高等学校研究生系列教材

数值分析基础教程

李庆扬　编

高等教育出版社

内容提要

本书是根据工程硕士"数值分析"课程教学基本要求和同等学力人员申请硕士学位全国统一考试"数值分析"大纲编写的.

主要内容有数值计算的误差,方程求根,解线性方程组的直接法与迭代法,插值与最小二乘法,数值积分,常微分方程数值解.每章附有习题其解答均在与该书配套的《数值分析复习与考试指导》(已由高等教育出版社出版)书中给出.书后还附有计算实验题.

本书适合作为工程硕士及同等学力人员申请硕士学位"数值分析"课程教材,也适合作为一般理工科研究生教材,还可以供科技人员学习参考.

图书在版编目(CIP)数据

数值分析基础教程/李庆扬编 .—北京:高等教育出版社,2001,5
ISBN 7-04-009850-4

Ⅰ.数… Ⅱ.李… Ⅲ.数值分析-研究生-教材
Ⅳ.0241

中国版本图书馆 CIP 数据核字(2001)第 10473 号

责任编辑 雷顺加 **封面设计** 王凌波 **责任绘图** 尹文军
版式设计 史新薇 **责任校对** 殷 然 **责任印制** 宋克学

数值分析基础教程
李庆扬 编

出版发行	高等教育出版社		
社 址	北京市东城区沙滩后街 55 号	**邮政编码**	100009
电 话	010-64054588	**传 真**	010-64014048
网 址	http://www.hep.edu.cn http://www.hep.com.cn		
经 销	新华书店北京发行所		
排 版	高等教育出版社照排中心		
印 刷	北京二二〇七工厂		
开 本	787×1092 1/16	**版 次**	2001 年 5 月第 1 版
印 张	9	**印 次**	2001 年 5 月第 1 次印刷
字 数	200 000	**定 价**	14.60 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换.

前　言

本书是根据工程硕士“数值分析”课程教学基本要求和同等学力人员申请硕士学位全国统一考试大纲编写的.根据此要求编写的《数值分析复习与考试指导》(以下简称《指导》),已由高等教育出版社于2000年6月出版.由于没有找到与该书相配套的《数值分析》教材,因此该书所列参考书目中没有一本是完全合适的,其中由李庆扬、王能超、易大义编、华中理工大学出版社出版的《数值分析》与大纲较为接近,但它的内容偏多、偏深,不完全适合工程硕士研究生的教学要求.因此,编写一本适合工程硕士教学要求,并与《指导》一书相匹配的数值分析教材是很有必要的.

作者近年多次为工程硕士研究生讲授“数值分析”课程,长期讲授各种不同要求的“数值分析”课程,写过多种计算数学教材并编写了《指导》一书,积累了一些经验,能较好地掌握教学要求,因而编写的这本教材内容与《指导》一书完全配套,采用的符号也尽可能统一,书中习题也基本上采用了《指导》给出的习题.另外,为使学生更好地掌握数值计算方法,必须进行适当的上机实验,书后给出的数值实验题,供学生选做.

本书适合工程硕士及同等学力人员申请硕士学位的“数值分析”课程教材,也适合作为一般理工科院校研究生“数值分析”课程的教材(全部讲授完约需64学时),还可供科技人员学习参考.

本书编写得到高等教育出版社计算机编辑室的大力支持,在此表示衷心的感谢.希望使用本书的广大读者和教师,对本书缺点和不足之处提出批评指正.

作者

2001年2月

目　录

第一章 绪 论

1.1 “数值分析”研究对象与特点

“数值分析”是计算数学的一个主要部分.而计算数学是数学科学的一个分支,它研究用计算机求解数学问题的数值计算方法及其软件实现.计算数学几乎与数学科学的一切分支有联系,它利用数学领域的成果发展了新的更有效的算法及其理论,反过来很多数学分支都需要探讨和研究适用于计算机的数值方法.因此,“数值分析”内容十分广泛.但本书作为“数值分析”基础,只介绍科学与工程计算中最常用的基本数值方法,包括线性方程组与非线性方程求根、插值与最小二乘拟合、数值积分与常微分方程数值解法等.这些都是计算数学中最基础的内容.

近几十年来由于计算机的发展及其在各技术科学领域的应用推广与深化,新的计算性学科分支纷纷兴起,如计算力学、计算物理、计算化学、计算经济学等等,不论其背景与含义如何,要用计算机进行科学计算都必须建立相应的数学模型,并研究其适合于计算机编程的计算方法.因此,计算数学是各种计算性科学的联系纽带和共性基础,是一门兼有基础性、应用性和边缘性的数学学科.

计算数学作为数学科学的一个分支,当然具有数学科学的抽象性与严密科学性的特点,但它又具有广泛的应用性和边缘性特点.

现代科学发展依赖于理论研究、科学实验与科学计算三种主要手段,它们相辅相成,互相独立,可以互相补充又都不可缺少,作为三种科学研究手段之一的科学计算是一门工具性、方法性、边缘性的新学科,发展迅速,它的物质基础是计算机(包括其软硬件系统),其理论基础主要是计算数学.

计算数学与计算工具发展密切相关,在计算机出现以前,数值计算方法只能计算规模小的问题,并且也没形成单独的学科,只有在计算机出现以后,数值计算才得以迅速发展并成为数学科学中一个独立学科——计算数学.当代计算能力的大幅度提高既来自计算机的进步,也来自计算方法的进步,计算机与计算方法的发展是相辅相成、互相促进的.计算方法的发展启发了新的计算机体系结构,而计算机的更新换代也对计算方法提出了新的标准和要求.例如为在计算机上求解大规模的计算问题、提高计算效率,诞生并发展了并行计算机.自计算机诞生以来,经典的计算方法业已经历了一个重新评价、筛选、改造和创新的过程,与此同时,涌现了许多新概念、新课题和能充分发挥计算机潜力、有更大解题能力的新方法,这就构成了现代意义下的计算数学.这也是数值分析的研究对象与特点.

概括地说,数值分析是研究适合于在计算机上使用的实际可行、理论可靠、计算复杂性好的

数值计算方法.具体说就是:

第一,面向计算机,要根据计算机特点提供实际可行的算法,即算法只能由计算机可执行的加减乘除四则运算和各种逻辑运算组成.

第二,要有可靠的理论分析,数值分析中的算法理论主要是连续系统的离散化及离散型方程数值求解.有关基本概念包括误差、稳定性、收敛性、计算量、存储量等,这些概念是刻画计算方法的可靠性、准确性、效率以及使用的方便性.

第三,要有良好的复杂性及数值试验,计算复杂性是算法好坏的标志,它包括时间复杂性(指计算时间多少)和空间复杂性(指占用存储单元多少).对很多数值问题使用不同算法,其计算复杂性将会大不一样,例如对20阶的线性方程组若用代数中的Cramer法则作为算法求解,其乘除法运算次数需要9.7×10^{20}次,若用每秒运算1亿次的计算机计算也要30万年,这是无法实现的,而用"数值分析"中介绍的Gauss消去法求解,其乘除法运算次数只需3 060次,这说明选择算法的重要性.当然有很多数值方法不可能事先知道其计算量,故对所有数值方法除理论分析外,还必须通过数值试验检验其计算复杂性.本课程虽然只着重介绍数值方法及其理论,一般不涉及具体的算法设计及编程技巧,但作为基本要求仍希望读者能适当做一些计算机上的数值试验,它对加深算法的理解是很有好处的.

1.2 数值计算的误差

1.2.1 误差来源与分类

用计算机求解科学计算问题,首先由物理模型转化为数学模型时将会产生模型误差,还有许多物理量如温度、重量、长度等等都由观测得到,显然也会产生误差,这称为观测误差.所有这些不属于数学问题产生的误差,均不在"数值分析"讨论的范围内.本书只研究数值求解数学问题时产生的误差,它们主要有以下三类.

第一类是截断误差或方法误差,它是指将数学问题转化为数值计算问题时产生的误差,通常是用有限过程近似无限过程时产生的误差.例如,计算$f(x)=e^x$的值,用Taylor公式展开前4项

$$e^x\approx1+x+\frac{x^2}{2}+\frac{x^3}{6}=p_3(x)$$

当$|x|<1$时其截断误差为

$$R_3(x)=e^x-p_3(x)=\frac{1}{4!}e^{\xi}x^4,\quad \xi\text{在}0\text{与}x\text{之间}$$

截断误差将结合有关数值方法进行讨论,数值方法的误差估计指的就是这类误差.

第二是舍入误差,数值计算时由于计算是有限位的,所以原始数据、中间结果和最后结果都要舍入,这就产生舍入误差.在十进制运算中一般采用四舍五入.例如$\frac{1}{3}$写成0.333 3,$\pi\approx3.141\ 6$等等,都有舍入误差.

第三类是输入数据误差,称为初始误差,这些误差对计算也将造成影响,但分析初始误差与对舍入误差分析相似,因此可将它归入第二类.

由于对大规模数值计算问题舍入误差目前尚无有效方法进行定量估计,所以我们主要进行定性分析.但对误差估计的基本概念及较简单的数值运算误差估计还需作简单介绍.

1.2.2 误差与有效数字

定义 2.1 设准确值 x 的近似值为 x^*,则 $\varepsilon = x - x^*$ 称为近似值 x^* 的绝对误差,简称误差,$\varepsilon_r = \dfrac{\varepsilon}{x}$ 称为近似值 x^* 的相对误差.

绝对误差可正可负,一般说 ε 的准确值很难求出,往往只能求 $|\varepsilon|$ 的一个上界 δ,即 $|\varepsilon| = |x - x^*| \leqslant \delta(x^*)$,称为 x^* 的误差限.相对误差 ε_r 当 $x=0$ 时没有意义,且准确值 x 往往未知,故常用 $\dfrac{x - x^*}{x^*}$ 作为相对误差,而称 $\delta_r(x^*) = \dfrac{\delta(x^*)}{|x^*|}$ 为相对误差限.

例 1.1 已知 $\pi = 3.141\,592\,6\cdots$,若取近似数为 $x^* = 3.14$,则 $\varepsilon = \pi - x^* = 0.001\,592\,6\cdots$,$|\varepsilon| \leqslant 0.002 = \delta(x^*)$,为 x^* 的误差限,而相对误差限 $\delta_r(x^*) = \dfrac{\delta(x^*)}{3.14} < 0.007$.

通常在 x 有多位数字时,若取前有限位数的数字作近似值,都采用四舍五入原则,例如,$x = \pi$ 取 3 位 $x^* = 3.14$,$\varepsilon \leqslant 0.002$;取 5 位 $x^* = 3.141\,6$,$\varepsilon \leqslant 0.000\,01$ 它们的误差限都不超过近似数 x^* 末位数的半个单位,即

$$|\pi - 3.14| \leqslant \frac{1}{2} \times 10^{-2}, \quad |\pi - 3.141\,6| \leqslant \frac{1}{2} \times 10^{-4}$$

定义 2.2 设 x^* 是 x 的一个近似数,表示为

$$x^* = \pm 10^k \times 0.a_1 a_2 \cdots a_n \tag{1.2.1}$$

每个 $a_i (i = 1,2\cdots,n)$ 均为 0,1,…,9 中的一个数字,$a_1 \neq 0$,如果 $|x - x^*| \leqslant \dfrac{1}{2} \cdot 10^{k-n}$,则称 x^* 近似 x 有 n 位有效数字.

例如,用 3.14 近似 π 有 3 位有效数字,用 3.141 6 近似 π 有 5 位有效数字.

显然,近似数的有效位数越多,相对误差限就越小,反之也对.

定理 2.1 设 x 的近似数 x^* 表示为式(1.2.1),如果 x^* 具有 n 位有效数字,则其相对误差限为

$$\frac{|x - x^*|}{|x^*|} \leqslant \delta_r(x^*) = \frac{1}{2a_1} \times 10^{-(n-1)} \tag{1.2.2}$$

反之,若

$$\frac{|x - x^*|}{|x^*|} \leqslant \frac{1}{2(a_1 + 1)} \times 10^{-n+1} \tag{1.2.3}$$

则 x^* 至少具有 n 位有效数字.

证明 由式(1.2.1)可得

$$a_1\times10^{k-1}\leqslant|x^*|\leqslant(a_1+1)\times10^{k-1}$$

所以当 x^* 有 n 位有效数字时

$$\varepsilon_r(x^*)=\frac{|x-x^*|}{|x^*|}\leqslant\frac{\frac{1}{2}\times10^{k-n}}{a_1\times10^{k-1}}=\frac{1}{2a_1}\times10^{-(n-1)}$$

反之,由式(1.2.3)有

$$|x-x^*|=|x^*|\varepsilon_r(x^*)\leqslant(a_1+1)\times10^{k-1}\times\frac{1}{2(a_1+1)}\times10^{-n+1}$$

$$=\frac{1}{2}\times10^{k-n}$$

故 x^* 有 n 位有效数字. 证毕.

例 1.2 下列近似数有几位有效数字? 其相对误差限是多少?

(1) $x=e\approx2.718\,28=x^*$,(2) $x=0.030\,021\approx0.030\,0=x^*$.

解(1) 由 $|e-2.718\,28|\leqslant\frac{1}{2}\times10^{-5}$,因 $k=1$,故 $n=6$ 有 6 位有效数字.因 $a_1=2$,相对误差限 $\delta_r(x^*)=\frac{1}{4}\times10^{-5}$.

(2) $|x-0.030\,0|\leqslant\frac{1}{2}\times10^{-4}$,因 $k=-1$,故 $n=3$,即有 3 位有效数字,由 $a_1=3$ 知 $\delta_r(x^*)\leqslant\frac{1}{6}\times10^{-2}$.

1.2.3 函数计算的误差估计

设一元函数 $f(x)$具有二阶导数,自变量 x 的一个近似值 x^*,$f(x)$的近似值为 $f(x^*)$,用 $f(x)$在 x^* 点的 Taylor 展开估计误差,可得

$$|f(x)-f(x^*)|\leqslant|f'(x^*)(x-x^*)|+\frac{1}{2}|f''(\xi)(x-x^*)^2|$$

其中 ξ 在 x 与 x^* 之间,如果 $f'(x^*)\neq0$,$|f''(\xi)|$ 与 $|f'(x^*)|$ 有相同数量级,而 $\delta(x^*)\geqslant|x-x^*|$很小,则可得

$$\delta f(x^*)\approx|f'(x^*)|\delta(x^*),\delta_r f(x^*)\approx\left|\frac{f'(x^*)}{f(x^*)}\right|\delta(x^*)\tag{1.2.4}$$

分别为 $f(x^*)$的一个近似误差限与相对误差限.

如果 f 为多元函数,自变量为 $x_1,\cdots,x_n$,其近似值为 $x_1^*,\cdots,x_n^*$,则类似于一元函数可用多元函数 $f(x_1,x_2,\cdots,x_n)$的 Taylor 展开,取一阶近似得误差限

$$\delta f(x_1^*,\cdots,x_n^*)\approx\sum_{i=1}^{n}\left|\frac{\partial f(x_1^*,\cdots,x_n^*)}{\partial x_i}\right|\delta(x_i^*)\tag{1.2.5}$$

及相对误差限

$$\delta_r f(x_1^*,\cdots,x_n^*)\approx\sum_{i=1}^{n}\left|\frac{\partial f(x_1^*,\cdots,x_n^*)}{\partial x_i}\right|\frac{\delta(x_i^*)}{|f(x_1^*,\cdots,x_n^*)|}\tag{1.2.6}$$

若把(1.2.5)用到两个或多个数的算术运算中,则可得到近似数 x_1^* 及 x_2^* 的四则运算误差估计:

$$\delta(x_1^* \pm x_2^*) = \delta(x_1^*) + \delta(x_2^*) \tag{1.2.7}$$

$$\delta(x_1^* x_2^*) = |x_1^*|\delta(x_2^*) + |x_2^*|\delta(x_2^*) \tag{1.2.8}$$

$$\delta\left(\frac{x_1^*}{x_2^*}\right) = \frac{|x_1^*|\delta(x_2^*) + |x_2^*|\delta(x_1^*)}{|x_2^*|^2}, x_2^* \neq 0 \tag{1.2.9}$$

例 1.3 已测得某场地长 l 的值为 $l^* = 110$ m,宽 d 的值为 $d^* = 80$ m,已知 $|l - l^*| \leqslant 0.2$ m, $|d - d^*| \leqslant 0.1$ m,试求面积 $S = ld$ 的绝对误差限与相对误差限.

解 因 $S = ld$, $\frac{\partial S}{\partial l} = d$, $\frac{\partial S}{\partial d} = l$,由(1.2.5)知

$$\delta(S^*) = \left|\left(\frac{\partial S}{\partial l}\right)^*\right|\delta(l^*) + \left|\left(\frac{\partial S}{\partial d}\right)^*\right|\delta(d^*)$$

其中 $\left(\frac{\partial S}{\partial l}\right)^* = d^* = 80$ m, $\left(\frac{\partial S}{\partial d}\right)^* = l^* = 110$ m, $\delta(l^*) = 0.2$ m, $\delta(d^*) = 0.1$ m,从而有

$$\delta(S^*) = 80 \times 0.2 + 110 \times 0.1 = 27 \text{ m}^2$$

相对误差限为

$$\delta_r(S^*) = \frac{\delta(S^*)}{|S^*|} = \frac{27}{80 \times 110} = 0.31\%$$

1.3 误差定性分析与避免误差危害

上面给出的误差估计方法只对运算量很少的情形适用,对大规模数值计算的舍入误差估计目前尚无有效的方法做出定量估计,为了确保数值计算结果的正确性,应对数值计算问题进行定性分析,以保证其舍入误差不会影响计算的精度,这就是本节要讨论的问题.

1.3.1 病态问题与条件数

对一个数值问题,往往由于问题本身而使计算结果相对误差很大,这种问题就是病态问题.例如计算函数值 $f(x)$,若 x 的近似值为 x^*,其相对误差为 $\frac{x - x^*}{x}$,函数值 $f(x^*)$ 的相对误差为 $\frac{f(x) - f(x^*)}{f(x)}$,它们相对误差之比的绝对值为

$$\left|\frac{[f(x) - f(x^*)]/f(x)}{(x - x^*)/x}\right| \approx \left|\frac{xf'(x)}{f(x)}\right| = C_p \tag{1.3.1}$$

C_p 称为计算函数值 $f(x)$ 的条件数,如果 C_p 很大,将引起函数值 $f(x^*)$ 的相对误差很大,出现这种情况时,就认为问题是病态的.例如 $f(x) = x^n$, $f'(x) = nx^{n-1}$,则 $C_p = n$,它表示相对误差可能放大 n 倍.如 $n = 10$,有 $f(1) = 1$, $f(1.02) \approx 1.24$,若 $x = 1$, $x^* = 1.02$,则自变量相对误差为 2%,而函数值 $f(1.02)$ 的相对误差为 24%,这时就认为问题是病态的.一般情况下若条件数

$C_p \geqslant 10$,则认为是问题病态,C_p 越大病态越严重.

其他计算问题也要分析是否病态,例如解线性方程组,如果输入数据有微小误差,引起解的误差绝对值很大,就认为是病态方程组.

例 1.4 求解方程组

$$\begin{cases} x + \alpha y = 1 \\ \alpha x + y = 0 \end{cases}$$

解 当 $\alpha = 1$,系数矩阵奇异,方程无解,当 $\alpha \neq 1$,解为 $x = \dfrac{1}{1-\alpha^2}$, $y = \dfrac{\alpha}{1-\alpha^2}$,当 $\alpha \approx 1$,若输入数据 α 有误差,则解的误差很大,例如,当 $\alpha = 0.99$,解 $x \approx 50.25$,当 α 有误差,$\alpha^* = 0.991$,则解 $x^* \approx 55.81$,误差 $|x - x^*| \approx 5.56$ 很大,问题病态.对此例中 $x = (1-\alpha^2)^{-1}$ 应用式(1.3.1)求 C_p 得

$$C_p = \left| \frac{\alpha x'(\alpha)}{x(\alpha)} \right| = \left| \frac{2\alpha^2}{1-\alpha^2} \right|$$

当 $\alpha = 0.99$ 时 $C_p \approx 100$,故问题病态.只当 $|\alpha| \ll 1$ 时问题才为良态.

1.3.2 算法的数值稳定性

一个数值方法进行计算时,由于原始数据有误差,在计算中这些误差会传播,有时误差增长很快使计算结果误差很大,影响了结果不可靠.

定义 3.1 一个算法如果原始数据有扰动(即误差),而计算过程舍入误差不增长,则称此算法是数值稳定的.否则,若误差增长则称算法不稳定.

例 1.5 计算积分 $I_n = \int_0^1 x^n e^{x-1} dx$, $n = 0, 1, \cdots$.

当 $n = 0$ 时,$I_0 = 1 - e^{-1}$,对 I_n 用分部积分法得

$$I_n = 1 - nI_{n-1}, n = 1, 2, \cdots \tag{1.3.2}$$

若计算 I_0 时取 $e^{-1} \approx 0.367\,9$ 由(1.3.2)依次计算 $I_1, \cdots, I_9$ 考察其计算结果是否正确.

解 由 $I_0 = 1 - e^{-1} \approx 0.632\,1 = I_0^*$,故有误差

$\varepsilon_0 = I_0 - I_0^*$,由式(1.3.2),计算 $I_n^* = 1 - nI_{n-1}^*$, $n = 1, 2, \cdots, 9$ 得

$$\begin{aligned} &I_1^* = 0.367\,9, && I_2^* = 0.264\,2, && I_3^* = 0.207\,4, \\ &I_4^* = 0.170\,4, && I_5^* = 0.148\,0, && I_6^* = 0.112\,0, \\ &I_7^* = 0.216\,0, && I_8^* = -0.728\,0, && I_9^* = 7.552. \end{aligned}$$

显然,结果不正确,因为 $I_n > 0$,而 $I_8^* < 0$.实际上,$\varepsilon_n = I_n - I_n^* = -n(I_{n-1} - I_{n-1}^*) = \cdots = (-1)^n n! \varepsilon_0$,当 n 增大时 ε_n 是递增的,且 I_9^* 的误差达到 $-9!\varepsilon_0$,是严重失真的.它表明式(1.3.2)给出的算法是不稳定的.如果在式(1.3.2)中将算法改为

$$I_{n-1}=\frac{1}{n}(1-I_n),n=9,8,\cdots,2,1 \tag{1.3.3}$$

由于 $\frac{1}{n+1}e^{-1}\leqslant I_n\leqslant\frac{1}{n+1}$，取 $I_n\approx\frac{1}{2}\frac{1}{n+1}(e^{-1}+1)$

当 $n=9,I_9=\frac{1}{20}(1+e^{-1})\approx0.068\ 4=\tilde{I}_9$，再由式(1.3.3)，可求出 $\tilde{I}_8=0.103\ 5,\cdots,\tilde{I}_2=0.264\ 3$，$\tilde{I}_1=0.267\ 3,\tilde{I}_0=0.632\ 1$，此时 $\tilde{\varepsilon}_{n-1}=I_{n-1}-\tilde{I}_{n-1}=-\frac{1}{n}(I_n-\tilde{I}_n),|\tilde{\varepsilon}_0|=\frac{1}{n!}|\tilde{\varepsilon}_n|$，计算是稳定的.

数值不稳定的算法是不能使用的.

1.3.3 避免误差危害的若干原则

数值计算中除了要分清问题是否病态和算法的数值稳定性外，还应尽量避免误差危害.通常运算中应注意以下若干原则：

(1) 避免用绝对值很小的数做除法.

(2) 避免两个相近数相减，以免有效数字损失.

(3) 注意运算次序，防止大数"吃掉"小数，如多个数相加减，应按绝对值由小到大的次序运算.

(4) 简化计算步骤，尽量减少运算次数.

为了说明以上原则，下面给出一些例题.

例 1.6 求 $x^2-16x+1=0$ 的小正根.

解 方程的两根为 $x_1=8-\sqrt{63}$ 及 $x_2=8+\sqrt{63}$，小正根 $x_1=8-\sqrt{63}\approx8-7.94=0.06$ 只有一位有效数字.为避免两相近数相减可改用 $x_1=8-\sqrt{63}=\frac{1}{8+\sqrt{63}}\approx\frac{1}{15.94}\approx0.062\ 7$ 仍有三位有效数字.

例 1.7 求 $x=1-\cos 2^\circ\approx1-0.999\ 4=0.000\ 6$，若改用

$1-\cos 2^\circ=\frac{(\sin 2^\circ)^2}{1+\cos 2^\circ}\approx\frac{(0.034\ 90)^2}{1.999\ 4}\approx6.092\times10^{-4}$，具有四位有效数字.

例 1.8 计算多项式

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

的值，若直接计算 $a_kx^k(k=1,\cdots,n)$ 再逐项相加，需进行 $1+2+\cdots+n=\frac{n(n+1)}{2}$ 次乘法和 n 次加法，若采用以下算法：

$$S_n=a_n,S_k=xS_{k+1}+a_k,k=n-1,n-2,\cdots,1,0 \tag{1.3.4}$$

只需 n 次乘法和 n 次加法，则得 $p(x)=S_0$，此算法称为秦九韶算法，也称霍纳(Horner)算法.(秦九韶于 1247 年提出此算法.比霍纳 1819 年提出此算法早 500 多年).

习 题 一

1. 设 $x>0$，x^* 的相对误差限为 δ，求 $f(x)=\ln x$ 的误差限.

2. 下列各数都是经过四舍五入得到的近似值，试指出它们有几位有效数字，并给出其误差限与相对误差限.

$$x_1^*=1.102\,1,\quad x_2^*=0.031,\quad x_3^*=560.40$$

3. 下列公式如何计算才比较准确？

(1) $\dfrac{e^{2x}-1}{2}$，$|x|\ll 1$

(2) $\displaystyle\int_N^{N+1}\frac{1}{1+x^2}\mathrm{d}x$，$N\gg 1$

(3) $\sqrt{x+\dfrac{1}{x}}-\sqrt{x-\dfrac{1}{x}}$，$|x|\gg 1$

4. 序列 $\{y_n\}$ 满足递推关系 $y_n=10y_{n-1}-1$，$n=1,2,\cdots$，若 $y_0=\sqrt{2}\approx 1.41$，计算到 y_{10} 时误差有多大？这个计算数值稳定吗？

5. 计算 $x=(\sqrt{2}-1)^6$，取 $\sqrt{2}\approx 1.41$，直接计算 x 和利用以下等式

$$\frac{1}{(\sqrt{2}+1)^6},\quad (3-2\sqrt{2})^3,\quad \frac{1}{(3+2\sqrt{2})^3},\quad 99-70\sqrt{2}$$

计算，哪一个最好？

第二章　方程求根

2.1　方程求根与二分法

2.1.1　引言

单个变量的方程

$$f(x)=0 \tag{2.1.1}$$

求根是数值计算经常遇到的问题.当 $f(x)$ 为一般连续函数时,称式(2.1.1)为超越方程,如果 f 为多项式

$$f(x)=p(x)=a_0x^n+a_1x^{n-1}+\cdots+a_{n-1}x+a_n=0 \tag{2.1.2}$$

若 $a_0\neq0$,$p(x)$ 为 n 次多项式,此时方程(2.1.1)称为代数(或多项式)方程.如果 x^*(实数或复数)使 $f(x^*)=0$,则称 x^* 为方程(2.1.1)的根,若 $f(x)=(x-x^*)^mg(x)$,$m\geqslant1$,且 $g(x^*)\neq0$,当 $m>1$ 时,称 x^* 为方程(2.1.1)的 m 重根或称 x^* 是 f 的 m 重零点.若 x^* 是 f 的 m 重零点,且 g 充分光滑,则 $f(x^*)=f'(x^*)=\cdots=f^{(m-1)}(x^*)=0$.当 f 为式(2.1.2)表示的代数多项式时,根据代数基本定理可知方程(2.1.1)有 n 个根(含复根,m 重根为 m 个根),对 $n=2$ 的代数方程

$$ax^2+bx+c=0$$

它的根可由公式表示为

$$x_{1,2}=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

而当 $n=3,4$ 时方程的根虽可用公式表示,但表达式太复杂,一般不用,当 $n\geqslant5$ 已没有直接用公式表达的求根算法.因此对 $n\geqslant3$ 的代数方程求根方法与一般超越方程(2.1.1)一样都采用迭代方法求根,设 $f\in C[a,b]$(表示 f 在区间 $[a,b]$ 上连续),若有 $f(a)f(b)<0$,则 $f(x)=0$ 在区间 $[a,b]$ 上至少有一个实根,$[a,b]$ 称为有根区间,通常可用逐次搜索法求得方程(2.1.1)的有根区间.

例 2.1　求方程 $f(x)=x^3-11.1x^2+38.8x-41.77=0$ 的有根区间.

解　根据有根区间定义,对方程的根进行搜索计算,结果如下表:

x	0	1	2	3	4	5	6
$f(x)$符号	−	·	+	+	−	−	+

方程的三个有根区间为[1,2],[3,4],[5,6].

2.1.2 二分法

设 $f\in C[a,b]$,且$[a,b]$为有根区间,取中点 $x_0=\dfrac{a+b}{2}$,将它分为两半,检查$f(x_0)$与 $f(a)$是否同号,若是,说明根 x^* 仍在 x_0 右侧,取 $a_1=x_0$, $b_1=b$,否则取 $a_1=a$,$b_1=x_0$,得到新的有根区间$[a_1,b_1]$长度仅为$[a,b]$的一半(见图 2-1).重复以上过程,即取 $x_1=\dfrac{a_1+b_1}{2}$,将$[a_1,b_1]$再分半,确定根在 x_1 的哪一侧,得到新区间$[a_2,b_2]$,其长度为$[a_1,b_1]$的一半,从而可得一系列有根区间

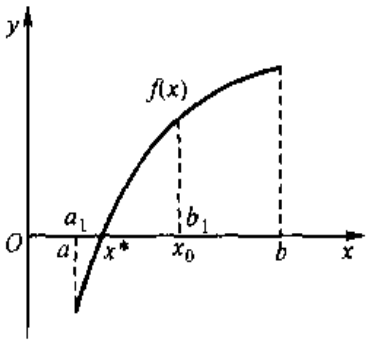

图 2-1

$$[a,b]\supset[a_1,b_1]\supset[a_2,b_2]\supset\cdots\supset[a_n,b_n]\supset\cdots$$

其中每一个区间长度都是前一个区间长度的一半,因此,$[a_n,b_n]$的长度为

$$b_n-a_n=\frac{b-a}{2^n}$$

且 $\lim\limits_{n\to\infty}x_n=\lim\limits_{n\to\infty}\dfrac{a_n+b_n}{2}=x^*$,$x_n$ 即为方程(2.1.1)的根 x^* 的一个足够精确的近似根,且误差为

$$|x_n-x^*|\leqslant\frac{b_n-a_n}{2}=\frac{b-a}{2^{n+1}}\tag{2.1.3}$$

以上过程称为解方程的二分法.它计算简单且收敛性有保证,但收敛较慢,通常可用于求迭代法的一个足够好的初始近似.

例 2.2 求方程 $f(x)=x^3-x-1=0$ 在区间[1.0,1.5]内的一个实根,要求准确到小数点后第二位.

解 这里 $a=1.0$,$b=1.5$,而 $f(a)<0$,$f(b)>0$,故在[1.0,1.5]中有根,由式(2.1.3)知

$|x_n-x^*|\leqslant\dfrac{b-a}{2^{n+1}}=\dfrac{1}{2^{n+2}}\leqslant\dfrac{1}{2}\times10^{-2}$,即 $2^{n+1}\geqslant10^2$,当 $n=6$ 时得 $|x_6-x^*|\leqslant0.005$,已达到精度要求,各次计算结果见表 2-1.

表 2-1

n	a_n	b_n	x_n	$f(x_n)$的符号
0	1.0	1.5	1.25	-
1	1.25		1.375	+
2		1.375	1.312 5	-
3	1.312 5		1.343 8	+
4		1.343 8	1.328 2	+
5		1.328 2	1.320 4	-
6	1.320 4		1.324 3	-

故 $x_6=1.324$ 为方程的近似根，误差不超过 0.005.

2.2 迭代法及其收敛性

2.2.1 不动点迭代法

求方程(2.1.1)的根时将方程改写为等价形式

$$x=\varphi(x) \tag{2.2.1}$$

若 x^* 满足 $x^*=\varphi(x^*)$，则称 x^* 为 φ 的一个不动点，x^* 也是方程(2.1.1)的一个根，如果 φ 连续，可构造迭代法

$$x_{k+1}=\varphi(x_k),\quad k=0,1,\cdots \tag{2.2.2}$$

称为不动点迭代法，φ 称为迭代函数.若给定初始近似 x_0，由式(2.2.2)逐次迭代得到序列$\{x_k\}$，如果 $\lim\limits_{k\to\infty}x_k=x^*$，则由式(2.2.2)两端取极限，得

$$x^*=\lim_{k\to\infty}x_{k+1}=\lim_{k\to\infty}\varphi(x_k)=\varphi(\lim_{k\to\infty}x_k)=\varphi(x^*)$$，即 x^* 为 φ 的不动点.

从几何图象(参见图 2－2)可知，φ 的不动点就是直线 $y=x$ 与曲线 $y=\varphi(x)$ 的交点 P^*，横坐标 x^* 即为不动点，从它的某个初始近似 x_0 出发，在曲线 $y=\varphi(x)$ 确定一点 P_0，引平行 x 轴直线，与直线 $y=x$ 交于点 Q_1，其横坐标即为 x_1，由式(2.2.2)逐次求得 $x_1,x_2,\cdots$，即为如图 2－2 所示点 $P_1,P_2\cdots$的横坐标.

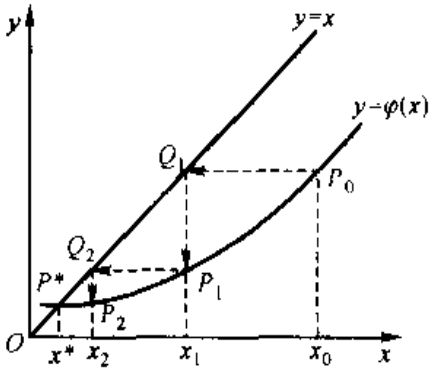

图 2－2

例 2.3 求方程 $f(x)=x^3-x-1=0$ 在 $x_0=1.5$ 附近的根.

解 若将方程改写为 $x=x^3-1$，构造迭代法

$$x_{k+1}=x_k^3-1,k=0,1,\cdots \tag{2.2.3}$$

由 $x_0=1.5,x_1=2.375,x_2=12.39,\cdots$，可知，$x_k$ 显然不收敛.

若将方程改为 $x=(x+1)^{1/3}$，构造迭代法

$$x_{k+1}=(x_k+1)^{1/3},k=0,1,\cdots \tag{2.2.4}$$

计算结果见表 2－2.

表 2－2

k	x_k	k	x_k
0	1.5	5	1.324 76
1	1.357 21	6	1.324 73
2	1.330 86	7	1.324 72
3	1.325 88	8	1.324 72
4	1.324 94		

从结果看它是收敛的，且在6位有效数字时 $x_7=x_8=1.324\ 72$ 即为根 x^* 的近似值.

例题表明构造迭代法(2.2.2)，必须使迭代序列$\{x_k\}$收敛，才能求得方程(2.2.1)的解 x^*. 下面给出解的存在唯一性和迭代收敛性定理.

定理 2.1 设 $\varphi\in C[a,b]$，如果对 $\forall x\in[a,b]$ 有 $a\leqslant\varphi(x)\leqslant b$，且存在常数 $L\in(0,1)$ 使

$$|\varphi(x)-\varphi(y)|\leqslant L|x-y|,\quad \forall x,y\in[a,b] \tag{2.2.5}$$

则 φ 在区间$[a,b]$上存在唯一不动点 x^*，且由式(2.2.2)生成的迭代序列$\{x_k\}$对任何 $x_0\in[a,b]$ 收敛于 x^*，并有误差估计

$$|x_k-x^*|\leqslant\frac{L^k}{1-L}|x_1-x_0| \tag{2.2.6}$$

证明 先证明不动点存在性，记 $f(x)=x-\varphi(x)$，由定理条件有 $f(a)=a-\varphi(a)\leqslant 0$及 $f(b)=b-\varphi(b)\geqslant 0$，若有一等号成立，则 $f(a)=0$ 或 $f(b)=0$，即 φ 有不动点，否则必有 $f(a)f(b)<0$，因 $f(x)=x-\varphi(x)\in C[a,b]$故必有 $x^*\in[a,b]$，使 $f(x^*)=x^*-\varphi(x^*)=0$，$x^*$即为 φ 的不动点.

再证明唯一性，设 $x_1^*,x_2^*\in C[a,b]$都是 φ 的不动点，且 $x_1^*\neq x_2^*$，则由式(2.2.5)得

$$|x_1^*-x_2^*|=|\varphi(x_1^*)-\varphi(x_2^*)|\leqslant L|x_1^*-x_2^*|<|x_1^*-x_2^*|$$

与假设矛盾，这表明 $x_1^*=x_2^*$，即不动点是唯一的.

下面证明由式(2.2.2)生成的迭代序列$\{x_k\}$收敛于 φ 的唯一不动点 x^*，由于 $\varphi(x)\in[a,b]$，故$\{x_k\}\in[a,b]$，再由式(2.2.5)有

$$|x_k-x^*|=|\varphi(x_{k-1})-\varphi(x^*)|\leqslant L|x_{k-1}-x^*|\leqslant\cdots\leqslant L^k|x_0-x^*|$$

因 $0<L<1$，故 $\lim\limits_{k\to\infty}|x_k-x^*|=0$，即 $\lim\limits_{k\to\infty}x_k=x^*$，

再利用式(2.2.5)考察

$$\begin{aligned}|x_{k+p}-x_k|&=|x_{k+p}-x_{k+p-1}+x_{k+p-1}-\cdots+x_{k+1}-x_k|\\&\leqslant|x_{k+p}-x_{k+p-1}|+|x_{k+p-1}-x_{k+p-2}|+\cdots+|x_{k+1}-x_k|\\&\leqslant(L^{p-1}+L^{p-2}+\cdots+L+1)|x_{k+1}-x_k|\\&=\frac{1-L^p}{1-L}|x_{k+1}-x_k|<\frac{L^k}{1-L}|x_1-x_0|\end{aligned}$$

上式中令 $p\to\infty$ 则得式(2.2.6). 定理证毕.

推论 若 $\varphi\in C^1[a,b]$(表示 φ 在$[a,b]$上一阶导数连续)，则定理2.1中的条件式(2.2.5)可改为

$$\max_{a\leqslant x\leqslant b}|\varphi'(x)|\leqslant L<1 \tag{2.2.7}$$

则定理2.1中结论全部成立.

实际上，由微分中值定理可得，对 $\forall x,y\in[a,b]$有

$$|\varphi(x)-\varphi(y)|=|\varphi'(\xi)(x-y)|\leqslant\max_{a\leqslant\xi\leqslant b}|\varphi'(\xi)||x-y|\leqslant L|x-y|$$

故式(2.2.5)成立.

以后使用时，如果 $\varphi'(x)$连续都可用式(2.2.7)代替式(2.2.5). 根据定理2.1可验证例2.3

中迭代(2.2.4)收敛,因迭代函数 $\varphi(x)=(x+1)^{1/3}$,在[1,2]中 $\varphi(x)\in(\sqrt[3]{2},\sqrt[3]{3})\subset[1,2]$,且 $\varphi'(x)=\frac{1}{3}(x+1)^{-2/3}$,$\max\limits_{1\leqslant x\leqslant 2}|\varphi'(x)|\leqslant\frac{1}{3}\frac{1}{\sqrt[3]{4}}\leqslant 0.21<1$,故迭代序列$\{x_k\}$收敛.

而对迭代法(2.2.3),因 $\varphi(x)=x^3-1,\varphi'(x)=3x^2$ 在区间[1,2]中$|\varphi'(x)|>1$,故迭代(2.2.3)是发散的.

2.2.2 局部收敛性与收敛阶

定理 2.1 给出了迭代法(2.2.2)在区间$[a,b]$上的收敛性,称为全局收敛性,下面讨论 φ 在不动点 x^* 附近的收敛性问题.

定义 2.1 设 φ 在某区间 I 有不动点 x^*,若存在 x^* 的一个邻域 $S=\{|x-x^*|<\delta\}\subset I$,对 $\forall x_0\in S$,迭代法(2.2.2)生成的序列$\{x_k\}\subset S$,且收敛于 x^*,则称迭代序列(2.2.2)局部收敛.

定理 2.2 设 x^* 为 φ 的不动点,$\varphi'(x)$在 x^* 的邻域 S 连续,且$|\varphi'(x^*)|<1$,则迭代法(2.2.2)局部收敛.

