但它不是自开始的, 需要借助于其他方法提供开始值, 它们适用于求解函数 $f(x, y)$ 较复杂的问题.

预测—校正公式比同阶的多步法每次仅多计算一次函数值, 就可得到更精确的结果, 并且稳定性好.

对于一般问题而言, 阶数越高的方法不一定计算结果就越精确. 对具体问题要进行具体分析. 若能确定问题真解的高阶导数存在, 就选用高阶方法; 否则, 宜用低阶方法.

选取步长 h 的大小, 既要考虑到节省计算量, 又要减少局部截断误差以保证计算结果的精度. 选取 h , 应使 h 落在绝对稳定区域.

常微分方程初值问题的求解是一个逐步计算的过程. 每一步计算产生的误差都会对以后各步计算产生影响, 所以求解应选用绝对稳定性较好的方法. 通常隐式法的绝对稳定性都比同类同阶的显式法好.

一阶常微分方程组初值问题是用单个方程的方法求解的, 只需把函数改为向量函数即可. 而高阶常微分方程则是化为一阶微分方程组的初值问题来求解.

刚性方程组在工程技术领域有着广泛的应用, 具体求解有一定的难度, 其理论和解法内容很多, 并且有现成的数学软件可供使用, 具体可参见文献 [19], [27].

求解边值问题的有限差分法是用导数的近似式代替导数将边值问题进行离散化处理, 最终归结为求解一个三对角线性方程组. 解此方程组得到原方程的数值解.

打靶法实质上是将求解边值问题转化为求解一系列的初值问题. 当 $|y(b, m_k) - \beta| < \varepsilon$ 时, 取 $y(x_i, m_k)$ ($i = 1, 2, \dots, n$) 作为边值问题的数值解.

习 题

9.1 取 $h = 0.1$, 用欧拉法、后退欧拉法、中点法、梯形法求解下列初值问题:

- (1) $y' = 1 - y$, $y(0) = 0$, $0 \leq x \leq 1$;
- (2) $y' = xy^2$, $y(0) = 1$, $0 \leq x \leq 1$;
- (3) $y' = y - 2x/y$, $y(0) = 1$, $0 \leq x \leq 1$;
- (4) $y' = x - y + 1$, $y(0) = 1$, $0 \leq x \leq 1$.

9.2 用数值积分法导出欧拉法、后退欧拉法、中点法的计算公式.

9.3 利用求 $f'(x_0)$ 三点数值微分公式, 导出求解初值问题 (9.0.1) 的计算公式.

9.4 利用改进欧拉法求解习题 9.1.

9.5 利用标准的 4 级 4 阶 R-K 法求解习题 9.1.

9.6 利用修正汉明预测—校正法求解习题 9.1.

9.7 利用待定系数法确定下列求解公式中的系数, 使其阶数尽可能高, 并导出截断误差表示式:

- (1) $y_{i+1} = \alpha_0 y_i + \alpha_1 y_{i-1} + \beta h f_{i+1}$;
- (2) $y_{i+1} = y_i + h(\beta_0 f_i + \beta_1 f_{i-1})$;
- (3) $y_{i+1} = y_{i-3} + h(\beta_0 f_i + \beta_1 f_{i-1} + \beta_2 f_{i-2})$.

9.8 仿照修正汉明预测—校正公示的推导, 对下列预测—校正公式进行修正:

- (1) 预测: 米尔恩公式 (9.1.28), 校正: 辛普森公式 (9.1.7);
- (2) 预测: $k = 3$ 的亚当斯显式公式, 校正: $k = 3$ 的亚当斯隐式公式.

9.9 求下列公式的绝对稳定区域:

- (1) 梯形法; (2) 改进欧拉法.

9.10 对于初值问题

$$\begin{cases} y' = -100(y - x^2) + 2x, & 0 \leq x \leq 1, \\ y(0) = 0. \end{cases}$$

- (1) 用欧拉法求解, 步长 h 在什么范围内取值, 计算过程稳定;
- (2) 若用梯形公式计算, 步长有没有限制.

9.11 将下列高阶微分方程化为一阶微分方程组初值问题:

$$(1) \begin{cases} y'' = y'(1 - y^2) - y, \\ y(x_0) = y_0, y'(x_0) = y'_0; \end{cases}$$

$$(2) \begin{cases} y''' = y'' - 2y' + y - x + 1, \\ y(x_0) = y_0, y'(x_0) = y'_0, y''(x_0) = y''_0. \end{cases}$$

9.12 用差分法求解以下边值问题:

(1) 取 $h = 0.5$,

$$\begin{cases} y'' = (1 + x^2)y, & -1 < x < 1, \\ y(-1) = y(1) = 1; \end{cases}$$

(2) 取 $h = 0.25$,

$$\begin{cases} y'' = 6x + y, & 0 < x < 1, \\ y(0) = y(1) = 0. \end{cases}$$

计 算 实 习

9.1 选用标准 4 级 4 阶 R-K 法或修正汉明预测—校正法求解下列初值问题, 取步长 $h = 0.1, 0.01, 0.001$, 计算 $y(x)$ 在 $x = 1$ 处的近似值:

$$(1) y' = xy^2 + y, \quad y(0) = 1;$$

$$(2) y' = e^x y^2 - 2y, \quad y(0) = 1;$$

$$(3) y' = x^2 + y^2, \quad y(0) = 1.$$

9.2 用标准 4 级 4 阶 R-K 法求解

$$\begin{cases} y''' = y'' + y' - y + 2x - 3, \\ y(0) = -1, y'(0) = 3, y''(0) = 2. \end{cases}$$

取步长 $h = 0.05$, 计算 $y(1)$ 的近似值, 并与解析解 $y(x) = xe^x + 2x - 1$ 作比较.

9.3 (1) 写出打靶法的计算步骤;

(2) 取 $h = 0.1$, 用打靶法求解以下边值问题 (其中求解初值问题用标准 4 级 4 阶 R-K 法):

$$(i) \begin{cases} y'' = -(1+x)y' + 2y + (1-x^2)e^{-x}, & 0 < x < 1, \\ y(0) = y(1) = 0; \end{cases}$$

$$(ii) \begin{cases} y'' - (1+x^2)y = -1, & -1 < x < 1, \\ y(-1) = y(1) = 0. \end{cases}$$

第 10 章 偏微分方程的数值解法

本章讨论求解偏微分方程定解问题的数值解法, 主要介绍有限差分方法和有限元方法.

10.1 椭圆型边值问题

本节先讨论一个典型的椭圆型方程边值问题 —— 泊松方程第一边值问题

$$\begin{cases} \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y), & (x, y) \in \Omega, \\ u = g(x, y), & (x, y) \in \Gamma = \partial\Omega, \end{cases} \quad (10.1.1)$$

其中 $\Omega = \{(x, y) | 0 < x < a, 0 < y < b\}$, $f(x, y)$, $g(x, y)$ 为已知函数, $u = u(x, y)$ 为未知函数.

10.1.1 差分方程的建立

为了数值求解 (10.1.1), 取两族等距的平行线 (称为网格线)

$$x = x_0 + ih, \quad i = 0, \pm 1, \pm 2, \dots,$$

$$y = y_0 + j\tau, \quad j = 0, \pm 1, \pm 2, \dots,$$

计算网格线交点处解 $u(x_i, y_j)$ 的近似值 u_{ij} . 网格线交点 (x_i, y_j) 称为网点. 包含于 Ω 内及其边界上的网点称为节点, 它们的全体称为网域. 所有节点可分为三类:

① 相邻 4 网点都在 Ω 内, 这种节点称为内部节点, 其全体记为 Ω_1 ; ② 既非内部节点, 又不在边界 Γ 上的点, 这种节点称为网域边界点, 其全体记为 Ω_2 ; ③ 恰恰位于区域边界的节点称为边界节点, 其全体记为 Ω_3 . 对第一边值问题, Γ 上解的值已知.

1. 五点差分格式

利用泰勒展开, 对内部节点 (i, j) ,

$$\begin{aligned} & \frac{1}{h^2} [u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)] + \frac{1}{\tau^2} [u(x_i, y_{j+1}) \\ & \quad - 2u(x_i, y_j) + u(x_i, y_{j-1})] - f(x_i, y_j) \\ &= \frac{h^2}{12} \frac{\partial^4 u(\xi_i, y_j)}{\partial x^4} + \frac{\tau^2}{12} \frac{\partial^4 u(x_i, \eta_j)}{\partial y^4}, \end{aligned}$$

略去右边误差项, 得如下的一个差分方程:

$$\Delta_h u_{ij} = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) + \frac{1}{\tau^2}(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) = f_{i,j}. \quad (10.1.2)$$

差分方程 (10.1.2) 称为五点差分格式, 略去的项

$$R(x_i, y_j) = \frac{h^2}{12} \frac{\partial^4 u(\xi_i, y_j)}{\partial x^4} + \frac{\tau^2}{12} \frac{\partial^4 u(x_i, \eta_j)}{\partial y^4},$$

称为差分格式的截断误差. 显然, 差分格式 (10.1.2) 的截断误差为 $O(h^2) + O(\tau^2)$.