证明 由 $\varphi'(x)$的连续性,$\exists S=\{|x-x^*|<\delta\}$使

$$\max_{x\in S}|\varphi'(x)|\leqslant L<1 \quad \text{并有}$$

$$|\varphi(x)-x^*|=|\varphi(x)-\varphi(x^*)|\leqslant L|x-x^*|<\delta$$

所以对 $\forall x\in S$,有 $\varphi(x)\in S$,故 φ 在区间 $S=[x^*-\delta,x^*+\delta]$满足定理 2.1 的条件.故由式(2.2.2)生成的序列$\{x_k\}$对 $\forall x_0\in S$ 均收敛于 x^*.证毕.

注意局部收敛性定理是假定 φ 的不动点 x^* 存在时得到的,它只要求$|\varphi'(x^*)|<1$,于是其应用较定理 2.1 简单,并且还可判断不同迭代序列收敛的快慢.

例 2.4 构造不同迭代法求 $x^2-3=0$ 的根 $x^*=\sqrt{3}$.

解 (1) $x_{k+1}=\frac{3}{x_k},k=0,1,\cdots,\varphi(x)=\frac{3}{x},\varphi'(x)=-\frac{3}{x^2},\quad \varphi'(x^*)=-1$,不满足定理2.2 条件.

(2) $x_{k+1}=x_k-\frac{1}{4}(x_k^2-3),k=0,1,\cdots,\varphi(x)=x-\frac{1}{4}(x^2-3),\varphi'(x)=1-\frac{1}{2}x,\varphi'(x^*)=\varphi'(\sqrt{3})=1-\frac{\sqrt{3}}{2}=0.134<1$.收敛.

(3) $x_{k+1}=\frac{1}{2}\left(x_k+\frac{3}{x_k}\right),k=0,1,\cdots,\varphi(x)=\frac{1}{2}\left(x+\frac{3}{x}\right),\varphi'(x)=\frac{1}{2}\left(1-\frac{3}{x^2}\right),\varphi'(x^*)=\varphi'(\sqrt{3})=0$.收敛.

若取 $x_0=2$,分别用上述三种迭代计算,结果见表 2-3.$\sqrt{3}=1.732\ 050\ 8\cdots$,从表 2-3 看到迭代法(1)不收敛,迭代法(2)和迭代法(3)收敛,在迭代法(3)中,$\varphi'(x^*)=0$ 收敛最快.

定义 2.2 设序列$\{x_k\}$收敛到 x^*,记误差 $\varepsilon_k=x_k-x^*$,若存在实数 $p\geqslant 1$ 及 $\alpha>0$,使

表 2-3

k	x_k	迭代法(1)	迭代法(2)	迭代法(3)
0	x_0	2	2	2
1	x_1	1.5	1.75	1.75
2	x_2	2	1.734 75	1.732 143
3	x_3	1.5	1.732 361	1.732 051
⋮	⋮	⋮	⋮	⋮

$$\lim_{k\to\infty}\frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^p}=\alpha \tag{2.2.8}$$

则称序列$\{x_k\}$是 p 阶收敛的，α 称为渐近误差常数，当 $p=1$ 时 $0<\alpha<1$，称为线性收敛. 若 $\alpha=0$ 称为超 p 阶收敛，$p>1$ 称为超线性收敛，$p=2$ 称为平方收敛.

定理 2.3 设 x^* 为 φ 的不动点，整数 $p>1$，$\varphi^{(p)}(x)$在 x^* 的邻域连续，且满足

$$\varphi'(x^*)=\cdots=\varphi^{(p-1)}(x^*)=0,\text{而 }\varphi^{(p)}(x^*)\neq 0 \tag{2.2.9}$$

则由迭代法(2.2.2)生成的序列$\{x_k\}$在 x^* 的邻域是 p 阶收敛的，并有

$$\lim_{k\to\infty}\frac{\varepsilon_{k+1}}{\varepsilon_k^p}=\frac{\varphi^{(p)}(x^*)}{p!} \tag{2.2.10}$$

证明 由于 $p>1$，故 $\varphi'(x^*)=0$，由定理 2.2 可知$\{x_k\}$局部收敛，对 x^* 邻域中的初始近似 $x_0\neq x^*$，由(2.2.2)迭代到 $x_k\neq x^*$，将 $\varphi(x_k)$在 x^* 处按 Taylor 展开，得

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

$$\frac{\varphi^{(p)}(\xi)}{p!}(x_k-x^*)^p,\qquad \xi\text{ 在 }x_k\text{ 与 }x^*\text{ 之间}$$

由式(2.2.9)得

$$x_{k+1}-x^*=\varphi(x_k)-\varphi(x^*)=\frac{\varphi^{(p)}(\xi)}{p!}(x_k-x^*)^p,\qquad \xi\text{ 在 }x^*\text{ 与 }x_k\text{ 之间}$$

即

$$\frac{\varepsilon_{k+1}}{\varepsilon_k^p}=\frac{\varphi^{(p)}(\xi)}{p!}$$

由 $\varphi^{(p)}(x)$的连续性，上式取极限 $k\to\infty$ 则得式(2.2.10).

根据此定理的结论，对例 2.4 中迭代法(3)的 $\varphi'(x^*)=0$，而 $\varphi''(x)=\frac{6}{x^3}$，$\varphi''(x^*)=\frac{2}{\sqrt{3}}\neq 0$，故知 $p=2$，即该迭代序列是二阶收敛的.

2.3 Steffensen 加速迭代法

不动点迭代(2.2.2)通常只有线性收敛，有时甚至不收敛，为加速迭代法的收敛性通常可采

用 Steffensen 加速迭代.

设 $x^*=\varphi(x^*)$是 φ 的不动点,记 $\varepsilon_k=x_k-x^*$,利用中值定理有

$$\frac{\varepsilon_{k+1}}{\varepsilon_k}=\frac{x_{k+1}-x^*}{x_k-x^*}=\frac{\varphi(x_k)-\varphi(x^*)}{x_k-x^*}=\varphi'(\xi_k),\xi_k \text{ 在 } x^* \text{ 与 } x_k \text{ 之间通常 } \varphi'(\xi_k) \text{依赖于 } k,\text{若}$$

$\varphi'(x)$变化不大,设 $\varphi'(\xi_k)\approx C$,于是有

$$x_{k+1}-x^*\approx C(x_k-x^*)$$

$$x_{k+2}-x^*\approx C(x_{k+1}-x^*)$$

从上两式消去 C,则得

$$\frac{x_{k+2}-x^*}{x_{k+1}-x^*}\approx\frac{x_{k+1}-x^*}{x_k-x^*}\text{或}(x_{k+2}-x^*)(x_k-x^*)\approx(x_{k+1}-x^*)^2$$

解得

$$x^*\approx\frac{x_{k+2}x_k-x_{k+1}^2}{x_{k+2}-2x_{k+1}+x_k}=x_k-\frac{(x_{k+1}-x_k)^2}{x_{k+2}-2x_{k+1}+x_k}$$

若记

$$\bar{x}_{k+1}=x_k-\frac{(x_{k+1}-x_k)^2}{x_{k+2}-2x_{k+1}+x_k},k=0,1,\cdots \tag{2.3.1}$$

用序列$\{\bar{x}_k\}$作为不动点 x^* 的新近似,一般说,它比迭代法(2.2.2)收敛更快,实际上迭代法(2.3.1)可改为

$$\begin{cases}y_k=\varphi(x_k),z_k=\varphi(y_k)\\ x_{k+1}=x_k-\dfrac{(y_k-x_k)^2}{z_k-2y_k+x_k},k=0,1,\cdots\end{cases} \tag{2.3.2}$$

称为 Steffensen 迭代法,它是将原不动点迭代(2.2.2)计算两次合并成一步得到,可改为另一种不动点迭代法

$$x_{k+1}=\psi(x_k),k=0,1,\cdots \tag{2.3.3}$$

其中

$$\psi(x)=x-\frac{[\varphi(x)-x]^2}{\varphi(\varphi(x))-\varphi(x)+x} \tag{2.3.4}$$

并有以下局部收敛定理.

定理 3.1 若 x^* 为(2.3.4)定义的函数 ψ 的不动点,则 x^* 为 φ 的不动点.反之,若 x^* 是 φ 的不动点,设 $\varphi''(x)$连续,且 $\varphi'(x^*)\neq 1$,则 x^* 是 ψ 的不动点且迭代法(2.3.3)是二阶收敛的.(证明见[3])

例 2.5 在例 2.3 中求方程 $f(x)=x^3-x-1=0$ 的第一种迭代(2.2.3)是不收敛的,试用 Steffensen 迭代法求解.

解 式(2.2.3)中 $\varphi(x)=x^3-1$,现用式(2.3.2)迭代,结果见表 2-4.

它是收敛的,此例表明,即使对(2.2.2)不收敛的迭代法,用 Steffensen 加速迭代(2.3.2)仍可能收敛.

表 2-4

k	x_k	y_k	z_k
0	1.5	2.375 00	12.396 5
1	1.416 29	1.840 92	5.238 88
2	1.355 65	1.491 40	2.317 28
3	1.328 95	1.347 10	1.444 35
4	1.324 80	1.325 18	1.327 14
5	1.324 72	1.324 72	

2.4 Newton 迭代法

2.4.1 Newton 法及其收敛性

求方程(2.1.1)的根 x^*,如果已知它的一个近似 x_k,可利用 Taylor 展开式求出 $f(x)$ 在 x_k 附近的线性近似,即

$$f(x)=f(x_k)+f'(x_k)(x-x_k)+\frac{f''(\xi)}{2!}(x-x_k)^2,\xi \text{ 在 } x \text{ 与 } x_k \text{ 之间}$$

忽略余项,则得方程(2.1.1)的近似

$$f(x)\approx f(x_k)+f'(x_k)(x-x_k)=0$$

右端为 x 的线性方程,若 $f'(x_k)\neq 0$,则解 $x=x_k-\dfrac{f(x_k)}{f'(x_k)}$,记作 x_{k+1},它可作为 $f(x)=0$ 的解 x^* 的新近似,即

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)},k=0,1,\cdots \tag{2.4.1}$$

称为解方程(2.1.1)的 Newton 法. 在几何上求方程 $f(x)=0$ 的解 x^*,即求曲线 $y=f(x)$ 与 x 轴交点 x^*. 若已知 x^* 的一个近似 x_k,通过点 $(x_k,f(x_k))$ 作曲线 $y=f(x)$ 的切线,它与 x 轴交点为 x_{k+1},作为 x^* 的新近似,如图 2-3 所示.

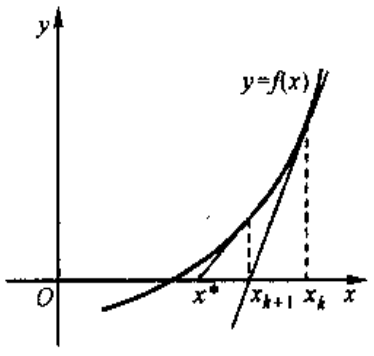

图 2-3

关于 Newton 法收敛性有以下的局部收敛定理.

定理 4.1 设 x^* 是 $f(x)=0$ 的一个根,$f(x)$ 在 x^* 附近二阶导数连续,且 $f'(x^*)\neq 0$,则 Newton 法(2.4.1)具有二阶收敛,且

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(x_k-x^*)^2}=\frac{1}{2}\frac{f''(x^*)}{f'(x^*)} \tag{2.4.2}$$

证明 由式(2.4.1)知迭代函数 $\varphi(x)=x-\dfrac{f(x)}{f'(x)}$，$\varphi'(x)=\dfrac{f(x)f''(x)}{[f'(x)]^2}$，$\varphi'(x^*)=0$，而 $\varphi''(x^*)=\dfrac{f''(x^*)}{f'(x^*)}\neq 0$，由定理 2.3 可知，Newton 迭代(2.4.1)具有二阶收敛，由式(2.2.10)可得到式(2.4.2)．证毕．

定理表明 Newton 法收敛很快，但 x_0 在 x^* 附近时才能保证迭代序列收敛．有关 Newton 法半局部收敛性与全局收敛定理．此处不再讨论．

例 2.6 用 Newton 法求方程 $xe^x-1=0$ 的根．

解 $f(x)=xe^x-1$，$f'(x)=(x+1)e^x$，Newton 迭代为

$$x_{k+1}=x_k-\frac{x_k-e^{-x_k}}{x_k+1},k=0,1,\cdots$$

取 $x_0=0.5$，$x_1=0.571\ 02$，$x_2=0.567\ 16$，$x_3=0.567\ 14$，x_3 即为根 x^* 的近似，它表明 Newton 法收敛很快．

例 2.7 设 $a>0$，求平方根$\sqrt{a}$的过程可化为解方程 $f(x)=x^2-a=0$．若用 Newton 法求解，由式(2.4.1)得

$$x_{k+1}=\frac{1}{2}\left(x_k+\frac{a}{x_k}\right) \tag{2.4.3}$$

这是在计算机上作开方运算的一个实际有效的方法，它每步迭代只做一次除法和一次加法再做一次移位即可，计算量少，又收敛很快，对 Newton 法我们已证明了它的局部收敛性，对式(2.4.3)可证明对任何 $x_0>0$ 迭代法都是收敛的，因为当 $0<x_0<\sqrt{a}$ 时有

$$x_1-\sqrt{a}=\frac{(x_0-\sqrt{a})^2}{2x_0}>0$$

即 $x_1>\sqrt{a}$，而对任意 $x_k>\sqrt{a}$，也可验证 $x_{k+1}>\sqrt{a}$，即从 $k=1$ 开始 $x_k>\sqrt{a}$，且

$$x_{k+1}-x_k=\frac{(a-x_k^2)}{2x_k}<0$$

所以$\{x_k\}$从 $k=1$ 起是一个单调递减有下界的序列，$\{x_k\}$有极限 x^*．在式(2.4.3)中令 $k\to\infty$ 可得 $x^*=\sqrt{a}$，这就说明了只要 $x_0>0$，迭代(2.4.3)总收敛到$\sqrt{a}$，且是二阶收敛．

在例 2.4 的迭代法(3)中，用式(2.4.3)求$\sqrt{3}$只迭代 3 次就得到$\sqrt{3}=1.732\ 051$，具有 7 位有效数字．

2.4.2 Newton 下山法

Newton 法是一种局部收敛方法，通常要求初始近似 x_0 在解 x^* 附近才保证迭代序列收敛．为扩大收敛范围，使对任意 x_0 迭代序列收敛，通常可引入参数，并将 Newton 迭代(2.4.1)改为

$$x_{k+1}=x_k-\lambda_k\frac{f(x_k)}{f'(x_k)},k=0,1,\cdots \tag{2.4.4}$$

其中 $0<\lambda_k\leqslant 1$，称为下山因子，式(2.4.4)称为 Newton 下山法. 通常可选择 λ_k 使 $|f(x_{k+1})|<|f(x_k)|$，计算时可取 $\lambda_k=1,\frac{1}{2},\frac{1}{4},\cdots$直到满足要求为止. 由此得到的序列 $|x_k|$ 由于满足下山条件 $|f(x_{k+1})|<|f(x_k)|$，故它是收敛的，但它只是线性收敛.

例 2.8 用 Newton 下山法求 $f(x)=x^3-x-1=0$ 的解，取 $x_0=0.6$，计算精确到 10^{-5}.

解 由于 $f(x)=x^3-x-1$，$f'(x)=3x^2-1$，由式(2.4.4)得 Newton 下山法为 $x_{k+1}=x_k-\lambda_k\frac{x_k^3-x_k-1}{3x_k^2-1}$，若 $x_0=1.5$ 用 Newton 法($\lambda_k=1$)迭代 3 步则求得解 x^* 的近似 $x_3=1.324\ 72$. 现用 $x_0=0.6$，用 $\lambda_k=1$，则得 $x_1=17.9$，且 $f(0.6)=-1.384$，而 $f(x_1)=5\ 716.439$，$|f(x_1)|>|f(x_0)|$ 不满足下山条件. 通过试算，当 $\lambda_0=\frac{1}{32}$ 时，

$x_1=1.140\ 625$，$f(x_1)=-0.656\ 643$ 满足 $|f(x_1)|<|f(x_0)|$.

以下计算 $x_2,x_3,\cdots$时参数 $\lambda_1=\lambda_2=\cdots=1$，且

$$x_2=1.361\ 81,\quad f(x_2)=0.186\ 64,\quad x_3=1.326\ 28,$$
$$f(x_3)=0.006\ 67,\quad x_4=1.324\ 72,\quad f(x_4)=0.000\ 008\ 6.$$

2.4.3 重根情形

当 $f'(x^*)=0$，则 x^* 为方程(2.1.1)的重根，此时 $f(x)=(x-x^*)^m g(x)$，$g(x^*)\neq 0$，Newton 法的迭代函数 $\varphi(x)=x-\frac{f(x)}{f'(x)}$，$\varphi'(x^*)=1-\frac{1}{m}\neq 0(m\geqslant 2)$ 且有 $|\varphi'(x^*)|<1$，故 Newton 法仍收敛，但只是线性收敛.

若迭代函数改为 $\varphi(x)=x-m\frac{f(x)}{f'(x)}$，则 $\varphi'(x^*)=0$，故迭代法

$$x_{k+1}=x_k-m\frac{f(x_k)}{f'(x_k)},k=0,1,\cdots \tag{2.4.5}$$

具有二阶收敛.

对重根还可构造另一种迭代法，令 $\mu(x)=\frac{f(x)}{f'(x)}$ 若 x^* 是 $f(x)=0$ 的 m 重根，则

$$\mu(x)=\frac{(x-x^*)g(x)}{mg(x)+(x-x^*)g'(x)}$$

所以 x^* 是 $\mu(x)=0$ 的单根，对它用 Newton 法，迭代函数为

$$\varphi(x)=x-\frac{\mu(x)}{\mu'(x)}=x-\frac{f(x)f'(x)}{[f'(x)]^2-f(x)f''(x)}$$

从而可构造迭代法

$$x_{k+1}=x_k-\frac{f(x_k)f'(x_k)}{[f'(x_k)]^2-f(x_k)f''(x_k)},k=0,1,\cdots \tag{2.4.6}$$

它也是二阶收敛的.

例 2.9 方程 $x^4-4x^2+4=0$ 的根 $x^*=\sqrt{2}$ 是二重根，试用 Newton 法及(2.4.5)、(2.4.6)三种迭代法各计算 3 步.

解 $f(x)=(x^2-2)^2, f'(x)=4x(x^2-2), f''(x)=4(3x^2-2)$

方法(1)：Newton 迭代，$x_{k+1}=x_k-\dfrac{(x_k^2-2)}{4x_k}, k=0,1,\cdots$

方法(2)：迭代法(2.4.5)，$x_{k+1}=x_k-\dfrac{x_k^2-2}{2x_k}, k=0,1,\cdots$

方法(3)：迭代法(2.4.6)，$x_{k+1}=x_k-\dfrac{x_k(x_k^2-2)}{x_k^2+2}, k=0,1,\cdots$

三种方法均取 $x_0=1.5$ 计算结果如下：

	方法(1)	方法(2)	方法(3)
x_1	1.458 333 333	1.416 666 667	1.411 764 706
x_2	1.436 607 143	1.414 215 686	1.414 211 438
x_3	1.425 497 619	1.414 213 562	1.414 213 562

方法(2)与方法(3)均达到 10^{-9} 精确度，而方法(1)只有线性收敛，要达到相同精度需迭代 30 次.

2.4.4 离散 Newton 法(割线法)

求解方程(2.1.1)的 Newton 法(2.4.1)要计算 $f'(x_k)$，如果 $f(x)$ 导数计算不方便，通常可用计算函数差商近似，即

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

将它代入式(2.4.1)则得离散 Newton 法：

$$x_{k+1}=x_k-\frac{f(x_k)}{f(x_k)-f(x_{k-1})}(x_k-x_{k-1}), k=1,2,\cdots \tag{2.4.7}$$

这种迭代法与式(2.2.2)不同，它要给出 x_0, x_1 两个初始近似，才能逐次计算出 $x_2, x_3, \cdots$. 因此称为多点(两点)迭代，迭代(2.4.7)称为割线法，其几何意义是，用曲线 $y=f(x)$ 上两点 $(x_{k-1}, f(x_{k-1}))$，$(x_k, f(x_k))$ 的割线与 x 轴交点作为 $f(x)=0$ 根的新近似，即 $f(x)\approx\dfrac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}(x-x_k)+f(x_k)=0$ 的根 x，记作 x_{k+1}，它就是方程(2.1.1)根 x^* 的新近似，如图 2-4 所示.

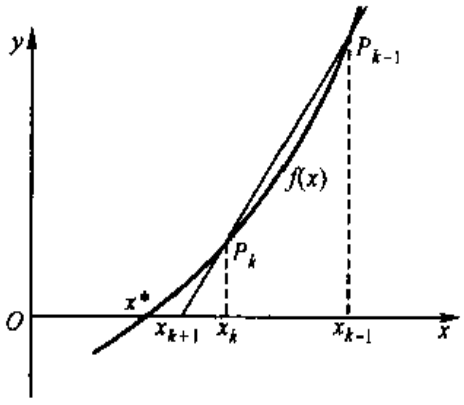

图 2-4

由于割线法与单点迭代法(2.2.2)不同，其收敛性要复

杂一些.但可以证明割线法(2.4.7)是超线性收敛的,且收敛阶 $p=\frac{1+\sqrt{5}}{2}\approx 1.618$,故割线法收敛也是很快的.

习 题 二

1. 用二分法求方程 $x^2-x-1=0$ 的正根,使误差小于 0.05.

2. 求方程 $x^3-x^2-1=0$ 在 $x_0=1.5$ 附近的一个根,将方程改写成下列等价形式,并建立相应迭代公式.

(1) $x=1+\frac{1}{x^2}$,迭代公式 $x_{k+1}=1+\frac{1}{x_k^2}$.

(2) $x^3=1+x^2$,迭代公式 $x_{k+1}=(1+x_k^2)^{\frac{1}{3}}$.

(3) $x^2=\frac{1}{x-1}$,迭代公式 $x_{k+1}=\frac{1}{\sqrt{x_k-1}}$.

试分析每种迭代公式的收敛性,并选取一种收敛最快的方法求具有 4 位有效数字的近似根.

3. 设方程 $12-3x+2\cos x=0$ 的迭代法

$$x_{k+1}=4+\frac{2}{3}\cos x_k$$

(1) 证明对 $\forall x_0\in\mathbf{R}$,均有 $\lim\limits_{k\to\infty}x_k=x^*$,其中 x^* 为方程的根.

(2) 取 $x_0=4$,求此迭代法的近似根,使误差不超过 10^{-3},并列出各次迭代值.

(3) 此迭代法收敛阶是多少?证明你的结论.

4. 给定函数 $f(x)$,设对一切 x,$f'(x)$存在,而且 $0<m\leqslant f'(x)\leqslant M$.证明对 $0<\lambda<\frac{2}{M}$的任意常数 λ,迭代法 $x_{k+1}=x_k-\lambda f(x_k)$均收敛于方程 $f(x)=0$ 的根.

5. 用 Steffensen 方法计算第 2 题中(2)、(3)的近似根,精确到 10^{-5}.

6. 用 Newton 法求下列方程的根,计算准确到 4 位有效数字.

(1) $f(x)=x^3-3x-1=0$ 在 $x_0=2$ 附近的根.

(2) $f(x)=x^2-3x-e^x+2=0$ 在 $x_0=1$ 附近的根.

7. 应用 Newton 法于方程 $x^3-a=0$,求立方根$\sqrt[3]{a}$的迭代公式,并讨论其收敛性.

8. 把单步迭代法

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

看成 Newton 法的一种修正,设 f 有二阶连续导数,$f(x^*)=0$,$f'(x^*)\neq 0$,试证明这种迭代法二阶收敛.

9. $\varphi(x)=x-p(x)f(x)-q(x)f^2(x)$,试确定函数 $p(x)$和 $q(x)$,使求解 $f(x)=0$ 且以 φ 为迭代函数的迭代法至少三阶收敛.

10. 用(2.4.1),(2.4.5),(2.4.6)三种不同迭代法求方程 $f(x)=\left(\sin x-\frac{x}{2}\right)^2=0$ 的一个近似根,准确到 10^{-5},初始值 $x_0=\frac{\pi}{2}$,并比较结果优劣.

第三章　解线性方程组的直接法

3.1　引言与矩阵一些基础知识

3.1.1　引言

线性方程组求解是科学计算中最常遇到的问题，如在应力分析、电路分析、分子结构、测量学中都会遇到解线性方程组问题. 在很多广泛应用的数学问题的数值方法中，如三次样条、最小二乘法、微分方程边值问题的差分法与有限元法也都涉及到求解线性方程组.

本章讨论 n 个变量 n 个线性方程的方程组求解，其表达式为

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

通常用向量矩阵表示，则上述方程可写成

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{3.1.1}$$

其中

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix},\ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},\ \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

并记做 $\boldsymbol{A} \in \mathbf{R}^{n\times n}$，$\boldsymbol{x}, \boldsymbol{b} \in \mathbf{R}^n$，分别表示 $\boldsymbol{A}$ 为 $n \times n$ 阶实矩阵，$\boldsymbol{x}, \boldsymbol{b}$ 为 n 维实向量. 根据线性代数知识可知 $\boldsymbol{A}$ 非奇异，即 $\det \boldsymbol{A} \neq 0$，方程组(3.1.1)有唯一解，并可用 Cramer 法则将解用公式表示出来，但由于计算量太大，因此不能用于实际求解. 本章要讨论的直接方法是将方程组(3.1.1)转化为三角方程组，再通过回代求三角方程组的解. 理论上，直接法可在有限步内求得方程的精确解，但由于数值运算有舍入误差，因此实际在计算机上求得的数值解仍是近似解，仍要对它进行误差分析. 解线性方程组的另一类方法是迭代法，将在下一章讨论.

下面给出本章和下章需要的一些矩阵基础知识.

3.1.2　矩阵特征值与谱半径

定义 1.1　设 $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n\times n}$，若存在一个数 λ(实数或复数)和非零向量 $\boldsymbol{x} = (x_1, \cdots, x_n)^{\mathrm{T}}$

$\in \mathbf{R}^n$,使

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \tag{3.1.2}$$

则称 λ 为 $\boldsymbol{A}$ 的特征值,$\boldsymbol{x}$ 为 $\boldsymbol{A}$ 对应 λ 的特征向量,$\boldsymbol{A}$ 的全体特征值称为 $\boldsymbol{A}$ 的谱,记作 $\sigma(\boldsymbol{A})$,即 $\sigma(\boldsymbol{A}) = \{\lambda_1, \cdots, \lambda_n\}$.

$$\rho(\boldsymbol{A}) = \max_{\lambda \in \sigma(\boldsymbol{A})} |\lambda| \tag{3.1.3}$$

称为 $\boldsymbol{A}$ 的谱半径.

由式(3.1.2)知,λ 可使齐次方程

$$(\lambda \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$$

有非零解,故系数行列式 $\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = 0$,即

$$p(\lambda) = \det(\lambda \boldsymbol{I} - \boldsymbol{A}) = \begin{vmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \vdots & \vdots & & \vdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{vmatrix}$$

$$= \lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n = 0 \tag{3.1.4}$$

$p(\lambda)$称为特征多项式,方程(3.1.4)称为特征方程,在复数域中有 n 个根 $\lambda_1, \lambda_2, \cdots, \lambda_n$,故

$$p(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n)$$

由行列式(3.1.4)展开可知:

$$-c_1 = \lambda_1 + \lambda_2 + \cdots + \lambda_n = \sum_{i=1}^{n} a_{ii} = \operatorname{tr} \boldsymbol{A}$$

$$c_n = (-1)^n \lambda_1 \cdots \lambda_n = (-1)^n \det \boldsymbol{A}$$

于是,矩阵 $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n \times n}$ 的 n 个特征值 $\lambda_1, \cdots, \lambda_n$ 是它的特征方程(3.1.4)的 n 个根,$\boldsymbol{A}$ 的迹 $\operatorname{tr} \boldsymbol{A}$ 有

$$\operatorname{tr} \boldsymbol{A} = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i \tag{3.1.5}$$

$$\det \boldsymbol{A} = \lambda_1 \lambda_2 \cdots \lambda_n \tag{3.1.6}$$

此外,$\boldsymbol{A}$ 的特征值 λ 和特征向量 $\boldsymbol{x}$ 还有如下性质:

(1) $\boldsymbol{A}^{\mathrm{T}}$ 与 $\boldsymbol{A}$ 有相同的特征值 λ 及相同的特征向量 $\boldsymbol{x}$.

(2) 若 $\boldsymbol{A}$ 非奇异,则 $\boldsymbol{A}^{-1}$ 的特征值为 λ^{-1},特征向量为 $\boldsymbol{x}$.

(3) 相似矩阵 $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ 有相同特征多项式.

例 3.1 求 $\boldsymbol{A} = \begin{bmatrix} 1 & -2 & 2 \\ -2 & -2 & 4 \\ 2 & 4 & -2 \end{bmatrix}$ 的特征值及谱半径.

解 $\boldsymbol{A}$ 的特征方程为

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = \begin{vmatrix} \lambda - 1 & 2 & -2 \\ 2 & \lambda + 2 & -4 \\ -2 & -4 & \lambda + 2 \end{vmatrix} = \lambda^3 + 3\lambda^2 - 24\lambda + 28$$

$$=(\lambda-2)^2(\lambda+7)=0$$

故 $\boldsymbol{A}$ 的特征值为 $\lambda_1=\lambda_2=2$，$\lambda_3=-7$. 谱半径为 $\rho(\boldsymbol{A})=7$.

3.1.3 对称正定矩阵

定义 1.2 设 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$，如果 $\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}$，即 $a_{ij}=a_{ji}$，$i,j=1,2,\cdots,n$，则称 $\boldsymbol{A}$ 为对称矩阵，若还满足对于 $\forall\, x\neq\mathbf{0}, x\in\mathbf{R}^n$，$(x,Ax)=x^{\mathrm{T}}Ax>0$，则称 $\boldsymbol{A}$ 为对称正定矩阵，如果对 $\forall\, x\neq\mathbf{0}$ 有 $(x,Ax)\geqslant 0$，则称 $\boldsymbol{A}$ 为半正定矩阵.

当 $\boldsymbol{A}$ 为对称时，$\boldsymbol{A}$ 的特征值皆为实数，且有 n 个线性无关的特征向量.

对称正定矩阵还有以下重要性质：

(1) $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 对称正定，则 $\boldsymbol{A}$ 非奇异，且 $\boldsymbol{A}^{-1}$ 也对称正定；

(2) $\boldsymbol{A}$ 对称正定的充要条件是，$\boldsymbol{A}$ 的所有特征值 $\lambda_i>0$，$i=1,2,\cdots,n$；

(3) $\boldsymbol{A}$ 对称正定，则 $\boldsymbol{A}$ 的对角元素 $a_{ii}>0$，$i=1,\cdots,n$；

(4) $\boldsymbol{A}$ 对称正定的充要条件是 $\boldsymbol{A}$ 的所有顺序主子式

$$\Delta_i=\det\boldsymbol{A}_i=\begin{vmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} \end{vmatrix}>0,\quad i=1,2,\cdots,n$$

以上性质可直接由定义证明，证明略.

3.1.4 正交矩阵与初等矩阵

定义 1.3 若 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 且 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I}$，则称 $\boldsymbol{A}$ 为正交矩阵.

由定义知 $\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}^{-1}$，且 $\boldsymbol{A}^{\mathrm{T}}$ 与 $\boldsymbol{A}^{-1}$ 均为正交阵，且有 $|\det\boldsymbol{A}|=1$.

定义 1.4 设 $\boldsymbol{u},\boldsymbol{v}\in\mathbf{R}^n$，$\sigma\in\mathbf{R}$，$\sigma\neq 0$. 则

$\boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\sigma)=\boldsymbol{I}-\sigma\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$，$\boldsymbol{I}$ 为单位矩阵，称为(实)初等矩阵.

显然 $\boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\sigma)\in\mathbf{R}^{n\times n}$，例如 $n=3$，$\boldsymbol{u}=(u_1,u_2,u_3)^{\mathrm{T}}$，$\boldsymbol{v}=(v_1,v_2,v_3)^{\mathrm{T}}$，则

$$\boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\sigma)=\begin{bmatrix} 1-\sigma u_1v_1 & -\sigma u_1v_2 & -\sigma u_1v_3 \\ -\sigma u_2v_1 & 1-\sigma u_2v_2 & -\sigma u_2v_3 \\ -\sigma u_3v_1 & -\sigma u_3v_2 & 1-\sigma u_3v_3 \end{bmatrix}$$

设 $\sigma,\tau\in\mathbf{R}$，则

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\sigma)\boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\tau)&=(\boldsymbol{I}-\sigma\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})(\boldsymbol{I}-\tau\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})\\&=\boldsymbol{I}-(\sigma+\tau-\sigma\tau\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\end{aligned}$$

若 σ 已知，选 $\tau\in\mathbf{R}$，使 $\sigma+\tau-\sigma\tau\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}=0$，则当 $\sigma\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}\neq 1$ 时，令

$$\tau=\frac{\sigma}{\sigma\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}-1}\tag{3.1.7}$$

则有

$$\boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\sigma)^{-1}=\boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\tau)\tag{3.1.8}$$

此外,还有

$$\det \boldsymbol{E}(\boldsymbol{u},\boldsymbol{v};\sigma)=1-\sigma\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u} \tag{3.1.9}$$

下面给出两个常用的初等矩阵.

例 3.2 初等排列矩阵 $\boldsymbol{I}_{ij}$. 令 $\sigma=1,\boldsymbol{u}=\boldsymbol{v}=\boldsymbol{e}_i-\boldsymbol{e}_j$

则

$$\boldsymbol{I}_{ij}=\boldsymbol{E}(\boldsymbol{e}_i-\boldsymbol{e}_j,\boldsymbol{e}_i-\boldsymbol{e}_j;1)=\boldsymbol{I}-(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^{\mathrm{T}}$$

$$=\begin{bmatrix}
1 & & & \vdots & & & & \vdots & & & & \\
& \ddots & & \vdots & & & & \vdots & & & & \\
& & 1 & \vdots & & & & \vdots & & & & \\
\cdots & \cdots & \cdots & 0 & \cdots & \cdots & \cdots & 1 & \cdots & \cdots & \cdots & i \\
& & & \vdots & \ddots & & & \vdots & & & & \\
& & & \vdots & & 1 & & \vdots & & & & \\
& & & \vdots & & & \ddots & \vdots & & & & \\
\cdots & \cdots & \cdots & 1 & \cdots & \cdots & \cdots & 0 & \cdots & \cdots & \cdots & j \\
& & & \vdots & & & & \vdots & 1 & & & \\
& & & \vdots & & & & \vdots & & \ddots & & \\
& & & \vdots & & & & \vdots & & & 1 & \\
& & & i & & & & j & & & &
\end{bmatrix}$$

矩阵 $\boldsymbol{I}_{ij}$ 也称初等置换矩阵,它由单位矩阵 $\boldsymbol{I}$ 交换 i 行与 j 行得到,它有以下性质:

(1) $\boldsymbol{I}_{ij}^{-1}=\boldsymbol{I}_{ij}^{\mathrm{T}}=\boldsymbol{I}_{ij}$,故 $\boldsymbol{I}_{ij}$ 为正交矩阵.

(2) $\det \boldsymbol{I}_{ij}=-1$.

(3) $\boldsymbol{I}_{ij}\boldsymbol{A}$ 是将 $\boldsymbol{A}$ 的 i,j 行互换,$\boldsymbol{A}\boldsymbol{I}_{ij}$ 是将 $\boldsymbol{A}$ 的 i,j 列互换.

例 3.3 初等下三角矩阵. 设 $\sigma=-1,\boldsymbol{u}=\boldsymbol{l}_j=(0,\cdots,0,l_{j+1,j},\cdots,l_{n,j})^{\mathrm{T}}\in\mathbf{R}^n,\boldsymbol{v}=\boldsymbol{e}_j$,则记 $\boldsymbol{L}_j(\boldsymbol{l}_j)=\boldsymbol{E}(\boldsymbol{l}_j,\boldsymbol{e}_j;-1)$称为指标为 j 的初等下三角矩阵,即

$$\boldsymbol{L}_j=\boldsymbol{L}_j(\boldsymbol{l}_j)=\boldsymbol{I}+\boldsymbol{l}_j\boldsymbol{e}_j^{\mathrm{T}}=\begin{bmatrix}
1 & & & \\
& \ddots & & \\
& & 1 & \\
& & l_{j+1,j} & 1 & \\
& & \vdots & & \ddots \\
& & l_{n,j} & & & 1
\end{bmatrix} \tag{3.1.10}$$

矩阵 $\boldsymbol{L}_j(\boldsymbol{l}_j)$ 有以下性质:

(1) $\boldsymbol{L}_j(\boldsymbol{l}_j)^{-1}=\boldsymbol{I}-\boldsymbol{l}_j\boldsymbol{e}_j^{\mathrm{T}}=\boldsymbol{L}_j(-\boldsymbol{l}_j)$.

(2) $\det \boldsymbol{L}_j(\boldsymbol{l}_j)=1$.

(3) $\boldsymbol{L}=\boldsymbol{L}_1(l_1)\boldsymbol{L}_2(l_2)\cdots\boldsymbol{L}_{n-1}(l_{n-1})=\begin{bmatrix}1&&&&\\l_{21}&1&&&\\\vdots&l_{32}&1&&\\\vdots&\vdots&&\ddots&\\l_{n1}&l_{n2}&\cdots&\cdots&1\end{bmatrix}$，为单位下三角矩阵. 初等下三角矩阵 $\boldsymbol{L}_j=\boldsymbol{L}_j(l_j)$ 也称为 Gauss 变换矩阵.

3.2 Gauss 消去法

3.2.1 Gauss 顺序消去法

Gauss 消去法就是将方程组(3.1.1)通过$(n-1)$步消元，将(3.1.1)转化为上三角方程组

$$\begin{bmatrix}a_{11}^{(1)}&a_{12}^{(1)}&\cdots&a_{1n}^{(1)}\\&a_{22}^{(2)}&\cdots&a_{2n}^{(2)}\\&&\ddots&\vdots\\&&&a_{nn}^{(n)}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}=\begin{bmatrix}b_1^{(1)}\\b_2^{(2)}\\\vdots\\b_n^{(n)}\end{bmatrix}\tag{3.2.1}$$

再回代求此方程组的解.

下面记增广矩阵$[\boldsymbol{A}^{(1)}\mid\boldsymbol{b}^{(1)}]=[\boldsymbol{A}\mid\boldsymbol{b}]$，即

$$[\boldsymbol{A}^{(1)}\mid\boldsymbol{b}^{(1)}]=\left[\begin{array}{cccc|c}a_{11}^{(1)}&a_{12}^{(1)}&\cdots&a_{1n}^{(1)}&b_1^{(1)}\\a_{21}^{(1)}&a_{22}^{(1)}&\cdots&a_{2n}^{(1)}&b_2^{(1)}\\\vdots&\vdots&&\vdots&\vdots\\a_{n1}^{(1)}&a_{n2}^{(1)}&\cdots&a_{nn}^{(1)}&b_n^{(1)}\end{array}\right]$$

第 1 步　设 $a_{11}^{(1)}\neq0$，计算 $l_{i1}=\dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}$，$i=2,3,\cdots,n$，记为 $\boldsymbol{l}_1=(0,l_{21},\cdots,l_{n1})^{\mathrm{T}}$，若用 $-l_{i1}$ 乘 $[\boldsymbol{A}^{(1)}\mid\boldsymbol{b}^{(1)}]$某一行加到某 i 行，可消去 $a_{i1}^{(1)}(i=2,3,\cdots,n)$，用 Gauss 变换矩阵表示

$$\boldsymbol{L}_1=\boldsymbol{I}+\boldsymbol{l}_1\boldsymbol{e}_1^{\mathrm{T}},\ \boldsymbol{L}_1^{-1}=\boldsymbol{I}-\boldsymbol{l}_1\boldsymbol{e}_1^{\mathrm{T}}$$

令
$$[\boldsymbol{A}^{(2)}\mid\boldsymbol{b}^{(2)}]=\boldsymbol{L}_1^{-1}[\boldsymbol{A}^{(1)}\mid\boldsymbol{b}^{(1)}]=[\boldsymbol{L}_1^{-1}\boldsymbol{A}^{(1)}\mid\boldsymbol{L}_1^{-1}\boldsymbol{b}^{(1)}]$$

其中，$a_{ij}^{(2)}=a_{ij}^{(1)}-l_{i1}a_{1j}^{(1)}$，$b_i^{(2)}=b_i^{(1)}-l_{i1}b_1^{(1)}$，$i,j=2,3,\cdots,n$.