在 $h = \tau$ 的情形下, 差分格式 (10.1.2) 可简化为

$$\begin{aligned} & u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} \\ & = h^2 f_{i,j}. \end{aligned} \quad (10.1.3)$$

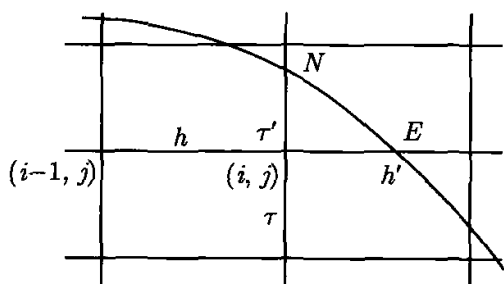


图 10.1

对 Ω_2 中的节点 (i, j) , 近似代替微分方程 (10.1.1) 中偏导数, 也可得到差分方程, 应用节点不等距的数值微分公式. 例如, 对图 10.1 中的节点 (i, j) 可用

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} &= \frac{2}{h+h'} \left[\frac{u(E) - u(x_i, y_j)}{h'} - \frac{u(x_i, y_j) - u(x_{i-1}, y_j)}{h} \right] \\ &+ \frac{1}{3}(h-h') \frac{\partial^3 u(\xi_i, y_j)}{\partial x^3} - \frac{hh'}{12} \frac{\partial^4 u(\xi_i^*, y_j)}{\partial x^4}, \end{aligned}$$

由此可得差分方程

$$\frac{2}{h+h'} \left(\frac{u_E - u_{ij}}{h'} - \frac{u_{ij} - u_{i-1,j}}{h} \right) + \frac{2}{\tau+\tau'} \left(\frac{u_N - u_{ij}}{\tau'} - \frac{u_{ij} - u_{i,j-1}}{\tau} \right) = f_{i,j}, \quad (10.1.4)$$

其截断误差

$$\begin{aligned} R(x_i, y_j) &= -\frac{1}{3}(h-h') \frac{\partial^3 u(\xi_i, y_j)}{\partial x^3} + \frac{hh'}{12} \frac{\partial^4 u(\xi_i^*, y_j)}{\partial x^4} \\ &- \frac{1}{3}(\tau-\tau') \frac{\partial^3 u(x_i, \eta_j)}{\partial y^3} + \frac{\tau\tau'}{12} \frac{\partial^4 u(x_i, \eta_j^*)}{\partial y^4}. \end{aligned}$$

例 10.1.1 设 Ω 为图 10.2 中 L 型区域

$OABCDEO$, 取 $h = \tau = \frac{1}{4}$, 用五点差分格式解

$$\begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, & (x, y) \in \Omega, \\ u|_{x=0} = y(1-y), \\ u|_{y=0} = x(1-x), \\ u = 0, & (x, y) \in ABCDE. \end{cases}$$

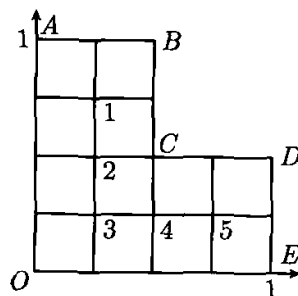


图 10.2

解 由图 10.1 知, 只有节点 1, 2, 3, 4, 5 为内部节点, 并且由方程和区域的对称性得 $u_1 = u_5$, $u_2 = u_4$, 在节点 1, 2, 3 处, 按差分格式 (10.1.2) 得

$$\begin{cases} \frac{3}{16} + u_2 - 4u_1 = 0, \\ \frac{4}{16} + u_1 + u_3 - 4u_2 = 0, \\ \frac{3}{16} + \frac{3}{16} + 2u_2 - 4u_3 = 0, \end{cases}$$

解得

$$u_1 = u_5 \approx 0.0769, \quad u_2 = u_4 \approx 0.1202, \quad u_3 \approx 0.1538.$$

解毕.

一般情况下, 由差分格式 (10.1.2) 确定的线性代数方程组中, 每个方程最多只涉及 5 个未知量, 故系数矩阵是稀疏矩阵, 并且对角占优. 因此, 方程组可用迭代法求解.

2. 九点差分格式

为了提高差分格式的精度, 下面给出九点差分格式. 为描述简单, 令 $h = \tau$, 于是得如下的九点差分格式:

$$\begin{aligned} \Delta_h u_{ij} = & 4(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}) + (u_{i+1,j+1} + u_{i-1,j-1} \\ & + u_{i-1,j+1} + u_{i+1,j-1}) - 20u_{i,j} = 6h^2 f_{i,j} + \frac{1}{2}h^4 \Delta f. \end{aligned} \quad (10.1.5)$$

显然, 差分格式 (10.1.5) 的截断误差为 $O(h^4)$, 即九点差分格式是 4 阶精度的格式.

10.1.2 差分解的误差估计与收敛性

类似于常微分方程边值问题的推导, 这里也可以利用极值原理导出差分解的收敛性和误差估计式.

定理 10.1.1 (极值原理) 设 u_{ij} 是定义在 $\bar{\Omega}_h = \Omega_h \cup \partial\Omega_h$ 上的函数,

- (1) 如果 $\Delta_h u_{ij} \geq 0$ ($\forall (x_i, y_j) \in \Omega_h$), 则有 $\max_{\Omega_h} u_{ij} \leq \max_{\partial\Omega_h} u_{ij}$;
- (2) 如果 $\Delta_h u_{ij} \leq 0$ ($\forall (x_i, y_j) \in \Omega_h$), 则有 $\min_{\Omega_h} u_{ij} \leq \min_{\partial\Omega_h} u_{ij}$.

证明 用反证法证明 (1). 设在 Ω_h 内存在一点 (x_{i_0}, y_{j_0}) 及一个常数 M , 使得 $u(x_{i_0}, y_{j_0}) = M$ 且 $M \geq u_{ij}$ ($(x_i, y_j) \in \Omega_h$) 和 $M > u_{ij}$ ($(x_i, y_j) \in \partial\Omega_h$). 考虑五点差分

$$\Delta_h u_{i_0, j_0} = \frac{1}{h^2}(u_{i_0+1, j_0} - 2u_{i_0, j_0} + u_{i_0-1, j_0}) + \frac{1}{\tau^2}(u_{i_0, j_0+1} - 2u_{i_0, j_0} + u_{i_0, j_0-1}),$$

由假定 $\Delta_h u_{i_0, j_0} \geq 0$ 有

$$M = u_{i_0, j_0} \leq \frac{1}{\frac{1}{h^2} + \frac{1}{\tau^2}} \left(\frac{u_{i_0+1, j_0} + u_{i_0-1, j_0}}{2h^2} + \frac{u_{i_0, j_0+1} + u_{i_0, j_0-1}}{2\tau^2} \right).$$

但 $M \geq u_{ij} ((x_i, y_j) \in \Omega_h)$, 因此,

$$u_{i_0+1, j_0} = u_{i_0-1, j_0} = u_{i_0, j_0+1} = u_{i_0, j_0-1} = M.$$

把点 $(x_{i_0+1}, y_{j_0}), (x_{i_0-1}, y_{j_0}), (x_{i_0}, y_{j_0+1}), (x_{i_0}, y_{j_0-1})$ 中任一点按上述证明可以得出同样结论. 如果上述点的邻点中有边界点, 则在这些点上同样有 $M = u_{ij}$, 继续这个过程就可推出

$$M = u_{ij}, \quad \forall (x_i, y_j) \in \bar{\Omega}_h.$$

这与假设 $M > u_{ij} ((x_i, y_j) \in \partial\Omega_h)$ 矛盾, (1) 得证.

类似地证明可得 (2). 证毕.

下面讨论收敛性.

定理 10.1.2 设 u_{ij} 是定义在 $\bar{\Omega}_h$ 上的函数, 则有

$$\max_{\Omega_h} |u_{ij}| \leq \max_{\partial\Omega_h} |u_{ij}| + \frac{a^2}{2} \max_{\Omega_h} |\Delta_h u_{ij}|, \quad (10.1.6)$$

其中 a 为矩形区域 Ω 的 x 方向的边长.

证明 引进函数 $\phi(x, y) = \frac{1}{2}x^2$, 则在 $\bar{\Omega}_h$ 上有

$$0 \leq \phi_{ij} \leq \frac{a^2}{2}, \quad \Delta_h \phi_{ij} = 1.$$

定义函数 $u_{ij}^\pm = \pm u_{ij} + \phi_{ij} \max_{\Omega_h} |\Delta_h u_{ij}|$ 则

$$\Delta_h u_{ij}^\pm = \pm \Delta_h u_{ij} + \max_{\Omega_h} |\Delta_h u_{ij}| \geq 0.$$

应用定理 10.1.1 有

$$\max_{\Omega_h} u_{ij}^\pm \leq \max_{\partial\Omega_h} u_{ij}^\pm \leq \max_{\partial\Omega_h} \left(\pm u_{ij} + \frac{a^2}{2} \max_{\Omega_h} |\Delta_h u_{ij}| \right),$$

由此得到

$$u_{ij}^\pm \leq \max_{\partial\Omega_h} (\pm u_{ij}) + \frac{a^2}{2} \max_{\Omega_h} |\Delta_h u_{ij}|.$$

又 $\pm u_{ij} \leq u_{ij}^\pm$, 因此,

$$\pm u_{ij} \leq \max_{\partial\Omega_h} (\pm u_{ij}) + \frac{a^2}{2} \max_{\Omega_h} |\Delta_h u_{ij}| \leq \max_{\partial\Omega_h} |u_{ij}| + \frac{a^2}{2} \max_{\Omega_h} |\Delta_h u_{ij}|,$$

从而 (10.1.6) 成立. 证毕.

定理 10.1.3 设第一边值问题 (10.1.1) 的解在 $\bar{\Omega}$ 上有 4 阶连续的偏导数, 则五点差分格式 (10.1.2) 收敛, 并且有误差估计式

$$\max_{\Omega_h} |u_{ij} - u(x_i, y_j)| \leq K(h^2 + \tau^2). \quad (10.1.7)$$

证明 令 $\varepsilon_{ij} = \Delta_h u(x_i, y_j) - \Delta u(x_i, y_j)$, 因为 $\Delta u(x_i, y_j) = f(x_i, y_j)$, 从而得

$$\Delta_h u(x_i, y_j) = \varepsilon_{ij} + f(x_i, y_j).$$

又 $\Delta_h u_{ij} = f_{ij}$, 两式相减得

$$\Delta_h(u_{ij} - u(x_i, y_j)) = -\varepsilon_{ij}, \quad (x_i, y_j) \in \Omega_h.$$

在边界上有 $u_{ij} - u(x_i, y_j) = 0$ ($(x_i, y_j) \in \partial\Omega_h$). 对函数 $u_{ij} - u(x_i, y_j)$ 应用定理 10.1.2, 得到

$$\max_{\Omega_h} |u_{ij} - u(x_i, y_j)| \leq \frac{a^2}{2} \max_{\Omega_h} |\varepsilon_{ij}|. \quad (10.1.8)$$

利用泰勒展开有

$$\begin{aligned} u(x_i \pm h, y_j) &= u(x_i, y_j) \pm h \frac{\partial}{\partial x} u(x_i, y_j) + \frac{h^2}{2} \frac{\partial^2}{\partial x^2} u(x_i, y_j) \\ &\quad \pm \frac{h^3}{3!} \frac{\partial^3}{\partial x^3} u(x_i, y_j) + \frac{h^4}{4!} \frac{\partial^4}{\partial x^4} u(x_i + \theta_{\pm} h, y_j), \end{aligned}$$

其中 $|\theta_{\pm}| \leq 1$. 于是

$$\frac{1}{h^2} (u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)) - \frac{\partial^2}{\partial x^2} u(x_i, y_j) = \frac{h^2}{12} \frac{\partial^4}{\partial x^4} u(x_i + \theta_1 h, y_j),$$

其中 $|\theta_1| \leq 1$. 同样地有

$$\frac{1}{\tau^2} (u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})) - \frac{\partial^2}{\partial y^2} u(x_i, y_j) = \frac{\tau^2}{12} \frac{\partial^4}{\partial y^4} u(x_i, y_j + \theta_2 \tau),$$

其中 $|\theta_2| \leq 1$. 于是

$$\varepsilon_{ij} = \frac{1}{12} \left(h^2 \frac{\partial^4}{\partial x^4} u(x_i + \theta_1 h, y_j) + \tau^2 \frac{\partial^4}{\partial y^4} u(x_i, y_j + \theta_2 \tau) \right).$$

利用不等式 (10.1.8), 得到 (10.1.7) 成立. 证毕.

10.1.3 一般二阶椭圆型方程边值问题

设有更一般的二阶椭圆型方程边值问题

$$\begin{cases} a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial y^2} + c \frac{\partial u}{\partial x} + d \frac{\partial u}{\partial y} + eu = f, & (x, y) \in \Omega, \\ \alpha \frac{\partial u}{\partial n} + \beta u + \gamma = 0, & (x, y) \in \Gamma = \partial\Omega, \end{cases} \quad (10.1.9)$$

其中 $a, b, c, d, e, f, \alpha, \beta, \gamma$ 都为 (x, y) 的已知函数, 并且 $a \geq a_0 > 0, b \geq b_0 > 0, e \leq 0, a_0, b_0$ 为常数, $\frac{\partial u}{\partial n}$ 为外法线方向 n 的方向导数. 类似于泊松问题的推导, 可得差分方程

$$\begin{aligned} & \frac{a_{ij}}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) + \frac{b_{ij}}{\tau^2}(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) \\ & + \frac{c_{ij}}{2h}(u_{i+1,j} - u_{i-1,j}) + \frac{d_{ij}}{2\tau}(u_{i,j+1} - u_{i,j-1}) + e_{ij}u_{ij} = f_{ij}. \end{aligned} \quad (10.1.10)$$

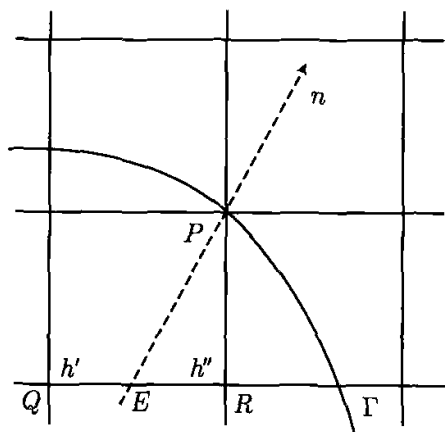


图 10.3

对边界条件 $\alpha \frac{\partial u}{\partial n} + \beta u + \gamma = 0$, 类似地有差分方程

$$\begin{aligned} & \frac{\alpha_{ij}}{l} \left(u_{ij} - \frac{h''u_{i-1,j-1} + h'u_{i,j-1}}{h} \right) \\ & + \beta_{ij}u_{ij} + \gamma_{ij} = 0, \end{aligned} \quad (10.1.11)$$

其中 h', h'' 如图 10.3 所示, $l = |PE|$.

由差分格式 (10.1.10), (10.1.11) 确定的线性代数方程组的系数矩阵是稀疏矩阵, 并且对角占优. 因此, 方程组可用迭代法求解.

10.2 抛物型方程初、边值问题

本节讨论一个典型的抛物型方程 —— 热传导方程的第一初边值问题

$$\begin{cases} \frac{\partial u}{\partial t} - a^2 \frac{\partial^2 u}{\partial x^2} = f(x, t), & 0 < x < l, 0 < t < T, \\ u(x, 0) = \varphi(x), & 0 \leq x \leq l, \\ u(0, t) = \mu_1(t), & u(l, t) = \mu_2(t), 0 \leq t \leq T, \end{cases} \quad (10.2.1)$$

其中 $\varphi(x), \mu_1(t), \mu_2(t)$ 为已知的连续函数, 并且满足相容性条件

$$\varphi(0) = \mu_1(0), \quad \varphi(l) = \mu_2(0).$$

10.2.1 差分方程的建立与求解

令 $h = l/m, \tau = T/n$, 作两族平行线

$$x = x_i = ih, \quad i = 0, 1, 2, \dots, m, \quad t = t_j = j\tau, \quad j = 0, 1, 2, \dots, n.$$

两族平行线的交点称为节点. 显然, 由边界条件知

$$\begin{cases} u_{i0} = u(x_i, 0) = \varphi(x_i), & i = 0, 1, 2, \dots, m, \\ u_{0j} = u(0, t_j) = \mu_1(t_j), & j = 0, 1, 2, \dots, n, \\ u_{mj} = u(l, t_j) = \mu_2(t_j), & j = 0, 1, 2, \dots, n. \end{cases}$$

1. 最简显式差分格式

在内部节点 (x_i, t_j) 处, $\frac{\partial u}{\partial t}$ 用向前欧拉数值微分公式, 从而得差分方程

$$\frac{1}{\tau}(u_{i,j+1} - u_{ij}) - \frac{a^2}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) = f_{i,j}. \quad (10.2.2)$$

它们称为最简显式差分格式.

2. 最简隐式差分格式

在内部节点 (x_i, t_j) 处, $\frac{\partial u}{\partial t}$ 用向后欧拉数值微分公式, 从而得差分方程

$$\frac{1}{\tau}(u_{ij} - u_{i,j-1}) - \frac{a^2}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) = f_{i,j}, \quad (10.2.3)$$

称为最简隐式差分格式. 由于 u_{i0} 已知, 按此式可逐层算出 $(u_{0j}, u_{1j}, \dots, u_{mj}) = u_j$ ($j = 0, 1, 2, \dots, n$). 当然, 由 (10.2.3), 用 u_j 计算 u_{j+1} 时, 需解三对角线性代数方程组, 可用追赶法.

3. 六点对称格式

将 (10.2.2), (10.2.3) 相加, 除以 2 得

$$\begin{aligned} \frac{1}{2\tau}(u_{i,j+1} - u_{i,j}) - \frac{a^2}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j} + u_{i+1,j+1} \\ - 2u_{i,j+1} + u_{i-1,j+1}) = \frac{1}{2}(f_{i,j} + f_{i,j+1}). \end{aligned} \quad (10.2.4)$$

它称为六点对称格式, 也是隐式差分格式. 由 u_j 计算 u_{j+1} 时, 可用追赶法.

4. Richardson 格式

如果在内部节点 (x_i, t_j) 处, $\frac{\partial u}{\partial t}$ 改用三点数值微分公式, 则可得差分方程

$$\frac{1}{2\tau}(u_{i,j+1} - u_{i,j-1}) - \frac{a^2}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) = f_{i,j}. \quad (10.2.5)$$

它称为 Richardson 格式, 其截断误差为

$$R(x_i, t_j) = \frac{\tau^2}{6} \frac{\partial^3 u(x_i, \eta_j)}{\partial t^3} - \frac{a^2 h^2}{12} \frac{\partial^4 u(\xi_i, t_j)}{\partial x^4}.$$

例 10.2.1 用最简显式和隐式差分格式解

$$\begin{cases} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, & 0 < x < \pi, 0 < t, \\ u(x, 0) = \varphi(x), & 0 \leq x \leq \pi, \\ u(0, t) = u(\pi, t) = 0, & t \geq 0, \end{cases}$$

其中

$$\varphi(x) = \begin{cases} x, & 0 \leq x \leq \frac{\pi}{2}, \\ \pi - x, & \frac{\pi}{2} \leq x \leq \pi. \end{cases}$$

取 $h = \frac{\pi}{20}, \tau = \frac{5}{9}$.

解 由题设的对称性知

$$u(\pi - x, t) = u(x, t), \quad 0 \leq x \leq \frac{\pi}{2},$$

故

$$u_{20-i,j} = u_{ij}, \quad i = 0, 1, \dots, 20.$$

按显式差分格式 (10.2.2) 计算可得表 10.1.

表 10.1

j	u_{1j}	u_{2j}	u_{3j}	u_{4j}	u_{5j}	$u_{10,j}$
1	0.157	0.314	0.471	0.628	0.785	1.396
5	0.157	0.314	0.471	0.628	0.785	1.234
10	0.156	0.314	0.461	0.626	0.738	1.230
15	0.159	0.289	0.479	0.544	0.789	0.806
20	0.117	0.353	0.326	0.710	0.471	1.387
40	-2.12	4.677	-6.28	9.127	-10.11	16.55

按隐式差分格式 (10.2.3) 计算可得表 10.2.

表 10.2

j	u_{1j}	u_{2j}	u_{3j}	u_{4j}	u_{5j}	$u_{10,j}$
1	0.157	0.314	0.471	0.628	0.785	1.473
5	0.157	0.314	0.470	0.625	0.778	1.290
10	0.155	0.310	0.463	0.611	0.753	1.163
15	0.152	0.301	0.447	0.587	0.717	1.067
20	0.146	0.289	0.427	0.558	0.667	0.987
40	0.115	0.228	0.335	0.434	0.523	0.742

问题的真解为

$$u(x, t) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)^2} e^{-(2k+1)^2 t} \sin(2k+1)x,$$

其值如表 10.3 所示.