一般地，假定已完成了$(k-1)$步消元，即已将$[\boldsymbol{A}^{(1)}\mid\boldsymbol{b}^{(1)}]$转化为以下形式：

$$[\boldsymbol{A}^{(k)}\mid\boldsymbol{b}^{(k)}]=\left[\begin{array}{cccc|c}a_{11}^{(1)}&a_{12}^{(1)}&\cdots&a_{1n}^{(1)}&b_1^{(1)}\\&a_{22}^{(2)}&\cdots&a_{2n}^{(2)}&b_2^{(2)}\\&\ddots&&\vdots&\vdots\\&&a_{kk}^{(k)}&\cdots a_{kn}^{(k)}&b_k^{(k)}\\&&\vdots&\vdots&\vdots\\&&a_{nk}^{(k)}&\cdots a_{nn}^{(k)}&b_n^{(k)}\end{array}\right]$$

第 k 步，假定 $a_{kk}^{(k)}\neq 0$，计算

$$l_{ik}=\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}},i=k+1,k+2,\cdots,n \tag{3.2.2}$$

记 $\boldsymbol{l}_k=(0,\cdots,0,l_{k+1k},\cdots,l_{nk})^{\mathrm{T}}$，$\boldsymbol{L}_k=\boldsymbol{I}+\boldsymbol{l}_k\boldsymbol{e}_k^{\mathrm{T}}$，$\boldsymbol{L}_k^{-1}=\boldsymbol{I}-\boldsymbol{l}_k\boldsymbol{e}_k^{\mathrm{T}}$，则

$$[\boldsymbol{A}^{(k+1)}\mid\boldsymbol{b}^{(k+1)}]=[\boldsymbol{L}_k^{-1}\boldsymbol{A}^{(k)}\mid\boldsymbol{L}_k^{-1}\boldsymbol{b}^{(k)}],$$

其中

$$\begin{cases}a_{ij}^{(k+1)}=a_{ij}^{(k)}-l_{ik}a_{kj}^{(k)}, & i,j=k+1,\cdots,n\\ b_i^{(k+1)}=b_i^{(k)}-l_{ik}b_k^{(k)}, & i=k+1,\cdots,n\end{cases} \tag{3.2.3}$$

当 $k=1,2,\cdots,n-1$ 则可得到 $[\boldsymbol{A}^{(n)}\mid\boldsymbol{b}^{(n)}]$，即方程组(3.2.1).

直接回代解(3.2.1)得，

$$x_n=\frac{b_n^{(n)}}{a_{nn}^{(n)}},x_k=(b_k^{(k)}-\sum_{j=k+1}^{n}a_{kj}^{(k)}x_j)/a_{kk}^{(k)},k=n-1,n-2,\cdots,1 \tag{3.2.4}$$

并且有 $\det\boldsymbol{A}=a_{11}^{(1)}a_{22}^{(2)}\cdots a_{nn}^{(n)}\neq 0$，由以上顺序消去过程可得如下定理.

定理2.1 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 非奇异，则通过两行互换总可使 $a_{kk}^{(k)}\neq 0,k=1,2,\cdots,n-1$. 可将方程组(3.1.1)转化为(3.2.1)并求得方程组(3.1.1)的解为(3.2.4)，且有 $\det\boldsymbol{A}=a_{11}^{(1)}\cdots a_{nn}^{(n)}$.

如果不做行交换，则使 $a_{kk}^{(k)}\neq 0$ 的条件如下.

定理2.2 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 非奇异，且各阶顺序主子式 $\Delta_k=\det\boldsymbol{A}_k=\begin{vmatrix}a_{11}&\cdots&a_{1k}\\ \vdots& &\vdots\\ a_{k1}&\cdots&a_{kk}\end{vmatrix}\neq 0$，

则 $a_{kk}^{(k)}\neq 0,k=1,2,\cdots,n-1$.

证明 用归纳法，当 $k=1,\Delta_1=a_{11}\neq 0$，故 $a_{11}^{(1)}=a_{11}\neq 0$. 现假设 $(k-1)$ 成立，即 $\Delta_i\neq 0$，对 $i=1,2,\cdots,k-1$ 已推出 $a_{ii}^{(i)}\neq 0$，故 Gauss 消去法能进行 $(k-1)$ 步消元，$\boldsymbol{A}$ 已约化为 $\boldsymbol{A}^{(k)}$，即

$$\boldsymbol{A}_k=\begin{bmatrix}a_{11}^{(1)}&\cdots&a_{1k}^{(1)}\\ \vdots&\ddots&\vdots\\ a_{k1}^{(1)}&\cdots&a_{kk}^{(1)}\end{bmatrix}\longrightarrow\begin{bmatrix}a_{11}^{(1)}&\cdots&a_{1n}^{(1)}\\ &a_{22}^{(2)}\cdots&a_{2n}^{(2)}\\ &&\ddots&\vdots\\ &&&a_{kk}^{(k)}\end{bmatrix}=\boldsymbol{A}_k^{(k)}$$

$$\Delta_k=\det\boldsymbol{A}_k=a_{11}^{(1)}a_{22}^{(2)}\cdots a_{kk}^{(k)}\neq 0$$

故 $a_{kk}^{(k)}\neq 0$ 对 $k=1,2,\cdots,n$ 均成立，证毕.

在整个消去法消元过程中，k 从 1 到 $(n-1)$ 共需乘除法运算次数为

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

加减法次数为

$$\sum_{i=1}^{n-1}(n-k)(n+1-k)=\frac{n^3}{3}-\frac{n}{3}$$

回代过程中由公式(3.2.4)可知乘除法次数为$\frac{n(n+1)}{2}$，加减法次数为$\frac{n(n-1)}{2}$，于是 Gauss 消去法的乘除法总次数为$\frac{n^3}{3}+n^2-\frac{1}{3}n$，加减法次数为$\frac{n^3}{3}+\frac{n^2}{2}-\frac{5}{6}n$.

例 3.4 用 Gauss 消去法解方程组

$$\begin{cases}2x_1+x_2+x_3=4\\ x_1+3x_2+2x_3=6\\ x_1+2x_2+2x_3=5\end{cases}$$

并求 $\det \boldsymbol{A}$.

解 消元得

$$[\boldsymbol{A}\mid \boldsymbol{b}]=\left[\begin{array}{ccc|c}2&1&1&4\\1&3&2&6\\1&2&2&5\end{array}\right]\longrightarrow\left[\begin{array}{ccc|c}2&1&1&4\\0&5/2&3/2&4\\0&3/2&3/2&3\end{array}\right]\longrightarrow\left[\begin{array}{ccc|c}2&1&1&4\\&5/2&3/2&4\\&&3/5&3/5\end{array}\right]$$

再由(3.2.4)回代，得解 $x_1=x_2=x_3=1$，$\det \boldsymbol{A}=2\cdot\frac{5}{2}\cdot\frac{3}{5}=3$.

3.2.2 消去法与矩阵三角分解

上述 Gauss 消去法的消元过程从矩阵变换角度看，就是进行了$(n-1)$次的 Gauss 变换，即

$$\boldsymbol{L}_{n-1}^{-1}\cdots\boldsymbol{L}_2^{-1}\boldsymbol{L}_1^{-1}\boldsymbol{A}^{(1)}=\boldsymbol{A}^{(n)}=\boldsymbol{U}$$

若令 $\boldsymbol{L}^{-1}=\boldsymbol{L}_{n-1}^{-1}\boldsymbol{L}_{n-2}^{-1}\cdots\boldsymbol{L}_1^{-1}$，则 $\boldsymbol{L}=\boldsymbol{L}_1\boldsymbol{L}_2\cdots\boldsymbol{L}_{n-1}=\begin{bmatrix}1&&&\\l_{21}&1&&\\\vdots&l_{32}&\ddots&\\&\vdots&&\\l_{n1}&l_{n2}&\cdots&1\end{bmatrix}$，则由 $\boldsymbol{L}^{-1}\boldsymbol{A}=\boldsymbol{U}$，得

$$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U}$$

其中 $\boldsymbol{L}$ 为单位下三角矩阵，$\boldsymbol{U}$ 为上三角矩阵.

定理 2.3 设 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 非奇异，且 $\boldsymbol{A}$ 的顺序主子式 $\Delta_i\neq 0(i=1,2,\cdots,n-1)$，则存在唯一的单位下三角矩阵 $\boldsymbol{L}$ 和上三角矩阵 $\boldsymbol{U}$，使 $\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U}$.

证明 存在性已从上面 $\boldsymbol{A}$ 的 Gauss 变换中得到，下面只证唯一性. 假定 $\boldsymbol{A}$ 有两种不同的分解式 $\boldsymbol{A}=\boldsymbol{L}_1\boldsymbol{U}_1=\boldsymbol{L}_2\boldsymbol{U}_2$，其中 $\boldsymbol{L}_1,\boldsymbol{L}_2$ 为单位下三角矩阵 $\boldsymbol{U}_1,\boldsymbol{U}_2$ 为上三角矩阵，因 $\boldsymbol{A}$ 非奇异，故 $\boldsymbol{L}_1,\boldsymbol{L}_2,\boldsymbol{U}_1,\boldsymbol{U}_2$ 均非奇异，于是上式用 $\boldsymbol{L}_1^{-1}$ 左乘，用 $\boldsymbol{U}_2^{-1}$ 右乘，则得

$$\boldsymbol{U}_1\boldsymbol{U}_2^{-1}=\boldsymbol{L}_1^{-1}\boldsymbol{L}_2$$

因 $\boldsymbol{U}_2^{-1}$ 仍为上三角矩阵，则 $\boldsymbol{U}_1\boldsymbol{U}_2^{-1}$ 为上三角矩阵，而 $\boldsymbol{L}_1^{-1}$ 仍为单位下三角矩阵，故 $\boldsymbol{L}_1^{-1}\boldsymbol{L}_2=\boldsymbol{I}$，且 $\boldsymbol{U}_1\boldsymbol{U}_2^{-1}=\boldsymbol{I}$. 由此可得 $\boldsymbol{L}_2=\boldsymbol{L}_1,\boldsymbol{U}_2=\boldsymbol{U}_1$. 证毕.

将 $\boldsymbol{A}$ 分解为单位下三角矩阵 $\boldsymbol{L}$ 及上三角矩阵 $\boldsymbol{U}$ 的乘积 $\boldsymbol{A}=\boldsymbol{L}\boldsymbol{U}$，称为 $\boldsymbol{A}$ 的 Doolittle(杜里特尔)分解.

3.2.3 列主元消去法

在顺序消元过程中，只要 $a_{kk}^{(k)} \neq 0 (k=1,2,\cdots,n-1)$ 即可进行计算，但如果 $|a_{kk}^{(k)}|$ 很小，则将导致舍入误差增长，使解的误差很大，见下例.

例 3.5 用 Gauss 消去法求解方程组

$$\begin{cases} 0.000\ 1x_1 + 2x_2 = 1 \\ 2x_1 + 3x_2 = 2 \end{cases}$$

解 因 $\Delta_1 = 0.000\ 1 \neq 0, \Delta_2 = \det A = -3.999\ 7 \neq 0$，故方程有唯一解，且精确解为 $x^* = (0.250\ 018\ 75, 0.499\ 987\ 50)^T$.

若用 Gauss 消去法取四位有效数字计算，可得解 $\tilde{x} = (0.000\ 0, 0.500\ 0)^T$，$\tilde{x}$ 与 x^* 比较，误差很大，若将两个方程互换为

$$\begin{cases} 2x_1 + 3x_2 = 2 \\ 0.000\ 1x_1 + 2x_2 = 1 \end{cases}$$

仍用 Gauss 消去法求解，则得 $x = (0.250\ 0, 0.500\ 0)^T$，它有四位有效数字，即 $\| x^* - x \|_\infty \leqslant \frac{1}{2} \times 10^{-4}$.

这例子表明通过行交换可避免舍入误差增长，这就是列主元消去法的基本思想。其计算步骤是：

第 1 步，在 $A = A^{(1)} = (a_{ij}^{(1)})$ 中的第 1 列选主元，即 $|a_{i_1 1}| = \max\limits_{1 \leqslant i \leqslant n} |a_{i1}^{(1)}|$，$i_1$ 行为主元. 若 $i_1 > 1$，将 $[A^{(1)} | b^{(1)}]$ 的第 i_1 行与第 1 行互换，再按消元公式计算得到 $[A^{(2)} | b^{(2)}]$. 假定上述过程已进行 $(k-1)$ 步，得到 $[A^{(k)} | b^{(k)}]$.

第 k 步，在 $A^{(k)}$ 中第 k 列选主元，$|a_{i_k k}^{(k)}| = \max\limits_{k \leqslant i \leqslant n} |a_{ik}^{(k)}|$，若 $i_k > k$，则在 $[A^{(k)} | b^{(k)}]$ 中将 i_k 与 k 行互换（若 $i_k = k$ 则不动），再按公式 (3.2.2)、(3.2.3) 求出 $(A^{(k+1)} | b^{(k+1)})$. 对 $k = 1,2,\cdots,n-1$，重复以上过程则得 $[A^{(n)} | b^{(n)}]$，如果某个 k 出现主元 $|a_{kk}^{(k)}| = 0$（或 ≈ 0），则表明 $\det A = 0$，方程没有唯一解或严重病态，否则可由 (3.2.4) 求得解.

上述每步行交换后再消元相当于

$$L_k^{-1} I_{i_k k} [A^{(k)} | b^{(k)}] = [A^{(k+1)} | b^{(k+1)}], k = 1,2,\cdots,n-1 \tag{3.2.5}$$

其中 L_k^{-1} 是指标为 k 的初等下三角矩阵，$I_{i_k k}$ 为初等排列矩阵，$i_k = k$ 时 $I_{i_k k} = I$，表示不换行，经过 $(n-1)$ 步换行与消元，A 化为上三角矩阵 $A^{(n)} = U$，即

$$L_{n-1}^{-1} I_{i_{n-1} n-1} \cdots L_2^{-1} I_{i_2 2} L_1^{-1} I_{i_1 1} A = A^{(n)} = U$$

$$A = I_{i_1 1} L_1 I_{i_2 2} L_2 \cdots I_{i_{n-1} n-1} L_{n-1} U$$

它也表明当 A 非奇异时，存在排列矩阵 P（若干初等排列矩阵的乘积），使 $PA = LU$，其中 L 为单位下三角矩阵，其元素 $|l_{ij}| \leqslant 1$，U 为上三角矩阵.

例 3.6 用列主元消去法解 $Ax = b$，其中

$$[\boldsymbol{A}\,|\,\boldsymbol{b}]=\left[\begin{array}{ccc|c}-0.002 & 2 & 2 & 0.4\\ 1 & 0.781\,25 & 0 & 1.381\,6\\ 3.996 & 5.562\,5 & 4 & 7.417\,8\end{array}\right]$$

解 记$[\boldsymbol{A}^{(1)}\,|\,\boldsymbol{b}^{(1)}]=[\boldsymbol{A}\,|\,\boldsymbol{b}]$.第1步,选主元$a_{31}^{(1)}=3.996$,$i_1=3$,第3行与第1行互换.再由式(3.2.2)、(3.2.3)计算得

$$[\boldsymbol{A}^{(2)}\,|\,\boldsymbol{b}^{(2)}]=\left[\begin{array}{ccc|c}3.996 & 5.562\,5 & 4 & 7.417\,8\\ 0 & -0.610\,77 & -1.001\,0 & -0.474\,71\\ 0 & 2.002\,8 & 2.002\,0 & 0.403\,71\end{array}\right]$$

第2步,对$\boldsymbol{A}^{(2)}$选主元$a_{32}^{(2)}=2.002\,8$,$i_2=3$,将$[\boldsymbol{A}^{(2)}\,|\,\boldsymbol{b}^{(2)}]$中第3行与第2行交换,再消元得

$$[\boldsymbol{A}^{(3)}\,|\,\boldsymbol{b}^{(3)}]=\left[\begin{array}{ccc|c}3.996 & 5.562\,5 & 4 & 7.417\,8\\ 0 & 2.002\,8 & 2.002\,0 & 0.403\,71\\ 0 & 0 & -0.390\,47 & -0.35159\end{array}\right]$$

消元结束.由回代公式(3.2.4)求得解

$$\boldsymbol{x}=(1.927\,3,-0.698\,50,0.900\,43)^{\mathrm{T}}$$

此例的精确解为$\boldsymbol{x}^*=(1.927\,30,-0.698\,496,0.900\,423)^{\mathrm{T}}$,可见结果精度较高.若不选列主元Gauss消去法,求解得$\boldsymbol{x}=(1.930\,0,-0.686\,95,0.888\,88)^{\mathrm{T}}$,误差较大.

除列主元消去法外,还有一种消去法,是在$\boldsymbol{A}$的所有元素a_{ij}中选主元,称为全主元消去法.因计算量较大且应用列主元已能满足实际要求,故不再讨论.目前很多数学软件库都有列主元消去法,可直接调用.

3.3 直接三角分解法

3.3.1 Doolittle分解法

上节定理2.3已经证明当$\boldsymbol{A}\in\mathbf{R}^{n\times n}$非奇异,且$\Delta_i\neq0(i=1,2,\cdots,n-1)$,则$\boldsymbol{A}$可做$\boldsymbol{LU}$分解,即$\boldsymbol{A}=\boldsymbol{LU}$,其中$\boldsymbol{L}$为单位下三角矩阵,$\boldsymbol{U}$为上三角矩阵.现在可直接通过矩阵乘法求得$\boldsymbol{L}$及$\boldsymbol{U}$的元素,于是解方程(3.1.1)就转化为求解

$$\boldsymbol{LUx}=\boldsymbol{b} \tag{3.3.1}$$

若令$\boldsymbol{Ux}=\boldsymbol{y}$,则解方程(3.1.1)转化为求两个三角方程

$$\boldsymbol{Ly}=\boldsymbol{b}\quad\text{及}\quad\boldsymbol{Ux}=\boldsymbol{y} \tag{3.3.2}$$

下面直接用矩阵乘法求$\boldsymbol{U}$及$\boldsymbol{L}$的元素,由

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

直接得到

$$u_{1j}=a_{1j},j=1,2,\cdots,n$$
$$l_{i1}=\frac{a_{i1}}{u_{11}},i=2,3,\cdots,n \tag{3.3.3}$$

若已求得 $\boldsymbol{U}$ 的$(i-1)$行及 $\boldsymbol{L}$ 的$(i-1)$列,则由矩阵乘法有

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

可得

$$u_{ij}=a_{ij}-\sum_{k=1}^{i-1}l_{ik}u_{kj},j=i,i+1,\cdots,n \tag{3.3.4}$$

这就求得 $\boldsymbol{U}$ 的第 i 行元素,求 $\boldsymbol{L}$ 的第 i 列可由

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

若 $u_{jj}\neq0$,可得

$$l_{ij}=\frac{1}{u_{jj}}\left(a_{ij}-\sum_{k=1}^{j-1}l_{ik}u_{kj}\right),i=j+1,\cdots,n \tag{3.3.5}$$

计算规律是先由(3.3.3)求 $\boldsymbol{U}$ 的第 1 行和 $\boldsymbol{L}$ 的第 1 列元素,再由(3.3.4)求 $\boldsymbol{U}$ 的第 i 行($i=2,3,\cdots,n$)元素 u_{ij},再由(3.3.5)计算 $\boldsymbol{L}$ 的第 i 列元素 $l_{ij}(i=j+1,\cdots,n)$,求出 $\boldsymbol{L}$ 及 $\boldsymbol{U}$ 后再解方程(3.3.2),其计算公式为

$$y_1=b_1,y_i=b_i-\sum_{j=1}^{i-1}l_{ij}y_j,i=2,3,\cdots,n \tag{3.3.6}$$

$$x_n=\frac{y_n}{u_{nn}},x_i=\frac{1}{u_{ii}}\left(y_i-\sum_{j=i+1}^{n}u_{ij}x_j\right),i=n-1,\cdots,2,1 \tag{3.3.7}$$

例 3.7 用 Doolittle 分解法求方程

$$\begin{bmatrix}2&1&1\\1&3&2\\1&2&2\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}4\\6\\5\end{bmatrix}$$

的解.

解 先用公式(3.3.3)~(3.3.5)求出 $u_{11}=2,u_{12}=u_{13}=1,l_{21}=1/2,l_{31}=1/2,u_{22}=5/2,u_{23}=3/2,l_{32}=3/5,u_{33}=3/5$,于是

$$\boldsymbol{L}=\begin{bmatrix}1&&\\1/2&1&\\1/2&3/5&1\end{bmatrix},\quad \boldsymbol{U}=\begin{bmatrix}2&1&1\\&5/2&3/2\\&&3/5\end{bmatrix},$$

再由(3.3.6)求得 $\boldsymbol{Ly}=\boldsymbol{b}$ 的解 $\boldsymbol{y}=(4,4,3/5)^{\mathrm{T}}$

由(3.3.7)求得 $\boldsymbol{Ux}=\boldsymbol{y}$ 的解 $\boldsymbol{x}=(1,1,1)^{\mathrm{T}}$.

3.3.2 Cholesky 分解与平方根法

当 $A\in\mathbf{R}^{n\times n}$ 对称正定时，A 的顺序主子式 $\Delta_i>0(i=1,\cdots,n)$，故由定理 2.3 知，$A=LU$ 的分解存在且唯一，其中 L 为单位下三角矩阵，U 为上三角矩阵，且 $u_{ii}\neq0$.

定理3.1 若 $A\in\mathbf{R}^{n\times n}$ 对称，且 A 的顺序主子式 $\Delta_i\neq0(i=1,2,\cdots,n)$，则 A 可唯一分解为 $A=LDL^{\mathrm{T}}$，其中 L 为单位下三角矩阵，D 为对角矩阵 $D=\mathrm{diag}(d_1,\cdots,d_n)$.

证明 由定理 2.3 可知 $A=LU$，而 $u_{ii}\neq0$，故

$$U=\begin{bmatrix}u_{11}&u_{12}&\cdots&u_{1n}\\&u_{22}&\cdots&u_{2n}\\&&\ddots&\vdots\\&&&u_{nn}\end{bmatrix}=\begin{bmatrix}u_{11}&&&\\&u_{22}&&\\&&\ddots&\\&&&u_{nn}\end{bmatrix}\begin{bmatrix}1&\frac{u_{12}}{u_{11}}&\cdots&\frac{u_{1n}}{u_{11}}\\&1&\cdots&\frac{u_{2n}}{u_{22}}\\&&\ddots&\vdots\\&&&1\end{bmatrix}=D\tilde{U}$$

这里 $D=\mathrm{diag}(d_1,\cdots,d_n)=\mathrm{diag}(u_{11},\cdots,u_{nn})$，$A=LD\tilde{U}$，$\tilde{U}$ 为单位上三角矩阵. 因 $A=A^{\mathrm{T}}=\tilde{U}^{\mathrm{T}}(DL^{\mathrm{T}})$. 由 $A=LU$ 的分解唯一性，得 $\tilde{U}^{\mathrm{T}}=L$，于是有 $A=LDL^{\mathrm{T}}$. 证毕.

定理 3.2 若 $A\in\mathbf{R}^{n\times n}$ 对称正定，则存在唯一的对角元为正的下三角矩阵 L，使 A 分解为

$$A=LL^{\mathrm{T}} \tag{3.3.8}$$

这种分解称为 Cholesky 分解.

证明 由定理 3.1 可知 $A=L_1DL_1^{\mathrm{T}}$，这里 L_1 为单位下三角矩阵，$D=\mathrm{diag}(d_1,d_2,\cdots,d_n)$，$d_i=u_{ii}$. 由于 A 的顺序主子式 $\Delta_k=u_{11}\cdots u_{kk}=d_1\cdots d_k(k=1,\cdots,n)$，因 A 正定，故 $\Delta_k>0(k=1,2,\cdots,n)$，可推出 $d_k>0(k=1,2,\cdots,n)$. 若记 $D^{1/2}=\mathrm{diag}(d_1^{1/2},\cdots,d_n^{1/2})$，于是有 $A=L_1D^{1/2}D^{1/2}L_1^{\mathrm{T}}$. 若 $L=L_1D^{1/2}$，则 L 为下三角矩阵，且对角元为正，故有 $A=(L_1D^{1/2})(L_1D^{1/2})^{\mathrm{T}}=LL^{\mathrm{T}}$，即为(3.3.8). 证毕.

利用 Cholesky 分解式(3.3.8)将求方程组(3.1.1)的解转化为求方程 $LL^{\mathrm{T}}x=b$ 的解. 令 $L^{\mathrm{T}}x=y$，则得

$$Ly=b \quad 及 \quad L^{\mathrm{T}}x=y \tag{3.3.9}$$

根据矩阵乘法，由 $A=LL^{\mathrm{T}}$，求 L 的元素

$$A=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\vdots&\vdots&&\vdots\\a_{n1}&a_{n2}&\cdots&a_{nn}\end{bmatrix}=\begin{bmatrix}l_{11}&&&\\l_{31}&l_{22}&&\\\vdots&\vdots&\ddots&\\l_{n1}&l_{n2}&\cdots&l_{nn}\end{bmatrix}\begin{bmatrix}l_{11}&l_{21}&\cdots&l_{n1}\\&l_{22}&\cdots&l_{n2}\\&&\ddots&\vdots\\&&&l_{nn}\end{bmatrix}$$

得

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

当 $i=j$ 有

$$l_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2\right)^{1/2}, j = 1,2,\cdots,n \tag{3.3.10}$$

当 $i>j$,得

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}}{l_{jj}}, i = j+1,\cdots,n \tag{3.3.11}$$

注意当 $j=1$ 时有 $l_{11}=\sqrt{a_{11}}$,$l_{i1}=\dfrac{a_{i1}}{l_{11}}$,对 $j=2,3,\cdots,n$ 由(3.3.10),(3.3.11)逐列求得 $\boldsymbol{L}$ 的元素 l_{ij},这就是 $\boldsymbol{A}$ 的 Cholesky 分解,然后再解(3.3.9)的两个三角方程组,得

$$y_1 = b_1/l_{11}, y_i = \left(b_i - \sum_{k=1}^{i-1} l_{ik}y_k\right)/l_{ii}, i = 2,3,\cdots,n \tag{3.3.12}$$

及

$$x_n = y_n/l_{nn}, x_i = \left(y_i - \sum_{k=i+1}^{n} l_{ki}x_k\right)/l_{ii}, i = n-1,\cdots,1 \tag{3.3.13}$$

这就是对称正定方程组的平方根法.其工作量约为 Doolittle 分解法的一半.另外,由于

$$\sum_{k=1}^{j} l_{jk}^2 = a_{jj}(j=1,2,\cdots,n),\text{而 } l_{jj}>0(j=1,2,\cdots,n)$$

故有 $\max\limits_{\substack{1\leqslant j\leqslant n\\ 1\leqslant k\leqslant j}} |l_{jk}| \leqslant \max\limits_{1\leqslant j\leqslant n} \sqrt{a_{jj}}$

这表明分解过程中矩阵 $\boldsymbol{L}$ 中元素 l_{jk} 的数量级不增长,因此平方根法计算是数值稳定的.

例 3.8 用平方根法求以下方程组的解.

$$\begin{bmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -0.5 \\ 1.25 \end{bmatrix}$$

解 先验证系数矩阵 $\boldsymbol{A}$ 对称正定,对称显然,$\Delta_1=4>0$,$\Delta_2=\begin{vmatrix} 4 & -1 \\ -1 & 4.25 \end{vmatrix}=16>0$,

$\Delta_3=\det \boldsymbol{A}=16>0$,故 $\boldsymbol{A}$ 对称正定,可用 Cholesky 分解(3.3.10),(3.3.11)计算,求得

$$l_{11}=2,\ l_{21}=-0.5,\ l_{31}=0.5,\ l_{22}=2,\ l_{32}=1.5,\ l_{33}=1$$

即 $\boldsymbol{L}=\begin{bmatrix} 2 & & \\ -0.5 & 2 & \\ 0.5 & 1.5 & 1 \end{bmatrix}$,求解 $\boldsymbol{Ly}=\boldsymbol{b}$.由公式(3.3.12)得 $\boldsymbol{y}=(3,0.5,-1)^{\mathrm{T}}$,再由 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{y}$ 的公式(3.3.13)求得 $\boldsymbol{x}=(2,1,-1)^{\mathrm{T}}$.

Cholesky 分解法要用到开方运算,为避免开方运算,可将 $\boldsymbol{A}$ 分解为 $\boldsymbol{A}=\boldsymbol{LDL}^{\mathrm{T}}$(其中 $\boldsymbol{L}$ 为单位下三角矩阵),再分别解方程组 $\boldsymbol{Ly}=\boldsymbol{b}$ 及 $\boldsymbol{DL}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{y}$ 或 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{D}^{-1}\boldsymbol{y}$,这种方法称为改进平方根法,可作为练习自行推导.

3.3.3 三对角方程组的追赶法

在许多科学计算问题中,常常所要求解的方程组为三对角方程组,即

$$\boldsymbol{Ax}=\boldsymbol{f} \tag{3.3.14}$$

其中

$$\mathrm{A}=\begin{bmatrix} b_1 & c_1 & & \\ a_2 & b_2 & c_2 & \\ & \ddots & \ddots & c_{n-1} \\ & & a_n & b_n \end{bmatrix}, \boldsymbol{f}=\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} \tag{3.3.15}$$

并满足条件

$$\begin{cases} |b_1|>|c_1|>0 \\ |b_i|\geqslant|a_i|+|c_i|, a_ic_i\neq 0, i=2,3,\cdots,n-1 \\ |b_n|>|a_n|>0 \end{cases} \tag{3.3.16}$$

称 $\boldsymbol{A}$ 为对角占优的三对角矩阵，对这种简单方程可通过对 $\boldsymbol{A}$ 的三角分解建立计算量更少的求解公式.现将 $\boldsymbol{A}$ 分解为下三角矩阵 $\boldsymbol{L}$ 及单位上三角矩阵 $\boldsymbol{U}$ 的乘积.即 $\boldsymbol{A}=\boldsymbol{LU}$，其中

$$\boldsymbol{L}=\begin{bmatrix} \alpha_1 & & & \\ \gamma_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \gamma_n & \alpha_n \end{bmatrix}, \boldsymbol{U}=\begin{bmatrix} 1 & \beta_1 & & \\ & 1 & \ddots & \\ & & \ddots & \beta_{n-1} \\ & & & 1 \end{bmatrix} \tag{3.3.17}$$

直接用矩阵乘法公式可得到

$$a_i=\gamma_i, i=2,3,\cdots,n$$
$$b_1=\alpha_1, b_i=\gamma_i\beta_{i-1}+\alpha_i, i=2,3,\cdots,n$$
$$c_i=\alpha_i\beta_i, i=1,2,\cdots,n-1$$

于是有

$$\begin{cases} \gamma_i=a_i, \alpha_1=b_1, \alpha_i=b_i-a_i\beta_{i-1}, i=2,3,\cdots,n \\ \beta_i=c_i/\alpha_i, i=1,\cdots,n-1 \end{cases} \tag{3.3.18}$$

由此可见将 $\boldsymbol{A}$ 分解为 $\boldsymbol{L}$ 及 $\boldsymbol{U}$，只需计算 $\{\alpha_i\}$ 及 $\{\beta_i\}$ 两组数，然后解 $\boldsymbol{Ly}=f$，计算公式为

$$y_1=f_1/\alpha_1, y_i=(f_i-a_iy_{i-1})/\alpha_i, i=2,3,\cdots,n \tag{3.3.19}$$

再解 $\boldsymbol{Ux}=\boldsymbol{y}$ 则得

$$x_n=y_n,\ x_i=y_i-\beta_ix_{i+1},\ i=n-1,\cdots,1 \tag{3.3.20}$$

整个求解过程是先由(3.3.18)及(3.3.19)求 $\{\alpha_i\}$，$\{\beta_i\}$ 及 $\{y_i\}$，这时 $i=1,\cdots,n$ 是“追”的过程，再由(3.3.20)求出 $\{x_i\}$，这时 $i=n,\cdots,1$ 是往回“赶”，故求解方程组(3.3.14)的整个过程称为追赶法.它只用 $(5n-4)$ 次乘除法运算，计算量只是 $O(n)$，而通常方程组求解计算量为 $O(n^3)$，另外，追赶法计算 $\{\alpha_i\}$，$\{\beta_i\}$ 是数值稳定的，因为在式(3.3.16)表示的条件下，可证明

$$0<|\beta_i|<1 \quad 及 \quad 0<|\alpha_i|<|a_i|+|b_i|, \quad i=1,2,\cdots,n$$

所以，追赶法是一种计算量少及数值稳定的好算法.

3.4 向量和矩阵范数

3.4.1 内积与向量范数

为了研究方程组 $Ax=b$ 解的误差和迭代法收敛性,需对向量 $x\in\mathbf{R}^n$ 及矩阵 $\mathbf{A}\in\mathbf{R}^{n\times n}$ 的"大小"引进一种度量,就要定义范数,它是向量"长度"概念的直接推广,通常用 $\mathbf{R}^n$ 表示 n 维实向量空间,$\mathbf{C}^n$ 表示 n 维复向量空间.

定义 4.1 设 $x,y\in\mathbf{R}^n$(或 $\mathbf{C}^n$),$x=(x_1,\cdots,x_n)^{\mathrm{T}}$,$y=(y_1,\cdots,y_n)^{\mathrm{T}}$,实数$(x,y)=\sum_{i=1}^{n}x_iy_i$

或复数$(x,y)=y^{\mathrm{H}}x=\sum_{i=1}^{n}x_i\bar{y}_i$($y^{\mathrm{H}}=\bar{y}^{\mathrm{T}}$,$\bar{y}$为 y 的共轭),称为向量 x 与 y 的数量积也称内积.

非负实数 $\|x\|_2=(x,x)^{1/2}=\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}$,称为向量 x 的欧氏范数或 2-范数.

定理 4.1 设 $x,y\in\mathbf{R}^n$(或 $\mathbf{C}^n$)则内积有以下性质:

(1) $(x,x)\geqslant 0$,当且仅当 $x=0$ 时等号成立;

(2) $(\alpha x,y)=\alpha(x,y)$,$\alpha\in\mathbf{R}^1$ 或$(x,\alpha y)=\alpha(x,y)$,$\alpha\in\mathbf{C}^1$;

(3) $(x,y)=(y,x)$,或$(x,y)=\overline{(y,x)}$,$x,y\in\mathbf{C}^n$;

(4) $(x+y,z)=(x,z)+(y,z)$,$x,y,z\in\mathbf{R}^n$;

(5) $|(x,y)|\leqslant\|x\|_2\|y\|_2$. (3.4.1)

称为 Cauch-Schwarz 不等式.

(6) $\|x+y\|_2\leqslant\|x\|_2+\|y\|_2$,称为三角不等式.

定义 4.2 向量 $x\in\mathbf{R}^n$ 的某个实值函数 $N(x)$,记作 $\|x\|$,若满足下列条件:

(1) $\|x\|\geqslant 0$,当且仅当 $x=0$ 时等号成立(正定性);

(2) $\|\alpha x\|=|\alpha|\,\|x\|$,$\alpha\in\mathbf{R}$(齐次性);

(3) $\|x+y\|\leqslant\|x\|+\|y\|$(三角不等式);

则称 $N(x)=\|x\|$ 是 $\mathbf{R}^n$ 上的一个向量范数.

对于 $\|x\|_2=\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}$,由内积性质可知它满足定义 4.2 的三个条件,故它是一种向量范数.此外还有以下几种常用的向量范数.

$$\|x\|_\infty=\max_{1\leqslant i\leqslant n}|x_i|\text{(称为}\infty\text{-范数)}$$

$$\|x\|_1=\sum_{i=1}^{n}|x_i|\text{(称为 1-范数)}$$

容易验证 $\|x\|_\infty$及 $\|x\|_1$ 均满足定义 4.2 的三个条件.更一般的还可定义

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

但只有 $p=1,2,\infty$ 时的三种范数是常用的向量范数.

例如给定 $\boldsymbol{x}=(1,2,-3)^{\mathrm{T}}$,则可求出 $\|\boldsymbol{x}\|_1=6$, $\|\boldsymbol{x}\|_2=\sqrt{14}$, $\|\boldsymbol{x}\|_\infty=3$.

定理 4.2 设 $N(\boldsymbol{x})=\|\boldsymbol{x}\|$ 是 $\mathbf{R}^n$ 上任一种向量范数,则 $N(\boldsymbol{x})$ 是向量 $\boldsymbol{x}$ 的分量 $x_1,x_2,\cdots,x_n$ 的连续函数.

定理 4.3 设 $\|\cdot\|_s$ 与 $\|\cdot\|_t$ 是 $\mathbf{R}^n$ 上任意两种向量范数,则存在常数 $c_1\geqslant c_2>0$,使

$$c_1\|\boldsymbol{x}\|_s\geqslant\|\boldsymbol{x}\|_t\geqslant c_2\|\boldsymbol{x}\|_s \tag{3.4.2}$$

不等式称为向量范数等价性.

以上两定理证明可见[2],[3].

3.4.2 矩阵范数

矩阵 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 可看成 $n\times n$ 维向量,如果直接将向量的 2-范数用于矩阵 $\boldsymbol{A}$,则可定义

$$F(\boldsymbol{A})=\|\boldsymbol{A}\|_F=\Big(\sum_{i,j=1}^{n}a_{ij}^2\Big)^{1/2} \tag{3.4.3}$$

称为矩阵 $\boldsymbol{A}$ 的 Frobenius 范数,简称 F-范数.它显然满足向量范数的三条性质,但由于矩阵还有乘法运算,因此矩阵范数的定义中应增加新条件.

定义 4.3 如果 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 的某个非负实函数 $N(\boldsymbol{A})$,记作 $\|\boldsymbol{A}\|$,满足条件:

(1) $\|\boldsymbol{A}\|\geqslant 0$,当且仅当 $\boldsymbol{A}=\boldsymbol{0}$(零矩阵)时等号成立;

(2) $\|\alpha\boldsymbol{A}\|=|\alpha|\,\|\boldsymbol{A}\|$, $\alpha\in\mathbf{R}$;

(3) $\|\boldsymbol{A}+\boldsymbol{B}\|\leqslant\|\boldsymbol{A}\|+\|\boldsymbol{B}\|$, $\boldsymbol{A},\boldsymbol{B}\in\mathbf{R}^{n\times n}$;

(4) $\|\boldsymbol{AB}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$, $\boldsymbol{A},\boldsymbol{B}\in\mathbf{R}^{n\times n}$;

则称 $N(\boldsymbol{A})=\|\boldsymbol{A}\|$ 为 $\mathbf{R}^{n\times n}$ 上的一种矩阵范数.

显然 $\|\boldsymbol{A}\|_F$ 满足定义中的四个条件[(3),(4)两条均可由 Cauchy-Schwarz 不等式(3.4.1)证明],故 $\|\boldsymbol{A}\|_F$ 是一种矩阵范数.

除矩阵自身的运算外,在解方程中矩阵乘向量的运算即 $\boldsymbol{Ax}$,也是必不可少的.因此要求所引进的范数应满足条件:

$$\|\boldsymbol{Ax}\|_v\leqslant\|\boldsymbol{A}\|_v\|\boldsymbol{x}\|_v \tag{3.4.4}$$

上式称为相容性条件.为使引进的矩阵范数满足条件(3.4.4),我们给出以下定义.

定义 4.4 设 $\boldsymbol{x}\in\mathbf{R}^n$, $\boldsymbol{A}\in\mathbf{R}^{n\times n}$,当给定向量范数 $\|\cdot\|_v$ 时可定义

$$\|\boldsymbol{A}\|_v=\max_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\|\boldsymbol{Ax}\|_v}{\|\boldsymbol{x}\|_v}=\max_{\|\boldsymbol{x}\|_v=1}\|\boldsymbol{Ax}\|_v \tag{3.4.5}$$

称为矩阵的算子范数或从属范数.

定理 4.4 设 $\boldsymbol{x}\in\mathbf{R}^n$, $\boldsymbol{A}\in\mathbf{R}^{n\times n}$, $\|\boldsymbol{x}\|_v$ 是 $\mathbf{R}^n$ 上的一种向量范数,则由(3.4.5)定义的 $\|\boldsymbol{A}\|_v$ 是一种矩阵范数,且满足相容性条件(3.4.4).