表 10.3

$u(x_i, t_j)$	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 10$
$j = 1$	0.157	0.314	0.471	0.619	0.785	1.438
$j = 5$	0.157	0.314	0.471	0.626	0.781	1.275
$j = 10$	0.156	0.311	0.464	0.613	0.755	1.153
$j = 15$	0.152	0.303	0.449	0.588	0.717	1.060
$j = 20$	0.146	0.289	0.427	0.558	0.667	0.980
$j = 40$	0.115	0.227	0.333	0.432	0.520	0.737

解毕.

比较这些表可见, 隐式格式计算结果和真解相差不大, 显式格式计算结果却随着 j 的增加而越来越偏离真解. 产生这一现象的原因是显式格式在 $\tau = 5/9$ 时不稳定.

10.2.2 差分格式的稳定性

现在研究差分格式的稳定性.

定义 10.2.1 令 $u_j = (u_{0j}, u_{1j}, \dots, u_{mj})$. 设初值 u_0 有误差而变为 u_0^* , 按某种差分格式算出的 u_j 也有误差变为 u_j^* , 如果存在常数 M , 使得

$$\|u_j - u_j^*\| \leq M \|u_0 - u_0^*\|$$

成立, 则称此差分格式稳定; 否则, 称为不稳定.

显式差分格式 (10.2.2) 是条件稳定的.

定理 10.2.1 显式差分格式 (10.2.2) 是稳定的充分必要条件是 $r = \frac{a^2 \tau}{h^2} \leq \frac{1}{2}$.

证明 因为 u_j^* 满足 (10.2.2), 故

$$\frac{1}{\tau}(u_{i,j+1}^* - u_{ij}^*) - \frac{a^2}{h^2}(u_{i+1,j}^* - 2u_{ij}^* + u_{i-1,j}^*) = f_{i,j}.$$

将此式与 (10.2.2) 相减, 令 $e_j = u_j - u_j^*$, 则得

$$e_{i,j+1} = r e_{i-1,j} + (1 - 2r) e_{i,j} + r e_{i+1,j}. \quad (10.2.6)$$

当 $r \leq \frac{1}{2}$ 时,

$$\begin{aligned}\|e_{j+1}\|_{\infty} &= \max_i |e_{i,j+1}| \leq r \max_i |e_{i-1,j}| + (1-2r) \max_i |e_{ij}| + r \max_i |e_{i+1,j}| \\ &= [r + (1-2r) + r] \max_i |e_{ij}| = \max_i |e_{ij}| = \|e_j\|_{\infty},\end{aligned}$$

所以

$$\|e_{j+1}\|_{\infty} \leq \|e_j\|_{\infty} \leq \cdots \leq \|e_1\|_{\infty} \leq \|e_0\|_{\infty},$$

于是当 $r \leq \frac{1}{2}$ 时, (10.2.2) 稳定; 当 $r > \frac{1}{2}$ 时, 方程 (10.2.6) 必有形式为 $e_{ij} = \alpha^i \beta^j \delta$ 的解, 其中 α, β 满足

$$\alpha^i \beta^{j+1} \delta = r \alpha^{i-1} \beta^j \delta + (1-2r) \alpha^i \beta^j \delta + r \alpha^{i+1} \beta^j \delta,$$

即

$$\beta = \frac{r}{\alpha} + (1-2r) + \alpha r.$$

由此可见, 令 $\alpha = -1, \beta = 1-4r$, 则差分方程 (10.2.6) 有解

$$e_{ij} = (-1)^i (1-4r)^j \delta, \quad \delta > 0.$$

由于 $4r-1 > 1$, 所以当 $j \rightarrow \infty$ 时,

$$\|e_j\|_{\infty} = \max_i |e_{ij}| = (4r-1)^j \delta = (4r-1)^j \|e_0\|_{\infty} \rightarrow \infty.$$

这说明当 $r > \frac{1}{2}$ 时, 差分格式 (10.2.2) 不稳定. 证毕.

对隐式差分格式 (10.2.3) 对应的齐次方程 (即当 $f(x, t) = 0$ 时) 可证极值原理: 差分解在第 $j+1$ 层的值 $u_{i,j+1}$ 必介于两端与第 j 层值 u_{ij} 的极值之间, 即

$$\min\{u_{0,j+1}, u_{ij}, u_{m,j+1}\} \leq u_{i,j+1} \leq \max\{u_{0,j+1}, u_{ij}, u_{m,j+1}\}.$$

定理 10.2.2 隐式差分格式 (10.2.3) 无条件稳定.

证明 令 $e_{ij} = u_{ij} - u_{ij}^*$, 则得

$$-re_{i-1,j+1} + (1+2r)e_{i,j+1} - re_{i+1,j+1} = e_{i,j}. \quad (10.2.7)$$

注意到 $e_{0j} = e_{mj} = 0$, 由极值原理得

$$\min_i e_{ij} \leq e_{i,j+1} \leq \max_i e_{ij},$$

从而

$$\|e_{j+1}\|_{\infty} = \max_i |e_{i,j+1}| \leq \max_i |e_{ij}| = \|e_j\|_{\infty},$$

所以

$$\|e_{j+1}\|_{\infty} \leq \|e_j\|_{\infty} \leq \cdots \leq \|e_1\|_{\infty} \leq \|e_0\|_{\infty}.$$

证毕.

类似地可以证明, 六点对称格式 (10.2.4) 无条件稳定. Richardson 格式 (10.2.5) 恒不稳定, 因而尽管 Richardson 格式的截断误差很小, 但不实用.

10.2.3 差分解的误差估计与收敛性

定理 10.2.3 隐式差分格式 (10.2.3) 的解 u_{ij} 和真解 $u(x_i, t_j)$ 间有如下的误差估计式:

$$|u(x_i, t_j) - u_{ij}| \leq \left(\frac{\tau}{2} M_2 + \frac{a^2 h^2}{12} M_4 \right) t_j, \quad (10.2.8)$$

其中 M_2 为 $|\partial^2 u / \partial t^2|$ 在区域 Ω 的上界, M_4 为 $|\partial^4 u / \partial t^4|$ 在区域 Ω 的上界.

证明 令 $e_{ij} = u(x_i, t_j) - u_{ij}$, 则得

$$\frac{1}{\tau}(e_{i,j+1} - e_{ij}) - \frac{a^2}{h^2}(e_{i+1,j+1} - 2e_{i,j+1} + e_{i-1,j+1}) = R_{ij},$$

其中

$$R_{ij} = -\frac{\tau}{2} \frac{\partial^2 u(x_i, \eta_{j+1})}{\partial t^2} - \frac{a^2 h^2}{12} \frac{\partial^4 u(\xi_i, t_{j+1})}{\partial x^4},$$

从而

$$e_{i,j+1} = e_{ij} + \tau R_{ij} + r(e_{i+1,j+1} - 2e_{i,j+1} + e_{i-1,j+1}).$$

类似极值原理得

$$\min\{0, e_{ij} + \tau R_{ij}\} \leq e_{i,j+1} \leq \max\{0, e_{ij} + \tau R_{ij}\}.$$

于是注意到 $|R_{ij}| \leq \frac{\tau}{2} M_2 + \frac{a^2 h^2}{12} M_4 = F$, 则有

$$\begin{aligned} \|e_{j+1}\|_{\infty} &\leq \|e_j\|_{\infty} + \tau F \leq \|e_{j-1}\|_{\infty} + 2\tau F \leq \cdots \\ &\leq \|e_0\|_{\infty} + (j+1)\tau F = (j+1)\tau F = t_{j+1} F. \end{aligned}$$

将 $j+1$ 改为 j , 便证明了 (10.2.8). 证毕.

当 $t_j \leq T$ 时, 由 (10.2.8) 可得差分解的又一误差估计式

$$|u(x_i, t_j) - u_{ij}| \leq \left(\frac{\tau}{2} M_2 + \frac{a^2 h^2}{12} M_4 \right) T.$$

由此显然, 在 $|\partial^2 u / \partial t^2|$, $|\partial^4 u / \partial t^4|$ 有界的情况下, 当 $\tau \rightarrow 0$, $h \rightarrow 0$ 时, $u_{ij} \rightarrow u(x_i, t_j)$, 即差分解收敛于真解.

10.3 双曲型方程混合问题

本节研究双曲型方程求解的差分方法.

10.3.1 一阶双曲型方程

首先考虑如下的一阶双曲型方程混合问题:

$$\begin{cases} \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = f(x, t), & x \in \mathbf{R}, t > 0, \\ u(x, 0) = u_0, & x \in \mathbf{R}. \end{cases} \quad (10.3.1)$$

1. 迎风格式

和 10.2 节一样, 作两族平行线, 用向前欧拉差商代替导数, 并且关于 x 的偏导数用在特征线 $(x - at = \xi)$ 方向一侧的单边差商来代替, 在内部节点 (x_i, t_j) 处得 (10.3.1) 的迎风格式

$$\begin{cases} \frac{u_{i,j+1} - u_{ij}}{\tau} + a \frac{u_{i,j} - u_{i-1,j}}{h} = f_{ij}, & a > 0, \\ \frac{u_{i,j+1} - u_{ij}}{\tau} + a \frac{u_{i+1,j} - u_{i,j}}{h} = f_{ij}, & a < 0. \end{cases} \quad (10.3.2)$$

容易证明, 迎风差分格式 (10.3.2) 稳定的充分必要条件是 $\frac{|a|\tau}{h} \leq 1$.

2. Lax-Friedrichs 格式

关于空间的偏导数用中心差分格式, 可得到如下格式:

$$\frac{u_{i,j+1} - \frac{1}{2}(u_{i+1,j} + u_{i-1,j})}{\tau} + a \frac{u_{i+1,j} - u_{i-1,j}}{2h} = f_{ij}. \quad (10.3.3)$$

这称为 Lax-Friedrichs 格式, 其截断误差是 $O(\tau + h^2) + O\left(\frac{h^2}{\tau}\right)$, 并且容易证明,

Lax-Friedrichs 格式 (10.3.3) 稳定的充分必要条件是 $\frac{|a|\tau}{h} \leq 1$.

3. Lax-Wendroff 格式

前面讨论的迎风格式和 Lax-Friedrichs 格式是一阶精度的差分格式. Lax 和 Wendroff 给出了如下的格式:

$$u_{i,j+1} = u_{ij} - \frac{a\tau}{2h}(u_{i+1,j} - u_{i-1,j}) + \frac{a^2\tau^2}{2h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) + f_{ij}. \quad (10.3.4)$$

这称为 Lax-Wendroff 格式. 它是二阶精度的差分格式, 并且容易证明, Lax-Wendroff 格式 (10.3.4) 稳定的条件是

$$\frac{|a|\tau}{h} \leq 1.$$

4. 蛙跳格式

考虑如下的三层格式:

$$\frac{u_{i,j+1} - u_{i,j-1}}{2\tau} + a \frac{u_{i+1,j} - u_{i-1,j}}{2h} = f_{ij}. \quad (10.3.5)$$

这个差分格式称为蛙跳格式, 其截断误差是 $O(\tau^2 + h^2)$, 并且容易证明, 蛙跳格式 (10.3.5) 稳定的条件是 $\frac{|a|\tau}{h} < 1$, 而当 $\frac{|a|\tau}{h} = 1$ 时, 蛙跳格式不稳定.

10.3.2 一阶常系数双曲型方程组

考虑如下的一阶常系数线性方程组:

$$\begin{cases} \frac{\partial u}{\partial t} + A \frac{\partial u}{\partial x} = 0, & x \in \mathbf{R}, t > 0, \\ u(x, 0) = u_0, & x \in \mathbf{R}, \end{cases} \quad (10.3.6)$$

其中 $u = u(x, t) = [u_1(x, t), u_2(x, t), \dots, u_p(x, t)]^T$, $A \in \mathbf{R}^{p \times p}$ 为常数矩阵.

1. Lax-Friedrichs 格式

上一节中, 双曲型方程的 Lax-Friedrichs 格式直接推广到方程组, 可得到如下格式:

$$\frac{u_{i,j+1} - \frac{1}{2}(u_{i+1,j} + u_{i-1,j})}{\tau} + A \frac{u_{i+1,j} - u_{i-1,j}}{2h} = 0. \quad (10.3.7)$$

这是一阶精度的差分格式, 并且容易证明, Lax-Friedrichs 格式 (10.3.7) 稳定的条件是

$$\frac{\tau}{h} \rho(A) \leq 1, \quad (10.3.8)$$

其中 $\rho(A)$ 为矩阵 A 的谱半径.

2. Lax-Wendroff 格式

单个双曲型方程的方法推广到方程组, 可得到如下 Lax-Wendroff 格式:

$$u_{i,j+1} = u_{ij} - \frac{\tau}{2h} A(u_{i+1,j} - u_{i-1,j}) + \frac{\tau^2}{2h^2} A^2(u_{i+1,j} - 2u_{ij} + u_{i-1,j}). \quad (10.3.9)$$

它是二阶精度的差分格式, 同样容易证明, Lax-Wendroff 格式 (10.3.4) 稳定的条件也是 (10.3.8).

10.3.3 二阶双曲型方程

本节讨论一个典型的双曲型方程——波动方程的混合问题

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, t), & 0 \leq x \leq l, 0 \leq t \leq T, \\ u|_{t=0} = \varphi(x), \frac{\partial u}{\partial t} \Big|_{t=0} = \psi(x), & 0 \leq x \leq l, \\ u|_{x=0} = \mu_1(t), u|_{x=l} = \mu_2(t), & 0 \leq t \leq T. \end{cases} \quad (10.3.10)$$

和 10.2 节一样, 作两族平行线, 并且由边界条件知

$$\begin{cases} u_{i0} = u(x_i, 0) = \varphi(x_i), & i = 0, 1, 2, \dots, m, \\ \frac{\partial u(x_i, 0)}{\partial t} = \psi(x_i), & i = 0, 1, 2, \dots, m, \\ u_{0j} = u(0, t_j) = \mu_1(t_j), & j = 0, 1, 2, \dots, n, \\ u_{mj} = u(l, t_j) = \mu_2(t_j), & j = 0, 1, 2, \dots, n. \end{cases}$$

1. 显式差分格式

对 $\frac{\partial u(x_i, 0)}{\partial t} = \psi(x_i)$, 应用数值微分公式得

$$\frac{u(x_i, t_1) - u(x_i, 0)}{\tau} - \frac{\tau}{2} \frac{\partial^2 u(x_i, \eta_0)}{\partial t^2} = \psi(x_i).$$

略去误差项得

$$u_{i1} = u_{i0} + \tau \psi_i, \quad \psi_i = \psi(x_i).$$

在内部节点 (x_i, t_j) 处, 对 $\frac{\partial^2 u}{\partial t^2}, \frac{\partial^2 u}{\partial x^2}$ 应用中心差商来逼近, 从而得差分方程

$$\frac{1}{\tau^2}(u_{i,j-1} - 2u_{ij} + u_{i,j+1}) - \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) = f_{i,j}. \quad (10.3.11)$$

它称为显式格式.

定理 10.3.1 显式差分格式 (10.3.11) 稳定的条件是 $s = \frac{\tau}{h} < 1$.

证明 令 $u_{ij} = v_{jn} e^{in\omega x_i}$, 代入显式格式 (10.3.11) 对应的齐次方程

$$u_{i,j+1} = s^2 u_{i-1,j} + 2(1-s^2)u_{ij} + s^2 u_{i+1,j} - u_{i,j-1}.$$

约去公因式 $e^{in\omega x_i}$ 得

$$v_{j+1,n} = [s^2 e^{-in\omega h} + 2(1-s^2) + s^2 e^{in\omega h}]v_{jn} - v_{j-1,n}$$

或

$$v_{j+1,n} - 2 \left(1 - 2s^2 \sin \frac{n\omega h}{2} \right) v_{jn} + v_{j-1,n} = 0,$$

它对应特征方程

$$\lambda^2 - 2 \left(1 - 2s^2 \sin \frac{n\omega h}{2} \right) \lambda + 1 = 0.$$

它的两根之积为 1, 而判别式

$$\Delta = 16s^2 \sin^2 \frac{n\omega h}{2} \left(s^2 \sin^2 \frac{n\omega h}{2} - 1 \right).$$

由此可见

(1) 当 $s < 1$ 时, $\Delta < 0$, 两个根 λ_1, λ_2 都是复数, $|\lambda_1| = |\lambda_2| = 1 < 1 + c\tau$, 格式稳定;

(2) 当 $s = 1$ 时, 对个别 n 有 $\sin \frac{n\omega h}{2} = 0$, $\lambda_1 = \lambda_2 = 1$ 是重根, 但对一般 n 情况同上, 格式不稳定;

(3) 当 $s > 1$ 时, 两个根 λ_1, λ_2 都是实数, 因为 $\lambda_1 \lambda_2 = 1$, 必有 $|\lambda_1| > 1$ 或 $|\lambda_2| > 1$, 格式不稳定, 证毕.

2. 隐式差分格式

对 $\frac{\partial^2 u}{\partial x^2}$, 应用不同的数值微分公式得

$$\frac{\partial^2 u(x_i, t_{j+1})}{\partial x^2} \approx \frac{1}{h^2} (u(x_{i+1}, t_{j+1}) - 2u(x_i, t_{j+1}) + u(x_{i-1}, t_{j+1}))$$

和

$$\frac{\partial^2 u(x_i, t_{j-1})}{\partial x^2} \approx \frac{1}{h^2} (u(x_{i+1}, t_{j-1}) - 2u(x_i, t_{j-1}) + u(x_{i-1}, t_{j-1})),$$

两式相加除以 2, 代入 (10.3.10) 得差分方程

$$\begin{aligned} & \frac{1}{2h^2} (u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} + u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1}) \\ & - \frac{1}{\tau^2} (u_{i,j-1} - 2u_{i,j} + u_{i,j+1}) = f_{i,j}. \end{aligned} \quad (10.3.12)$$

它称为隐式格式.

类似于定理 10.3.1 的推导, 隐式差分格式 (10.3.12) 对应特征方程

$$\lambda^2 + \frac{2}{1 + 2s^2 \sin^2 \frac{n\omega h}{2}} \lambda + 1 = 0.$$

由于 $s = \tau/h > 0$, 特征方程判别式 $\Delta < 0$, 两个根 λ_1, λ_2 都是复数, $|\lambda_1| = |\lambda_2| = 1 < 1 + c\tau$, 格式稳定. 故有如下定理:

定理 10.3.2 隐式差分格式 (10.3.12) 无条件稳定.

10.4 有限元法

本节介绍求解偏微分方程定解问题的另一种数值方法——有限元方法. 这种方法特别适合几何、物理上比较复杂的问题. 作为有限元方法的初步学习, 为简单起见, 本节只讨论二阶椭圆边值问题的求解.

10.4.1 变分原理

以二维线性问题为例, 考虑如下问题:

$$\begin{cases} -\operatorname{div}(\operatorname{grad} u) = -\frac{\partial}{\partial x} \left(p \frac{\partial u}{\partial x} \right) - \frac{\partial}{\partial y} \left(p \frac{\partial u}{\partial y} \right) = f(x, y), & (x, y) \in \Omega, \\ u|_{\Gamma_1} = 0, \\ \left(p \frac{\partial u}{\partial n} + \sigma u \right) \Big|_{\Gamma_2} = g, \end{cases} \quad (10.4.1)$$

其中 Ω 为 $\mathbf{R}^2$ 中的连通区域, 其边界 $\partial\Omega = \Gamma_1 \cup \Gamma_2$ 分段光滑, n 为 $\partial\Omega$ 的外法线方法, $p(x, y)$ 一阶连续可导且 $p(x, y) \geq p_0 > 0$, $\sigma(x, y)$ 连续且 $\sigma(x, y) \geq 0$.