证明 因 $\|\boldsymbol{Ax}\|_v$ 是 $\mathbf{R}^n$ 中有界闭集

$$D=\{x \mid x=(x_1,\cdots,x_n)^{\mathrm{T}}, \|x\|_v=1\}$$

上的连续函数，故 $\|Ax\|_v$ 在 D 上有最大值，即 $\exists x_0 \in D$ 使 $\|Ax_0\|_v=\max\limits_{\|x\|_v=1}\|Ax\|_v=\|A\|_v$，而对 $\forall x\in\mathbf{R}^n, x\neq 0$，令 $\bar{x}=\dfrac{x}{\|x\|_v}$

故
$$\frac{\|Ax\|_v}{\|x\|_v}=\left\|A\frac{x}{\|x\|_v}\right\|_v=\|Ax\|_v, x\in D$$

所以
$$\|A\|_v=\max_{x\neq 0}\frac{\|Ax\|_v}{\|x\|_v}\geqslant\frac{\|Ax\|_v}{\|x\|_v}$$

从而当 $x\neq\mathbf{0}$ 时(3.4.4)成立，而 $x=\mathbf{0}$ 时(3.4.4)显然也成立.

下面验证由(3.4.5)定义的 $\|A\|_v$ 满足矩阵定义的四个条件. 条件(1),(2)显然成立. 下面只需证(3),(4). 由(3.4.5)及(3.4.4)有

$$\|A+B\|_v=\max_{\|x\|_v=1}\|(A+B)x\|_v\leqslant\max_{\|x\|_v=1}(\|Ax\|+\|Bx\|)$$
$$\leqslant\max_{\|x\|_v=1}\|Ax\|_v+\max_{\|x\|_v=1}\|Bx\|_v=\|A\|_v+\|B\|_v$$

故(3)成立，又因

$$\|ABx\|_v\leqslant\|A\|_v\|Bx\|_v\leqslant\|A\|_v\|B\|_v\|x\|_v$$

故有 $\|AB\|_v=\max\limits_{x\neq 0}\dfrac{\|ABx\|_v}{\|x\|_v}\leqslant\|A\|_v\|B\|_v$. 证毕.

定理 4.5 设 $x\in\mathbf{R}^n, A\in\mathbf{R}^{n\times n}$ 则

$$\|A\|_\infty=\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}| \quad (A\text{ 的行范数})$$

$$\|A\|_1=\max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}| \quad (A\text{ 的列范数})$$

$$\|A\|_2=\sqrt{\rho(A^{\mathrm{T}}A)} \quad (A\text{ 的 2 范数})$$

这里 $\rho(\cdot)$ 为矩阵的谱半径.

证明 这里只给出 $\|A\|_1$ 的证明，其余可见[2]. 设 $x=(x_1,\cdots,x_n)^{\mathrm{T}}, A=(a_{ij})\in\mathbf{R}^{n\times n}$，由

$$\|Ax\|_1=\sum_{i=1}^{n}\left|\sum_{j=1}^{n}a_{ij}x_j\right|\leqslant\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}||x_j|=\sum_{j=1}^{n}\sum_{i=1}^{n}|a_{ij}||x_j|$$
$$\leqslant\left(\max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|\right)\sum_{j=1}^{n}|x_j|=\mu\|x\|_1$$

其中 $\mu=\max\limits_{1\leqslant j\leqslant n}\sum\limits_{i=1}^{n}|a_{ij}|=\sum\limits_{i=1}^{n}|a_{ij_0}|$. 故对 $\forall x\in\mathbf{R}^n, x\neq\mathbf{0}$ 均有

$\dfrac{\|Ax\|_1}{\|x\|_1}\leqslant\mu$. 下面证明 $\exists x^0\neq\mathbf{0}$，使得 $\dfrac{\|Ax^0\|_1}{\|x^0\|_1}=\mu$.

现设 $x^0=(x_1^0,\cdots,x_n^0)^{\mathrm{T}}$，其中 $x_j^0=\begin{cases}1, \text{当 } j=j_0.\\ 0, j\neq j_0.\end{cases}$

显然 $\| x^0 \|_1 = 1$，且 $\| Ax^0 \|_1 = \sum\limits_{i=1}^{n} \left| \sum\limits_{j=1}^{n} a_{ij}^0 x_j^0 \right| = \sum\limits_{i=1}^{n} |a_{i_0}| = \mu$. 证毕.

从定理可以看出，计算 $\| A \|_\infty$ 及 $\| A \|_1$ 较容易，而计算 $\| A \|_2$ 时因为要求 $A^{\mathrm{T}}A$ 的特征值，所以较为困难. 但当 A 对称时，有 $\| A \|_2 = \sqrt{\rho(A^{\mathrm{T}}A)} = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)} = \sqrt{\lambda_{\max}(A^2)} = \sqrt{\lambda_{\max}^2(A)} = \rho(A)$

例 3.9 已知 $A = \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix}$，求 $\| A \|_\infty$，$\| A \|_F$，$\| A \|_2$，$\| A \|_1$.

解 $\| A \|_\infty = 7$，$\| A \|_1 = 6$，$\| A \|_F = \sqrt{30} \approx 5.477$.

$$A^{\mathrm{T}}A = \begin{bmatrix} 10 & -14 \\ -14 & 20 \end{bmatrix}, \quad \det(\lambda I - A^{\mathrm{T}}A) = \begin{vmatrix} \lambda - 10 & 14 \\ 14 & \lambda - 20 \end{vmatrix} = \lambda^2 - 30\lambda + 4 = 0$$

$$\lambda_{1,2} = 15 \pm \sqrt{221} \qquad \| A \|_2 = \sqrt{29.866} \approx 5.465$$

定理 4.6 对任何 $A \in \mathbf{R}^{n\times n}$，$\| \cdot \|$ 为任一种从属范数则

$$\rho(A) \leqslant \| A \| \tag{3.4.6}$$

反之，对任意 $\varepsilon > 0$，至少存在一种从属范数 $\| \cdot \|_\varepsilon$，使

$$\| A \|_\varepsilon \leqslant \rho(A) + \varepsilon \tag{3.4.7}$$

证明 只证定理前一半，后半证明可见[3].

设 λ 为 A 的特征值，则 $\exists x \neq \mathbf{0}$，使 $Ax = \lambda x$，由(3.4.4)有 $\| Ax \| \leqslant \| A \| \| x \|$ 及 $\| Ax \| = |\lambda| \| x \|$

得 $$|\lambda| \leqslant \| A \| \quad \text{即} \quad \rho(A) = \max|\lambda| \leqslant \| A \|$$

证毕.

注意，当 A 对称时，$\rho(A) = \| A \|_2$，表明(3.4.6)可取等号.

定理 4.7 （矩阵范数等价性）对 $\mathbf{R}^{n\times n}$ 上的任两种范数 $\| \cdot \|_s$ 及 $\| \cdot \|_t$，存在常数 $c_1 \geqslant c_2 > 0$，使

$$c_1 \| A \|_s \geqslant \| A \|_t \geqslant c_2 \| A \|_s \tag{3.4.8}$$

证明略.

例 3.10 证明

$$\frac{1}{\sqrt{n}} \| A \|_F \leqslant \| A \|_2 \leqslant \| A \|_F \tag{3.4.9}$$

解 因 $A^{\mathrm{T}}A$ 对称半正定，$\lambda_i(A^{\mathrm{T}}A) \geqslant 0$，利用特征值性质(3.1.5)及 $\| A \|_2$ 和 $\| A \|_F$ 的定义，有

$$\| A \|_2^2 = \lambda_{\max}(A^{\mathrm{T}}A) \leqslant \lambda_1(A^{\mathrm{T}}A) + \cdots + \lambda_n(A^{\mathrm{T}}A) = \mathrm{tr}(A^{\mathrm{T}}A)$$

$$= \sum_{j=1}^{n} a_{1j}^2 + \sum_{j=1}^{n} a_{2j}^2 + \cdots + \sum_{j=1}^{n} a_{nj}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 = \| A \|_F$$

另一方面

$$\| A \|_2^2 = \lambda_{\max}(A^{\mathrm{T}}A) \geqslant \frac{1}{n}[\lambda_1(A^{\mathrm{T}}A) + \cdots + \lambda_n(A^{\mathrm{T}}A)] = \frac{1}{n}\| A \|_F^2$$

于是 $\| A \|_2 \geqslant \frac{1}{\sqrt{n}} \| A \|_F$，证毕.

定理 4.8 设 $B \in \mathbf{R}^{n\times n}$，$\| B \| < 1$ 则 $I \pm B$ 非奇异，且

$$\| (I \pm B)^{-1} \| \leqslant \frac{1}{1 - \| B \|} \tag{3.4.10}$$

证明 用反证法. 假定$(I + B)$奇异，则齐次方程$(I \pm B)x = 0$ 有非零解 $x \in \mathbf{R}^n$，即 $\exists\, x^0 \neq 0$ 使

$$x^0 = \pm Bx^0 \quad \text{故} \quad \frac{\| Bx^0 \|}{\| x^0 \|} = 1$$

而 $\| B \| = \max\limits_{x\neq 0} \frac{\| Bx \|}{\| x \|} \geqslant \frac{\| Bx^0 \|}{\| x^0 \|} = 1$ 与 $\| B \| < 1$ 的假设矛盾，故$(I + B)$非奇异.

又 $(I \pm B)(I \pm B)^{-1} = I$，即$(I \pm B)^{-1} \pm B(I \pm B)^{-1} = I$，

得 $(I \pm B)^{-1} = I \mp B(I \pm B)^{-1}$，取范数 $\| (I \pm B)^{-1} \| \leqslant 1 + \| B \| \, \| (I \pm B)^{-1} \|$，于是得(3.4.10). 证毕.

3.5 误差分析与病态方程组

3.5.1 矩阵条件数与扰动方程组误差界

在解方程组 $Ax = b$ 时，由于各种原因，A 或 b 往往有误差，从而使得解也产生误差.

例 3.11 方程组

$$\begin{bmatrix} 2 & 6 \\ 2 & 6.000\,1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 8.000\,1 \end{bmatrix}$$

的准确解为 $x^* = (1.1)^{\mathrm{T}}$，当 A 与 b 有微小变化时，如变为方程

$$\begin{bmatrix} 2 & 6 \\ 2 & 5.999\,9 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 8.000\,2 \end{bmatrix}$$

则准确解变为 $\tilde{x} = (10, -2)^{\mathrm{T}}$，它表明 A，b 的微小扰动引起方程解 x 的很大变化，这就是病态方程.

定义 5.1 求解线性方程组 $Ax = b$ 时，若 A 或 b 有微小扰动 $\| \delta A \|$ 或 $\| \delta b \|$ 时，解 x 的误差 $\| \delta x \|$ 很大，则称此方程组为病态方程组，相应的系数矩阵 A 称为病态矩阵. 反之，若此时 $\| \delta x \|$ 很小，则称方程组为良态方程组，矩阵 A 为良态矩阵.

注意方程组是否病态与用什么数值方法无关，它是由方程自身性质决定的. 在例 3.11 中因为行列式 $\det A = 2 \cdot 10^{-4} \approx 0$，因此出现病态. 但有时 A 从表面上看性质很好，也可能是病态的.

例 3.12 方程组 $Ax = b$ 表示为

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

它的准确解 $x^* = (1,1,1,1)^T$，A 对称正定且 $\det A = 1$. 表面看性质“较好”，但若对右端 b 作微小变化，如方程改为 $A(x+\delta x) = (32.1, 22.9, 33.1, 30.9)^T$，则解变为 $x+\delta x = (9.2, -12.6, 4.5, -1.1)^T$. 这里 b 的相对误差大约只有 $\frac{1}{200}$，但解的相对误差却很大，故 A 也是病态矩阵.

那么如何判断 A 是否病态？先给出如下定义.

定义 5.2 设 $A \in \mathbf{R}^{n\times n}$ 非奇异，$\|\cdot\|$ 为矩阵的任一种从属范数，则

$$\mathrm{Cond}(A) = \|A\|\,\|A^{-1}\| \tag{3.5.1}$$

称为矩阵 A 的条件数.

从定义看到矩阵条件数依赖于范数的选取，如范数为 2 - 范数，则记为 $\mathrm{Cond}(A)_2 = \|A\|_2\|A^{-1}\|_2 = \sqrt{\frac{\lambda_{\max}(A^TA)}{\lambda_{\min}(A^TA)}}$. 同理有 $\mathrm{Cond}(A)_\infty$，$\mathrm{Cond}(A)_1$ 等等.

条件数有以下性质：

(1) $\mathrm{Cond}(A)\geqslant 1$，$\mathrm{Cond}(A) = \mathrm{Cond}(A^{-1})$；

(2) $\mathrm{Cond}(\alpha A) = \mathrm{Cond}(A)$，$\alpha \in \mathbf{R}^1$，$\alpha \neq 0$；

(3) U 为正交矩阵，则

$$\mathrm{Cond}(U)_2 = 1,\ \mathrm{Cond}(A)_2 = \mathrm{Cond}(UA)_2 = \mathrm{Cond}(AU)_2;$$

(4) 若 λ_1 与 λ_n 为 A 的按模最大与最小特征值，则

$$\mathrm{Cond}(A)\geqslant \frac{|\lambda_1|}{|\lambda_n|}.$$

若 A 对称，则 $\mathrm{Cond}(A)_2 = \frac{|\lambda_1|}{|\lambda_n|}$.

下面给出扰动方程组解的误差分析. 先考察 b 有扰动 δb，则扰动方程为

$$A(x+\delta x) = b + \delta b,\ b \neq 0$$

由于 $Ax = b$，故得 $A\delta x = \delta b$，$\delta x = A^{-1}\delta b$，于是 $\|\delta x\| \leqslant \|A^{-1}\|\,\|\delta b\|$，再由 $Ax = b$，有 $\|A\|\,\|x\| \geqslant \|b\|$，即 $\frac{1}{\|x\|} \leqslant \frac{\|A\|}{\|b\|}$，故得

$$\frac{\|\delta x\|}{\|x\|} \leqslant \|A^{-1}\|\,\|A\|\frac{\|\delta b\|}{\|b\|} = \mathrm{Cond}(A)\frac{\|\delta b\|}{\|b\|} \tag{3.5.2}$$

下面再研究方程 $Ax = b$，当 A 有扰动 δA 时，其解 $x+\delta x$ 的误差分析. 此时扰动方程为

$$(A+\delta A)(x+\delta x) = b$$

因 $Ax = b$，故有

$$(A+\delta A)\delta x = (-\delta A)x \tag{3.5.3}$$

因 A^{-1}存在，$A+\delta A=A(I+A^{-1}\delta A)$

若假定 $\|A^{-1}\delta A\|\leqslant\|A^{-1}\|\,\|\delta A\|<1$ 则由定理 4.8 可知$(I+A^{-1}\delta A)$非奇异，并有

$$\|(I+A^{-1}\delta A)^{-1}\|\leqslant\frac{1}{1-\|A^{-1}\|\,\|\delta A\|}$$

由(3.5.3)可得

$$\delta x=-(A+\delta A)^{-1}(\delta A)x=(I+A^{-1}\delta A)^{-1}A^{-1}(-\delta A)x$$

即
$$\|\delta x\|\leqslant\|(I+A^{-1}\delta A)^{-1}\|\,\|A^{-1}\|\,\|\delta A\|\,\|x\|$$

故有

$$\frac{\|\delta x\|}{\|x\|}\leqslant\frac{\|A^{-1}\|\,\|\delta A\|}{1-\|A^{-1}\|\,\|\delta A\|}=\frac{\mathrm{Cond}(A)\dfrac{\|\delta A\|}{\|A\|}}{1-\mathrm{Cond}(A)\dfrac{\|\delta A\|}{\|A\|}}\tag{3.5.4}$$

综合(3.5.2)与(3.5.4)的结果可得到如下定理.

定理 5.1 设 $A\in\mathbf{R}^{n\times n}$，$b\neq0$，$x$ 是方程组 $Ax=b$ 的解，$x+\delta x$ 是扰动方程组$(A+\delta A)(x+\delta x)=b+\delta b$ 的解.如果 $\|A^{-1}\|\,\|\delta A\|<1$，则有误差估计

$$\frac{\|\delta x\|}{\|x\|}\leqslant\frac{\mathrm{Cond}(A)}{1-\mathrm{Cond}(A)\dfrac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)\tag{3.5.5}$$

此定理包含了(3.5.2)及(3.5.4)两种特例.当 $\delta A=0$ 则得(3.5.2)式；当 $\delta b=0$ 时则得(3.5.4)式.实际使用时，由于误差 $\|\delta A\|$ 及 $\|\delta b\|$ 很小，并且一般是可以控制的，故定理中的条件 $\|A^{-1}\|\,\|\delta A\|<1$ 是可以成立的.

从(3.5.5)看到，当 A 的条件数 Cond(A)很大时，解的相对误差 $\dfrac{\|\delta x\|}{\|x\|}$ 也很大，故方程组为病态.在例 3.11 中 $\|A\|_\infty=8.000\,1$，而

$$A^{-1}=\begin{bmatrix}30\,000.5 & -30\,000\\ -10\,000 & 10\,000\end{bmatrix},\quad\|A^{-1}\|_\infty=60\,000.5$$

于是 $\mathrm{Cond}(A)_\infty=\|A\|_\infty\|A^{-1}\|_\infty\approx480\,010$

条件数很大，故方程是严重病态的.

例 3.13 Hibert 矩阵是一个著名的病态矩阵，记作

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

它是一个对称正定矩阵，当 $n\geqslant3$ 时它是病态矩阵.例如 $n=3$，$\|H_3\|_\infty=\dfrac{11}{6}$，

$$\boldsymbol{H}_3^{-1}=\begin{bmatrix}9 & -36 & 30\\ -36 & 192 & -180\\ 30 & -180 & 180\end{bmatrix},\ \|\boldsymbol{H}_3^{-1}\|_\infty=408$$

故 $\mathrm{Cond}(\boldsymbol{H}_3)_\infty=748$.

另外还有 $\mathrm{Cond}(\boldsymbol{H}_4)_\infty=2.84\times10^4$, $\mathrm{Cond}(\boldsymbol{H}_6)_\infty=2.9\times10^7$. 等等

因此 $\boldsymbol{H}_n$ 是严重的病态矩阵,且 n 越大 $\mathrm{Cond}(\boldsymbol{H}_n)$越大.

例 3.14 在例 3.12 的方程组中可算出 $\boldsymbol{A}$ 的特征值$\lambda_1\approx30.288\ 7$, $\lambda_2\approx3.858$, $\lambda_3\approx0.843\ 1$, $\lambda_4\approx0.010\ 15$,故

$$\mathrm{Cond}(\boldsymbol{A})_2=\frac{\lambda_1}{\lambda_4}\approx2\ 984$$

例中 $\delta\boldsymbol{b}=(0.1,-0.1,0.1,-0.1)^{\mathrm{T}}$, $\boldsymbol{b}=(32,23,33,31)^{\mathrm{T}}$, $\delta\boldsymbol{x}=(8.2,-13.6,3.5,-2.1)^{\mathrm{T}}$,实际相对误差是

$$\frac{\|\delta\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}=\frac{16.396\ 95}{2}=8.198\ 5$$

而根据(3.5.2)的误差估计为

$$\frac{\|\delta\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}\leqslant\mathrm{Cond}(\boldsymbol{A})_2\frac{\|\delta\boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2}\approx9.943$$

这与实际相差不大,即相对误差放大了将近 3 000 倍.故方程为病态方程组.

定理 5.2 设方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$, $\boldsymbol{b}\neq0$,若实际求得解为 $\bar{\boldsymbol{x}}$,则

$$\frac{\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|}\leqslant\mathrm{Cond}(\boldsymbol{A})\frac{\|\boldsymbol{b}-\boldsymbol{A}\bar{\boldsymbol{x}}\|}{\|\boldsymbol{b}\|}\tag{3.5.6}$$

证明 记剩余 $\boldsymbol{r}=\boldsymbol{b}-\boldsymbol{A}\bar{\boldsymbol{x}}\neq0$,则 $\boldsymbol{A}(\boldsymbol{x}-\bar{\boldsymbol{x}})=\boldsymbol{r}$, $\boldsymbol{x}-\bar{\boldsymbol{x}}=\boldsymbol{A}^{-1}\boldsymbol{r}$, $\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|\leqslant\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{r}\|$,又 $\frac{1}{\|\boldsymbol{x}\|}\leqslant\frac{\|\boldsymbol{A}\|}{\|\boldsymbol{b}\|}$故有 $\frac{\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|}{\|\boldsymbol{b}\|}\|\boldsymbol{r}\|=\mathrm{Cond}(\boldsymbol{A})\frac{\|\boldsymbol{b}-\boldsymbol{A}\bar{\boldsymbol{x}}\|}{\|\boldsymbol{b}\|}$. 证毕.

这是关于方程解的事后误差估计,它表明如果方程组病态,即使剩余 $\|\boldsymbol{r}\|$ 很小,解的相对误差仍可能很大.

3.5.2 病态方程组的解法

如果 $\boldsymbol{A}$ 的条件数 $\mathrm{Cond}(\boldsymbol{A})\gg1$,则 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 为病态方程,但计算 $\mathrm{Cond}(\boldsymbol{A})$时需要求 $\boldsymbol{A}^{-1}$,计算量很大,相当于解方程组,在实际中常可通过求解过程直观地判断方程组的病态性质,如果解方程时出现下述情况之一,则可能是"病态"方程组.

(1) 在列主元消去法中出现小主元;

(2) 在计算过程中行或列几乎线性相关或三角分解中对角元出现近似零的元素;

(3) 矩阵 A 的元素数量级相差很大且无规律;

(4) 剩余 $\boldsymbol{r}=\boldsymbol{A}\bar{\boldsymbol{x}}-\boldsymbol{b}$ 很小,而解 $\|\bar{\boldsymbol{x}}\|$ 很大,又达不到精度要求.

对病态方程组求解可采用以下措施:

(1) 采用高精度运算,减轻病态影响,例如用双倍字长运算.

(2) 用预处理方法改善 A 的条件数,即选择非奇矩阵 $P, Q \in \mathbf{R}^{n\times n}$,使 $PAQ(Q^{-1}x) = Pb$ 与 $Ax = b$ 等价,而 $\tilde{A} = PAQ$ 的条件数比 A 改善,则求 $\tilde{A}\tilde{x} = \tilde{b} = Pb$ 的解 $\tilde{x} = Q^{-1}x$,即 $x = Q\tilde{x}$ 为原方程的解.计算时可选择 P, Q 为对角矩阵或三角矩阵.

(3) 平衡方法,当 A 中元素的数量级相差很大,可采用行均衡或列均衡的方法改善 A 的条件数.设 $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ 非奇异,计算 $s_i = \max\limits_{1\leqslant j\leqslant n} |a_{ij}|\ (i = 1, 2, \cdots, n)$,令 $D = \mathrm{diag}\left(\frac{1}{s_1}, \frac{1}{s_2}, \cdots, \frac{1}{s_n}\right)$,于是求 $Ax = b$ 等价于求 $DAx = Db$,或 $\tilde{A}x = \tilde{b}$.这时 $\tilde{A} = DA$ 的条件数可得到改善,这就是行均衡法.

例 3.15 给定方程组 $Ax = b$ 为

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

A 的条件数 $\mathrm{Cond}(A)_\infty \approx 10^4$,若用行均衡法可取 $D = \mathrm{diag}(10^{-4}, 1)$,则平衡后的方程 $\tilde{A}x = \tilde{b}$ 为

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

$\mathrm{Cond}(\tilde{A})_\infty \approx 4$,用三位有效数字的列主元消去法求解得 $x = (1.00, 1.00)^{\mathrm{T}}$.

习 题 三

1. 用 Gauss 消去法求解下列方程组.

(1)
$$\begin{cases} \frac{1}{4}x_1 + \frac{1}{5}x_2 + \frac{1}{6}x_3 = 9 \\ \frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = 8 \\ \frac{1}{2}x_1 + x_2 + 2x_3 = 8 \end{cases}$$

(2)
$$\begin{bmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \\ 5 \\ -5 \end{bmatrix}$$

2. 用列主元消去法求解方程组

$$\begin{cases} 12x_1 - 3x_2 + 3x_3 = 15 \\ -18x_1 + 3x_2 - x_3 = -15 \\ x_1 + x_2 + x_3 = 6 \end{cases}$$

并求出系数矩阵 A 的行列式 $\det A$ 的值.

3. 设 $A = (a_{ij}) \in \mathbf{R}^{n\times n}$ 是对称正定矩阵,经过 Gauss 消去法一步后 A 约化为

$$\begin{bmatrix} a_{11} & a_1^{\mathrm{T}} \\ 0 & A_2 \end{bmatrix}$$

其中 $A_2=(a_{ij}^{(2)})\in\mathbf{R}^{(n-1)\times(n-1)}$，证明

(1) A 的对角元素 $a_{ii}>0$，$i=1,2,\cdots,n$.

(2) A_2 也对称正定.

(3) $a_{ii}^{(2)}\leqslant a_{ii}$，$i=2,3,\cdots,n$.

4. 用 Doolittle 分解法求习题 1(2)方程组的解.

5. 下述矩阵能否作 Doolittle 分解，若能分解，分解式是否唯一？

$$A=\begin{bmatrix}1&2&3\\2&4&1\\4&6&7\end{bmatrix},B=\begin{bmatrix}1&1&1\\2&2&1\\3&3&1\end{bmatrix},C=\begin{bmatrix}1&2&6\\2&5&15\\6&15&46\end{bmatrix}$$

6. 设 $A\in\mathbf{R}^{n\times n}$ 为非奇异阵，若 A 有唯一分解 $A=LU$(其中 L 为单位下三角矩阵，U 为上三角矩阵)，试证 A 的顺序主子式 $\Delta_i\neq0(i=1,2,\cdots,n-1)$.

7. 设 L 为非奇异下三角矩阵.

(1) 列出逐次代入求解 $Lx=f$ 的公式.

(2) 上述求解过程需要多少次乘除法运算.

(3) 给出求 L^{-1} 的计算公式.

8. 用追赶法解三对角方程组 $Ax=b$，其中

$$A=\begin{bmatrix}2&-1&0&0&0\\-1&2&-1&0&0\\0&-1&2&-1&0\\0&0&-1&2&-1\\0&0&0&-1&2\end{bmatrix},b=\begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}$$

9. 用平方根法解方程组

$$\begin{bmatrix}16&4&8\\4&5&-4\\8&-4&22\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}-4\\3\\10\end{bmatrix}$$

10. 设 $x\in\mathbf{R}^n$，证明(1) $\|x\|_\infty\leqslant\|x\|_1\leqslant n\|x\|_\infty$

(2) $\|x\|_\infty\leqslant\|x\|_2\leqslant\sqrt{n}\|x\|_\infty$

11. 设 $A\in\mathbf{R}^{n\times n}$ 对称正定，记 $\|x\|_A=(Ax,x)^{\frac{1}{2}}$，$\forall x\in\mathbf{R}^n$，证明 $\|x\|_A$ 是 $\mathbf{R}^n$ 上的一种向量范数.

12. 设

$$A=\begin{bmatrix}0.6&0.5\\0.1&0.3\end{bmatrix}$$

计算 A 的行范数，列范数及 F－范数.

13. 设 $\|x\|$ 为 $\mathbf{R}^n$ 上任一种范数，$P\in\mathbf{R}^{n\times n}$ 是非奇异的，定义 $\|x\|_P=\|Px\|$，证明 $\|A\|_P=\|PAP^{-1}\|$.

14. 给出 $A=\begin{bmatrix}2&-1&0\\-1&2&-1\\0&-1&2\end{bmatrix}$，求 $\mathrm{Cond}(A)_2$.

15. 求下面两个方程组的解，并利用矩阵的条件数估计 $\dfrac{\|\delta x\|}{\|x\|}$.

$$\begin{bmatrix} 240 & -319 \\ -179 & 240 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 3 \\ 4 \end{bmatrix},\text{即 } \boldsymbol{Ax}=\boldsymbol{b}$$

$$\begin{bmatrix} 240 & -319.5 \\ -179.5 & 240 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 3 \\ 4 \end{bmatrix},\text{即}(\boldsymbol{A}+\delta\boldsymbol{A})(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}$$

第四章　解线性方程组的迭代法

4.1　迭代法及其收敛性

4.1.1　向量序列及矩阵序列的极限

定义 1.1　设 $\mathbf{R}^n$ 中的向量序列 $\{\boldsymbol{x}^{(k)}\}_0^{\infty}$，$\boldsymbol{x}^{(k)}=(x_1^{(k)},\cdots,x_n^{(k)})^{\mathrm{T}}$，如果存在 $\boldsymbol{x}=(x_1,\cdots,x_n)^{\mathrm{T}}\in\mathbf{R}^n$，使

$$\lim_{k\to\infty}x_i^{(k)}=x_i,\ i=1,2,\cdots,n$$

则称向量序列 $\{\boldsymbol{x}^{(k)}\}_0^{\infty}$ 收敛于 $\boldsymbol{x}$，记作 $\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}$.

同样，矩阵序列 $\{\boldsymbol{A}^{(k)}\}_0^{\infty}\in\mathbf{R}^{n\times n}$，$\boldsymbol{A}^{(k)}=(a_{ij})\in\mathbf{R}^{n\times n}$，若有 $\lim\limits_{k\to\infty}a_{ij}^{(k)}=a_{ij}$，$i,j=1,2,\cdots,n$，则称矩阵序列 $\{\boldsymbol{A}^{(k)}\}_0^{\infty}$ 收敛于 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$，记作 $\lim\limits_{k\to\infty}\boldsymbol{A}^{(k)}=\boldsymbol{A}$.

概据定义，显然有

$$\lim_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}\quad\Leftrightarrow\quad\lim_{k\to\infty}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\|=\boldsymbol{0}$$

$$\lim_{k\to\infty}\boldsymbol{A}^{(k)}=\boldsymbol{A}\quad\Leftrightarrow\quad\lim_{k\to\infty}\|\boldsymbol{A}^{(k)}-\boldsymbol{A}\|=\boldsymbol{0}$$

定理 1.1　$\lim\limits_{k\to\infty}\boldsymbol{A}^{(k)}=\boldsymbol{0}$ 的充分必要条件是

$$\lim_{k\to\infty}\boldsymbol{A}^{(k)}\boldsymbol{x}=\boldsymbol{0},\ \forall\,\boldsymbol{x}\in\mathbf{R}^n\tag{4.1.1}$$

其中两个极限的右端分别指零矩阵与零向量.

证明　对任一种矩阵从属范数有

$$\|\boldsymbol{A}^{(k)}\boldsymbol{x}\|\leqslant\|\boldsymbol{A}^{(k)}\|\,\|\boldsymbol{x}\|$$

由 $\lim\limits_{k\to\infty}\boldsymbol{A}^{(k)}=\boldsymbol{0}\to\lim\limits_{k\to\infty}\|\boldsymbol{A}^{(k)}\|=\boldsymbol{0}$，故式(4.1.1)对 $\forall\,\boldsymbol{x}\in\mathbf{R}^n$ 成立. 反之，若取 $\boldsymbol{x}$ 为第 j 个坐标向量 $\boldsymbol{e}_j$，则 $\lim\limits_{k\to\infty}\boldsymbol{A}^{(k)}\boldsymbol{e}_j=\boldsymbol{0}$ 表示 $\boldsymbol{A}^{(k)}$ 第 j 列元素极限均为零；当 $j=1,2,\cdots,n$ 时则证明了 $\lim\limits_{k\to\infty}\boldsymbol{A}^{(k)}=\boldsymbol{0}$. 证毕.

定理 1.2　$\lim\limits_{k\to\infty}\boldsymbol{A}^k=\boldsymbol{0}$ 的充分必要条件是 $\rho(\boldsymbol{A})<1$，其中 $\rho(\cdot)$ 为谱半径.

证明　由于 $\rho(\boldsymbol{A}^k)=[\rho(\boldsymbol{A})]^k$，而 $\rho(\boldsymbol{A}^k)\leqslant\|\boldsymbol{A}^k\|$，当 $\lim\limits_{k\to\infty}\boldsymbol{A}^k=\boldsymbol{0}$ 时 $\lim\limits_{k\to\infty}\|\boldsymbol{A}^k\|=\boldsymbol{0}$，故有 $\lim\limits_{k\to\infty}\rho(\boldsymbol{A}^k)=\boldsymbol{0}$，即

$$\lim_{k\to\infty}[\rho(\boldsymbol{A})]^k=\boldsymbol{0},\text{故 }\rho(\boldsymbol{A})<1.$$

反之，当 $\rho(\boldsymbol{A})<1$ 时，由上章定理 4.6 可知，对任给 $\varepsilon>0$，存在 $\|\boldsymbol{A}\|_\varepsilon$，使 $\|\boldsymbol{A}\|_\varepsilon \leqslant \rho(\boldsymbol{A})+\varepsilon<1$，于是 $\lim\limits_{k\to\infty}\|\boldsymbol{A}^k\|_\varepsilon \leqslant \lim\limits_{k\to\infty}\|\boldsymbol{A}\|_\varepsilon^k=\boldsymbol{0}$ 从而 $\lim\limits_{k\to\infty}\boldsymbol{A}^k=\boldsymbol{0}$. 证毕.

定理 1.3 设 $\boldsymbol{B}\in\mathbf{R}^{n\times n}$，$\|\cdot\|$ 为任一种范数，则

$$\lim_{k\to\infty}\|\boldsymbol{B}^k\|^{1/k}=\rho(\boldsymbol{B}) \tag{4.1.2}$$

4.1.2 迭代法的构造

设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 非奇异，用迭代法解方程组

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b} \tag{4.1.3}$$

首先要构造迭代序列，通常可将方程改写为

$$\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f} \tag{4.1.4}$$

并由此构造迭代法

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{f},k=0,1,\cdots \tag{4.1.5}$$

其中 $\boldsymbol{B}\in\mathbf{R}^{n\times n}$ 称为迭代矩阵. 对任意给定的初始向量 $\boldsymbol{x}^{(0)}\in\mathbf{R}^n$，由(4.1.5)可求得向量序列 $\{\boldsymbol{x}^{(k)}\}_0^\infty$. 若 $\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^*$，则 $\boldsymbol{x}^*$ 就是方程(4.1.4)(或(4.1.3))的解.

定义 1.2 若迭代法(4.1.5)生成的序列 $\{\boldsymbol{x}^{(k)}\}$ 满足

$$\lim_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^* \qquad \forall\,\boldsymbol{x}^{(0)}\in\mathbf{R}^n$$

则称迭代法(4.1.5)是收敛的.

构造的迭代法(4.1.5)是否收敛，取决于迭代矩阵的性质，先看例题.

例 4.1 给定方程组

$$\begin{cases}8x_1-3x_2+2x_3=20\\4x_1+11x_2-x_3=33\\6x_1+3x_2+12x_3=36\end{cases}$$

它的精确解 $\boldsymbol{x}^*=(3,2,1)^{\mathrm{T}}$，可构造如下迭代法

$$\begin{cases}x_1^{(k+1)}=\dfrac{1}{8}(3x_2^{(k)}-2x_3^{(k)}+20)\\x_2^{(k+1)}=\dfrac{1}{11}(-4x_1^{(k)}+x_3^{(k)}+33)\\x_3^{(k+1)}=\dfrac{1}{12}(-6x_1^{(k)}-3x_2^{(k)}+36)\end{cases} \tag{4.1.6}$$

若写成式(4.1.5)的形式，则迭代矩阵 $\boldsymbol{B}$ 及 $\boldsymbol{f}$ 可表示为：

$$\boldsymbol{B}=\begin{bmatrix}0 & \dfrac{3}{8} & -\dfrac{1}{4}\\-\dfrac{4}{11} & 0 & \dfrac{1}{11}\\-\dfrac{1}{2} & -\dfrac{1}{4} & 0\end{bmatrix},\boldsymbol{f}=\begin{bmatrix}\dfrac{5}{2}\\3\\3\end{bmatrix}$$

若取 $x^{(0)}=(0,0,0)^{\mathrm{T}}$，按式(4.1.6)迭代 10 次可得 $x^{(10)}=(3,000\ 032,1.999\ 838,0.999\ 803)^{\mathrm{T}}$，误差 $\|x^{(10)}-x^*\|_\infty=0.000\ 197$. 它表明迭代序列(4.1.6)收敛.

对于方程组(4.1.3)，构造迭代法的一般原则是将 A 分解为

$$A=M-N \tag{4.1.7}$$

其中 M 非奇异且容易求 M^{-1}，则由(4.1.3)可得

$$x=M^{-1}Nx+M^{-1}b=Bx+f \tag{4.1.8}$$

其中

$$B=M^{-1}N=I-M^{-1}A,f=M^{-1}b \tag{4.1.9}$$

这样就得到与(4.1.3)等价的(4.1.8)，从而可构造(4.1.5)的迭代法，将 A 按不同方式分解为(4.1.7)，就可得到不同的迭代矩阵 B，从而得到不同的迭代法. 通常为使 M^{-1} 容易计算，可取 M 为对角矩阵，三角矩阵或三对角矩阵等. 在例 4.1 中，$M=D=\mathrm{diag}(a_{11},\cdots,a_{nn})$，$B=I-D^{-1}A$. 式(4.1.6)就是下节将要讨论的 Jacobi 迭代法.

4.1.3 迭代法的收敛性与收敛速度

下面讨论迭代法(4.1.5)的收敛性. 若 $\lim\limits_{k\to\infty}x^{(k)}=x^*$

则

$$x^*=Bx^*+f=(I-M^{-1}A)x^*+M^{-1}b$$

即 $M^{-1}(Ax^*-b)=0$，故 x^* 即为方程(4.1.3)的解.

令 $e^{(k)}=x^{(k)}-x^*$，由(4.1.5)减去等式 $x^*=Bx^*+f$，则得

$$e^{(k+1)}=x^{(k+1)}-x^*=B(x^{(k)}-x^*)=Be^{(k)}$$

由此递推得

$$e^{(k)}=B^ke^{(0)},k=1,2,\cdots \tag{4.1.10}$$

其中 $e^{(0)}=x^{(0)}-x^*$ 与 k 无关，所以 $\lim\limits_{k\to\infty}x^{(k)}=x^*$ 等价于

$$\lim_{k\to\infty}e^{(k)}=\lim_{k\to\infty}B^ke^{(0)}=0,\forall e^{(0)}\in\mathbf{R}^n$$

即 $\lim\limits_{k\to\infty}B^k=0$. 由定理 1.2 可知 $\lim\limits_{k\to\infty}B^k=0\Leftrightarrow\rho(B)<1$，于是有如下定理.

定理 1.4 迭代法(4.1.5)对 $\forall x^{(0)}\in\mathbf{R}^n$ 收敛的充分必要条件是

$$\rho(B)<1 \tag{4.1.11}$$

其中 $\rho(B)$ 为矩阵 B 的谱半径.

例 4.2 考察例 4.1 中迭代法(4.1.6)的收敛性.

解 由

$$B=\begin{bmatrix}0&\frac{3}{8}&-\frac{1}{4}\\-\frac{4}{11}&0&\frac{1}{11}\\-\frac{1}{2}&-\frac{1}{4}&0\end{bmatrix},\det(\lambda I-B)=\begin{vmatrix}\lambda&-\frac{3}{8}&\frac{1}{4}\\\frac{4}{11}&\lambda&-\frac{1}{11}\\\frac{1}{2}&\frac{1}{4}&\lambda\end{vmatrix}$$

可得 $\det(\lambda I-B)=\lambda^3+\frac{3}{88}\lambda+\frac{7}{176}=0$. 用方程求根方法可解得 $\lambda_1=-0.308\ 2,\lambda_{2,3}=0.154\ 1\pm$

0.324 5i,

$$\rho(\boldsymbol{B})=\max|\lambda_i|=0.359\ 2<1$$

故迭代法(4.1.6)收敛.由于计算 $\rho(\boldsymbol{B})$较困难,通常可利用 $\rho(\boldsymbol{B})\leqslant\|\boldsymbol{B}\|<1$ 判断迭代法的收敛性.在本例中由于 $\|\boldsymbol{B}\|_\infty=\frac{3}{4}<1$,故迭代法(4.1.6)收敛.

于是,迭代法(4.1.5)收敛的充分条件如下.