如果 $u(x, y) \in C^2(\Omega)$, 并且满足 (10.4.1), 则称 $u(x, y)$ 为问题 (10.4.1) 的古典解. 引入如下的内积和范数:

$$(u, v)_1 = \iint_{\Omega} (u_x v_x + u_y v_y + uv) dx dy,$$

$$\|u\|_1^2 = \iint_{\Omega} (u_x^2 + u_y^2 + u^2) dx dy,$$

从而得到希尔伯特空间 $H^1(\Omega)$, 再加上边界条件定义空间

$$H_0^1(\Omega) = \{v | v \in H^1(\Omega), v|_{\partial\Omega} = 0\},$$

再令

$$V = \{v | v \in H^1(\Omega), v|_{\Gamma_1} = 0\},$$

则 V 也是希尔伯特空间, 并且 $H_0^1(\Omega) \subset V \subset H^1(\Omega)$.

引进双线性泛函

$$B(u, v) = \iint_{\Omega} (pu_x v_x + pu_y v_y) dx dy + \int_{\Gamma_2} \sigma u v ds$$

和连续线性泛函

$$f(v) = \iint_{\Omega} f v dx dy + \int_{\Gamma_2} g v ds,$$

则得到原问题 (10.4.1) 相应的伽辽金变分问题

$$\begin{cases} \text{求 } u \in V, \text{ 使得} \\ B(u, v) = f(v), \quad \forall v \in V. \end{cases} \quad (10.4.2)$$

满足变分问题 (10.4.2) 的解称为原问题 (10.4.1) 的弱解, 将弱解所在的空间称为容许空间, 也称为试探空间, 将检验函数 v 所在的空间 V 称为检验空间. 由于变分问题 (10.4.2) 中已包含 Γ_2 上的条件, 因此, 边界 Γ_2 上的这类条件称为自然边界条件. 边界 Γ_1 上的条件称为本质边界条件, 也称为强加边界条件.

下面不加证明地给出问题 (10.4.1) 的古典解与弱解的关系.

定理 10.4.1 设 $u \in C^2(\Omega)$, 若 u 是 (10.4.1) 的古典解, 则 u 是 (10.4.2) 的弱解. 反之, 若 u 是 (10.4.2) 的弱解, 则 u 也是 (10.4.1) 的古典解.

伽辽金变分问题 (10.4.2) 解的存在性由 Lax-Milgram 定理保证.

定理 10.4.2 (Lax-Milgram 定理) 设 V 是一个希尔伯特空间, $B(u, v)$ 是 $V \times V$ 上的双线性泛函, 并且满足

- (1) 对称性: $B(u, v) = B(v, u) (\forall u, v \in V)$;
- (2) 连续性: $|B(u, v)| \leq M \|u\| \|v\| (\forall u, v \in V)$;
- (3) 强制性: $B(u, u) \geq \nu \|u\|^2 (\forall u \in V)$,

其中 M, ν 为不依赖于 u, v 的正常数, 而 f 为 V 上的连续线性泛函, 则伽辽金变分问题 (10.4.2) 存在唯一解 u^* , 并且有以下估计式:

$$\|u^*\| \leq \frac{1}{\nu} \|f\|_*,$$

其中 $\|\cdot\|_*$ 为 V 的对偶范数, 即 $\|f\|_* = \sup_{v \in V} \frac{(f, v)}{\|v\|}$.

引入二次泛函

$$J(v) = \frac{1}{2} B(v, v) - f(v),$$

则 (10.4.1) 的 Ritz 变分问题定义为

$$\begin{cases} \text{求 } u \in V, \text{ 使得} \\ J(u) = \min_{v \in V} J(v). \end{cases} \quad (10.4.3)$$

定理 10.4.3 设 V 是希尔伯特空间, $B(u, v)$ 是 $V \times V$ 上的双线性泛函, 并且满足定理 10.4.2 的条件, f 是 V 上的连续线性泛函, 则 (10.4.2) 和 (10.4.3) 两个问题中,

- (1) 任一个问题有解, 则解不多于一个;
- (2) 任一问题的解, 必是另一个问题的解;

(3) 如果 u^* 是它们的解, 则

$$J(v) - J(u^*) = \frac{1}{2}B(v - u^*, v - u^*), \quad \forall v \in V.$$

证明 先证明 (10.4.2) 的解是唯一的. 设 u_1, u_2 是 (10.4.2) 的解, 并且 $w = u_1 - u_2$. 由于

$$B(u_1, v) = f(v), \quad \forall v \in V,$$

$$B(u_2, v) = f(v), \quad \forall v \in V,$$

所以有

$$B(w, v) = 0, \quad \forall v \in V.$$

令 $v = w$, 则有 $B(w, w) = 0$. 由 B 的正定性可知 $w = 0$, 即 $u_1 = u_2$.

其次, 证明问题 (10.4.3) 的解也是 (10.4.2) 的解. 设 u 是 (10.4.3) 的解, 并设 $v \in V$, 并且 λ 为任一实数, 则

$$\begin{aligned} J(u + \lambda v) &= \frac{1}{2}B(u, u) + \lambda B(u, v) + \frac{1}{2}\lambda^2 B(v, v) - f(u) - \lambda f(v) \\ &\geq J(u) = \frac{1}{2}B(u, u) - f(u), \end{aligned}$$

即

$$\frac{1}{2}\lambda^2 B(v, v) + \lambda(B(u, v) - f(v)) \geq 0.$$

由 λ 的任意性, 故一定成立

$$B(u, v) = f(v), \quad \forall v \in V,$$

即 u 也是 (10.4.2) 的解.

最后, 证明问题 (10.4.2) 的解也是 (10.4.3) 的解. 设 u 是 (10.4.2) 的解, 则

$$\begin{aligned} J(v) - J(u) &= J(u + (v - u)) - J(u) \\ &= \frac{1}{2}B(u, u) + \frac{1}{2}B(v - u, v - u) - f(u) \\ &\quad + B(u, v - u) - f(v - u) - J(u). \end{aligned}$$

由于 $v - u \in V$ 及 u 是 (10.4.2) 的解得

$$B(u, v - u) - f(v - u) = 0,$$

由此可知

$$J(v) - J(u) = \frac{1}{2}B(v - u, v - u).$$

根据 B 的强制性得到

$$J(v) \geq J(u), \quad \forall v \in V,$$

即 u 也是 (10.4.3) 的解. 证毕.

在力学中, $J(u)$ 对应系统的总势能, 故称它为能量泛函. Ritz 变分形式对应最小位能原理, 而伽辽金变分形式对应虚功原理. 因此, 定理 10.4.3 说明, 在一定条件下, 它们反映了同一物理定律.

10.4.2 伽辽金逼近解

考察变分形式 (10.4.2) 的近似解法. 设 V_h 是 V 的有限维子空间, 当 $h \rightarrow 0$ 时, V_h 的维数无限增加, 并且满足逼近性质. 则 (10.4.2) 的伽辽金逼近解 $u_h \in V_h$, 使得

$$B(u_h, v) = f(v), \quad \forall v \in V_h. \quad (10.4.4)$$

设 $\{\varphi_i\}_{i=1}^n$ 是 V_h 的基函数系, 则有

$$u_h = \sum_{i=1}^n a^i \varphi_i, \quad v = \sum_{i=1}^n b^i \varphi_i.$$

将 u_h, v 代入 (10.4.4) 得

$$\sum_{i=1}^n b^i \left[\sum_{j=1}^n a^j B(\varphi_j, \varphi_i) - f(\varphi_i) \right] = 0,$$

由 v 的任意性有

$$\sum_{j=1}^n a^j B(\varphi_j, \varphi_i) = f(\varphi_i), \quad i = 1, 2, \dots, n.$$

令 $K_{ij} = B(\varphi_j, \varphi_i)$, $F_i = f(\varphi_i)$, 得到

$$\sum_{j=1}^n K_{ij} a^j = F_i, \quad i = 1, 2, \dots, n. \quad (10.4.5)$$

(10.4.5) 是对应于 (10.4.4) 的线性代数方程组, 称为有限元方程组, $(K_{ij})_{n \times n}$ 称为有限元刚度矩阵.

定理 10.4.4 设 $B(u, v)$ 是 $V \times V$ 上的双线性泛函, 并且满足定理 10.4.2 的条件, f 是 V 上的连续线性泛函, 则

- (1) 伽辽金逼近序列唯一存在;
- (2) 设 u 是 (10.4.1) 的解, u_h 是 (10.4.4) 的解, 则有

$$B(u - u_h, u - u_h) \leq B(u - v, u - v), \quad \forall v \in V_h, \quad (10.4.6)$$

$$B(u_h, u_h) \leq B(u, u), \quad (10.4.7)$$

$$B(u - u_h, v) = 0, \quad \forall v \in V_h; \quad (10.4.8)$$

(3) 伽辽金逼近解有如下先验估计:

$$\|u - u_h\|_V \leq \frac{M}{\nu} \inf_{v \in V_h} \|u - v\|. \quad (10.4.9)$$

证明 利用双线性形式 B 的强制性, 得到刚度矩阵对称正定, 从而有限元方程组 (10.4.5) 的解存在唯一.

由式 (10.4.2), (10.4.4) 得

$$B(u - u_h, v) = 0, \quad \forall v \in V_h,$$

式 (10.4.8) 得证.

令 $e = u - u_h$, 则 $\forall v \in V_h$ 有

$$\begin{aligned} B(u - v, u - v) &= B(u - u_h + u_h - v, u - u_h + u_h - v) \\ &= B(e, e) + 2B(e, u_h - v) + B(u_h - v, u_h - v) \\ &= B(e, e) + B(u_h - v, u_h - v). \end{aligned}$$

由 $B(u, v)$ 的强制性可得式 (10.4.6), 并且在上式中令 $v = 0$ 得

$$B(u, u) = B(e, e) + B(u_h, u_h),$$

从而推出式 (10.4.7).

再利用 $B(u, v)$ 的强制性,

$$\begin{aligned} \nu \|u - u_h\|^2 &\leq B(u - u_h, u - u_h) = B(u - u_h, u - v + v - u_h) \\ &= B(u - u_h, u - v) + B(u - u_h, v - u_h) \\ &\leq B(u - u_h, u - v) \leq M \|u - u_h\| \|u - v\|, \end{aligned}$$

从而得到式 (10.4.9). 证毕.