定理 1.5 对迭代法(4.1.5),如果迭代矩阵 $\boldsymbol{B}$ 的某种范数 $q=\|\boldsymbol{B}\|<1$,则对 $\forall x^0\in\mathbf{R}^n$ 及 $f\in\mathbf{R}^n$,迭代序列 $\{x^{(k)}\}_0^\infty$ 均收敛于 x^*,且有误差估计

$$\|x^{(k)}-x^*\|\leqslant\frac{q^k}{1-q}\|x^{(1)}-x^{(0)}\| \tag{4.1.12}$$

证明 由于 $\rho(\boldsymbol{B})\leqslant\|\boldsymbol{B}\|<1$,故由定理 1.4 得 $\{x^{(k)}\}_0^\infty$ 收敛于 x^*.又由(4.1.5)知

$$\|x^{(k)}-x^*\|\leqslant\|\boldsymbol{B}\|\,\|x^{(k-1)}-x^*\|\leqslant\|\boldsymbol{B}\|\,\|x^{(k-1)}-x_{(k)}\|+\|\boldsymbol{B}\|\,\|x^{(k)}-x^*\|$$

于是 $$\|x^{(k)}-x^*\|\leqslant\frac{\|\boldsymbol{B}\|}{1-\|\boldsymbol{B}\|}\|x^{(k)}-x^{(k-1)}\|\leqslant\cdots\leqslant\frac{\|\boldsymbol{B}\|^k}{1-\|\boldsymbol{B}\|}\|x^{(1)}-x^{(0)}\|$$

即为(4.1.12),证毕.

注意,定理只给出迭代序列(4.1.5)收敛的充分条件,即使条件 $\|\boldsymbol{B}\|<1$ 对任何范数都不成立,迭代序列仍可能收敛.

例 4.3 设 $x^{(k+1)}=\boldsymbol{B}x^{(k)}+f$,其中 $\boldsymbol{B}=\begin{bmatrix}0.9 & 0\\ 0.3 & 0.8\end{bmatrix}$,$f=\begin{bmatrix}1\\ 2\end{bmatrix}$,讨论迭代序列 $\{x^{(k)}\}$ 的收敛性.

解 显然 $\|\boldsymbol{B}\|_\infty=1.1$,$\|B\|_1=1.2$,$\|B\|_2=1.043$,$\|B\|_F=\sqrt{1.54}$

表明 $\boldsymbol{B}$ 的范数均大于 1,但由于 $\rho(\boldsymbol{B})=0.9<1$,故由定理 1.4 知此迭代序列 $\{x^{(k)}\}$ 是收敛的.

下面考察迭代法(4.1.5)的收敛速度.由(4.1.10)得

$$\|e^{(k)}\|\leqslant\|\boldsymbol{B}^k\|\,\|e^{(0)}\|,\ \forall e^{(0)}\neq\mathbf{0}$$

于是

$$\frac{\|e^{(k)}\|}{\|e^{(0)}\|}\leqslant\|\boldsymbol{B}^k\| \tag{4.1.13}$$

根据算子范数定义可知

$$\|\boldsymbol{B}^k\|=\max_{e^{(0)}\neq0}\frac{\|\boldsymbol{B}^ke^{(0)}\|}{\|e^{(0)}\|}=\max_{e^{(0)}\neq0}\frac{\|e^{(k)}\|}{\|e^{(0)}\|}$$

所以 $\|\boldsymbol{B}^k\|$ 是迭代 k 次后误差向量 $e^{(k)}$ 的范数与初始误差向量 $e^{(0)}$ 的范数之比的最大值.若要求迭代 k 次后

$$\|e^{(k)}\|\leqslant\varepsilon\|e^{(0)}\|,\text{即}\frac{\|e^{(k)}\|}{\|e^{(0)}\|}\leqslant\|\boldsymbol{B}^k\|\leqslant\varepsilon$$

这里 $\varepsilon=10^{-s}$ 是一个小数,通常 $\varepsilon\ll1$,所以 $\|\boldsymbol{B}^k\|^{1/k}<1$,由 $\|\boldsymbol{B}^k\|^{1/k}\leqslant\varepsilon^{\frac{1}{k}}$,两边取对数可得

$$k\geqslant\frac{-\ln\varepsilon}{-\ln\|\boldsymbol{B}^k\|^{1/k}}=\frac{s\ln10}{-\ln\|\boldsymbol{B}^k\|^{1/k}} \tag{4.1.14}$$

它表明迭代步数 k 与 $-\frac{1}{k}\ln\|\boldsymbol{B}^k\|$ 成反比，即

$$R_k(\boldsymbol{B})=-\ln\|\boldsymbol{B}^k\|^{1/k} \tag{4.1.15}$$

愈大，迭代次数 k 愈少．于是可定义(4.1.15)式中的量 $R_k(\boldsymbol{B})$ 为迭代法的平均收敛速度．它依赖于所取的范数，若利用定理 1.3 的(4.1.2)式则有

$$\lim_{k\to\infty}R_k(\boldsymbol{B})=\lim_{k\to\infty}(-\ln\|\boldsymbol{B}^k\|^{1/k})=-\ln\rho(\boldsymbol{B})$$

定义 1.3 $R(\boldsymbol{B})=-\ln\rho(\boldsymbol{B})$ 称为迭代法(4.1.5)的渐近收敛速度．

显然 $R(\boldsymbol{B})=\lim_{k\to\infty}R_k(\boldsymbol{B})$，它与 $\boldsymbol{B}$ 取何种范数无关．由于迭代法(4.1.5)收敛，故 $\rho(\boldsymbol{B})<1$，$\rho(\boldsymbol{B})$ 越小，$-\ln\rho(\boldsymbol{B})$ 越大，迭代法收敛越快，且当迭代次数 k 满足 $k\geqslant\frac{s\ln 10}{R(\boldsymbol{B})}$ 时，有 $\frac{\|e^{(k)}\|}{\|e^{(0)}\|}\leqslant 10^{-s}$．

例 4.4 对例 4.1 中的迭代序列(4.1.6)要使相对误差 $\frac{\|e^{(k)}\|}{\|e^{(0)}\|}\leqslant 10^{-5}$，至少要迭代几次？

解 因(4.1.6)中迭代矩阵 $\boldsymbol{B}$ 的谱半径 $\rho(\boldsymbol{B})=0.359\,2$，$R(\boldsymbol{B})=-\ln\rho(\boldsymbol{B})=1.023\,876$，$k\geqslant\frac{5\ln 10}{R(\boldsymbol{B})}\approx 11.99$，因此取 $k=12$，即为所求．

4.2 Jacobi 迭代法与 Gauss－Seidel 迭代法

4.2.1 Jacobi 迭代法

将方程组(4.1.3)中系数矩阵 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 分解为

$$\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U} \tag{4.2.1}$$

其中 $\boldsymbol{D}=\mathrm{diag}(a_{11},\cdots,a_{nn})$ 为 $\boldsymbol{A}$ 的对角矩阵，

$$\boldsymbol{L}=-\begin{bmatrix}0&&&\\a_{21}&0&&\\\vdots&\vdots&\ddots&\\a_{n1}&a_{n2}&\cdots&a_{nn-1}&0\end{bmatrix},\boldsymbol{U}=-\begin{bmatrix}0&a_{12}&\cdots&a_{1n}\\&0&\cdots&a_{2n}\\&&\ddots&\vdots\\&&&0&a_{n-1n}\\&&&&0\end{bmatrix} \tag{4.2.2}$$

$-\boldsymbol{L}$，$-\boldsymbol{U}$ 分别为 $\boldsymbol{A}$ 的严格下三角矩阵与 $\boldsymbol{A}$ 的严格上三角矩阵．假定 $a_{ii}\neq 0(i=1,2,\cdots,n)$，则 $\boldsymbol{D}$ 非奇异．取 $\boldsymbol{M}=\boldsymbol{D}$，$\boldsymbol{N}=\boldsymbol{L}+\boldsymbol{U}$，则得

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}_J\boldsymbol{x}^{(k)}+\boldsymbol{f},k=0,1,\cdots \tag{4.2.3}$$

其中

$$\boldsymbol{B}_J=\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A},\boldsymbol{f}=\boldsymbol{D}^{-1}\boldsymbol{b} \tag{4.2.4}$$

(4.2.3)称为解方程组的 Jacobi 迭代法，简称 J 法．计算时可写成如下分量形式：

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), i = 1,2,\cdots,n \tag{4.2.5}$$

例 4.1 中的迭代(4.1.6)就是 Jacobi 迭代法.将定理 1.4 与定理 1.5 用于 J 法就有如下定理.

定理 2.1 J 法收敛的充分必要条件是 $\rho(\boldsymbol{B}_J)<1$,收敛的充分条件是任一种范数 $\|\boldsymbol{B}_J\|<1$.

4.2.2 Gauss-Seidel 迭代法

在 J 法的计算公式(4.2.5)中,计算 $x_i^{(k+1)}$ 时前面 $i-1$ 个值 $x_1^{(k+1)},\cdots,x_{i-1}^{(k+1)}$ 均已算出,如果将这些新值代替旧值,则(4.2.5)变成如下迭代公式

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), i = 1,2,\cdots,n \tag{4.2.6}$$

称为 Gauss-Seidel 迭代法,简称 GS 法.

在计算机中计算与 J 法相似,只需用新值 $x_i^{(k+1)}$ 取代 $x_i^{(k)}$ 即可,它只用一组长度为 n 的工作单元存放解向量 $\boldsymbol{x}$ 即可.迭代公式(4.2.6)用矩阵表示为

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

或

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

于是 GS 法可表示为

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{G}\boldsymbol{x}^{(k)} + \boldsymbol{f}_G \tag{4.2.7}$$

其中

$$\boldsymbol{G} = (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U} = \boldsymbol{I} - (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{A}, \boldsymbol{f}_G = (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{b} \tag{4.2.8}$$

矩阵 $\boldsymbol{G}$ 称为 GS 法迭代矩阵.对比式(4.1.7)中 $\boldsymbol{A}$ 的分解有 $\boldsymbol{M}_G = \boldsymbol{D}-\boldsymbol{L}$, $\boldsymbol{N}_G = \boldsymbol{U}$,根据定理 1.4 与 1.5 有如下定理.

定理 2.2 GS 法(4.2.7)收敛的充分必要条件是迭代矩阵 $\boldsymbol{G}$ 的谱半径 $\rho(\boldsymbol{G})<1$,充分条件是对任一种范数 $\|\boldsymbol{G}\|<1$.

例 4.5 用 GS 法求例 4.1 中方程组的解.

解 GS 法的迭代公式为

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{8}(20 + 3x_2^{(k)} - 2x_3^{(k)}) \\ x_2^{(k+1)} = \dfrac{1}{11}(33 - 4x_1^{(k+1)} + x_3^{(k)}) \\ x_3^{(k+1)} = \dfrac{1}{12}(36 - 6x_1^{(k+1)} - 3x_2^{(k+1)}) \end{cases}$$

取 $\boldsymbol{x}^{(0)} = (0,0,0)^{\mathrm{T}}$,迭代 5 次得到

$$\boldsymbol{x}^{(5)} = (2.999\,843, 2.000\,072, 1.000\,061)^{\mathrm{T}}$$

$$\|\boldsymbol{e}^{(5)}\|_\infty = \|\boldsymbol{x}^{(5)} - \boldsymbol{x}^*\|_\infty = 0.000\,157$$

这个结果与用 J 法迭代 10 次的误差相当.在一定条件下,若 J 法与 GS 法均收敛,则 GS 法比 J 法

约快一倍,但也可能 J 法收敛而 GS 法不收敛或相反.

4.2.3 J 法与 GS 法的收敛性

定理 2.1 及 2.2 已给出一般情形下 J 法及 GS 法的收敛性条件,但它们的计算较复杂,特别是对 GS 法求迭代矩阵 $\boldsymbol{G}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}$ 时要计算$(\boldsymbol{D}-\boldsymbol{L})^{-1}$,很不方便,下面将针对两类经常遇到的特殊矩阵,讨论解方程组(4.1.3)时 J 法及 GS 法的收敛性,它可直接由给定的系数矩阵 $\boldsymbol{A}$ 来判断收敛性,为此先给出定义

定义 2.1 若 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$满足

$$|a_{ii}|\geqslant\sum_{\substack{j=1\\j\neq i}}^{n}|a_{ij}|,\ i=1,2,\cdots,n \tag{4.2.9}$$

若对所有 $i=1,2,\cdots,n$ 上面的严格不等式成立则称 $\boldsymbol{A}$ 为严格对角占优矩阵;若式(4.2.9)中至少有一个严格不等式成立则称 $\boldsymbol{A}$ 为弱对角占优矩阵.

定义 2.2 假定 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$,$n\geqslant 2$,若存在排列矩阵 $\boldsymbol{P}\in\mathbf{R}^{n\times n}$,使

$$\boldsymbol{P}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}=\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{0} & \boldsymbol{A}_{22}\end{bmatrix} \tag{4.2.10}$$

其中 $\boldsymbol{A}_{11}\in\mathbf{R}^{r\times r}$,$\boldsymbol{A}_{22}\in\mathbf{R}^{(n-r)\times(n-r)}$$(1\leqslant r<n)$则称 $\boldsymbol{A}$ 可约,否则,若不存在排列阵 $\boldsymbol{P}$,使(4.2.10)成立,则称 $\boldsymbol{A}$ 不可约.

当 $\boldsymbol{A}$ 为可约时,显然式(4.1.3)可化为两个低阶方程组求解.$\boldsymbol{A}$ 可约就是 $\boldsymbol{A}$ 通过行列互换总可转化为式(4.2.10)的形式.例如方程组

$$\begin{bmatrix}2 & 1 & 0\\ 1 & 1 & 0\\ 0 & 1 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}3\\ 2\\ 1\end{bmatrix}$$

就是可约的,求解时可先解

$$\begin{cases}2x_1+x_2=3\\ x_1+x_2=2\end{cases}$$

得解为 $x_1=x_2=1$,再代入最后一个方程 $x_2+x_3=1$,求得解 $x_3=0$.

若取

$$\boldsymbol{P}=\begin{bmatrix}0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{bmatrix},\text{则 }\boldsymbol{P}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}=\left[\begin{array}{c|cc}1 & 1 & 0\\ \hline 0 & 1 & 1\\ 0 & 1 & 2\end{array}\right]$$

它就是(4.2.10)的形式.

定理 2.3 若 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$为严格对角占优矩阵或不可约弱对角占优矩阵,则 $a_{ii}\neq 0(i=1,\cdots,n)$,且 $\boldsymbol{A}$ 非奇异.

证明 这里只证第一部分.假定 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$严格对角占优,则由定义可知$|a_{ii}|>\sum_{j\neq i}|a_{ij}|$

$\geqslant 0$ 故有 $a_{ii}\neq 0$,对 $i=1,\cdots,n$ 成立.再用反证法证明 $\boldsymbol{A}$ 非奇异.假定 $\boldsymbol{A}$ 奇异,则 $\exists\, \boldsymbol{x}\in\mathbf{R}^n$,$\boldsymbol{x}\neq 0$ 使 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}$,

记 $|x_k|=\max\limits_{1\leqslant i\leqslant n}|x_i|\neq 0$,于是 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}$ 中第 k 个方程为

$$a_{kk}x_k=-\sum_{\substack{j=1\\ j\neq k}}^{n}a_{kj}x_j$$

从而有

$$|a_{kk}|\leqslant\sum_{j\neq k}|a_{kj}|\left|\frac{x_j}{x_k}\right|\leqslant\sum_{j\neq k}|a_{kj}|$$

与 $\boldsymbol{A}$ 严格对角占优矛盾,故 $\boldsymbol{A}$ 非奇异,即 $\det \boldsymbol{A}\neq 0$.

定理第二部分的证明可见[3].

根据本定理可知,方程组(4.1.3)的解存在唯一,且解方程的 J 法及 GS 法均成立.

定理 2.4 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 为严格对角占优矩阵或不可约弱对角占优矩阵,则解方程组(4.1.3)的 J 法及 GS 法均收敛.

证明 这里只证明当 $\boldsymbol{A}$ 为不可约弱对角占优矩阵时 GS 法收敛,其余三种情况的证明相似,可留作练习.下面证明 $\rho(\boldsymbol{G})<1$ 即可,其中 $\boldsymbol{G}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}$.

设 $\boldsymbol{G}$ 的一个特征值为 λ,若 $\lambda\geqslant 1$ 则由 $\boldsymbol{G}$ 的特征方程得

$$\det[\lambda\boldsymbol{I}-(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}]=\det(\boldsymbol{D}-\boldsymbol{L})^{-1}\cdot\det[\lambda(\boldsymbol{D}-\boldsymbol{L}-\lambda^{-1}\boldsymbol{U})]=0$$

因 $a_{ii}\neq 0(i=1,2,\cdots,n)$故 $\det(\boldsymbol{D}-\boldsymbol{L})^{-1}\neq 0$,于是

$$\det(\boldsymbol{D}-\boldsymbol{L}-\lambda^{-1}\boldsymbol{U})=0 \tag{4.2.11}$$

而矩阵 $(\boldsymbol{D}-\boldsymbol{L}-\lambda^{-1}\boldsymbol{U})$ 与 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$ 的零元素与非零元素的位置完全相同,故 $(\boldsymbol{D}-\boldsymbol{L}-\lambda^{-1}\boldsymbol{U})$ 也不可约.又因 $|\lambda|\geqslant 1$,故 $(\boldsymbol{D}-\boldsymbol{L}-\lambda^{-1}\boldsymbol{U})$ 也是弱对角占优矩阵,由定理 2.3 知 $\det(\boldsymbol{D}-\boldsymbol{L}-\lambda^{-1}\boldsymbol{U})\neq 0$,这与(4.2.11)矛盾,故 $|\lambda|\leqslant 1$,即 $\rho(\boldsymbol{G})<1$,于是 GS 法收敛.证毕.

下面给出 $\boldsymbol{A}$ 为对称正定矩阵时,解方程(4.1.3)的 J 法及 GS 法的收敛性结论.

定理 2.5 若 $\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{n\times n}$ 对称正定,则

(1) 解方程组(4.1.3)的 GS 法收敛.

(2) 若 $2\boldsymbol{D}-\boldsymbol{A}$ 也对称正定,则 J 法也收敛.

本定理(1)为下面定理 3.2 的特例,定理(2)的证明可见[3].

例 4.6 方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 中,$A=\begin{bmatrix}1 & a & a\\ a & 1 & a\\ a & a & 1\end{bmatrix}$,

证明当 $-1/2<a<1$ 时 GS 法收敛,而 J 法只在 $-1/2<a<1/2$ 时才收敛.

解 只要证 $-1/2<a<1$ 时 $\boldsymbol{A}$ 正定即可,由顺序主子式 $\Delta_1=1>0$,$\Delta_2=\begin{vmatrix}1 & a\\ a & 1\end{vmatrix}=1-a^2>0$,得 $|a|<1$.而 $\Delta_3=\det\boldsymbol{A}=1+2a^3-3a^2=(1-a)^2(1+2a)>0$ 得 $a>-1/2$ 于是得到 $-1/2<a<1$ 时 $\Delta_1>0,\Delta_2>0,\Delta_3>0$,故 $\boldsymbol{A}$ 对称正定,GS 法收敛.

对 J 法，由于迭代矩阵 $\boldsymbol{B}_J=\begin{bmatrix}0 & -a & -a\\ -a & 0 & -a\\ -a & -a & 0\end{bmatrix}$，

$$\det(\lambda \boldsymbol{I}-\boldsymbol{B}_J)=\lambda^3-3\lambda a^2+2a^3=(\lambda-a)^2(\lambda+2a)=0$$

$\rho(\boldsymbol{B})=|2a|<1$，即 $|a|<1/2$ 是 J 法收敛的充要条件. 故 J 法只在 $|a|<1/2$ 时才收敛.

当 $a=0.8$ 时，GS 法收敛，而 $\rho(\boldsymbol{B}_J)=1.6>1$，J 法不收敛，此时 $2\boldsymbol{D}-\boldsymbol{A}$ 不是正定的.

注意，求方程 $\boldsymbol{Ax}=\boldsymbol{b}$ 时如原方程换行后满足 J 法和 GS 法的收敛性条件，则应按变换后的方程构造 J 法与 GS 法. 例如，方程组

$$\begin{cases}3x_1-10x_2=-7\\ 9x_1-4x_2=5\end{cases}$$

两方程互换为

$$\begin{cases}9x_1-4x_2=5\\ 3x_1-10x_2=-7\end{cases}$$

即将 $\boldsymbol{A}=\begin{bmatrix}3 & -10\\ 9 & -4\end{bmatrix}$ 变换为 $\tilde{\boldsymbol{A}}=\begin{bmatrix}9 & -4\\ 3 & -10\end{bmatrix}$

显然 $\tilde{\boldsymbol{A}}$ 为严格对角占优矩阵，故对它构造的 J 法与 GS 法均收敛.

4.3 逐次超松弛迭代法

4.3.1 SOR 迭代公式

逐次超松弛(Successive Over Relaxation)迭代法，简称 SOR 迭代法，它是在 GS 法基础上为提高收敛速度，采用加权平均而得到的新算法，设解方程(4.1.3)的 GS 法记为

$$\bar{x}_i^{(k+1)}=\frac{1}{a_{ii}}\left(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\right),i=1,2,\cdots,n \tag{4.3.1}$$

再由 $x_i^{(k)}$ 与 $\bar{x}_i^{(k+1)}$ 加权平均得

$$x_i^{(k+1)}=(1-\omega)x_i^{(k)}+\omega\bar{x}_i^{(k+1)}=x_i^{(k)}+\omega(\bar{x}_i^{(k+1)}-x_i^{(k)}),i=1,2,\cdots,n$$

这里 $\omega>0$ 称为松弛参数，将(4.3.1)代入则得

$$x_i^{(k+1)}=(1-\omega)x_i^{(k)}+\frac{\omega}{a_{ii}}\left(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\right),i=1,2,\cdots,n \tag{4.3.2}$$

称为 SOR 迭代法，$\omega>0$ 称为松弛因子，当 $\omega=1$ 时(4.3.2)即为 GS 法，将(4.3.2)写成矩阵形式，则得

$$\boldsymbol{Dx}^{(k+1)}=(1-\omega)\boldsymbol{Dx}^{(k)}+\omega(\boldsymbol{b}+\boldsymbol{Lx}^{(k+1)}+\boldsymbol{Ux}^{(k)})$$

即

$$(\boldsymbol{D}-\omega\boldsymbol{L})\boldsymbol{x}^{(k+1)}=[(1-\omega)\boldsymbol{D}+\omega\boldsymbol{U}]\boldsymbol{x}^{(k)}+\omega\boldsymbol{b}$$

于是得 SOR 迭代的矩阵表示

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{G}_\omega \boldsymbol{x}^{(k)} + \boldsymbol{f}_\omega \tag{4.3.3}$$

其中

$$\boldsymbol{G}_\omega = (\boldsymbol{D} - \omega\boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D} + \omega\boldsymbol{U}] \tag{4.3.4}$$

$$\boldsymbol{f}_\omega = \omega(\boldsymbol{D} - \omega\boldsymbol{L})^{-1}\boldsymbol{b}$$

按(4.1.7)分解,有 $\boldsymbol{M}_\omega = \frac{1}{\omega}(\boldsymbol{D} - \omega\boldsymbol{L})^{-1}$, $\boldsymbol{N}_\omega = \frac{1}{\omega}[(1-\omega)\boldsymbol{D} + \omega\boldsymbol{U}]$.

例 4.7 给定方程组

$$\begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 24 \\ 30 \\ -24 \end{bmatrix}$$

精确解 $\boldsymbol{x}^* = (3,4,-5)^{\mathrm{T}}$,用 SOR 法求解,分别取 $\omega = 1$ 及 $\omega = 1.25$.

解 用 SOR 迭代公式(4.3.2)可得

$$\begin{cases} x_1^{(k+1)} = (1-\omega)x_1^{(k)} + \dfrac{\omega}{4}(24 - 3x_2^{(k)}) \\ x_2^{(k+1)} = (1-\omega)x_2^{(k)} + \dfrac{\omega}{4}(30 - 3x_1^{(k+1)} + x_3^{(k)}) \\ x_3^{(k+1)} = (1-\omega)x_3^{(k)} + \dfrac{\omega}{4}(-24 + x_2^{(k+1)}) \end{cases}$$

取 $\boldsymbol{x}^{(0)} = (1,1,1)^{\mathrm{T}}$,迭代 7 次后分别为

$\omega = 1$, $\boldsymbol{x}^{(7)} = (3.013\,411\,0, 3.988\,824\,1, -5.002\,794\,0)^{\mathrm{T}}$

$\omega = 1.25$, $\boldsymbol{x}^{(7)} = (3.000\,049\,8, 4.000\,258\,6, -5.000\,348\,6)^{\mathrm{T}}$

若要精确到小数后 7 位,对 $\omega = 1$(即 GS 法)需迭代 34 次,而对 $\omega = 1.25$ 的 SOR 法,只需迭代 14 次.它表明松弛因子 ω 选择的好坏,对收敛速度影响很大.

4.3.2 SOR 迭代法收敛性

根据迭代法收敛性定理,SOR 法收敛的充分必要条件为 $\rho(\boldsymbol{G}_\omega) < 1$,收敛的充分条件为 $\|\boldsymbol{G}_\omega\| < 1$,但要计算 $\boldsymbol{G}_\omega$ 比较复杂,通常都不用此结论,而直接根据方程组的系数矩阵 $\boldsymbol{A}$ 判断 SOR 迭代收敛性,下面先给出收敛必要条件.

定理 3.1 设 $\boldsymbol{A} = (a_{ij}) \in \mathbf{R}^{n\times n}$, $a_{ii} \neq 0 (i = 1,2,\cdots,n)$,则解方程 $\boldsymbol{Ax} = \boldsymbol{b}$ 的 SOR 迭代法收敛的必要条件是 $0 < \omega < 2$.

证明 由 SOR 迭代矩阵 $\boldsymbol{G}_\omega$ 的表达式(4.3.4)

$$\boldsymbol{G}_\omega = (\boldsymbol{D} - \omega\boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D} + \omega\boldsymbol{U}]$$

于是

$$\det \boldsymbol{G}_\omega = \det(\boldsymbol{D} - \omega\boldsymbol{L})^{-1}\det[(1-\omega)\boldsymbol{D} + \omega\boldsymbol{U}] = \det\boldsymbol{D}^{-1}\det[(1-\omega)\boldsymbol{D}] = (1-\omega)^n$$

另一方面,设 $\boldsymbol{G}_\omega$ 的特征值为 $\lambda_1,\cdots,\lambda_n$,由特征根性质,有

$$\rho(\boldsymbol{G}_\omega)=\max_{1\leqslant i\leqslant n}|\lambda_i|\geqslant|\lambda_1\cdots\lambda_n|^{1/n}=|\det\boldsymbol{G}_\omega|^{1/n}=|1-\omega|$$

若 SOR 法收敛,则 $\rho(\boldsymbol{G}_\omega)<1$,由 $|1-\omega|\leqslant\rho(\boldsymbol{G}_\omega)<1$,则得 $0<\omega<2$. 证毕.

定理 3.2 若 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 对称正定,且 $0<\omega<2$,则解 $\boldsymbol{Ax}=\boldsymbol{b}$ 的 SOR 迭代法(4.3.3)对 $\forall\boldsymbol{x}^0\in\mathbf{R}^n$ 迭代收敛.

证明 设 $\boldsymbol{G}_\omega$ 的特征值为 λ(可能是复数),对应特征向量 $\boldsymbol{x}\neq\boldsymbol{0}$,由(4.3.4)得

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

因 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$ 为实对称矩阵,故 $\boldsymbol{L}^{\mathrm{T}}=\boldsymbol{U}$,上式两边与 $\boldsymbol{x}$ 作内积,得

$$(1-\omega)(\boldsymbol{Dx},\boldsymbol{x})+\omega(\boldsymbol{Ux},\boldsymbol{x})=\lambda[(\boldsymbol{Dx},\boldsymbol{x})-\omega(\boldsymbol{Lx},\boldsymbol{x})]\tag{4.3.5}$$

因 $\boldsymbol{A}$ 正定,故 $\boldsymbol{D}$ 也正定,记 $(\boldsymbol{Dx},\boldsymbol{x})=p>0$. 又记 $(\boldsymbol{Lx},\boldsymbol{x})=\alpha+\mathrm{i}\beta$,由复内积性质得

$$(\boldsymbol{Ux},\boldsymbol{x})=(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x},\boldsymbol{x})=(\overline{\boldsymbol{Lx},\boldsymbol{x}})=\alpha-\mathrm{i}\beta$$

于是由(4.3.5)有

$$\lambda=\frac{(1-\omega)p+\omega\alpha-\mathrm{i}\omega\beta}{p-\omega\alpha-\mathrm{i}\omega\beta}$$

$$|\lambda|^2=\frac{[p-\omega(p-\alpha)]^2+\omega^2\beta^2}{(p-\omega\alpha)^2+\omega^2\beta^2}$$

由于 $\boldsymbol{A}$ 正定及 $0<\omega<2$,故

$$(\boldsymbol{Ax},\boldsymbol{x})=(\boldsymbol{Dx},\boldsymbol{x})-(\boldsymbol{Lx},\boldsymbol{x})-(\boldsymbol{Ux},\boldsymbol{x})=p-2\alpha>0$$

于是 $$[p-\omega(p-\alpha)]^2-(p-\omega\alpha)^2=p\omega(2-\omega)(2\alpha-p)<0$$

它表明 $|\lambda|^2$ 的分子小于分母,故有 $|\lambda|^2<1$,即

$\rho(\boldsymbol{G}_\omega)<1$,从而 SOR 迭代法收敛. 证毕.

注:当 $\omega=1$ 时 SOR 法即为 GS 法,故 GS 法也收敛,此即为定理 2.5(1)的结论.

对于 SOR 迭代法,松弛因子的选择对收敛速度影响较大,关于最优松弛因子 ω_{b} 研究较为复杂,且已有不少理论结果. 下面只给出一种简单且便于使用的结论.

定理 3.3 设 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 为对称正定的三对角矩阵,$\boldsymbol{B}_{\mathrm{J}}$ 是解方程(4.1.3)的 J 法迭代矩阵,若 $\rho(\boldsymbol{B}_{\mathrm{J}})<1$,记 $\mu=\rho(\boldsymbol{B}_{\mathrm{J}})$,则 SOR 法的最优松弛因子 ω_{b} 为

$$\omega_{\mathrm{b}}=\frac{2}{1+\sqrt{1-\mu^2}}\tag{4.3.6}$$

且

$$\rho(\boldsymbol{G}_\omega)=\begin{cases}\left[\omega\mu+\sqrt{\omega^2\mu^2-4(\omega-1)}\right]^2/4 & 0<\omega\leqslant\omega_{\mathrm{b}}\\ \omega-1 & \omega_{\mathrm{b}}\leqslant\omega<2\end{cases}\tag{4.3.7}$$

根据定理,$\rho(\omega_{\mathrm{b}})=\min\rho(\boldsymbol{G}_\omega)$,如图 4-1 所示. 由(4.3.7)可知,当 $\omega=1$,$\rho(\boldsymbol{G})=\mu^2=\rho^2(\boldsymbol{B}_{\mathrm{J}})$时,收敛速度为

$$R(\boldsymbol{G})=-\ln\rho(\boldsymbol{G})=-2\ln\rho(\boldsymbol{B}_{\mathrm{J}})=2R(\boldsymbol{B}_{\mathrm{J}}).$$

说明 GS 法比 J 法快一倍.

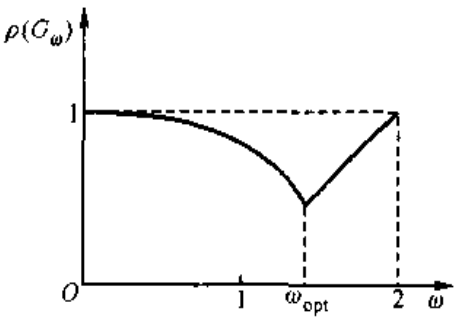

图 4-1

例 4.8 对例 4.7 中的方程组,用 SOR 迭代法求最优松弛因子 ω_b,并研究其收敛速度.

解 由于

$$\boldsymbol{A}=\begin{bmatrix}4 & 3 & 0\\ 3 & 4 & -1\\ 0 & -1 & 4\end{bmatrix}$$

是对称正定的三对角矩阵,SOR 迭代收敛.

$$\boldsymbol{B}_J=\begin{bmatrix}0 & -0.75 & 0\\ -0.75 & 0 & 0.25\\ 0 & 0.25 & 0\end{bmatrix},\det(\lambda\boldsymbol{I}-\boldsymbol{B}_J)=\lambda^3-\frac{5}{8}\lambda=0$$

故 $\rho(\boldsymbol{B}_J)=\sqrt{5/8}\approx 0.790$,$\rho(\boldsymbol{G})=0.625$,而 SOR 最优松弛因子

$$\omega_b=\frac{2}{1+\sqrt{1-\rho^2(\boldsymbol{B}_J)}}=\frac{2}{1+\sqrt{0.375}}\approx 1.24$$

故 $\rho(\boldsymbol{G}_{\omega_b})\approx 0.24$. 若要使误差 $\|\boldsymbol{e}^{(k)}\|_\infty\leqslant 10^{-7}\|\boldsymbol{e}^{(0)}\|_\infty$,由

$$R(\boldsymbol{G}_{\omega_b})=-\ln\rho(\boldsymbol{G}_{\omega_b})\approx-\ln 0.24\approx 1.427\ 1$$

$$k\geqslant\frac{7\ln 10}{R(\boldsymbol{G}_{\omega_b})}\approx 11.294\ 4,\text{取 } k=12 \text{ 即可.}$$

例 4.7 中取 $\omega=1.25$ 已近似 $\omega_b\approx 1.24$,故它收敛很快,实际计算时迭代 14 次可达到小数后 7 位精度.

对 $\omega=1$ 的 GS 法,由 $R(\boldsymbol{G})=-\ln\rho(\boldsymbol{G})=-\ln 0.625\approx 0.470$ 达到与 SOR 法的同样精度. 迭代次数 $k\geqslant\frac{7\ln 10}{R(\boldsymbol{G})}\approx 34.294$,故 $k\approx 34$ 与实际计算结果相符.

习 题 四

1. 设 $\boldsymbol{A}=\begin{bmatrix}0 & 0\\ 2 & 0\end{bmatrix}$,证明即使 $\|\boldsymbol{A}\|_1,\|\boldsymbol{A}\|_\infty>1$,级数 $\boldsymbol{I}+\boldsymbol{A}+\boldsymbol{A}^2+\cdots+\boldsymbol{A}^k+\cdots$ 也收敛.

2. 证明对于任意的矩阵 $\boldsymbol{A}$,序列

$$\boldsymbol{I},\boldsymbol{A},\frac{1}{2!}\boldsymbol{A}^2,\frac{1}{3!}\boldsymbol{A}^3,\frac{1}{4!}\boldsymbol{A}^4,\cdots$$

收敛于零矩阵.

3. 方程组

$$\begin{cases}5x_1+2x_2+x_3=-12\\ -x_1+4x_2+2x_3=20\\ 2x_1-3x_2+10x_3=3\end{cases}$$

(1) 考查用 Jacobi 法和 GS 法解此方程组的收敛性.

(2) 写出用 J 法及 GS 法解此方程组的迭代公式并以 $x^{(0)}=(0,0,0)^T$ 计算到 $\|x^{(k+1)}-x^{(k)}\|_\infty<10^{-4}$

为止.

4. 设方程组

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases} \quad (a_{11}, a_{22} \neq 0)$$

证明解此方程的 Jacobi 迭代法与 Gauss－Seidel 迭代法同时收敛或发散.

5. 下列两个方程组 $Ax = b$,若分别用 J 法及 GS 法求解,是否收敛?

$$(1)\ A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix} \qquad (2)\ A = \begin{bmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{bmatrix}$$

6. 设 $A = \begin{bmatrix} 10 & a & 0 \\ b & 10 & b \\ 0 & a & 5 \end{bmatrix}$, $\det A \neq 0$,用 a, b 表示解方程组 $Ax = f$ 的 J 法及 GS 法收敛的充分必要条件.

7. 用迭代公式 $x^{(k+1)} = x^{(k)} + \alpha(Ax^{(k)} - b)$求解方程组 $Ax = b$,问取什么实数 α 可使迭代收敛? 证明 $\alpha = -0.4$ 收敛最快.其中 $A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}$, $b = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$.

8. J 法的一种改进——JOR 迭代法为

$$x^{(k+1)} = B_\omega x^{(k)} + \omega D^{-1} b$$

其中 $B_\omega = \omega B + (1 - \omega) I$, B 为 J 法迭代矩阵,证明若 J 法收敛,则 JOR 方法对 $0 < \omega \leqslant 1$ 收敛.

9. 设 $A = \begin{bmatrix} 3 & 0 & -2 \\ 0 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix}$,若用 Jacobi 方法与 GS 方法解方程 $Ax = b$ 时,如果收敛,试比较哪种方法收敛快?

10. 用 SOR 方法解方程组(分别取 $\omega = 1.03, \omega = 1, \omega = 1.1$)

$$\begin{cases} 4x_1 - x_2 = 1 \\ -x_1 + 4x_2 - x_3 = 4 \\ -x_2 + 4x_3 = -3 \end{cases}$$

精确解 $x^* = \left(\frac{1}{2}, 1, -\frac{1}{2}\right)^T$,要求当 $\| x^* - x^{(k)} \|_\infty < 5 \times 10^{-6}$ 时迭代终止,并对每一个 ω 值确定迭代次数.

11. 对上题写出 SOR 迭代矩阵 G_ω,并求出最优松弛因子及渐近收敛速度,并求 J 法与 GS 法的渐近收敛速度.若要使 $\| x^* - x^{(k)} \|_\infty \leqslant 5 \times 10^{-6}$ 那么 J 法 GS 法和 SOR 法各需迭代多少次?

12. 填空题

(1) $A = \begin{bmatrix} a & 10 \\ 0 & \frac{1}{2} \end{bmatrix}$ 要使 $\lim\limits_{k \to \infty} A^k = 0$, a 应满足____.

(2) 已知方程组 $\begin{bmatrix} 1 & 2 \\ 0.32 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$,则解此方程组的 Jacobi 迭代法是否收敛____.它的渐近收敛速度 $R(B) =$ ____.

(3) 设方程组 $Ax = b$,其中 $A = \begin{bmatrix} 2 & -1 \\ 1 & 1.5 \end{bmatrix}$ 其 J 法的迭代矩阵是_____.GS 法的迭代矩阵是_____.

(4) 用 GS 法解方程组 $\begin{cases} x_1 + ax_2 = 4 \\ 2ax_1 + x_2 = -3 \end{cases}$，其中 a 为实数，方法收敛的充要条件是 a 满足_____.

(5) 给定方程组 $\begin{bmatrix} 1 & -a \\ -a & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$，$a$ 为实数. 当 a 满足_____，且 $0<\omega<2$ 时 SOR 迭代法收敛.

第五章　插值与最小二乘法

5.1　插值问题与插值多项式

实际问题中若给定函数 $y=f(x)$是区间$[a,b]$上的一个列表函数$(x_i,y_i)(i=0,1,\cdots,n)$，如果 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$，且 $f(x)$在区间$[a,b]$上是连续的，要求用一个简单的，便于计算的解析表达式 $p(x)$在区间$[a,b]$上近似 $f(x)$，使

$$p(x_i)=y_i,\quad i=0,1,\cdots,n \tag{5.1.1}$$

就称 $p(x)$为 $f(x)$的插值函数，点 $x_0,\cdots,x_n$ 称为插值节点，包含插值节点的区间$[a,b]$称为插值区间.

通常 $p(x)\in\Phi_n=\mathrm{Span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$，其中 $\varphi_i(x)(i=0,1,\cdots,n)$是一组在$[a,b]$上线性无关的函数族，$\Phi_n$ 表示 $\varphi_0,\varphi_1,\cdots,\varphi_n$ 组成的函数空间，$p(x)\in\Phi_n$ 表示为

$$p(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x) \tag{5.1.2}$$

这里 $a_i(i=0,1,\cdots,n)$是$(n+1)$个待定常数，它可根据条件(5.1.1)确定. 当 $\varphi_k(x)=x^k(k=0,1,\cdots,n)$时，$p(x)\in H_n$，$H_n$ 表示次数不超过 n 次的多项式集合，$H_n=\mathrm{Span}\{1,x,\cdots,x^n\}$，此时

$$p(x)=a_0+a_1x+\cdots+a_nx^n \tag{5.1.3}$$

称为插值多项式，如果 $\varphi_i(x)(i=0,1,\cdots,n)$为三角函数，则 $p(x)$为三角插值，同理还有分段多项式插值，有理插值等等. 由于计算机上只能使用+、−、×、÷运算，故常用的 $p(x)$就是多项式、分段多项式或有理分式，本章着重讨论多项式插值及分段多项式插值，其他插值问题不讨论.

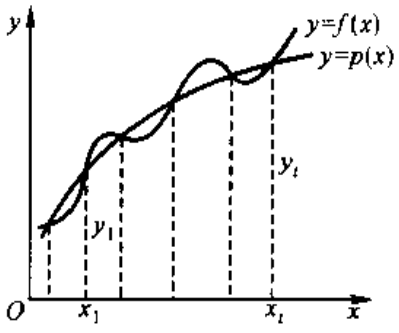

图 5-1

从几何上看，插值问题就是求过 $n+1$ 个点$(x_i,y_i)(i=0,1,\cdots,n)$的曲线 $y=p(x)$，使它近似于已给函数 $y=f(x)$，如图 5-1 所示.

插值法是一种古老的数学方法，它来自生产实践. 早在一千多年前，我国科学家在研究历法时就应用了线性插值与二次插值，但它的基本理论却是在微积分产生以后才逐步完善的，其应用也日益广泛. 特别是由于计算机的使用和航空、造船、精密机械加工等实际问题的需要，使插值法在理论上和实践上得到进一步发展. 尤其是近几十年发展起来的样条(Spline)插值，获得了极为广泛的应用，并成为计算机图形学的基础.

本章主要讨论如何求插值多项式、分段插值函数、三次样条插值、插值多项式的存在唯一性

及误差估计等.此外,还讨论列表函数的最小二乘曲线拟合问题与正交多项式.

5.2 Lagrange 插值

5.2.1 线性插值与二次插值

最简单的插值问题是已知两点$(x_0,f(x_0))$及$(x_1,f(x_1))$,通过此两点的插值多项式是一条直线,即两点式

$$L_1(x)=\frac{x-x_1}{x_0-x_1}f(x_0)+\frac{x-x_0}{x_1-x_0}f(x_1) \tag{5.2.1}$$

显然$L_1(x_0)=f(x_0),L_1(x_1)=f(x_1)$,满足插值条件,所以$L_1(x)$就是线性插值.若记$l_0(x)=\frac{x-x_1}{x_0-x_1},l_1(x)=\frac{x-x_0}{x_1-x_0}$则称$l_0(x),l_1(x)$为$x_0$与$x_1$的线性插值基函数.如图5-2所示.

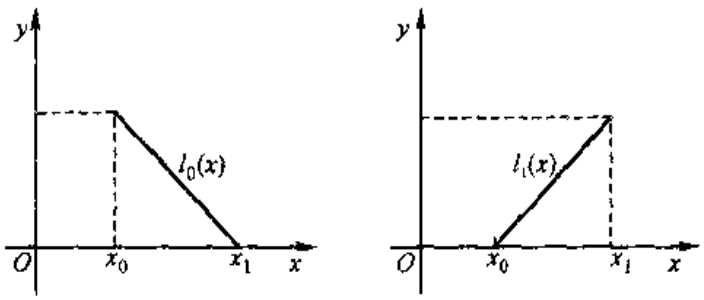

图5-2

于是 $$L_1(x)=l_0(x)f(x_0)+l_1(x)f(x_1)$$

当$n=2$,已给三点$(x_0,f(x_0)),(x_1,f(x_1)),(x_2,f(x_2))$,

$$l_0(x)=\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)},l_1(x)=\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)},l_2(x)=\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$

称为关于点x_0,x_1,x_2的二次插值基函数,它满足

$$l_i(x_j)=\begin{cases}1,j=i\\0,j\neq i,\end{cases}\quad i,j=0,1,2 \tag{5.2.2}$$

$y=l_i(x)(i=0,1,2)$的图形见图5-3.它们是满足(5.2.2)的二次插值多项式.满足条件$L_2(x_i)=f(x_i)(i=0,1,2)$的二次插值多项式$L_2(x)$可表示为

$$L_2(x)=l_0(x)f(x_0)+l_1(x)f(x_1)+l_2(x)f(x_2) \tag{5.2.3}$$

$y=L_2(x)$的图形是通过三点$(x_i,f(x_i))(i=0,1,2)$的抛物线.

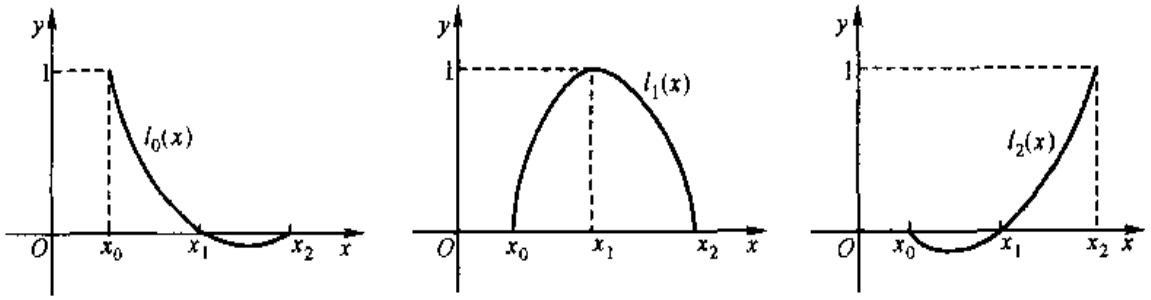

图 5-3

5.2.2 Lagrange 插值多项式

将 $n=1$ 及 $n=2$ 的插值推广到一般情形，考虑通过 $(n+1)$ 个点，$(x_i, f(x_i))(i=0,1,\cdots,n)$ 的插值多项式 $L_n(x)$，使

$$L_n(x_i)=f(x_i), i=0,1,\cdots,n \tag{5.2.4}$$

用插值基函数方法可得

$$L_n(x)=\sum_{i=0}^{n} l_i(x)f(x_i) \tag{5.2.5}$$

其中

$$l_i(x)=\frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}, i=0,1,\cdots,n \tag{5.2.6}$$

称为关于 $x_0, x_1, \cdots, x_n$ 的 n 次插值基函数，它满足条件

$$l_i(x_j)=\begin{cases}1, & j=i\\ 0, & j\neq i,\end{cases} \quad i,j=0,1,\cdots,n$$

显然(5.2.5)得到的插值多项式 $L_n(x)$ 满足条件(5.2.4)，则称 $L_n(x)$ 为 Lagrange(拉格朗日)插值多项式.

引入记号

$$\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n) \tag{5.2.7}$$

则 $$\omega'_{n+1}(x_i)=(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)$$

于是由(5.2.6)得到的 $l_i(x)$ 可改写为

$$l_i(x)=\frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)}$$

从而(5.2.4)中的 $L_n(x)$ 可改为表达式

$$L_n(x)=\sum_{i=0}^{n}\frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)}f(x_i) \tag{5.2.8}$$

并有以下关于插值多项式的存在唯一性结论.

定理 2.1 满足条件(5.2.4)的插值多项式 $L_n(x)\in H_n$ 是存在唯一的.

证明 存在性已由(5.2.5)给出的 $L_n(x)$证明,下面只需证明唯一性.用反证法,假定还有另一个 $P_n(x)\in H_n$ 使 $P_n(x_i)=f(x_i)(i=0,1,\cdots,n)$成立,于是有 $[L_n(x)-P_n(x)]\in H_n$ 且 $L_n(x_i)-P_n(x_i)=0(i=0,1,\cdots,n)$,它表明 n 次多项式 $L_n(x)-P_n(x)$有 $n+1$ 个根 $x_0,x_1,\cdots,x_n$ 这与代数基本定理 n 次多项式只有 n 个根矛盾,故 $L_n(x)\equiv P_n(x)$.证毕.

5.2.3 插值余项与误差估计

若插值区间为$[a,b]$,在$[a,b]$上有插值多项式 $L_n(x)\approx f(x)$,则称 $R_n(x)=f(x)-L_n(x)$为插值余项.

定理 2.2 设 $f(x)\in C^{n+1}[a,b]$(表示 $f(x)$在$[a,b]$上$(n+1)$阶导数连续),且节点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$,则满足条件(5.2.4)的插值多项式 $L_n(x)\in H_n$ 对 $\forall x\in[a,b]$有

$$R_n(x)=f(x)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x),a<\xi<b \tag{5.2.9}$$

这里 $\omega_{n+1}(x)$是(5.2.7)所定义的.

证明 由插值条件(5.2.4)可知 $R_n(x_i)=0(i=0,1,\cdots,n)$,故对任何 $x\in[a,b]$有

$$R_n(x)=K(x)(x-x_0)(x-x_1)\cdots(x-x_n)=K(x)\omega_{n+1}(x) \tag{5.2.10}$$

其中 $K(x)$是依赖于 x 的待定函数.将 $x\in[a,b]$看做区间$[a,b]$上任一固定点,作函数

$$\varphi(t)=f(t)-L_n(t)-K(x)(t-x_0)(t-x_1)\cdots(t-x_n),$$

显然 $\varphi(x_i)=0(i=0,1,\cdots,n)$,且 $\varphi(x)=0$,它表明 $\varphi(t)$在$[a,b]$上有 $n+2$ 个零点 $x_0,x_1,\cdots,x_n$ 及x,由 Rolle 定理可知 $\varphi'(t)$在$[a,b]$上至少有 $n+1$ 个零点.反复应用 Rolle 定理,可得 $\varphi^{(n+1)}(t)$在$[a,b]$上至少有一个零点 $\xi\in(a,b)$,使

$$\varphi^{(n+1)}(\xi)=f^{(n+1)}(\xi)-(n+1)!\ K(x)=0$$

即

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

代入(5.2.10)则得余项表达式(5.2.9).证毕.

注意定理中 $\xi\in(a,b)$依赖于 x 及点 $x_0,x_1,\cdots,x_n$,此定理只在理论上说明 ξ 存在,实际上 $f^{(n+1)}(\xi)$仍依赖于 x,即使 x 固定,ξ 也无法确定.因此,余项表达式(5.2.9)的准确值是算不出的,只能利用(5.2.9)式做截断误差估计,由

$$|f^{(n+1)}(\xi)|\leqslant\max_{a\leqslant x\leqslant b}|f^{(n+1)}(x)|\leqslant M_{n+1}$$

可得误差估计

$$|R_n(x)|\leqslant\frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)| \tag{5.2.11}$$

当 $n=1$ 时可得线性插值的误差估计

$$|R_1(x)|\leqslant\frac{M_2}{2!}|(x-x_0)(x-x_1)| \tag{5.2.12}$$

当 $n=2$ 时有二次插值的误差估计

$$|R_2(x)|\leqslant\frac{M_3}{3!}|(x-x_0)(x-x_1)(x-x_2)| \tag{5.2.13}$$

利用余项表达式(5.2.9),当 $f(x)=x^k(k\leqslant n)$ 时,由于 $f^{(n+1)}(x)=0$,于是有

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

即

$$\sum_{i=0}^{n}x_i^k l_i(x)=x^k,\quad k=0,1,\cdots,n \tag{5.2.14}$$

它表明当 $f(x)\in H_n$ 时,插值多项式 $L_n(x)$ 就是它自身,(5.2.14)也给出了插值基函数 $l_i(x)(i=0,1,\cdots,n)$ 的性质,特别当 $k=0$ 时有

$$\sum_{i=0}^{n}l_i(x)=1$$

例 5.1 已给 $\sin 0.32=0.314\ 567$,$\sin 0.34=0.333\ 487$,$\sin 0.36=0.352\ 274$,用线性插值及二次插值计算 $\sin 0.336\ 7$ 的近似值并估计误差.

解 由题意知被插函数为 $y=f(x)=\sin x$,给定插值点为 $x_0=0.32$,$y_0=0.314\ 567$,$x_1=0.34$,$y_1=0.333\ 487$,$x_2=0.36$,$y_2=0.352\ 274$. 由(5.2.1)知线性插值函数为

$$\begin{aligned}L_1(x)&=\frac{x-x_1}{x_0-x_1}y_0+\frac{x-x_0}{x_1-x_0}y_1\\&=\frac{x-0.34}{-0.02}\times 0.314\ 567+\frac{x-0.32}{0.02}\times 0.333\ 487\end{aligned}$$

当 $x=0.336\ 7$ 时

$$\sin 0.336\ 7\approx L_1(0.336\ 7)=\frac{0.336\ 7-0.34}{0.02}\times(-0.314\ 567)+$$

$$\frac{0.336\ 7-0.32}{0.02}\times 0.333\ 487$$

$$\approx 0.051\ 903\ 6+0.278\ 461\ 6\approx 0.330\ 365$$

其截断误差由(5.2.12)得

$$|R_1(x)|\leqslant\frac{M_2}{2}|(x-x_0)(x-x_1)|$$

其中 $M_2=\max\limits_{x_0\leqslant x\leqslant x_1}|f''(x)|$. 因 $f(x)=\sin x$,$f''(x)=-\sin x$,故

$$M_2=\max_{x_0\leqslant x\leqslant x_1}|\sin x|=\sin x_1\leqslant 0.333\ 5$$

于是

$$|R_1(0.336\ 7)|=|\sin 0.336\ 7-L_1(0.336\ 7)|$$

$$\leqslant\frac{1}{2}\times 0.333\,5\times 0.016\,7\times 0.003\,3\leqslant 0.92\times 10^{-5}$$

若用二次插值,在(5.2.3)中取 $n=2$,则得

$$L_2(x)=\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}y_0+\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}y_1+\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}y_2$$

$$\begin{aligned}\sin 0.336\,7\approx L_2(0.336\,7)&=\frac{0.768\,9\times 10^{-4}}{0.000\,8}\times 0.314\,567+\\&\frac{3.891\,1\times 10^{-4}}{0.000\,4}\times 0.333\,487+\\&\frac{-0.551\,1\times 10^{-4}}{0.000\,8}\times 0.352\,274\\&\approx 0.330\,374\end{aligned}$$

这个结果与6位有效数字的正弦函数表完全一样.其截断误差由(5.2.13)得

$$|R_2(x)|\leqslant\frac{M_3}{6}|(x-x_0)(x-x_1)(x-x_2)|$$

其中

$$M_3=\max_{x_0\leqslant x\leqslant x_2}|f'''(x)|=\max_{x_0\leqslant x\leqslant x_2}|\cos x|=\cos 0.32<0.950$$

于是

$$\begin{aligned}|R_2(0.336\,7)|&=|\sin 0.336\,7-L_2(0.336\,7)|\\&\leqslant\frac{1}{6}|0.950\times 0.016\,7\times 0.003\,3\times 0.023\,3|<0.204\times 10^{-6}\end{aligned}$$

例 5.2 设 $f\in C^2[a,b]$,试证

$$\max_{a\leqslant x\leqslant b}\left|f(x)-\left[f(a)+\frac{f(b)-f(a)}{b-a}(x-a)\right]\right|\leqslant\frac{1}{8}(b-a)^2\max_{a\leqslant x\leqslant b}|f''(x)|$$

解 由于 $f(x)$的线性插值 $L_1(x)=f(a)+\dfrac{f(b)-f(a)}{b-a}\cdot(x-a)$,

于是

$$\begin{aligned}&\max_{a\leqslant x\leqslant b}\left|f(x)-\left[f(a)+\frac{f(b)-f(a)}{b-a}(x-a)\right]\right|\\&=\max_{a\leqslant x\leqslant b}|f(x)-L_1(x)|\\&=\max_{a\leqslant x\leqslant b}\left|\frac{f''(\xi)}{2!}(x-a)(x-b)\right|\leqslant\frac{1}{2}\max_{a\leqslant x\leqslant b}|(x-a)(x-b)|\max_{a\leqslant x\leqslant b}|f''(x)|\\&=\frac{1}{8}(b-a)^2\max_{a\leqslant x\leqslant b}|f''(x)|\end{aligned}$$

例 5.3 证明 $\sum_{i=0}^{5}(x_i-x)^2l_i(x)=0$,其中 $l_i(x)$ 是关于点 $x_0,x_1,\cdots,x_5$ 的插值基函数.

解 $$\sum_{i=0}^{5}(x_i-x)^2l_i(x)=\sum_{i=0}^{5}(x_i^2-2x_ix+x^2)l_i(x)$$

$$= \sum_{i=0}^{5} x_i^2 l_i(x) - 2x \sum_{i=0}^{5} x_i l_i(x) + x^2 \sum_{i=0}^{5} l_i(x) = x^2 - 2x^2 + x^2 = 0$$

5.3 均差与 Newton 插值公式

5.3.1 均差及其性质

利用插值基函数求出 Lagrange 插值多项式(5.2.8),在理论上是很重要的,但用 $L_n(x)$计算 $f(x)$近似值却不大方便,特别当精度不够,需增加插值节点时,计算要全部重新进行.为此我们可以给出另一种便于计算的插值多项式 $N_n(x)$,它表达为

$$N_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)\cdots(x - x_{n-1}) \tag{5.3.1}$$

其中 $a_i(i = 0,1,\cdots,n)$为待定常数.显然,它可根据插值条件

$$N_n(x_i) = f(x_i), \quad i = 0,1,\cdots,n \tag{5.3.2}$$

直接得到,例如当 $x = x_0$ 时,得 $a_0 = f(x_0)$;当 $x = x_1$ 时,由(5.3.1)得 $N_n(x_1) = f(x_0) + a_1(x_1 - x_0) = f(x_1)$,得 $a_1 = \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}$.实际上 $y = N_1(x)$就是直线方程的点斜式.$N_1(x) = L_1(x)$,$N_n(x) \in H_n = \mathrm{Span}\{1, x - x_0, (x - x_0)(x - x_1), \cdots, (x - x_0)\cdots(x - x_{n-1})\}$.为了给出 $N_n(x)$的系数 $a_0, a_1, \cdots, a_n$ 的表达式,先引进以下定义.

定义 3.1 记 $f[x_m] = f(x_m)$为 f 的零阶均差,零阶均差的差商记为

$$f[x_0, x_m] = \frac{f[x_m] - f[x_0]}{x_m - x_0}$$

称为函数关于点 x_0, x_m 的一阶均差.一般地,记$(k-1)$阶均差的差商为

$$f[x_0, \cdots, x_{k-1}, x_m] = \frac{f[x_0, \cdots, x_{k-2}, x_m] - f[x_0, x_1, \cdots, x_{k-1}]}{x_m - x_{k-1}} \tag{5.3.3}$$

称为 f 关于点 $x_0, x_1, \cdots, x_{k-1}, x_m$ 的 k 阶均差.

均差有以下重要性质:

(1) 均差对称性.k 阶均差可表示为函数值 $f(x_0), f(x_1), \cdots, f(x_k)$的线性组合,即

$$f[x_0, x_1, \cdots, x_k] = \sum_{i=0}^{k} \frac{f(x_i)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_k)} \tag{5.3.4}$$

这个性质可用归纳法证明,见[3].(5.3.4)表明均差 $f[x_0, x_1 \cdots, x_k]$与节点排列次序无关,称为均差对称性.

(2) 如果 $f[x, x_0, \cdots, x_k]$是 x 的 m 次多项式,则 $f[x, x_0, \cdots, x_k, x_{k+1}]$是 x 的$(m-1)$次多项式.

证明 由均差定义可知

$$f[x,x_0,\cdots,x_k,x_{k+1}]=\frac{f[x,x_0,\cdots,x_k]-f[x_0,x_1,\cdots,x_{k+1}]}{x-x_{k+1}}$$

右端分子为 x 的 m 次多项式，且当 $x=x_{k+1}$ 时，此式为零，所以分子含有 $(x-x_{k+1})$ 的因子，与分母相约后得到 $(m-1)$ 次多项式.

(3) 若 $f\in C^n[a,b]$，并且 $x_i\in[a,b](i=0,1,\cdots,n)$ 互异，则有

$$f[x_0,x_1,\cdots,x_n]=\frac{f^{(n)}(\xi)}{n!},\text{其中 }\xi\in(a,b) \tag{5.3.5}$$

这公式可直接由 Rolle 定理证明(略).

其他均差性质可作为习题自己证明.均差计算可列均差表，见表 5-1.

表 5-1

x_k	$f(x_k)$	一阶均差	二阶均差	三阶均差	四阶均差
x_0	$\underline{f(x_0)}$				
x_1	$f(x_1)$	$\underline{f[x_0,x_1]}$			
x_2	$f(x_2)$	$f[x_1,x_2]$	$\underline{f[x_0,x_1,x_2]}$		
x_3	$f(x_3)$	$f[x_2,x_3]$	$f[x_1,x_2,x_3]$	$\underline{f[x_0,x_1,x_2,x_3]}$	
x_4	$f(x_4)$	$f[x_3,x_4]$	$f[x_2,x_3,x_4]$	$f[x_1,x_2,x_3,x_4]$	$\underline{f[x_0,x_1,x_2,x_3,x_4]}$
⋮	⋮	⋮	⋮	⋮	⋮

5.3.2 Newton 插值

根据均差定义，把 x 看成 $[a,b]$ 上一点，可得

$$f(x)=f(x_0)+f[x,x_0](x-x_0)$$

$$f[x,x_0]=f[x_0,x_1]+f[x,x_0,x_1](x-x_1)$$

……

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

只要把后一式代入前一式，就得到

$$\begin{aligned}f(x)&=f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2,](x-x_0)(x-x_1)+\cdots+\\&\quad f[x_0,x_1,\cdots,x_n](x-x_0)\cdots(x-x_{n-1})+f[x,x_0,\cdots,x_n]\omega_{n+1}(x)\\&=N_n(x)+R_n(x)\end{aligned}$$

其中

$$N_n(x)=f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+\cdots+$$
$$f[x_0,\cdots,x_n](x-x_0)\cdots(x-x_{n-1}) \tag{5.3.6}$$

$$R_n(x)=f(x)-N_n(x)=f[x,x_0,\cdots,x_n]\omega_{n+1}(x) \tag{5.3.7}$$

$\omega_{n+1}(x)$ 是由(5.2.7)定义的.由(5.3.6)确定的多项式 $N_n(x)$ 显然满足插值条件，且次数不

超过 n，它就是形如(5.3.1)的多项式，其系数为

$$a_k = f[x_0, \cdots, x_k], k = 0, 1, \cdots, n$$

我们称 $N_n(x)$为 Newton 均差插值多项式，系数 a_k 就是均差表 5－1 中加横线的各阶均差，它比 Lagrange 插值的计算量少，且便于程序设计.

(5.3.7) 为插值余项，由插值多项式的唯一性可知，它与(5.2.9)是等价的. 事实上，利用均差与导数关系式(5.3.5)，可由(5.3.7)推出(5.2.9). 但(5.3.7)更有一般性，它对 f 是由离散点给出的情形或 f 导数不存在时均适用.

例 5.4 给出 $f(x)$的函数表(见表 5－2)，求四次牛顿插值多项式，并由此计算 $f(0.596)$的近似值.

首先根据给定函数表造出均差表.

表 5－2

x_k	$f(x_k)$	一阶均差	二阶均差	三阶均差	四阶均差	五阶均差
0.40	0.410 75					
0.55	0.578 15	1.116 00				
0.65	0.696 75	1.186 00	0.280 00			
0.80	0.888 11	1.275 73	0.358 93	0.197 33		
0.90	1.026 52	1.384 10	0.433 48	0.213 00	0.031 34	
1.05	1.253 82	1.515 33	0.524 93	0.228 63	0.031 26	−0.000 12

从均差表看到四阶均差已近似于常数. 故取四次插值多项式 $N_4(x)$做近似即可.

$$N_4(x) = 0.410\,75 + 1.116(x - 0.4) + 0.28(x - 0.4)(x - 0.55) +$$
$$0.197\,33(x - 0.4)(x - 0.55)(x - 0.65) +$$
$$0.031\,34(x - 0.4)(x - 0.55)(x - 0.65)(x - 0.8)$$

于是

$$f(0.596) \approx N_4(0.596) = 0.631\,92$$

截断误差

$$|R_4(x)| \approx |f[x_0, \cdots, x_5]\omega_5(0.596)| \leqslant 3.63 \times 10^{-9}$$

这说明截断误差很小，可忽略不计.

5.4 差分与 Newton 前后插值公式

5.4.1 差分及其性质

当插值节点为等距节点 $x_k = x_0 + kh(k = 0, 1, \cdots, n)$时，称 h 为步长，此时均差及 Newton 均差插值多项式(5.3.6)均可简化.

定义 4.1 设 $f_k = f(x_k), x_k = x_0 + kh, k = 0, 1, \cdots, n$，记

$$\Delta f_k = f_{k+1} - f_k \tag{5.4.1}$$

$$\nabla f_k = f_k - f_{k-1} \tag{5.4.2}$$

分别称为 $f(x)$在 x_k 处以 h 为步长的一阶向前差分及一阶向后差分. 符号 Δ 及∇分别称为向前差分算子及向后差分算子.

利用一阶差分可定义二阶差分为

$$\Delta^2 f_k = \Delta f_{k+1} - \Delta f_k = f_{k+2} - 2f_{k+1} + f_k \quad \text{(二阶向前差分)}$$

$$\nabla^2 f_k = \nabla f_k - \nabla f_{k-1} - f_k - 2f_{k-1} + f_{k-2} \quad \text{(二阶向后差分)}$$

一般地，可定义 m 阶向前差分及 m 阶向后差分为

$$\Delta^m f_k = \Delta^{m-1} f_{k+1} - \Delta^{m-1} f_k, \nabla^m f_k = \nabla^{m-1} f_k - \nabla^{m-1} f_{k-1}$$

此外还可定义不变算子 I 及位移算子 E 为：

$$If_k = f_k, \quad Ef_k = f_{k+1} \tag{5.4.3}$$

于是，由 $\Delta f_k = f_{k+1} - f_k = Ef_k - If_k = (E - I)f_k$，可得

$$\Delta = E - I$$

同理可得 $$\nabla = I - E^{-1}.$$

由差分定义并应用算子符号运算可得下列基本性质.

性质 1 各阶差分均可用函数值表示. 例如

$$\Delta^n f_k = (E - I)^n f_k = \sum_{j=0}^{n} (-1)^j \binom{n}{j} E^{n-j} f_k$$

$$= \sum_{j=0}^{n} (-1)^j \binom{n}{j} f_{n+k-j} \tag{5.4.4}$$

$$\nabla^n f_k = (I - E^{-1})^n f_k = \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} E^{j-n} f_k$$

$$= \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} f_{k+j-n} \tag{5.4.5}$$

其中 $\binom{n}{j} = \dfrac{n(n-1)\cdots(n-j+1)}{j!}$ 为二项式展开系数.

性质 2 可用各阶差分表示函数值. 例如，可用向前差分表示 f_{n+k}，因为

$$f_{n+k} = E^n f_k = (I + \Delta)^n f_k = \left[\sum_{j=0}^{n} \binom{n}{j} \Delta^j \right] f_k$$

于是

$$f_{n+k} = \sum_{j=0}^{n} \left[\binom{n}{j} \Delta^j f_k \right] \tag{5.4.6}$$

性质 3 均差与差分有的关系. 由定义可知，向前差分

$$f[x_k, x_{k+1}] = \frac{f_{k+1} - f_k}{x_{k+1} - x_k} = \frac{\Delta f_k}{h}$$

$$f[x_k,x_{k+1},x_{k+2}]=\frac{f[x_{k+1},x_{k+2}]-f[x_k,x_{k+1}]}{x_{k+2}-x_k}=\frac{1}{2h^2}\Delta^2 f_k$$

一般地有

$$f[x_k,\cdots,x_{k+m}]=\frac{1}{m!}\frac{1}{h^m}\Delta^m f_k,\ m=1,2,\cdots,n \tag{5.4.7}$$

同理,对向后差分有

$$f[x_k,x_{k-1},\cdots,x_{k-m}]=\frac{1}{m!}\frac{1}{h^m}\nabla^m f_k \tag{5.4.8}$$

利用(5.4.7)及(5.3.5)又可得到

$$\Delta^n f_k=h^n f^{(n)}(\xi) \tag{5.4.9}$$

其中 $\xi\in(x_k,x_{k+n})$,这就是差分与导数的关系.差分的其他性质从略.

计算差分可列差分表,表 5-3 是向前差分表.

表 5-3

f_x	Δ	Δ^2	Δ^3	Δ^4
f_0	Δf_0	$\Delta^2 f_0$	$\Delta^3 f_0$	$\Delta^4 f_0$
f_1	Δf_1	$\Delta^2 f_1$	$\Delta^3 f_1$	⋮
f_2	Δf_2	$\Delta^2 f_2$	⋮	
f_3	Δf_3	⋮		
f_4	⋮			
⋮				

5.4.2 等距节点插值公式

将牛顿均差插值多项式(5.3.6)中各阶均差用相应差分代替,就可得到各种形式的等距节点插值公式.这里只推导常用的前插与后插公式.

如果有节点 $x_k=x_0+kh(k=0,1,\cdots,n)$,要计算 x_0 附近点 x 的函数 $f(x)$的值,可令 $x=x_0+th$,$0\leqslant t\leqslant 1$,于是

$$\omega_{k+1}(x)=\prod_{j=0}^{k}(x-x_j)=t(t-1)\cdots(t-k)h^{k+1}$$

将此式及(5.4.7)代入(5.3.6),则得

$$N_n(x_0+th)=f_0+t\Delta f_0+\frac{t(t-1)}{2!}\Delta^2 f_0+\cdots+\frac{t(t-1)\cdots(t-n+1)}{n!}\Delta^n f_0 \tag{5.4.10}$$

称为 Newton 前插公式,其余项由(5.2.9)得

$$R_n(x)=\frac{t(t-1)\cdots(t-n)}{(n+1)!}h^{n+1}f^{(n+1)}(\xi),\ \xi\in(x_0,x_n) \tag{5.4.11}$$

如果要用函数表示 x_n 附近的函数值 $f(x)$,此时应用牛顿插值公式(5.3.6),插值点应按 $x_n,x_{n-1},\cdots,x_0$ 的次序排列,有

$$N_n(x)=f(x_n)+f[x_n,x_{n-1}](x-x_n)+f[x_n,x_{n-1},x_{n-2}](x-x_n)(x-x_{n-1})+\cdots+$$
$$f[x_n,x_{n-1},\cdots,x_0](x-x_n)\cdots(x-x_1)$$

作变换 $x=x_n+th$ $(-1\leqslant t\leqslant 0)$,并利用公式(5.4.8),代入上式得

$$N_n(x_n+th)=f_n+t\nabla f_n+\frac{t(t+1)}{2!}\nabla^2 f_n+\cdots+\frac{t(t+1)\cdots(t+n-1)}{n!}\nabla^n f_n \tag{5.4.12}$$

称为 Newton 后插公式,其余项

$$R_n(x)=f(x)-N_n(x_n+th)$$
$$=\frac{t(t+1)\cdots(t+n)h^{n+1}f^{(n+1)}(\xi)}{(n+1)!},\xi\in(x_0,x_n) \tag{5.4.13}$$

例 5.5 设 $x_0=1.0,h=0.05$,给出 $f(x)=\sqrt{x}$ 在 $x_i=x_0+ih$ $(i=0,1,\cdots,6)$的值.试用三次等距节点插值公式求 $f(1.01)$及 $f(1.28)$的近似值.

解 本题只要构造出 f 的差分表,再按 Newton 前插公式及后插公式计算即可.$f(x)=\sqrt{x}$ 的差分表如下所示.

x_i	$f(x_i)$	$\Delta(\nabla)$	$\Delta^2(\nabla^2)$	$\Delta^3(\nabla^3)$
1.00	1.000 00			
		0.024 70		
1.05	1.024 70		−0.000 59	
		0.024 11		0.000 05
1.10	1.048 81		−0.000 54	
		0.023 57		0.000 04
1.15	1.072 38		−0.000 50	
		0.023 07		0.000 02
1.20	1.095 44		−0.000 48	
		0.022 59		0.000 03
1.25	1.118 03		−0.000 45	
		0.022 14		
1.30	1.140 17			

计算 $f(1.01)$可用 Newton 前插公式(5.4.10),此时用到差分表中的上半部分划波纹线的各阶差分值.

$$f(1.01)\approx N_3(1.01)=N_3\left(1.00+\frac{1}{5}\times h\right)$$
$$=1.000\,00+0.2\times 0.024\,70+\frac{1}{2}\times 0.2\times(0.2-1)(-0.000\,59)+$$
$$\frac{1}{6}\times 0.2\times(0.2-1)(0.2-2)(0.000\,05)$$
$$=1.004\,99$$

计算 $f(1.28)$要用 Newton 后插公式(5.4.12),它用到差分表下部分的差分∇(即下划直线的).

$$\begin{aligned} f(1.28) &\approx N_3(1.28) \\ &= N_3(1.30-0.4\times h) \\ &= 1.140\ 17+(-0.4)\times 0.022\ 14+\frac{1}{2}(-0.4)\times 0.6\times \\ &\quad (-0.000\ 45)+\frac{1}{6}(-0.4)\times 0.6\times 1.6\times 0.000\ 03 \\ &= 1.131\ 37 \end{aligned}$$

$f(1.01)$与 $f(1.28)$的 7 位有效数字分别为$\sqrt{1.01}=1.004\ 988$,$\sqrt{1.28}=1.131\ 371$,可见计算结果已相当精确.

5.5 Hermite 插值

不少问题不但要求在插值节点上函数值相等,而且还要求节点上导数值相等,有的甚至要求高阶导数值也相等,满足这种要求的插值多项式称为 Hermite 插值多项式.若给出的插值条件有$(m+1)$个则可造出 m 次插值多项式.建立 Hermite 插值多项式的方法仍可采用插值基函数和均差插值的方法,较常见的一类带导数插值的问题,是在给出节点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$ 上已知$f(x_i)=f_i$,$f'(x_i)=m_i(i=0,1,\cdots,n)$要求$H_{2n+1}(x)\in H_{2n+1}$,使

$$H_{2n+1}(x_i)=f_i,H'_{2n+1}(x_i)=m_i,i=0,1,\cdots,n \tag{5.5.1}$$

若用基函数方法表示可得

$$H_{2n+1}(x)=\sum_{i=0}^{n}[\alpha_i(x)f_i+\beta_i(x)m_i] \tag{5.5.2}$$

其中 $\alpha_i(x)$及 $\beta_i(x)(i=0,1,\cdots,n)$是关于点 $x_0,x_1,\cdots,x_n$ 的$(2n+1)$次 Hermite 插值基函数,它们为$(2n+1)$次多项式且满足条件

$$\begin{cases} \alpha_i(x_k)=\delta_{ik}=\begin{cases}0,i\neq k\\1,i=k\end{cases},\alpha_i'(x_k)=0 \\ \beta_i(x_k)=0,\beta_i'(x_k)=\delta_{ik},i,k=0,1,\cdots,n \end{cases} \tag{5.5.3}$$

若 $f(x)$在(a,b)上存在$(2n+2)$阶导数 $f^{(2n+2)}(x)$,则其插值余项为

$$R_{2n+1}(x)=f(x)-H_{2n+1}(x)=\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\omega_{n+1}^2(x) \tag{5.5.4}$$

其中 $\xi\in(a,b)$与 x 有关,$\omega_{n+1}(x)$由(5.2.7)表示.

下面只对 $n=1$ 的情形给出 $H_3(x)$的表达式.若插值节点为 x_k 及 x_{k+1},要求 $H_3(x)\in H_3$,使

$$\begin{cases} H_3(x_k)=f_k,H_3(x_{k+1})=f_{k+1} \\ H'_3(x_k)=m_k,H'_3(x_{k+1})=m_{k+1} \end{cases} \tag{5.5.5}$$

相应插值基函数为 $\alpha_k(x),\alpha_{k+1}(x),\beta_k(x),\beta_{k+1}(x)$,它们满足条件

$$\alpha_k(x_k)=1,\alpha_k(x_{k+1})=0,\alpha_k'(x_k)=\alpha_k'(x_{k+1})=0,$$
$$\alpha_{k+1}(x_k)=0,\alpha_{k+1}(x_{k+1})=1,\alpha_{k+1}'(x_k)=\alpha_{k+1}'(x_{k+1})=0,$$
$$\beta_k(x_k)=\beta_k(x_{k+1})=0,\beta_k'(x_k)=1,\beta_k(x_{k+1})=0,$$
$$\beta_{k+1}(x_k)=\beta_{k+1}(x_{k+1})=0,\beta_{k+1}'(x_k)=0,\beta_{k+1}(x_{k+1})=1$$

根据给出条件可令

$$\alpha_k(x)=(ax+b)\left(\frac{x-x_{k+1}}{x_k-x_{k+1}}\right)^2$$

显然

$$\alpha_k(x_{k+1})=\alpha'_k(x_{k+1})=0$$

再由 $\alpha_k(x_k)=ax_k+b=1$ 及 $\alpha_k'(x_k)=a+\frac{2}{x_k-x_{k+1}}(ax_k+b)=0$,

解得

$$a=-\frac{2}{x_k-x_{k+1}},b=1+\frac{2x_k}{x_k-x_{k+1}}$$

于是可得

$$\alpha_k(x)=\left(1+2\frac{x-x_k}{x_{k+1}-x_k}\right)\left(\frac{x-x_{k+1}}{x_k-x_{k+1}}\right)^2 \tag{5.5.6}$$

同理,可求得

$$\begin{cases}\alpha_{k+1}(x)-\left(1+2\dfrac{x-x_{k+1}}{x_k-x_{k+1}}\right)\left(\dfrac{x-x_k}{x_{k+1}-x_k}\right)^2\\ \beta_k(x)=(x-x_k)\left(\dfrac{x-x_{k+1}}{x_k-x_{k+1}}\right)^2\\ \beta_{k+1}(x)=(x-x_{k+1})\left(\dfrac{x-x_k}{x_{k+1}-x_k}\right)^2\end{cases} \tag{5.5.7}$$

于是满足条件(5.5.5)的 Hermite 插值多项式为

$$H_3(x)=\alpha_k(x)f_k+\alpha_{k+1}(x)f_{k+1}+\beta_k(x)m_k+\beta_{k+1}(x)m_{k+1} \tag{5.5.8}$$

它的插值余项为

$$R_3(x)=f(x)-H_3(x)=\frac{1}{4!}f^{(4)}(\xi)(x-x_k)^2(x-x_{k+1})^2,\xi\text{ 在 }x_k\text{ 与 }x_{k+1}\text{ 之间} \tag{5.5.9}$$

下面再给出一个典型的例子.

例 5.6 求 $p(x)\in H_3$,使 $p(x_i)=f(x_i)(i=0,1,2)$ 及 $p'(x_1)=f'(x_1)$ 的插值多项式及其余项表达式.

解 这里给出了四个条件故可造三次插值多项式 $p(x)\in H_3$,由 $p(x_i)=f(x_i)$ $(i=0,1,2)$,可用 Newton 均差插值,令

$$p(x)=f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+$$
$$a(x-x_0)(x-x_1)(x-x_2) \tag{5.5.10}$$

显然它满足条件 $p(x_i)=f(x_i)(i=0,1,2)$,a 为待定参数.

由 $p'(x_1)=f'(x_1)$可得

$$p'(x_1)=f[x_0,x_1]+f[x_0,x_1,x_2](x_1-x_0)+a(x_1-x_0)(x_1-x_2)=f'(x_1)$$

解得

$$a=\frac{1}{x_1-x_2}\left\{\frac{f'(x_1)-f[x_0,x_1]}{x_1-x_0}-f[x_0,x_1,x_2]\right\} \tag{5.5.11}$$

于是得到的插值多项式为(4.8)的 $p(x)$,其中 a 由(5.5.11)给出,它的余项表达式是

$$R_3(x)=f(x)-p(x)=\frac{1}{4!}f^{(4)}(\xi)(x-x_0)(x-x_1)^2(x-x_2) \tag{5.5.12}$$

其中 ξ 在 x_0 与 x_2 之间,而 $a\leqslant x_0<x_1<x_2\leqslant b$.

5.6 分段低次插值

5.6.1 多项式插值的收敛性问题

若在$[a,b]$上任给一组插值节点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$,假定 $f(x)\in C[a,b]$,按条件(5.2.4)造出 Lagrange 插值多项式 $L_n(x)\approx f(x)$,若 $n\to\infty(x_{i+1}-x_i\to 0,i=0,1,\cdots,n)$极限

$$\lim_{n\to\infty}L_n(x)=f(x),x\in(a,b) \tag{5.6.1}$$

就称插值多项式 $L_n(x)$收敛于 $f(x)$.但实际上甚至对各阶导数均存在的 $f(x)$也不能保证(5.6.1)成立,也就是插值多项式序列 $L_0(x),L_1(x),\cdots,L_n(x),\cdots$收敛性不成立,下面给出一个不收敛的例子.

例 5.7 设 $f(x)=\frac{1}{1+x^2}$在$[-5,5]$上取$(n+1)$个等距节点 $x_k=x_0+kh,h=\frac{10}{n}$,可造插值多项式

$$L_n(x)=\sum_{i=0}^{n}\frac{1}{1+x_i^2}\frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)}$$

记 $x_{n-1/2}=\frac{1}{2}(x_{n-1}+x_n)=5-\frac{5}{n}$,表 5-4 列出 $n=2,4,\cdots,20$ 的 $L_n(x_{n-1/2})$的计算结果及在 $x_{n-1/2}$处的误差 $R(x_{n-1/2})=f(x_{n-1/2})-L_n(x_{n-1/2})$.