10.4.3 单元及形状函数

有限元基函数是由形状函数生成的, 它关系到有限元方法的收敛性和精度. 应用有限元方法, 首先要进行有限元剖分, 即将区域 Ω 分割为若干个子区域 $\Omega_1, \Omega_2, \dots, \Omega_L$, 并称之为单元. 对每一个单元 Ω_e , 设置 N_d 个节点 x^j , 并建立形状函数族

$$\psi_i(x), \quad i = 1, 2, \dots, N_d,$$

满足

(1) 在单元外部

$$\psi_i(x) = 0;$$

(2) 在 Ω_e 上 $\psi_i(x)$ 充分光滑, 并且

$$\psi_i(x^j) = \delta_i^j, \quad i = 1, 2, \dots, Nd. \quad (10.4.10)$$

确定了形状函数, 就可以很方便地来构造有限元基函数, 也就是说, 基函数系 $\varphi_\Delta(x)$ 可以分片地构造. 在每一片里, 它由形状函数构成. 因此, 对应于节点 x^Δ 的基函数, 可以表示为

$$\varphi_\delta(x) = \sum_{i=1}^{m(x^\Delta)} \frac{\psi_{N_i}^{e_i}(x)}{m(x)}, \quad (10.4.11)$$

其中 $m(x)$ 为点 x 的重数, N_i 为在单元 Ω_{e_i} 上节点 x^Δ 的局部编码. 于是函数 $u(x)$ 的有限元插值

$$u_h(x) = \sum_{\Delta=1}^n u^\Delta \varphi_\Delta(x) = \sum_{\Delta=1}^n u^\Delta \sum_{i=1}^{m(x^\Delta)} \frac{\psi_{N_i}^{e_i}(x)}{m(x)}, \quad (10.4.12)$$

代入有限元展开式得

$$\sum_{\sigma=1}^n \sum_{e_i, e_j \in R_{\Delta\sigma}} B(\psi_{N_i}^{e_i}, \psi_{N_j}^{e_j}) u^\sigma = f \left(\sum_{e_i \in R_\Delta} (\psi_{N_i}^{e_i}) \right), \quad \Delta = 1, 2, \dots, n, \quad (10.4.13)$$

其中 R_Δ 为节点 x^Δ 的影响节点集, $R_{\Delta\sigma}$ 为节点 x^Δ 和 x^σ 的相互影响节点集. 记 $K_{N_i N_j}^{e_i} = B(\psi_{N_i}^{e_i}, \psi_{N_j}^{e_j})$, 称为单元刚度矩阵. 可见, 刚度矩阵是由单元刚度矩阵叠加生成的.

下面以三角形单元为例, 介绍具体的形状函数.

三角形单元形状简单、适应区域的能力强, 并且采用面积坐标后, 面元的形状函数生成简单、容易标准化, 因而被广泛应用.

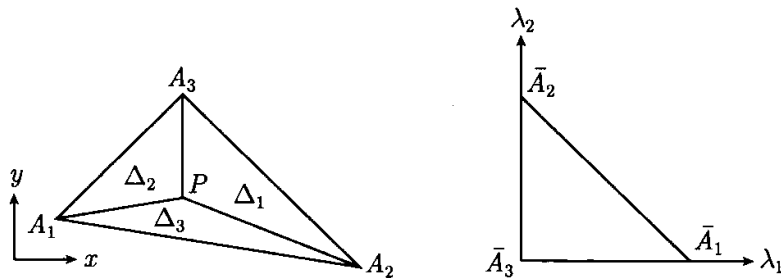


图 10.4 三角形元素面积坐标

如图 10.4 所示, 设任意三角形三个顶点 A_1, A_2, A_3 按逆时针排列, 三角形面积为 Δ . 取三角形内任一点 $P(x, y)$, 由点 P 向三个顶点 A_1, A_2, A_3 分别引直线, 将三角形分割成三个三角形 $PA_2A_3, PA_3A_1, PA_1A_2$, 分别记相应的面积为 $\Delta_1, \Delta_2, \Delta_3$. 令

$$\lambda_1 = \frac{\Delta_1}{\Delta}, \quad \lambda_2 = \frac{\Delta_2}{\Delta}, \quad \lambda_3 = \frac{\Delta_3}{\Delta}. \quad (10.4.14)$$

称 $\lambda_i (i = 1, 2, 3)$ 为 P 点的面积坐标. 三个顶点的面积坐标为

$$A_1(1, 0, 0), \quad A_2(0, 1, 0), \quad A_3(0, 0, 1).$$

若三个顶点的直角坐标为 $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, 并记

$$D = \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix},$$

$$a_i = y_j - y_k, \quad b_i = -(x_j - x_k), \quad c_i = \begin{vmatrix} x_j & x_k \\ y_j & y_k \end{vmatrix},$$

则面积坐标和直角坐标有如下关系:

$$\lambda_i = \frac{a_i x + b_i y + c_i}{|D|}, \quad i = 1, 2, 3, \quad (10.4.15)$$

$$\begin{cases} x = \sum_{i=1}^3 x_i \lambda_i, & y = \sum_{i=1}^3 y_i \lambda_i, \\ \lambda_1 + \lambda_2 + \lambda_3 = 1, & 0 \leq \lambda_i \leq 1. \end{cases} \quad (10.4.16)$$

下面给出三角形元在面积坐标下的形状函数. 取三角形三个顶点作为节点, 得到三角形线性元, 形状函数为

$$\psi_1(\lambda_1, \lambda_2, \lambda_3) = \lambda_1, \quad \psi_2(\lambda_1, \lambda_2, \lambda_3) = \lambda_2, \quad \psi_3(\lambda_1, \lambda_2, \lambda_3) = \lambda_3.$$

取三角形三个顶点和三条边的中点作为节点, 得到三角形二次元, 形状函数为

$$\psi_1 = \lambda_1(2\lambda_1 - 1), \quad \psi_2 = \lambda_2(2\lambda_2 - 1), \quad \psi_3 = \lambda_3(2\lambda_3 - 1),$$

$$\psi_4 = 4\lambda_2\lambda_3, \quad \psi_5 = 4\lambda_1\lambda_3, \quad \psi_6 = 4\lambda_1\lambda_2.$$

如果要求插值函数在 Ω 上有 $C^1(\Omega)$ 的连续性, 则可以考虑埃尔米特插值, 具体的形状函数参见文献 [17].

10.4.4 有限元求解步骤

综合前三节的描述, 有限元方法偏微分方程边值问题时, 大致经过如下几个步骤:

- (1) 寻找与原问题相适应的变分形式;
- (2) 对求解区域进行有限元剖分, 建立节点局部编码和总体编码的对应关系及节点坐标信息;
- (3) 建立有限元子空间, 即选择适当的剖分单元的形状函数;
- (4) 计算单元刚度矩阵、单元右端项, 并合成总刚度矩阵和总列阵;
- (5) 边界条件处理;
- (6) 有限元方程组求解;
- (7) 回到实际问题, 计算其他物理量.

例 10.4.1 考虑如下问题:

$$\begin{cases} -\frac{\partial}{\partial x} \left(\frac{\partial u}{\partial x} \right) - \frac{\partial}{\partial y} \left(\frac{\partial u}{\partial y} \right) + \alpha u = f, & (x, y) \in \Omega, \\ u = 0, & (x, y) \in \partial\Omega, \end{cases} \quad (10.4.17)$$

其中 $\alpha(x, y) = 5$, $f(x, y) = -\frac{1}{2} \sin(x + y)$, Ω 为图 10.5 中所示区域.

解 原问题相应的伽辽金变分问题为

$$\begin{cases} \text{求 } u \in V, \text{ 使得} \\ B(u, v) = f(v), \quad \forall v \in V, \end{cases}$$

其中

$$B(u, v) = \iint_{\Omega} (\nabla u \nabla v + \alpha uv) dx dy,$$

$$f(v) = \iint_{\Omega} f v dx dy.$$

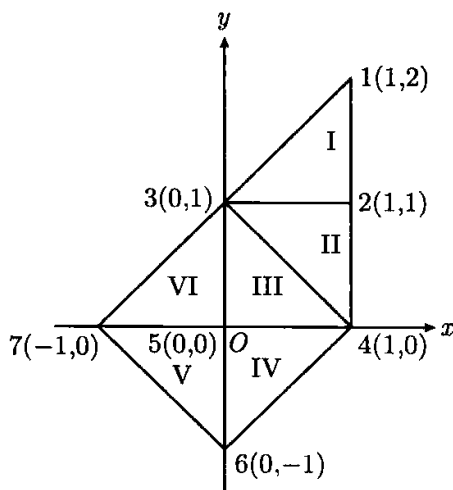


图 10.5 区域 Ω 剖分图

单元刚度矩阵和右端列阵分别为

$$K_{ij}^e = \iint_{\Delta_e} (\nabla \psi_i \nabla \psi_j + \alpha \psi_i \psi_j) dx dy,$$

$$F_i^e = \iint_{\Delta_e} f \psi_i dx dy.$$

用三角形单元剖分区域 Ω , 如图 10.5 所示, 采用线性三角形形状函数

$$\psi_1 = \lambda_1, \quad \psi_2 = \lambda_2, \quad \psi_3 = \lambda_3,$$

于是

$$K_{ij}^e = \iint_{\Delta_e} (\lambda_{ix} \lambda_{jx} + \lambda_{iy} \lambda_{jy} + \alpha \lambda_i \lambda_j) dx dy,$$

$$F_i^e = \iint_{\Delta_e} f \lambda_i dx dy.$$

由于

$$\lambda_i = \frac{a_i x + b_i y + c_i}{|D|},$$

从而得

$$\lambda_{ix} = \frac{a_i}{D}, \quad \lambda_{iy} = \frac{b_i}{D},$$

故

$$K_{ij}^e = \iint_{\Delta_e} \left(\frac{a_i a_j + b_i b_j}{D^2} + \alpha \lambda_i \lambda_j \right) dx dy,$$