表 5-4

n	$f(x_{n-1/2})$	$L_n(x_{n-1/2})$	$R(x_{n-1/2})$
2	0.137 931	0.759 615	-0.621 648
4	0.066 390	-0.356 826	0.423 216
6	0.054 463	0.607 879	-0.553 416
8	0.049 651	-0.831 017	0.880 668

续表

n	$f(x_{n-1/2})$	$L_n(x_{n-1/2})$	$R(x_{n-1/2})$
10	0.047 059	1.578 721	-1.531 662
12	0.045 440	-2.755 000	2.800 440
14	0.044 334	5.332 743	-5.288 409
16	0.043 530	-10.173 867	10.217 397
18	0.042 920	20.123 671	20.080 751
20	0.042 440	-39.952 449	39.994 889

可以看出随 n 的增加 $|R(x_{n-1/2})|$ 几乎成倍增加，这说明 $\{L_n(x)\}$ 在 $[-5,5]$ 上并不收敛. 当 $n=10$ 时，从 $y=L_{10}(x)$ 的图形(见图 5-4)也可看出它不收敛.

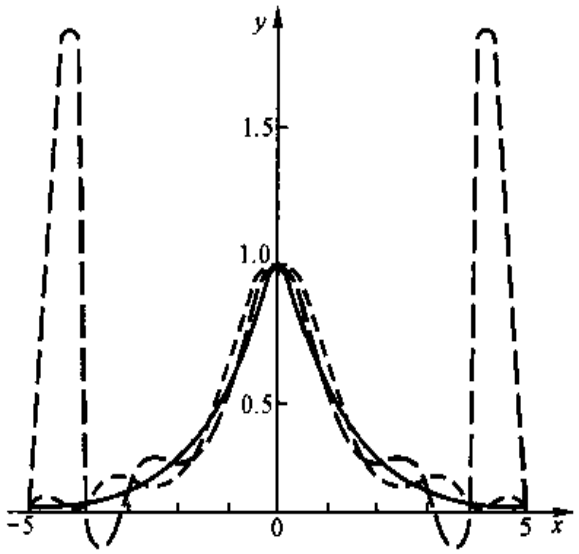

图 5-4

这个例子是 Runge 于 1901 年首先给出的，故把插值多项式不收敛的现象称为 Runge 现象. Runge 还证明了此例中 $|x|\leqslant C\approx 3.63$ 时，$\lim\limits_{n\to\infty} L_n(x)=f(x)$，但在 $|x|>C$ 时 $\{L_n(x)\}$ 发散.

由于高次插值收敛性没有保证，实际的计算稳定性也没保证. 因此当插值节点 n 较大时通常不采用高次多项式插值，而改用低次分段插值.

5.6.2 分段线性插值

设已知节点 $a=x_0<x_1<\cdots<x_n=b$ 上的函数值为 $f_0,f_1,\cdots,f_n$，$h_i=x_{i+1}-x_i$，$h=\max\limits_{0\leqslant i\leqslant n-1} h_i$，若一折线函数 $I_h(x)$ 满足条件

(1) $I_h(x)\in C[a,b]$；

(2) $I_h(x_i)=f_i$, $i=0,1,\cdots,n$;

(3) $I_h(x)$在每个小区间$[x_i,x_{i+1}](i=0,1,\cdots,n-1)$上为线性函数.

则称 $I_h(x)$为分段线性插值函数,I_h 在每个小区间上表示为

$$I_h(x)=\frac{x-x_{i+1}}{x_i-x_{i+1}}f_i+\frac{x-x_i}{x_{i+1}-x_i}f_{i+1},\quad x_i\leqslant x\leqslant x_{i+1} \tag{5.6.2}$$

在区间$[a,b]$上可表示为

$$I_h(x)=\sum_{i=0}^{n}f_i l_i(x) \tag{5.6.3}$$

其中

$$l_i(x)=\begin{cases}\dfrac{x-x_{i-1}}{x_i-x_{i-1}}, & x_{i-1}\leqslant x\leqslant x_i,\ i=0\ \text{略去}\\[2ex] \dfrac{x-x_{i+1}}{x_i-x_{i+1}}, & x_i\leqslant x\leqslant x_{i+1},\ i=n\ \text{略去}\\[2ex] 0, & x\in[a,b],x\notin[x_{i-1},x_{i+1}]\end{cases}$$

定理 6.1 若 $f(x)\in C[a,b]$,则当 $h\to 0$ 时 $I_h(x)$一致收敛于 $f(x)$.若 $f(x)\in C^2[a,b]$,则余项 $R(x)=f(x)-I_h(x)$有估计式

$$|R(x)|\leqslant\frac{Mh^2}{8},\quad M=\max_{a\leqslant x\leqslant b}|f''(x)| \tag{5.6.4}$$

5.6.3 分段三次 Hermite 插值

设函数 $f(x)$在节点 $a=x_0<x_1<\cdots<x_n=b$ 上的函数值为 $f_0,f_1,\cdots,f_n$,一阶导数值为 $f'_0,f'_1,\cdots,f'_n$,若 I_h 满足条件

(1) $I_h\in C^1[a,b]$;

(2) $I_h(x_i)=f_i,I'_h(x_i)=f'_i$, $(i=0,1,\cdots,n)$;

(3) 在每个子区间$[x_i,x_{i+1}](0\leqslant i\leqslant n-1)$上 I_h 是次数不大于 3 的多项式.

则称 I_h 是$f(x)$的分段三次 Hermite 插值函数.在每个子区间$[x_i,x_{i+1}]$上的表达式为

$$\begin{aligned}I_h(x)=&\left(1+2\frac{x-x_i}{x_{i+1}-x_i}\right)\left(\frac{x-x_{i+1}}{x_i-x_{i+1}}\right)^2 f_i+\left(1+2\frac{x-x_{i+1}}{x_i-x_{i+1}}\right)\left(\frac{x-x_i}{x_{i+1}-x_i}\right)^2 f_{i+1}+\\&(x-x_i)\left(\frac{x-x_{i+1}}{x_i-x_{i+1}}\right)^2 f'_i+(x-x_{i+1})\left(\frac{x-x_i}{x_{i+1}-x_i}\right)^2 f'_{i+1}\end{aligned} \tag{5.6.5}$$

在$[a,b]$上用插值基函数表示为

$$I_h(x)=\sum_{i=0}^{n}[\alpha_i(x)f_i+\beta_i(x)f'_i] \tag{5.6.6}$$

其中

$$\alpha_i(x)=\begin{cases}\left(1+2\dfrac{x-x_i}{x_{i-1}-x_i}\right)\left(\dfrac{x-x_{i-1}}{x_i-x_{i-1}}\right)^2, & x\in[x_{i-1},x_i],i=0\text{ 略去}\\ \left(1+2\dfrac{x-x_i}{x_{i+1}-x_i}\right)\left(\dfrac{x-x_{i+1}}{x_i-x_{i+1}}\right)^2, & x\in[x_i,x_{i+1}],i=n\text{ 略去}\\ 0, & x\notin[x_{i-1},x_{i+1}]\end{cases}$$

$$\beta_i(x)=\begin{cases}(x-x_i)\left(\dfrac{x-x_{i-1}}{x_i-x_{i-1}}\right)^2, & x\in[x_{i-1},x_i],i=0\text{ 略去}\\ (x-x_i)\left(\dfrac{x-x_{i+1}}{x_i-x_{i+1}}\right)^2, & x\in[x_i,x_{i+1}],i=n\text{ 略去}\\ 0, & x\notin[x_{i-1},x_{i+1}]\end{cases}$$

可以证明，若 $f(x)\in C[a,b]$，$h=\max\limits_{0\leqslant i\leqslant n-1}(x_{i+1}-x_i)$，则当 $h\to 0$ 时 $I_h(x)$一致收敛于 $f(x)$.

5.7 三次样条插值

5.7.1 三次样条函数

分段低次插值的优点是具有收敛性与稳定性，缺点是光滑性较差，不能满足实际需要. 例如高速飞机的机翼形线、船体放样形值线、精密机械加工等都要求有二阶光滑度，即二阶导数连续，通常三次样条(Spline)函数即可满足要求.

定义 7.1 设$[a,b]$上给出一组节点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$，若函数 $s(x)$满足条件

(1) $s(x)\in C^2[a,b]$；

(2) $s(x)$在每个小区间$[x_i,x_{i+1}](i=0,1,\cdots,n-1)$上是三次多项式.

则称 $s(x)$是节点 $x_0,x_1,\cdots,x_n$ 上的三次样条函数.

若 $s(x)$在节点上还满足插值条件

$$(3)\ s(x_i)=f_i,\quad i=0,1,\cdots,n \tag{5.7.1}$$

则称 $s(x)$为$[a,b]$上的三次样条插值函数.

例 5.8 设 $s(x)=\begin{cases}x^3+x^2, & 0\leqslant x\leqslant 1\\ 2x^3+ax^2+bx+c, & 1\leqslant x\leqslant 2\end{cases}$

是以 0,1,2 为节点的三次样条函数，则 a,b,c 应取何值?

解 因 $s(x)\in C^2[0,2]$，故在 $x_1=1$ 处由 $s(1),s'(1)$及 $s''(1)$连续，可得

$$\begin{cases}a+b+c+2=2\\ 2a+b+6=5\\ 2a+12=8\end{cases}$$

解得 $a=-2,b=3,c=-1$. 此时 $s(x)$是$[0,2]$上的三次样条函数.

由定义 7.1 可知 $s(x)$在每个小区间$[x_i,x_{i+1}]$上是三次多项式,它有四个待定系数,$[a,b]$中共有 n 个小区间,故待定的系数为 $4n$ 个,而由定义给出的条件 $s(x)\in C^2[a,b]$,在 $x_1,x_2,\cdots,x_{n-1}$这$(n-1)$个内点上应满足

$$\begin{cases} s(x_i-0)=s(x_i+0) \\ s'(x_i-0)=s'(x_i+0) \\ s''(x_i-0)=s''(x_i+0), i=1,2,\cdots,n-1 \end{cases} \tag{5.7.2}$$

它给出了 $3(n-1)$个条件,此外由插值条件(5.7.1)给出了$(n+1)$个条件,共有$(4n-2)$个条件,求三次样条插值函数 $s(x)$尚缺两个条件.为此要根据问题要求补充两种边界条件,它们分别是

问题Ⅰ $s'(x_0)=f'_0,\quad s'(x_n)=f'_n$ (5.7.3)

问题Ⅱ $s''(x_0)=f''_0,\quad s''(x_n)=f''_n$ (5.7.4)

或 $s''(x_0)=s''(x_n)=0$,称为自然边界条件.

问题Ⅲ 当 $f(x)$为周期函数,因 $f(x_0)=f(x_n)$,此时 $s(x_0)=s(x_n)=f(x_0)$,且 $s'(x_0+0)=s'(x_0-0),s''(x_0+0)=s''(x_0-0)$.这时 $s(x)$称为周期样条函数.

由此看到针对不同类型问题,补充相应边界条件后完全可以求得三次样条插值函数 $s(x)$.下面我们只就问题Ⅰ及问题Ⅱ介绍三弯矩方程及其解法.

5.7.2 三弯矩方程

设 $s(x)$在节点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$ 上的二阶导数值 $s''(x_i)=M_i(i=0,1,\cdots,n)$,$h_i=x_{i+1}-x_i$,$s(x)$在$[x_i,x_{i+1}]$上是三次多项式,故 $s''(x)$在$[x_i,x_{i+1}]$上是一次函数,可表示为

$$s''(x)=\frac{x_{i+1}-x}{h_i}M_i+\frac{x-x_i}{h_i}M_{i+1}$$

对此式积分两次,并利用 $s(x_i)=f_i,s(x_{i+1})=f_{i+1}$可确定积分常数,从而得到

$$s(x)=\frac{(x_{i+1}-x)^3}{6h_i}M_i+\frac{(x-x_i)^3}{6h_i}M_{i+1}+\frac{x_{i+1}-x}{h_i}\left(f_i-\frac{h_i^2}{6}M_i\right)+$$

$$\frac{x-x_i}{h_i}\left(f_{i+1}-\frac{h_i^2}{6}M_{i+1}\right),x\in[x_i,x_{i+1}] \tag{5.7.5}$$

这里 $M_i(i=0,1,\cdots,n)$是未知量,但它可利用条件(5.7.2)中 $s'(x_i-0)=s'(x_i+0)(i=1,2,\cdots,n-1)$得到关于 $M_0,M_1,\cdots,M_n$ 的方程组,由(5.7.5)对 $s(x)$求导得

$$s'(x)=-\frac{(x_{i+1}-x)^2}{2h_i}M_i+\frac{(x-x_i)^2}{2h_i}M_{i+1}+\frac{f_{i+1}-f_i}{h_i}-\frac{h_i}{6}(M_{i+1}-M_i),x\in[x_i,x_{i+1}] \tag{5.7.6}$$

由此可得

$$s'(x_i+0)=-\frac{h_i}{2}M_i+\frac{f_{i+1}-f_i}{h_i}-\frac{h_i}{6}(M_{i+1}-M_i) \tag{5.7.7}$$

当 $x\in[x_{i-1},x_i]$,类似(5.7.6)可得

$$s'(x)=-\frac{(x_i-x)^2}{2h_{i-1}}M_{i-1}+\frac{(x-x_{i-1})^2}{2h_{i-1}}M_i+\frac{f_i-f_{i-1}}{h_{i-1}}-\frac{h_{i-1}}{6}(M_i-M_{i-1})$$

于是

$$s'(x_i-0)=\frac{h_{i-1}}{2}M_i+\frac{f_i-f_{i-1}}{h_{i-1}}-\frac{h_{i-1}}{6}(M_i-M_{i-1}) \tag{5.7.8}$$

由 $s'(x_i-0)=s'(x_i+0)$,可得到

$$\mu_iM_{i-1}+2M_i+\lambda_iM_{i+1}=d_i,i=1,2,\cdots,n-1 \tag{5.7.9}$$

其中

$$\begin{cases}\mu_i=\dfrac{h_{i-1}}{h_{i-1}+h_i},\quad \lambda_i=1-\mu_i=\dfrac{h_i}{h_{i-1}+h_i}\\ d_i=f[x_{i-1},x_i,x_{i+1}],i=1,2,\cdots,n-1\end{cases} \tag{5.7.10}$$

(5.7.8)是关于 $M_0,M_1,\cdots,M_{n+1}$的$(n-1)$个方程,对问题Ⅰ,可由(5.7.3)补充两个方程,它们可由(5.7.7)当 $i=0$ 时及(5.7.8)当 $i=n$ 时得到,即

$$\begin{cases}2M_0+M_1=\dfrac{6}{h_0}(f[x_0,x_1]-f'_0)=d_0\\ M_{n-1}+2M_n=\dfrac{6}{h_{n-1}}(f'_n-f[x_{n-1},x_n])=d_n\end{cases} \tag{5.7.11}$$

将(5.7.9)与(5.7.11)合并则得到关于 $M_0,M_1,\cdots,M_n$ 的线性方程组,用矩阵形式表示为

$$\begin{bmatrix}2 & 1 & & & \\ \mu_1 & 2 & \lambda_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 2 & \lambda_{n-1}\\ & & & 1 & 2\end{bmatrix}\begin{bmatrix}M_0\\M_1\\\vdots\\M_{n-1}\\M_n\end{bmatrix}=\begin{bmatrix}d_0\\d_1\\\vdots\\d_{n-1}\\d_n\end{bmatrix} \tag{5.7.12}$$

这是关于 $M_0,M_1,\cdots,M_n$ 的三对角方程组.

对于问题Ⅱ,可直接由条件(5.7.4)得到

$$M_0=f''_0,M_n=f''_n$$

将它代入(5.7.9),并用矩阵形式表示为

$$\begin{bmatrix}2 & \lambda_1 & & \\ \mu_2 & 2 & \lambda_2 & \\ & \ddots & \ddots & \ddots\\ & & \mu_{n-2} & 2 & \lambda_{n-2}\\ & & & \mu_{n-1} & 2\end{bmatrix}\begin{bmatrix}M_1\\M_2\\\vdots\\M_{n-2}\\M_{n-1}\end{bmatrix}=\begin{bmatrix}d_1-\mu_1f''_0\\d_2\\\vdots\\d_{n-2}\\d_{n-1}-\lambda_{n-1}f''_n\end{bmatrix} \tag{5.7.13}$$

它是关于 $M_1, M_2, \cdots, M_{n-1}$ 的三对角方程组，不论是(5.7.12)还是(5.7.13)，它们中每个方程只与三个相邻的 M_i 相联系，而 M_i 在力学上表示细梁在 x_i 上的截面弯矩，故称(5.7.12)及(5.7.13)为三弯矩方程. 方程(5.7.12)及(5.7.13)的系数矩阵都是严格对角占优矩阵，它们可用追赶法求解. 得到 $M_0, M_1, \cdots, M_n$ 后，代入(5.7.5)，则得到$[a,b]$上的三次样条插值函数 $s(x)$.

例 5.9 设 $f(x)$为定义在$[0,3]$上的函数，插值节点为 $x_i = i, i = 0,1,2,3$，且 $f(x_0) = 0$，$f(x_1) = 0.5, f(x_2) = 2.0, f(x_3) = 1.5$. 当 $f'(x_0) = 0.2, f'(x_3) = -1$ 时，试求三次样条插值函数 $s(x)$，使其满足问题 Ⅰ 的边界条件(5.7.3).

解 根据三弯矩方程(5.7.12)，首先要求系数矩阵及右端项 $d_i (i = 0,1,2,3)$，由(5.7.10)及(5.7.11)可得

$$h_i = 1 (i = 0,1,2), \mu_1 = \mu_2 = \lambda_1 = \lambda_2 = 1/2$$

$$d_1 = 6f[x_0, x_1, x_2] = 3, d_2 = 6f[x_1, x_2, x_3] = -6$$

$$d_0 = \frac{6}{h_0}(f[x_0, x_1] - f'(x_0)) = 1.8, d_3 = \frac{6}{h_2}(f'(x_3) - f[x_2, x_3]) = -3$$

于是由(5.7.12)得三弯矩方程为

$$\begin{bmatrix} 2 & 1 & & \\ 0.5 & 2 & 0.5 & \\ & 0.5 & 2 & 0.5 \\ & & 1 & 2 \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \end{bmatrix} = \begin{bmatrix} 1.8 \\ 3 \\ -6 \\ -3 \end{bmatrix} \tag{5.7.14}$$

解此方程时可先消去 M_0, M_3 得

$$\begin{bmatrix} 3.5 & 1 \\ 1 & 3.5 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} = \begin{bmatrix} 5.1 \\ -10.5 \end{bmatrix}$$

解得 $M_1 = 2.52, M_2 = -3.72$，代入(5.7.14)得 $M_0 = -0.36, M_3 = 0.36$. 将 M_0, M_1, M_2, M_3 的值代入(5.7.5)可得三次样条函数

$$s(x) = \begin{cases} 0.48x^3 - 0.18x^2 + 0.2x, & x \in [0,1] \\ -1.04(x-1)^3 + 1.25(x-1)^2 + 1.28(x-1) + 0.5, & x \in [1,2] \\ 0.68(x-2)^3 - 1.86(x-2)^2 + 0.68(x-2) + 2.0, & x \in [2,3] \end{cases}$$

$y = s(x)$的图形见图 5－5.

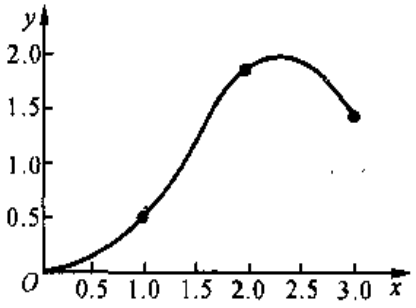

图 5－5

5.7.3 三次样条插值收敛性

定理 7.1 设 $f(x)\in C^4[a,b]$. $s(x)$为问题Ⅰ或问题Ⅱ的三次样条函数,则有估计式

$$\| f^{(k)}(x)-s^{(k)}(x)\|_\infty \leqslant C_k \| f^{(4)}(x)\|_\infty h^{4-k}, k=0,1,2 \tag{5.7.15}$$

其中 $h=\max\limits_{0\leqslant i\leqslant n-1} h_i, h_i=x_{i+1}-x_i \quad (i=0,1,\cdots,n-1), C_0=\dfrac{5}{384}, C_1=\dfrac{1}{24}, C_2=\dfrac{3}{8}, \| f(x)\|_\infty = \max\limits_{a\leqslant x\leqslant b}|f(x)|$.

定理证明见[3].

定理表明当 $h\to 0(n\to\infty)$时,$s(x), s'(x), s''(x)$分别一致收敛于 $f(x), f'(x), f''(x)$.

5.8 曲线拟合的最小二乘法

在科学实验数据处理中,往往要根据一组给定的实验数据$(x_i, y_i)(i=0,1,\cdots,m)$,求出自变量 x 与因变量 y 的函数关系 $y\approx s(x;a_0,\cdots,a_n)(n<m)$,这是 a_i 为待定参数,由于观测数据总有误差,且待定参数 a_i 的数量比给定数据点的数量少(即 $n<m$),因此它不同于插值问题.这类问题不要求 $y=s(x)\approx s(x;a_0,\cdots,a_n)$通过点$(x_i,y_i)(i=0,1,\cdots,m)$,而只要求在给定点 x_i 上的误差 $\delta_i=s(x_i)-y_i \quad (i=0,1,\cdots,m)$的平方和 $\sum\limits_{i=0}^{m}\delta_i^2$ 最小.当 $s(x)\in \mathrm{span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$时,即

$$s(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x) \tag{5.8.1}$$

这里 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\in C[a,b]$是线性无关的函数族,假定在$[a,b]$上给出一组数据$\{(x_i,y_i), i=0,1,\cdots,m\}$, $a\leqslant x_i\leqslant b$ 以及对应的一组权$\{\rho_i\}_0^m$,这里 $\rho_i>0$ 为权系数,要求 $s(x)=\mathrm{span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$使 $I(a_0,a_1,\cdots,a_n)$最小,其中

$$I(a_0,a_1,\cdots,a_n)=\sum_{i=0}^{m}\rho_i[s(x_i)-y_i]^2 \tag{5.8.2}$$

这就是最小二乘逼近,得到的拟合曲线为 $y=s(x)$,这种方法称为曲线拟合的最小二乘法.

(5.8.2)中 $I(a_0,a_1,\cdots,a_n)$实际上是关于 $a_0,a_1,\cdots,a_n$ 的多元函数,求 I 的最小值就是求多元函数 I 的极值,由极值必要条件,可得

$$\frac{\partial I}{\partial a_k}=2\sum_{i=0}^{m}\rho_i[a_0\varphi_0(x_i)+a_1\varphi_1(x_i)+\cdots+a_n\varphi_n(x_i)-y_i]\varphi_k(x_i)=0, k=0,1,\cdots,n \tag{5.8.3}$$

根据内积定义(见第三章)引入相应带权内积记号

$$\begin{cases}(\varphi_j,\varphi_k)=\sum\limits_{i=0}^{m}\rho_i\varphi_j(x_i)\varphi_k(x_i)\\ (y,\varphi_k)=\sum\limits_{i=0}^{m}\rho_i y_i\varphi_k(x_i)\end{cases} \tag{5.8.4}$$

则(5.8.3)可改写为

$$(\varphi_0,\varphi_k)a_0+(\varphi_1,\varphi_k)a_1+\cdots+(\varphi_n,\varphi_k)a_n=(y,\varphi_k),k=0,1,\cdots,n$$

这是关于参数 $a_0,a_1,\cdots,a_n$ 的线性方程组,用矩阵表示为

$$\begin{bmatrix}(\varphi_0,\varphi_0) & (\varphi_0,\varphi_1)\cdots(\varphi_0,\varphi_n)\\ (\varphi_1,\varphi_0) & (\varphi_1,\varphi_1)\cdots(\varphi_1,\varphi_n)\\ \vdots & \vdots \quad\quad \vdots\\ (\varphi_n,\varphi_0) & (\varphi_n,\varphi_1)\cdots(\varphi_n,\varphi_n)\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ \vdots\\ a_n\end{bmatrix}=\begin{bmatrix}(y,\varphi_0)\\ (y,\varphi_1)\\ \vdots\\ (y,\varphi_n)\end{bmatrix} \tag{5.8.5}$$

(5.8.5)称为法方程.当$\{\varphi_j(x);j=0,1,\cdots,n\}$线性无关,且在点集 $X=\{x_0,x_1,\cdots,x_m\}(m\geqslant n)$上至多只有 n 个不同零点,则称 $\varphi_0,\varphi_1,\cdots,\varphi_n$ 在 X 上满足 Haar 条件,此时(5.8.5)的解存在唯一(证明见[3]).记(5.8.5)的解为

$$a_k=a_k^*,k=0,1,\cdots,n$$

从而得到最小二乘拟合曲线

$$y=s^*(x)=a_0^*\varphi_0(x)+a_1^*\varphi_1(x)+\cdots+a_n^*\varphi_n(x) \tag{5.8.6}$$

可以证明对 $\forall(a_0,a_1,\cdots,a_n)^{\mathrm{T}}\in\mathbf{R}^{n+1}$,有

$$I(a_0^*,a_1^*,\cdots,a_n^*)\leqslant I(a_0,a_1,\cdots,a_n)$$

故(5.8.6)得到的 $s^*(x)$即为所求的最小二乘解.它的平方误差为

$$\|\delta\|_2^2=\sum_{i=0}^{m}\rho_i[s^*(x_i)-y_i]^2 \tag{5.8.7}$$

均方误差为

$$\|\delta\|_2=\sqrt{\sum_{i=0}^{m}\rho_i[s^*(x_i)-y_i]^2}$$

在最小二乘逼近中,若取 $\varphi_k(x)=x^k(k=0,1,\cdots,n)$,则 $s(x)\in\mathrm{span}\{1,x,\cdots,x^n\}$,表示为

$$s(x)=a_0+a_1x+\cdots+a_nx^n \tag{5.8.8}$$

此时关于系数 $a_0,a_1,\cdots,a_n$ 的法方程(5.8.5)是病态方程,通常当 $n\geqslant 3$ 时都不直接取 $\varphi_k(x)=x^k$ 作为基,其具体方法下节再讨论,下而只给出 $n=1$ 的例子.

例 5.10 已知一组实验数据如表所示.

x_i	1	2	3	4	5
y_i	4	4.5	6	8	8.5
ρ_i	2	1	3	1	1

试求最小二乘拟合曲线.

解 将所给数据在坐标纸上标出,如图 5-6 所示,说明它可用线性函数作曲线拟合,即选择形如 $s_1(x)=a_0+a_1x$ 作为拟合曲线.这里 $\varphi_0=1,\varphi_1=x,m=5,n=1$,故

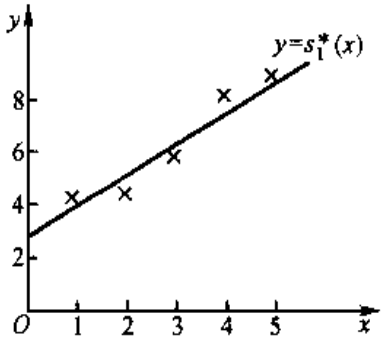

图 5－6

$$(\varphi_0,\varphi_0)=\sum_{i=0}^{4}\rho_i=8,(\varphi_1,\varphi_0)=(\varphi_0,\varphi_1)=\sum_{i=0}^{4}\rho_i x_i=22$$

$$(\varphi_1,\varphi_1)=\sum_{i=0}^{4}\rho_i x_i^2=74,(\varphi_0,y)=\sum_{i=0}^{4}\rho_i y_i=47,(\varphi_1,y)=\sum_{i=0}^{4}\rho_i x_i y_i=145.5$$

于是由(5.8.5)得法方程

$$\begin{cases}8a_0+22a_1=47\\22a_0+74a_1=145.5\end{cases}$$

解得

$$a_0=2.77,a_1=1.13$$

于是所求的最小二乘拟合曲线为

$$y=s_1^*(x)=2.77+1.13x$$

均方误差为 $\|\delta\|_2=0.942\,4$.

使用最小二乘逼近时,模型的选择是很重要的,通常模型 $y=s(x)$ 是由物理规律或数据分布情况确定的,不一定都是形如(5.8.1)的线性模型,但有的模型经过变换可化为线性模型,这些也应按线性模型处理,例如

$$y=ae^{bx},s(x)=ae^{bx}$$

它是指数函数,关于系数 a,b 并非线性,但对上式两端取对数得到

$$\ln y=\ln a+bx$$

令 $\bar{y}=\ln y,A=\ln a$,则上式转化为 $\bar{y}=A+bx$,它是线性模型,仍可按上面介绍的方法求 $y=s(x)$.

例 5.11 给定数据$(x_i,y_i)(i=0,1,2,3,4)$如下:

i	0	1	2	3	4
x_i	1.00	1.25	1.50	1.75	2.00
y_i	5.10	5.79	6.53	7.45	8.46

求 $y=ae^{bx}$ 的最小二乘拟合曲线.

解 $y=ae^{bx}$ 不是多项式,但两端取对数得 $\ln y=\ln a+bx$.若令 $\bar{y}=\ln y, A=\ln a$,则有 $\bar{y}=A+bx$,它是线性最小二乘拟合问题.可取 $\varphi_0(x)=1, \varphi_1(x)=x$,为求得 A, b,先将 (x_i, y_i) 化为 $(x_i, \bar{y}_i)$.转化后的数据表为

x_i	1.00	1.25	1.50	1.75	2.00
$\bar{y}_i$	1.629	1.756	1.876	2.008	2.135

根据最小二乘原理先求法方程系数

$$(\varphi_0,\varphi_0)=5,\quad (\varphi_0,\varphi_1)=\sum_{i=0}^{4}x_i=7.5$$

$$(\varphi_1,\varphi_1)=\sum_{i=0}^{4}x_i^2=11.875,\quad (\varphi_0,\bar{y})=\sum_{i=0}^{4}\bar{y}_i=9.404$$

$$(\varphi_1,\bar{y})=\sum_{i=0}^{4}x_i\bar{y}_i=14.422$$

故有法方程

$$\begin{cases}5A+7.50b=9.404\\7.50A+11.875b=14.422\end{cases}$$

解得 $A=1.122, b=0.505\,6, a=e^A=3.071$,于是得最小二乘拟合曲线

$$y=3.071e^{0.505\,6x}$$

5.9 正交多项式及其在最小二乘的应用

5.9.1 内积与正交多项式

将 $\mathbf{R}^n$ 空间向量的内积定义推广到连续函数空间 $C[a,b]$,就有

定义 9.1 设 $f(x)g(x)\in C[a,b]$,$\rho(x)$是$[a,b]$上的权函数记

$$(f,g)=\int_a^b f(x)g(x)\rho(x)\mathrm{d}x \tag{5.9.1}$$

称为函数 $f(x)$与 $g(x)$在$[a,b]$上的带权内积.

内积有以下性质:

(1) $(f,g)=(g,f)$;

(2) $(af,g)=a(f,g), a\in\mathbf{R}^1$;

(3) $(f_1+f_2,g)=(f_1,g)+(f_2,g)$;

(4) $(f,f)\geqslant 0$,当且仅当 $f\equiv 0$ 时等号成立.

定义 9.2 设 $f(x), g(x)\in C[a,b]$,$\rho(x)$为$[a,b]$上的权函数,若

$$(f,g)=\int_a^b f(x)g(x)\rho(x)\mathrm{d}x=0$$

则称 $f(x)$与 $g(x)$在$[a,b]$上带权 $\rho(x)$正交.若函数序列$\{\varphi_i(x)\}_0^\infty$ 在$[a,b]$上两两正交,即

$$(\varphi_i,\varphi_j)=\begin{cases}0, & i\neq j\\ A_j\neq 0, & i=j\end{cases}$$

则称$\{\varphi_i(x)\}_0^\infty$ 为正交函数族.

例 5.12 三角函数族 $1,\sin x,\cos x,\sin 2x,\cos 2x,\cdots$在$[-\pi,\pi]$上是正交函数族(权 $\rho(x)\equiv 1$).

实际上$(1,1)=\int_{-\pi}^{\pi}\mathrm{d}x=2\pi$,而

$$(\sin nx,\sin mx)=\int_{-\pi}^{\pi}\sin nx\sin mx\mathrm{d}x=\begin{cases}\pi, m=n\\ 0, m\neq n\end{cases}\quad n,m=1,2,\cdots$$

$$(\cos nx,\cos mx)=\int_{-\pi}^{\pi}\cos nx\cos mx\mathrm{d}x=\begin{cases}\pi, m=n\\ 0, m\neq n\end{cases}\quad n,m=1,2,\cdots$$

$$(\cos nx,\sin mx)=\int_{-\pi}^{\pi}\cos nx\sin mx\mathrm{d}x=0,\quad m,n=0,1,\cdots$$

定义 9.3 设 $\varphi_n(x)$是首项系数 $a_n\neq 0$ 的 n 次多项式,如果多项式序列$\{\varphi_n(x)\}_0^\infty$ 满足

$$(\varphi_i,\varphi_j)=\int_a^b\varphi_i(x)\varphi_j(x)\rho(x)\mathrm{d}x=\begin{cases}0, & i\neq j\\ A_j\neq 0, & i=j\end{cases}\tag{5.9.2}$$

则称多项式序列$\{\varphi_i(x)\}_0^\infty$ 为在$[a,b]$上带权 $\rho(x)$的正交多项式族,$\varphi_n(x)$称为$[a,b]$上带权 $\rho(x)$的 n 次正交多项式.

只要给定区间$[a,b]$及权函数 $\rho(x)$,均可由线性无关的一组基$\{1,x,x^2,\cdots,x^n,\cdots\}$,利用正交化构造出正交多项式$\{\varphi_n(x)\}_0^\infty$

$$\varphi_0(x)=1,\varphi_n(x)=x^n-\sum_{k=0}^{n-1}\frac{(x^n,\varphi_k)}{(\varphi_k,\varphi_k)}\varphi_k(x),\quad n=1,2,\cdots\tag{5.9.3}$$

这样构造的正交多项式有以下性质:

(1) $\varphi_n(x)$是最高项系数为 1 的 n 次多项式;

(2) 任何 n 次多项式 $p_n(x)\in H_n$,均可表示为 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$的线性组合;

(3) 当 $n\neq m$ 时,$(\varphi_n,\varphi_m)=0$,且 $\varphi_n(x)$与任一次数小于 n 的多项式正交;

(4) 递推关系

$$\varphi_{n+1}(x)=(x-a_n)\varphi_n(x)-\beta_n\varphi_{n-1}(x),\quad n=0,1,\cdots\tag{5.9.4}$$

其中

$$\varphi_0(x)=1,\varphi_{-1}(x)=0$$

$$a_n=\frac{(x\varphi_n,\varphi_n)}{(\varphi_n,\varphi_n)},\qquad n=0,1,\cdots$$

$$\beta_n=\frac{(\varphi_n,\varphi_n)}{(\varphi_{n-1},\varphi_{n-1})},\qquad n=1,2,\cdots$$

这里$(x\varphi_n,\varphi_n)=\int_a^b x\varphi_n^2(x)\rho(x)\mathrm{d}x$.

(5) 设$\{\varphi_n(x)\}_0^\infty$是在$[a,b]$上带权$\rho(x)$的正交多项式序列,则$\varphi_n(x)(n\geqslant 1)$的$n$个根都是单重实根,且都在区间$(a,b)$内.

以上性质的证明见[4].下面给出常见的面又十分重要的正交多项式.

5.9.2 Legendre 多项式

在区间$[-1,1]$上权函数$\rho(x)=1$的正交多项式称为 Legendre 多项式,其表达式为

$$P_0(x)=1,P_n(x)=\frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\{(x^2-1)^n\},\quad n=1,2,\cdots \tag{5.9.5}$$

$P_n(x)$的首项x^n的系数为$\frac{(2n)!}{2^n(n!)^2}$,记

$$\tilde{P}_0(x)=1,\tilde{P}_n(x)=\frac{n!}{(2n)!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\{(x^2-1)^n\},\quad n=1,2,\cdots \tag{5.9.6}$$

则$\tilde{P}_n(x)$是首项x^n系数为1的 Legendre 多项式.

Legendre 多项式有许多重要性质,其中较重要的有:

(1) 正交性

$$(P_n,P_m)=\int_{-1}^1 P_n(x)P_m(x)\mathrm{d}x=\begin{cases}0, & m\neq n\\ \dfrac{2}{2n+1}, & m=n\end{cases} \tag{5.9.7}$$

只要令$\varphi(x)=(x^2-1)^n$,则$\varphi^{(k)}(\pm 1)=0,k=0,1,\cdots,n-1$且$P_n(x)=\frac{1}{2^n n!}\varphi^{(n)}(x)$.设多项式$Q(x)\in H_n$,用分部积分得

$$\begin{aligned}\int_{-1}^1 P_n(x)Q(x)\mathrm{d}x&=\frac{1}{2^n n!}\int_{-1}^1 Q(x)\varphi^{(n)}(x)\mathrm{d}x\\&=-\frac{1}{2^n n!}\int_{-1}^1 Q'(x)\varphi^{(n-1)}(x)\mathrm{d}x\\&=\cdots=\frac{(-1)^n}{2^n n!}\int_{-1}^1 Q^{(n)}(x)\varphi(x)\mathrm{d}x\end{aligned}$$

当$Q(x)$为次数不超过$(n-1)$时$Q^{(n)}(x)=0$,于是有

$$\int_{-1}^1 P_n(x)P_m(x)\mathrm{d}x=0,\quad m\neq n$$

当$Q(x)=P_n(x)$,则$Q^{(n)}(x)=P_n^{(n)}(x)=\frac{(2n)!}{2^n n!}$,于是

$$\int_{-1}^1 P_n^2(x)\mathrm{d}x=\frac{(-1)^n(2n)!}{2^{2n}(n!)^2}\int_{-1}^1(x^2-1)^n\mathrm{d}x=\frac{2}{2n+1}$$

这就证明了(5.9.7)的正确性.

(2) 递推公式

$$(n+1)P_{n+1}(x)=(2n+1)xP_n(x)-nP_{n-1}(x),\quad n=1,2,\cdots \tag{5.9.8}$$

其中 $P_0(x)=1,P_1(x)=x$.

此公式可直接利用正交性证明.由(5.9.8)可得

$$P_2(x)=\frac{1}{2}(3x^2-1)$$

$$P_3(x)=\frac{1}{2}(5x^3-3x)$$

$$P_4(x)=\frac{1}{8}(35x^4-30x^2+3)$$

$$P_5(x)=\frac{1}{8}(63x^5-70x^3+15x)$$

.....

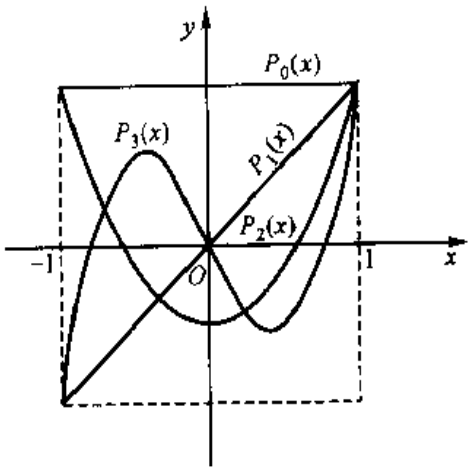

图 5-7

图 5-7 给出了 $P_0(x),P_1(x),P_2(x),P_3(x)$的图形.

(3) 奇偶性

$$P_n(-x)=(-1)^nP_n(x)$$

5.9.3 Chebyshev 多项式

在区间$[-1,1]$上权函数 $\rho(x)=\dfrac{1}{\sqrt{1-x^2}}$的正交多项式称为 Chebyshev 多项式,它可表示为

$$T_n(x)=\cos(n\arccos x),\quad n=0,1,\cdots \tag{5.9.9}$$

若令 $x=\cos\theta$,则 $T_n(x)=\cos n\theta,0\leqslant\theta\leqslant\pi$,这是 $T_n(x)$的参数表示.利用三角公式可将 $\cos n\theta$ 展成 $\cos\theta$ 的一个 n 次多项式,故(5.9.9)可视为 x 的 n 次多项式.下面给出 $T_n(x)$的主要性质:

(1) 正交性

$$(T_n,T_m)=\int_{-1}^{1}\frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\mathrm{d}x=\begin{cases}0, & m\neq n\\ \dfrac{\pi}{2}, & m=n\neq 0\\ \pi, & m=n=0\end{cases} \tag{5.9.10}$$

只要对积分做变换 $x=\cos\theta$,利用三角公式即可得到(5.9.10)的结果.