再应用一点数值求积公式得

$$K_{ij}^e = \frac{a_i a_j + b_i b_j}{2D^2} + \frac{|D|}{18} \alpha(x_0, y_0).$$

利用有限元剖分中建立的节点局部编码和总体编码的对应关系及节点坐标信息, 并对 F_i^e 建立类似的计算公式, 从而得有限元方程组

$$\begin{pmatrix} 0.7778 & -0.2222 & 0.2778 \\ -0.2222 & 2.5556 & -0.4444 & -0.2222 \\ 0.2778 & -0.4444 & 3.1111 & 0.5556 & -0.4444 & 0 & 0.2778 \\ & -0.2222 & 0.5556 & 2.3333 & -0.4444 & 0.2778 \\ & & -0.4444 & -0.4444 & 5.1111 & -0.4444 & -0.4444 \\ & & & 0.2778 & -0.4444 & 1.5556 & 0.2778 \\ & & & & 0.2778 & 0 & -0.4444 & 0.2778 & 1.5556 \end{pmatrix} \times \begin{pmatrix} u^1 \\ u^2 \\ u^3 \\ u^4 \\ u^5 \\ u^6 \\ u^7 \end{pmatrix} = \begin{pmatrix} -0.0758 \\ -0.1568 \\ -0.2083 \\ -0.1325 \\ 0 \\ 0.0515 \\ 0.0515 \end{pmatrix}.$$

解之得

$$\begin{aligned} u^1 &= -0.0989, & u^2 &= -0.0859, & u^3 &= -0.0646, & u^4 &= -0.0546, \\ u^5 &= -0.0041, & u^6 &= 0.0351, & u^7 &= 0.0372. \end{aligned}$$

对于 Ω 内任意一点 (x_0, y_0) , 均可用形状函数插值求出其上的解 $u(x_0, y_0)$. 例如, $(x_0, y_0) = (0.7, 1.2)$, 它在第 1 个元素, 故有

$$u_h(x_0, y_0) = u^1 \lambda_1(x_0, y_0) + u^2 \lambda_2(x_0, y_0) + u^3 \lambda_3(x_0, y_0) = -0.2494,$$

$$u_{hx} = -0.0213, \quad u_{hy} = -0.0130.$$

小 结

本章介绍了几种典型偏微分方程定解问题的差分解法及椭圆型方程的有限元方法. 差分法的基本思想是用数值微分近似代替导数, 得出真解近似满足的代数方程组 (差分方程), 解此方程组得到真解的近似值. 这种方法也适用于更复杂的问题. 然而, 要使算出的解确是真解的近似值, 所用差分方程必须稳定、收敛、便于求解. 在很多情况下, 收敛的充要条件是稳定. 显式格式更容易实现, 但是条件稳定; 隐式格式每一步需求解代数方程组, 但隐式格式无条件稳定.

有限元方法建立在变分原理的基础上, 是通过求解有限维空间的变分方程来实现的. 弱解降低了对解的光滑性要求, 并且有限元法对求解区域的适应性更强, 而且有限元方法得到的总刚度矩阵是对称、正定、稀疏的, 更适合大规模问题的求解.

习 题

10.1 求解下列椭圆型方程边值问题:

$$(1) \begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, & 0 < x < 4, 0 < y < 3, \\ u|_{x=0} = y(3-y), & u|_{x=4} = 0, \\ u|_{y=0} = \sin \frac{\pi x}{4}, & u|_{y=3} = 0, \end{cases}$$

取 $h = \tau = 1$;

$$(2) \begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -1, & -1 < x < 1, -1 < y < 1, \\ \left(\frac{\partial u}{\partial n} + u \right) \Big|_{\Gamma} = 0, \end{cases}$$

取 $h = \tau = \frac{2}{3}$;

$$(3) \begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -1, & 0 < x < 1, 0 < y < 1, \\ u|_{y=0} = u|_{y=1} = 1, \\ \left(\frac{\partial u}{\partial x} + u \right) \Big|_{x=0} = 0, & u|_{x=1} = 1, \end{cases}$$

取 $h = \tau = \frac{1}{3}$;

$$(4) \begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, & x^2 + y^2 < 4, \\ u|_{\Gamma} = x^2 y^2, \end{cases}$$

其中 Γ 为圆周 $x^2 + y^2 = 4$, 取 $h = \tau = 1$.

10.2 求解下列抛物型方程定解问题:

$$(1) \begin{cases} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, & 0 < x < 1, 0 < t, \\ u(x, 0) = 4x(1-x), & 0 \leq x \leq 1, \\ u(0, t) = u(1, t) = 0, & 0 \leq t, \end{cases}$$

取 $h = 0.2, r = 1/6$;

$$(2) \begin{cases} \frac{\partial u}{\partial t} = e^x \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial u}{\partial x} \right), & 0 < x < 1, 0 < t, \\ u(x, 0) = 4x(1-x), & 0 \leq x \leq 1, \\ u(0, t) = u(1, t) = 0, & 0 \leq t, \end{cases}$$

取 $h = 1/3, r = 1$;

$$(3) \begin{cases} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, & 0 < x < 1, 0 < t, \\ u(x, 0) = 4x(1-x), & 0 \leq x \leq 1, \\ u(0, t) = u(1, t) = 0, & 0 \leq t, \end{cases}$$

取 $h = 1/3, r = 1$.

10.3 证明六点对称格式 (10.2.4) 无条件稳定.

10.4 对下列热传导方程定解问题, 列出显式差分方程.

$$\begin{cases} \frac{\partial u}{\partial t} - a^2(x) \frac{\partial^2 u}{\partial x^2} = 0, & 0 < x < 1, 0 < t < T, 0 < a_0 < a(x), \\ u(x, 0) = \varphi(x), & 0 \leq x \leq 1, \\ u(0, t) = u(1, t) = 0, & 0 \leq t \leq T. \end{cases}$$

证明当 $a(x) \leq b$ 时, 若 $\frac{b^2 \tau}{h^2} \leq \frac{1}{2}$, 则所列差分格式稳定, 并且在 $\frac{\partial^2 u}{\partial t^2}, \frac{\partial^4 u}{\partial x^4}$ 有界的条件下, 差分收敛.

10.5 求解下列双曲型方程混合问题:

$$(1) \begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, & 0 < x < 1, 0 < t, \\ u(x, 0) = \sin(\pi x), \frac{\partial u(x, 0)}{\partial t} = x(1-x), & 0 \leq x \leq 1, \\ u(0, t) = u(1, t) = 0, & 0 \leq t, \end{cases}$$

取 $s = 1, h = 0.2$;

$$(2) \begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, & 0 < x < 1, 0 < t, \\ u(x, 0) = \frac{1}{2}x(1-x), \frac{\partial u(x, 0)}{\partial t} = 0, & 0 \leq x \leq 1, \\ u(0, t) = u(1, t) = 0, & 0 \leq t, \end{cases}$$

取 $\tau = h = 0.1$.

10.6 列出下列定解问题的显式差分格式, 并指出稳定性条件:

$$(1) \begin{cases} \frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x, t), & 0 < x < l, 0 < t, \\ u(x, 0) = \varphi(x), \frac{\partial u(x, 0)}{\partial t} = \psi(x), & 0 \leq x \leq l, \\ u(0, t) = \mu_1(t), u(l, t) = \mu_2(t), & 0 \leq t; \end{cases}$$

$$(2) \begin{cases} \frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + bu, & 0 < x < l, 0 < t, \\ u(x, 0) = \varphi(x), \frac{\partial u(x, 0)}{\partial t} = \psi(x), & 0 \leq x \leq l, \\ u(0, t) = \mu_1(t), u(l, t) = \mu_2(t), & 0 \leq t; \end{cases}$$

$$(3) \begin{cases} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, & 0 < x < l, 0 < t, \\ u(x, 0) = \varphi(x), & 0 \leq x \leq l. \end{cases}$$

10.7 迎风差分格式 (10.3.2) 稳定的条件是 $\frac{|a|\tau}{h} \leq 1$.

10.8 Lax-Friedrichs 格式 (10.3.3) 稳定的条件是 $\frac{|a|\tau}{h} \leq 1$.

10.9 Lax-Wendroff 格式 (10.3.4) 稳定的条件是 $\frac{|a|\tau}{h} \leq 1$.

10.10 蛙跳格式 (10.3.5) 稳定的条件是 $\frac{|a|\tau}{h} < 1$. 当 $\frac{|a|\tau}{h} = 1$ 时, 蛙跳格式不稳定.

10.11 用有限元方法求解下列边值问题:

(1) 设 Ω 为图 10.6 中的三角形 $A_1A_2A_3$, 假定 Ω 不再剖分,

$$\begin{cases} -2\frac{\partial^2 u}{\partial x^2} - 2\frac{\partial^2 u}{\partial y^2} + 24u = 9xy + 4, \\ \left(\frac{\partial u}{\partial n} + 6u\right)\Big|_{A_1 A_2 \cup A_2 A_3} = 0, \\ \left(\frac{\partial u}{\partial n}\right)\Big|_{A_1 A_3} = 0; \end{cases}$$

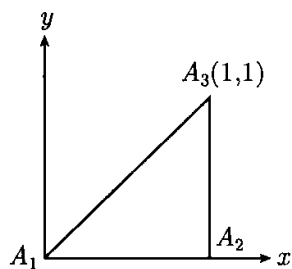


图 10.6 求解区域

(2) 设 Ω 为正方形, $|x| < 1, |y| < 1, \Gamma$ 为边界. 假定用对角线把 Ω 剖分成 4 个三角形,

$$\begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 1, & (x, y) \in \Omega, \\ u|_{\Gamma} = 0. \end{cases}$$

计 算 实 习

编程利用较小的 h 和 τ 求解习题 10.1, 习题 10.2, 习题 10.5, 习题 10.11.

(O-4470.0101)

科学出版中心 数理分社
电话: (010) 64033664
Email: math-phy@mail.sciencep.com

销售分类建议: 高等数学

www.sciencep.com

ISBN 978-7-03-032192-3



9 787030 321923 >

定 价: 65.00 元