(2) 递推公式

$$T_{n+1}(x)=2xT_n(x)-T_{n-1}(x),\quad n=1,2,\cdots \tag{5.9.11}$$

其中 $T_0(x)=1,\quad T_1(x)=x$.

由 $x=\cos\theta,T_{n+1}(x)=\cos(n+1)\theta$ 用三角公式

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

则得(5.9.11).由(5.9.11)可推出 $T_2(x)$到 $T_8(x)$如下:

$$T_2(x)=2x^2-1$$
$$T_3(x)=4x^3-3x$$
$$T_4(x)=8x^4-8x^2+1$$
$$T_5(x)=16x^5-20x^3+5x$$
$$T_6(x)=32x^6-48x^4+18x^2-1$$
$$T_7(x)=64x^7-112x^5+56x^3-7x$$
$$T_8(x)=128x^8-256x^6+160x^4-32x^2+1$$

图 5-8 给出了 $T_0(x),T_1(x),T_2(x),T_3(x)$的图形.

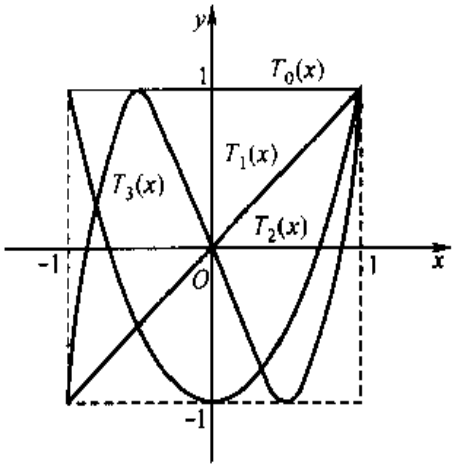

图 5-8

(3) 奇偶性

$$T_n(-x)=(-1)^nT_n(x)$$

(4) $T_n(x)$在$(-1,1)$内的 n 个零点为 $x_k=\cos\left(\frac{2k-1}{2n}\pi\right)(k=1,2,\cdots,n)$,在$[-1,1]$上有$(n+1)$个极点 $y_k=\cos\left(\frac{k}{n}\pi\right)(k=0,1,\cdots,n)$.

(5) $T_n(x)$的最高次幂 x^n 的系数为 2^{n-1},$n\geqslant 1$.

5.9.4 其他正交多项式

除上述两个最常用的正交多项式外,较重要的还有无穷区间的正交多项式,它们是:

(1) Laguerre 多项式

在区间$[0,\infty)$上,权函数 $\rho(x)=e^{-x}$的正交多项式称为 Laguerre 多项式,其表达式为

$$L_n(x)=e^x\frac{d^n}{dx^n}(x^ne^{-x}),\quad n=0,1,\cdots \tag{5.9.12}$$

它的递推公式为

$$L_{n+1}(x)=(1+2n-x)L_n(x)-n^2L_{n-1}(x),\quad n=1,2,\cdots \tag{5.9.13}$$

其中 $L_0(x)=1,L_1(x)=1-x$.正交性为

$$(L_n,L_m)=\int_0^\infty L_n(x)L_m(x)e^{-x}dx=\begin{cases}0, & m\neq n\\ (n!)^2, & m=n\end{cases}$$

(2) Hermite 多项式

在区间$(-\infty,\infty)$上,带权函数 $\rho(x)=e^{-x^2}$的正交多项式称为 Hermite 多项式,其表达式为

$$H_n(x)=(-1)^ne^{x^2}\frac{d^n}{dx^n}e^{-x^2} \tag{5.9.14}$$

它的递推公式为

$$H_{n+1}(x)=2xH_n(x)-2nH_{n-1}(x),\quad n=1,2,\cdots \tag{5.9.15}$$

其中 $H_0(x)=1,H_1(x)=2x$.正交性为

$$(\mathrm{H}_n,\mathrm{H}_m)=\int_{-\infty}^{\infty}\mathrm{H}_n(x)\mathrm{H}_m(x)\mathrm{e}^{-x^2}\mathrm{d}x=\begin{cases}0, & m\neq n\\ 2^n n!\sqrt{\pi}, & m=n\end{cases}.$$

5.9.5 用正交多项式作最小二乘拟合

在最小二乘拟合中若 $H_n=\mathrm{span}\{1,x,\cdots,x^n\}$，模型取为(5.8.8)时，由于法方程是病态方程，因此使用时应取$\{\varphi_k(x)\}$为关于给定点的正交多项式，可避免求解病态方程组.类似定义9.3给出以下定义.

定义 9.4 设给定拟合数据(x_i,y_i)及权 $\omega_i(i=0,1,\cdots,m)$可构造多项式$\{P_k(x)\}_0^n$，其中$P_k(x)\in H_k$，且

$$(P_j,P_k)=\sum_{i=0}^{m}\omega_i P_j(x_i)P_k(x_i)=\begin{cases}0, & j\neq k\\ A_k, & j=k\end{cases}\tag{5.9.16}$$

则称$\{P_k\}_0^n$是关于点集$\{x_i\}_0^m$.带权$\{\omega_i\}_0^m$的正交多项式族，$P_k(x)$为 k 次正交多项式.

根据定义，若令 $P_0(x)=1$，　$P_1(x)=(x-\alpha_1)P_0(x)$.

由递推关系得

$$P_{k+1}(x)=(x-\alpha_{k+1})P_k(x)-\beta_k P_{k-1}(x),\quad k=1,2,\cdots,n-1\tag{5.9.17}$$

利用正交性

$$(P_k,P_j)=\sum_{i=0}^{m}\omega_i P_k(x_i)P_j(x_i)=\begin{cases}0, & j\neq k\\ A_k>0, & j=k\end{cases},\quad j,k=0,1,\cdots,n$$

求得 α_{k+1}及 β_k 为

$$\begin{cases}\alpha_{k+1}=\dfrac{(xP_k,P_k)}{(P_k,P_k)}=\dfrac{\sum\limits_{i=0}^{m}\omega_i x_i P_k^2(x_i)}{\sum\limits_{i=0}^{m}\omega_i P_k^2(x_i)}\\ \beta_k=\dfrac{(P_k,P_k)}{(P_{k-1},P_{k-1})},\quad k=1,2,\cdots,n-1\end{cases}\tag{5.9.18}$$

令 $\varphi_k=P_k(k=0,1,\cdots,n)$，由法方程(5.8.5)可求得解

$$a_k=a_k^*=\frac{(y,P_k)}{(P_k,P_k)}=\frac{\sum\limits_{i=0}^{m}\omega_i y_i P_k(x_i)}{\sum\limits_{i=0}^{m}\omega_i P_k^2(x_i)}\tag{5.9.19}$$

从而得到最小二乘拟合曲线

$$y=s_n^*(x)=a_0^*P_0(x)+a_1^*P_1(x)+\cdots+a_n^*P_n(x)\tag{5.9.20}$$

它仍然是多项式函数，即 $s_n^*(x)\in H_n$.用计算机计算时求$\{P_k\}$系数 α_{k+1}及 β_k 与求$s_n^*(x)$系数 a_k^* 可同时进行.当 $k=0,1,\cdots,n$ 时若有 $\|\delta\|_2^2=\|y-s_n^*\|_2^2\leqslant\|y-s_{n+1}^*\|_2^2$ 时，计算停止，此时 $y=s_n^*(x)$即为所求.

习 题 五

1. 给定 $f(x)=\ln x$ 的数值表

x	0.4	0.5	0.6	0.7
$\ln x$	−0.916 291	−0.693 147	−0.510 826	−0.356 675

用线性插值与二次插值计算 $\ln 0.54$ 的近似值并估计误差限.

2. 在 $-4\leqslant x\leqslant 4$ 上给出 $f(x)=e^x$ 的等距节点函数表,若用二次插值法求 e^x 的近似值,要使误差不超过 10^{-6},函数表的步长 h 应取多少?

3. 设 x_j 为互异节点($j=0,1,\cdots,n$),$l_j(x)$为 n 次插值基函数.证明

(1) $\sum_{j=0}^{n} l_j(x)=1$

(2) $\sum_{j=0}^{n} l_j(0)x_j^k=\begin{cases}1, & k=0\\ 0, & k=1,2,\cdots,n\\ (-1)^n x_0x_1\cdots x_n, & k=n+1\end{cases}$

4. 若 $f(x)=x^7+x^4+3x+1$,求 $f[2^0,2^1,\cdots,2^7]$和 $f[2^0,2^1,\cdots,2^8]$.

5. 若 $f(x)=\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$,$x_i(i=0,1,\cdots,n)$互异,求 $f[x_0,x_1,\cdots,x_p]$的值,这里 $p\leqslant n+1$.

6. 求证 $\sum_{j=0}^{n-1}\Delta^2 y_j=\Delta y_n-\Delta y_0$.

7. 已知 $f(x)=\operatorname{sh} x$ 的函数表

x_i	0	0.20	0.30	0.50
$f(x_i)$	0	0.201 34	0.304 52	0.521 10

求出三次 Newton 均差插值多项式,计算 $f(0.23)$的近似值并用均差的余项表达式估计误差.

8. 给定 $f(x)=\cos x$ 的函数表

x_i	0	0.1	0.2	0.3	0.4	0.5	0.6
$f(x_i)$	1.000 00	0.995 00	0.980 07	0.955 34	0.921 06	0.877 58	0.825 34

用 Newton 等距插值公式计算 $\cos 0.048$ 及 $\cos 0.566$ 的近似值并估计误差.

9. 求一个次数不高于四次的多项式 $p(x)$,使它满足 $p(0)=p'(0)=0$,$p(1)=p'(1)=1$,$p(2)=1$.

10. 求多项式 p,使其满足条件 $p(x_i)=f(x_i)$,其中 $i=0,1$,$p'(x_0)=f'(x_0)$,$p''(x_0)=f''(x_0)$,并求余项表达式.

11. 给定数据

x_i	1	2
$f(x_i)$	2	3
$f'(x_i)$	0	-1

试构造 Hermite 插值多项式 $H_3(x)$并计算 $f(1.5)$.

12. 求 $f(x)=x^2$ 在$[a,b]$上的分段线性插值函数 $I_h(x)$,并估计误差.

13. 给定数据表

x_i	27.7	28	29	30
$f(x_i)$	4.1	4.3	4.1	3.0

试求三次样条插值函数 $s(x)$,使其满足边界条件

(1) $s'(27.7)=3.0$, $s'(30)=-4.0$

(2) $s''(27.7)=s''(30)=0$

14. 令 $s_n(x)=\frac{1}{n+1}T'_{n+1}(x),n\geqslant 0,s_n$ 称为第二类 Chebyshev 多项式,试求 s_n 的表达式,并证明$\{s_n\}$是$[-1,1]$上带权 $\rho(x)=\sqrt{1-x^2}$的正交多项式序列.

15. 已知点序列$\{x_i\}_{i=0}^{m}=\{-2,-1,0,1,2\}$和权数$\{\omega_i\}_{i=0}^{m}=\{0.5,1,1,1,1.5\}$,试用三项递推公式构造对应的正交多项式 $p_0(x),p_1(x),p_2(x)$.

16. 用最小二乘法求一个形如 $y=a+bx^2$ 的经验公式,使它拟合下列数据,并计算均方误差.

x_i	19	25	31	38	44
y_i	19.0	32.3	49.0	73.3	97.8

17. 填空题

(1) 满足条件 $p(0)=1,p(1)=p'(1)=0,p(2)=2$ 的插值多项式 $p(x)=$_____.

(2) $f(x)=2x^3+5$,则 $f[1,2,3,4]=$_____,$f[1,2,3,4,5]=$_____.

(3) 设 $x_i(i=0,1,2,3,4)$为互异节点,$l_i(x)$为对应的四次插值基函数,则 $\sum_{i=0}^{4}x_i^4l_i(0)=$_____, $\sum_{i=0}^{4}(x_i^4+2)l_i(x)=$_____.

(4) 设$\{\varphi_k(x)\}_{k=0}^{\infty}$是区间[0,1]上权函数为 $\rho(x)=x$ 的最高项系数为 1 的正交多项式序列,其中 $\varphi_0(x)=1$,则 $\int_0^1 x\varphi_k(x)\mathrm{d}x=$_____,$\varphi_2(x)=$_____.

第六章 数值积分

6.1 数值积分基本概念

6.1.1 引言

在区间$[a,b]$上求定积分

$$I(f)=\int_a^b f(x)\mathrm{d}x \tag{6.1.1}$$

是一个具有广泛应用的古典问题,从理论上讲,计算定积分可用 Newton－Leibniz 公式

$$\int_a^b f(x)\mathrm{d}x=F(b)-F(a) \tag{6.1.2}$$

其中 $F(x)$是被积函数 $f(x)$的原函数.但实际上有很多被积函数找不到用解析式子表达的原函数,例如$\int_0^1 \frac{\sin x}{x}\mathrm{d}x$, $\int_0^{\pi} \mathrm{e}^{\cos\theta}\mathrm{d}\theta$, $\int_0^1 \mathrm{e}^{-x^2}\mathrm{d}x$ 等等,表面看它们并不复杂,但却无法求得 $F(x)$.此外,有的积分即使能找到 $F(x)$表达式,但式子非常复杂,计算也很困难.还有的被积函数是列表函数,也无法用(6.1.2)的公式计算.而数值积分则只需计算 $f(x)$ 在节点 $x_i(i=0,1,\cdots,n)$上的值,计算方便且适合于在计算机上机械地实现.

本章将介绍常用的数值积分公式及其误差估计、求积公式的代数精确度、收敛性和稳定性以及 Romberg 求积法与外推原理等.

6.1.2 插值求积公式

根据定积分定义,对 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$ 及 $\forall\,\xi_i\in[x_{i-1},x_i]$都有 $I(f)=\int_a^b f(x)\mathrm{d}x=\lim\limits_{\Delta x_i\to 0}\sum\limits_{i=1}^{n} f(\xi_i)\Delta x_i$(极限存在)若不取极限,则积分 $I(f)$可近似表示为

$$I(f)=\int_a^b f(x)\mathrm{d}x\approx\sum_{i=0}^{n} A_i f(x_i)=I_n(f) \tag{6.1.3}$$

这里 $x_i(i=0,1,\cdots,n)$称为求积节点,A_i 与 f 无关,称为求积系数,(6.1.3)称为机械求积公式.

为了得到形如(6.1.3)的求积公式,可在$[a,b]$上用 Lagrange 插值多项式 $L_n(x)\approx f(x)$,则得

$$I(f)=\int_a^b f(x)\mathrm{d}x\approx\int_a^b L_n(x)\mathrm{d}x=\int_a^b\sum_{i=0}^{n}\left[\frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)}f(x_i)\right]\mathrm{d}x$$

$$=\sum_{i=0}^{n}A_if(x_i)$$

其中

$$A_i=\int_a^b\frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)}\mathrm{d}x \tag{6.1.4}$$

这里求积系数 A_i 由插值基函数 $l_i(x)$ 积分得到，它与 $f(x)$ 无关. 如果求积公式(6.1.3)中的系数由(6.1.4)给出，则称(6.1.3)为插值求积公式. 此时可由插值余项得到

$$R_n(f)=\int_a^b f(x)\mathrm{d}x-\sum_{i=0}^{n}A_if(x_i)=\int_a^b\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)\mathrm{d}x \tag{6.1.5}$$

这里 $\xi\in(a,b)$，(6.1.5)称为插值求积公式余项.

当 $n=1$ 时，$x_0=a$，$x_1=b$，此时

$$L_1(x)=\frac{x-b}{a-b}f(a)+\frac{x-a}{b-a}f(b)$$

由(6.1.4)可得

$$A_0=\int_a^b\frac{x-b}{a-b}\mathrm{d}x=\frac{b-a}{2},\quad A_1=\int_a^b\frac{x-b}{b-a}\mathrm{d}x=\frac{b-a}{2}$$

于是

$$I(f)=\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{2}[f(a)+f(b)] \tag{6.1.6}$$

称为梯形公式. 从几何上看它是梯形 $AabB$(见图 6-1)的面积近似曲线 $y=f(x)$ 下的曲边梯形面积 $\int_a^b f(x)\mathrm{d}x$，公式(6.1.6)的余项为

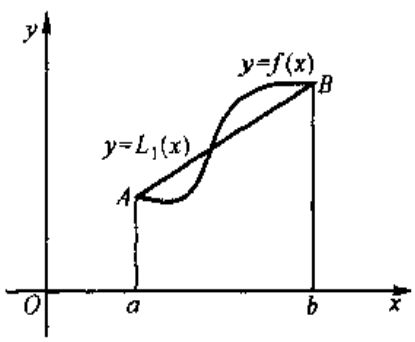

图 6-1

$$R_1(f)=\int_a^b\frac{f''(\xi)}{2}(x-a)(x-b)\mathrm{d}x,\xi\in(a,b) \tag{6.1.7}$$

6.1.3 求积公式的代数精确度

当被积函数 $f(x)\in H_n$ 即 f 为次数不超过 n 的代数多项式时，$f^{(n+1)}(\xi)=0$，故由(6.1.5)有 $R_n(f)=0$，它表明插值求积公式(6.1.3)精确成立. 对一般机械求积公式(6.1.3)，同样可以根据公式是否对 m 次多项式精确成立作为确定公式(6.1.3)中系数 A_i 及节点 $x_i(i=0,1,\cdots,n)$ 的一种方法. 在此先给出定义.

定义 1.1 一个求积公式(6.1.3)若对 $f(x)=1,x,\cdots,x^m$ 精确成立，而对 $f(x)=x^{m+1}$ 不精确成立，则称求积公式(6.1.3)具有 m 次代数精确度.

根据定义，当 $f(x)=1,x,\cdots,x^m$ 时公式(6.1.3)精确成立，故有等式

$$\left.\begin{aligned}
&\text{当 } f(x)=1 \text{ 时，} \sum_{i=0}^{n} A_i = \int_a^b \mathrm{d}x = b-a \\
&\text{当 } f(x)=x \text{ 时，} \sum_{i=0}^{n} A_i x_i = \int_a^b x\,\mathrm{d}x = \frac{1}{2}(b^2-a^2) \\
&\qquad\cdots\cdots \\
&\text{当 } f(x)=x^m \text{ 时，} \sum_{i=0}^{n} A_i x_i^m = \int_a^b x^m\,\mathrm{d}x = \frac{1}{m+1}(b^{m+1}-a^{m+1})
\end{aligned}\right\} \tag{6.1.8}$$

而
$$\sum_{i=0}^{n} A_i x_i^{m+1} \neq \int_a^b x^{m+1}\,\mathrm{d}x$$

(6.1.8)是关于系数 A_i 及节点 $x_i(i=0,1,\cdots,n)$的方程组，当节点 $x_0,x_1,\cdots,x_n$ 给定时，(6.1.8)取 $m=n$ 就是关于系数$A_0,A_1,\cdots,A_n$ 的线性方程组，求此方程组就可求得求积系数 $A_i(i=0,1,\cdots,n)$.

例如 $n=1$，取 $x_0=a,x_1=b$，求积公式为

$$I(f)=\int_a^b f(x)\mathrm{d}x \approx A_0 f(a)+A_1 f(b)$$

在(6.1.8)中令 $m=1$，可得

$$\begin{cases} A_0+A_1=b-a \\ A_0 a+A_1 b=\dfrac{1}{2}(b^2-a^2) \end{cases}$$

解得
$$A_0=A_1=\frac{1}{2}(b-a)$$

它就是梯形公式(6.1.6)的系数，它与用公式(6.1.4)算出的结果完全一样.对梯形公式(6.1.6)，当 $f(x)=x^2$ 时

$$A_0 a^2+A_1 b^2=\frac{b-a}{2}(a^2+b^2)\neq\int_a^b x^2\mathrm{d}x=\frac{1}{3}(b^3-a^3)$$

故求积公式(6.1.6)的代数精确度为一次.

对于具有$(n+1)$个节点的插值求积公式(6.1.3)，当 $f(x)=1,x,\cdots,x^n$ 时 $R_n(f)=0$，故公式精确成立，它至少有 n 次代数精确度.反之，若求积公式(6.1.3)至少有 n 次代数精确度，则它是插值求积公式，即(6.1.3)的求积系数 $A_i(i=0,1,\cdots,n)$一定可用(6.1.4)求出.实际上，此时对 $\forall f(x)\in H_n$ 求积公式(6.1.3)精确成立，若取 $f(x)$为插值基函数，即

$$f(x)=l_k(x)=\frac{\omega_{n+1}(x)}{(x-x_k)\omega'_{n+1}(x_k)}\in H_n$$

由(6.1.3)精确成立，可得

$$\int_a^b l_k(x)\mathrm{d}x=\int_a^b\frac{\omega_{n+1}(x)}{(x-x_k)\omega'_{n+1}(x_k)}\mathrm{d}x=\sum_{i=0}^{n}A_i l_k(x_i)=A_k,\quad k=0,1,\cdots,n$$

这就是(6.1.4)得到的插值求积公式系数.

定理 1.1 求积公式(6.1.3)是插值求积公式的充分必要条件是(6.1.3)至少具有 n 次代数精确度.

定理表明直接利用代数精确度概念,由(6.1.8)可求得插值求积公式.更一般地,含有被积函数 $f(x)$的导数 $f'(x_i)$的求积公式也同样可用代数精确度定义建立.如下例所示.

例 6.1 求积公式 $\int_0^1 f(x)\mathrm{d}x \approx A_0 f(0) + A_1 f(1) + B_0 f'(0)$,已知其余项表达式为 $R(f) = kf'''(\xi)$,$\xi \in (0,1)$.试确定系数 A_0, A_1 及 B_0,使该求积公式具有尽可能高的代数精确度,并给出代数精确度的次数及求积公式余项.

解 本题虽用到 $f'(0)$的值,但仍可用代数精确度定义确定参数 A_0, A_1 及 B_0.令 $f(x) = 1, x, x^2$,分别代入求积公式;令公式两端相等,则

当 $f(x) = 1$, $A_0 + A_1 = 1$

当 $f(x) = x$, $A_1 + B_0 = \dfrac{1}{2}$

当 $f(x) = x^2$, $A_1 = \dfrac{1}{3}$

解得 $A_0 = \dfrac{2}{3}, A_1 = \dfrac{1}{3}, B_0 = \dfrac{1}{6}$,于是有

$$\int_0^1 f(x)\mathrm{d}x \approx \frac{2}{3} f(0) + \frac{1}{3} f(1) + \frac{1}{6} f'(0)$$

再令 $f(x) = x^3$,此时 $\int_0^1 x^3 \mathrm{d}x = \dfrac{1}{4}$,而上式右端为 $\dfrac{1}{3}$,两端不等,则求积公式对 $f(x) = x^3$ 不精确成立,故它的代数精确度为二次.

为求余项可将 $f(x) = x^3$ 代入求积公式

$$\int_0^1 f(x)\mathrm{d}x = \frac{2}{3} f(0) + \frac{1}{3} f(1) + \frac{1}{6} f'(0) + kf'''(\xi), \xi \in (0,1)$$

当 $f(x) = x^3, f'(x) = 3x^2, f''(x) = 6x, f'''(x) = 6$,代入上式得

$$\frac{1}{4} = \int_0^1 x^3 \mathrm{d}x = \frac{1}{3} + 6k, \text{即 } 6k = \frac{1}{4} - \frac{1}{3} = -\frac{1}{12}, k = -\frac{1}{72}$$

所以余项 $R(f) = -\dfrac{1}{72} f'''(\xi), \xi \in (0,1)$.

6.1.4 求积公式的收敛性与稳定性

定义 1.2 若 $\lim\limits_{n\to\infty} \sum\limits_{i=0}^{n} [A_i f(x_i)] = \int_a^b f(x)\mathrm{d}x$,则称求积公式(6.1.3)是收敛的.

定义中 $n \to \infty$ 包含了 $\max\limits_{0 \leqslant i \leqslant n-1} (x_{i+1} - x_i) \to 0$ 通常都要求用于计算积分(6.1.1)的求积公式(6.1.3)是收敛的.本章后而给出的求积公式都必须先证明其收敛性.

稳定性是研究计算和式

$$I_n(f)=\sum_{i=0}^{n}[A_i f(x_i)]$$

当$f(x_i)$有误差δ_i时，$I_n(f)$的误差是否增长．现设$f(x_i)\approx\tilde{f}_i$，误差为$\delta_i=|f(x_i)-\tilde{f}_i|$$(i=0,1,\cdots,n)$．

定义 1.3 对任给$\varepsilon>0$，$\exists\delta>0$，只要$\delta_i=|f(x_i)-\tilde{f}_i|\leqslant\delta(i=0,1,\cdots,n)$，就有

$$|I_n(f)-I_n(\tilde{f})|\leqslant\varepsilon$$

则称求积公式(6.1.3)是(数值)稳定的．

定义表明只要被积函数$f(x)$的误差δ_i充分小，积分和式$I_n(f)$的误差限就可任意小，则(6.1.3)就是数值稳定的．

定理 1.2 若求积公式(6.1.3)的系数$A_i>0(i=0,1,\cdots,n)$则求积公式(6.1.3)是稳定的．

证明 由于$A_i>0(i=0,1,\cdots,n)$，$|f(x_i)-\tilde{f}_i|\leqslant\delta(i=0,1,\cdots,n)$，故有

$$|I_n(f)-I_n(\tilde{f})|=|\sum_{i=0}^{n}[A_i(f(x_i)-\tilde{f}_i)]|\leqslant\delta\sum_{i=0}^{n}A_i=\delta(b-a)$$

于是对$\forall\varepsilon>0$，$\exists\delta=\frac{\varepsilon}{b-a}$，只要$\delta_i=|f(x_i)-\tilde{f}_i|\leqslant\delta$，就有

$$|I_n(f)-I_n(\tilde{f})|\leqslant\delta(b-a)\leqslant\varepsilon$$

故求积公式(6.1.3)是稳定的．

6.2 梯形公式与 Simpson 求积公式

6.2.1 Newton－Cotes 公式与 Simpson 公式

在插值求积公式中，若求积节点$x_k=a+kh(k=0,1,\cdots,n)$，$h=\frac{b-a}{n}$，此时可将求积公式写成

$$I(f)=\int_a^b f(x)\mathrm{d}x\approx(b-a)\sum_{k=0}^{n}[C_k^{(n)}f(x_k)]\tag{6.2.1}$$

称为 Newton－Cotes 求积公式，其中系数$C_k^{(n)}$称作 Cotes 系数．按(6.1.4)式引入变换$x=a+th$，则有

$$C_k^{(n)}=\frac{h}{b-a}\int_0^n\prod_{\substack{j=0\\j\neq k}}^{n}\frac{t-j}{k-j}\mathrm{d}t=\frac{(-1)^{n-k}}{n\cdot k!\,(n-k)!}\int_0^n\prod_{\substack{j=0\\j\neq k}}^{n}(t-j)\mathrm{d}t\tag{6.2.2}$$

由于是多项式积分，计算不会有困难．例如$n=1$时，$C_0^{(1)}=\int_0^1(t-1)\mathrm{d}t=1/2$，$C_1^{(1)}=\int_0^1 t\mathrm{d}t=1/2$．这时求积公式就是我们熟悉的梯形公式(6.1.6)．$n=1\sim8$时的系数见表 6－1．从表中看到$n=8$时出现负数，稳定性没保证，所以一般只用$n\leqslant4$的公式．

表 6-1

n	$C_k^{(n)}$								
1	$\frac{1}{2}$	$\frac{1}{2}$							
2	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$						
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$					
4	$\frac{7}{90}$	$\frac{16}{45}$	$\frac{2}{15}$	$\frac{16}{45}$	$\frac{7}{90}$				
5	$\frac{19}{288}$	$\frac{25}{96}$	$\frac{25}{144}$	$\frac{25}{144}$	$\frac{25}{96}$	$\frac{19}{288}$			
6	$\frac{41}{840}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{34}{105}$	$\frac{9}{280}$	$\frac{9}{35}$	$\frac{41}{840}$		
7	$\frac{751}{17\,280}$	$\frac{3\,577}{17\,280}$	$\frac{1\,323}{17\,280}$	$\frac{2\,989}{17\,280}$	$\frac{2\,989}{17\,280}$	$\frac{1\,323}{17\,280}$	$\frac{3\,577}{17\,280}$	$\frac{751}{17\,280}$	
8	$\frac{989}{28\,350}$	$\frac{5\,888}{28\,350}$	$\frac{-928}{28\,350}$	$\frac{10\,496}{28\,350}$	$\frac{-4\,540}{28\,350}$	$\frac{10\,496}{28\,350}$	$\frac{-928}{28\,350}$	$\frac{5\,888}{28\,350}$	$\frac{989}{28\,350}$

当 $n=2$ 时，由(6.2.2)可得 $C_0^{(2)}=C_2^{(2)}=1/6, C_1^{(2)}=4/6$，求积公式为

$$I(f)=\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right] \tag{6.2.3}$$

称为 Simpson 求积公式.

对梯形公式(6.1.6)，已知它的代数精确度为一次，且它的余项已由(6.1.7)给出，若记 $T_1=\frac{b-a}{2}[f(a)+f(b)]$，则

$$R_1(f)=I(f)-T_1=\frac{1}{2}\int_a^b f''(\xi)(x-a)(x-b)\mathrm{d}x$$

由于 $(x-a)(x-b)\leqslant 0$，在 $[a,b]$ 上不变号，由积分中值定理得知 $\exists\,\eta\in(a,b)$，使

$$R_1(f)=\frac{f''(\eta)}{2}\int_a^b(x-a)(x-b)\mathrm{d}x=-\frac{f''(\eta)}{12}(b-a)^3,\ \eta\in(a,b) \tag{6.2.4}$$

这就是梯形公式(6.1.6)的截断误差.

对 Simpson 公式(6.2.3)可以证明它的代数精确度是三次，根据定理 1.1，显然(6.2.3)对 $f(x)=1,x,x^2$ 精确成立，再对 $f(x)=x^3$，左端为 $\int_a^b x^3\mathrm{d}x=\frac{1}{4}(b^4-a^3)$，右端为

$$\frac{b-a}{6}\left[a^3+4\left(\frac{a+b}{2}\right)+b^3\right]=\frac{b-a}{6}\frac{3}{2}[a^3+a^2b+ab^2+b^3]=\frac{1}{4}(b^4-a^4)$$

故(6.2.3)对 $f(x)=x^3$ 也精确成立. 而对 $f(x)=x^4$，公式(6.2.3)不精确成立. 故求积公式(6.2.3)的代数精确度是三次，即(6.2.3)对任何 $P(x)\in H_3$ 精确成立. 若令 $P(x)$ 满足条件

$$P(x_i)=f(x_i)(i=0,1,2),\quad P'(x_1)=f'(x_1)$$

这里 $x_0=a, x_1=\frac{a+b}{2}, x_2=b$，于是由(6.2.3)有

$$\int_a^b P(x)\mathrm{d}x = \frac{b-a}{6}\left[P(a)+4P\left(\frac{a+b}{2}\right)+P(b)\right]$$
$$= \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]$$

根据上章例 5.6 中(5.5.12)式有

$$f(x)-P(x)=\frac{f^{(4)}(\xi)}{4!}(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b), \xi\in(a,b)$$

上式两边积分，并记 $S_1=\frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]$，则得

$$R_2(f)=\int_a^b f(x)\mathrm{d}x - S_1$$
$$= \frac{1}{4!}\int_a^b f^{(4)}(\xi)(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x, \xi\in(a,b)$$

由于在区间$[a,b]$上$(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\leqslant 0$不变号，故由积分中值定理得

$$R_2(f)=\frac{1}{4!}f^{(4)}(\eta)\int_a^b (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x$$
$$= \frac{f^{(4)}(\eta)}{4!}\left[-\frac{(b-a)^5}{120}\right], \eta\in(a,b)$$

于是有

$$R_2(f)=-\frac{b-a}{180}\left(\frac{b-a}{2}\right)^4 f^{(4)}(\eta), \eta\in(a,b) \tag{6.2.5}$$

这就是 Simpson 求积公式(6.2.3)的余项，即截断误差.

6.2.2 复合梯形公式与复合 Simpson 公式

直接用梯形公式(6.1.6)及 Simpson 公式(6.2.3)计算积分 $I(f)$误差较大，为达到精度要求，通常可将$[a,b]$分为 n 个小区间，在小区间上应用梯形公式及 Simpson 公式即可达到要求. 为此取分点 $x_k=a+kh\ (k=0,1,\cdots,n), h=\frac{b-a}{n}$，在每个小区间$[x_k,x_{k+1}]$上用梯形公式(6.1.6)，则得

$$I(f)=\int_a^b f(x)\mathrm{d}x=\sum_{k=0}^{n-1}\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x\approx\sum_{k=0}^{n-1}\frac{h}{2}[f(x_k)+f(x_{k+1})]=T_n$$

或

$$I(f)\approx T_n=\frac{h}{2}\left[f(a)+2\sum_{k=1}^{n-1}f(x_k)+f(b)\right] \tag{6.2.6}$$

称为复合梯形公式. 根据定积分定义可知

$$\lim_{\substack{n\to\infty\\ h\to 0}} T_n=\lim_{h\to 0}\frac{1}{2}\left[\sum_{k=0}^{n-1}f(x_k)h+\sum_{k=1}^{n}f(x_k)h\right]=\int_a^b f(x)\mathrm{d}x=I(f)$$

故复合梯形公式(6.2.6)是收敛的,且(6.2.6)的求积系数 $A_k>0(k=0,1,\cdots,n)$,故它也是稳定的.

(6.2.6)的截断误差可由(6.2.4)得到

$$R_n(f)=I(f)-T_n=\sum_{k=0}^{n-1}\left[-\frac{h^3}{12}f''(\eta_k)\right]$$

$$=-\frac{h^2}{12}(b-a)\frac{1}{n}\sum_{k=0}^{n-1}f''(\eta_k),\eta_k\in(x_k,x_{k+1})$$

若 $f\in C^2[a,b]$,根据连续函数性质,$\exists\,\eta\in(a,b)$,使

$$\frac{1}{n}\sum_{k=0}^{n-1}f''(\eta_k)=f''(\eta)$$

于是得

$$R_n(f)=-\frac{b-a}{12}h^2f''(\eta),\quad \eta\in(a,b) \tag{6.2.7}$$

它表明复合梯形公式(6.2.6)的截断误差阶为 $R_n(f)=O(h^2)$

如果在每个小区间$[x_k,x_{k+1}]$上使用 Simpson 公式(6.2.3),则得

$$I(f)=\int_a^b f(x)\mathrm{d}x=\sum_{k=0}^{n-1}\int_{x_k}^{x_{k+1}}f(x)\mathrm{d}x$$

$$\approx\frac{h}{6}\left[f(a)+4\sum_{k=0}^{n-1}f(x_{k+1/2})+2\sum_{k=1}^{n-1}f(x_k)+f(b)\right]=S_n \tag{6.2.8}$$

称为复合 Simpson 公式,它的余项由(6.2.5)可得

$$R_n(f)=I(f)-S_n=-\frac{h}{180}\left(\frac{h}{2}\right)^4\sum_{k=0}^{n-1}f^{(4)}(\eta_k),\eta_k\in(x_k,x_{k+1})$$

$$=-\frac{b-a}{180}\left(\frac{h}{2}\right)^4f^{(4)}(\eta),\eta\in(a,b)$$

即

$$R_n(f)=-\frac{b-a}{2\,880}h^4f^{(4)}(\eta),\eta\in(a,b) \tag{6.2.9}$$

它表明 $R_n(f)=O(h^4)$.此外,还可证明

$$\lim_{n\to\infty}S_n=\int_a^b f(x)\mathrm{d}x$$

故复合 Simpson 公式(6.2.8)是收敛的,并且 $A_k>0(k=0,1,\cdots,n)$,故公式也是稳定的.

例 6.2 用 $n=8$ 的复合梯形公式及 $n=4$ 的复合 Simpson 公式,计算积分 $I=\int_0^1\frac{\sin x}{x}\mathrm{d}x$,并估计误差.

解 只要将区间[0,1]分为 8 等分,用公式(6.2.6)时取 $n=8,h=0.125$,对复合 Simpson 公式取 $n=4,h=0.25$.计算各分点 $x_k(k=0,1,\cdots,8)$的函数值 $f(x_k)=\frac{\sin x_k}{x_k}$.

k	x_k	$f(x_k)$	k	x_k	$f(x_k)$
0	0	1.000 000 0	5	0.625	0.936 155 6
1	0.125	0.997 397 8	6	0.75	0.908 851 6
2	0.25	0.989 615 8	7	0.875	0.877 192 5
3	0.375	0.976 726 7	8	1	0.841 470 9
4	0.5	0.958 851 0			

由公式(6.2.6)及(6.2.8)得

$$T_8 = \frac{1}{16}\left[f(0) + 2\sum_{k=1}^{7} f(x_k) + f(1)\right] = 0.945\ 690\ 9$$

$$\begin{aligned} S_4 = &\frac{1}{24}\times[f(0)+4\times(f(0.125)+f(0.375)+f(0.625)+f(0.875))+2\times \\ &(f(0.25)+f(0.5)+f(0.75))+f(1)] \\ =&0.946\ 083\ 2 \end{aligned}$$

为了估计误差,要求 $f(x)=\dfrac{\sin x}{x}$的高阶导数,由于

$$f(x)=\frac{\sin x}{x}=\int_0^1 \cos(xt)\mathrm{d}t$$

所以

$$f^{(k)}(x)=\int_0^1 \frac{\mathrm{d}^k}{\mathrm{d}x^k}\cos(xt)\mathrm{d}t=\int_0^1 t^k\cos\left(xt+\frac{k\pi}{2}\right)\mathrm{d}t$$

故

$$|f^{(k)}(x)|\leqslant\int_0^1 t^k\left|\cos\left(xt+\frac{k\pi}{2}\right)\right|\mathrm{d}t\leqslant\int_0^1 t^k\mathrm{d}t=\frac{1}{k+1}$$

由(6.2.7)得

$$\begin{aligned} |R_8(f)| = |I(f)-T_8| &= \left|-\frac{1}{12}h^2 f''(\eta)\right| \\ &\leqslant\frac{1}{12}\left(\frac{1}{8}\right)^2\frac{1}{3}=0.000\ 434 \end{aligned}$$

对 Simpson 公式,由(6.2.9)得

$$|R_4(f)|=|I(f)-S_4|\leqslant\frac{1}{2\ 880}\left(\frac{1}{4}\right)^4\frac{1}{5}=0.271\times10^{-6}$$

例 6.3 计算积分 $I=\int_0^1 \mathrm{e}^x\mathrm{d}x$,若用复合梯形公式,问区间[0,1]应分多少等分才能使截断误差不超过$\dfrac{1}{2}\times10^{-5}$?若改用复合 Simpson 公式,要达到同样精确度,区间[0,1]应分多少等分?

解 本题只要根据 T_n 及 S_n 的余项表达式(6.2.7)及(6.2.9)即可求出其截断误差应满足的精度.由于 $f(x)=\mathrm{e}^x$,$f''(x)=\mathrm{e}^x$,$f^{(4)}(x)=\mathrm{e}^x$,$b-a=1$,对复合梯形公式

$$|R_n(f)|=\left|-\frac{b-a}{12}h^2 f''(\eta)\right|\leqslant\frac{1}{12}\left(\frac{1}{n}\right)^2\mathrm{e}\leqslant\frac{1}{2}\times10^{-5}$$

即 $n^2 \geqslant \frac{1}{6}\mathrm{e} \times 10^5$，$n \geqslant 212.85$. 取 $n=213$，即将区间$[0,1]$分为 213 等分时，用复合梯形公式计算误差不超过$\frac{1}{2} \times 10^{-5}$. 而用复合 Simpson 公式，则要求

$$\begin{aligned} |R_n(f)| &= \left| -\frac{b-a}{2\,880} h^4 f^{(4)}(\eta) \right| \\ &\leqslant \frac{1}{2\,880}\left(\frac{1}{n}\right)^4 \mathrm{e} \leqslant \frac{1}{2} \times 10^{-5} \end{aligned}$$

即 $n^4 \geqslant \frac{\mathrm{e}}{144} \times 10^4$，$n \geqslant 3.706\,6$. 取 $n=4$，即将区间分为 8 等分，用 $n=4$ 的复合 Simpson 公式即可达到精确度$\frac{1}{2} \times 10^{-5}$.

6.3 外推原理与 Romberg 求积

6.3.1 复合梯形公式递推化与节点加密

在计算机上用等距节点求积公式时，若精度不够可以逐步加密节点. 设将区间$[a,b]$分为 n 等分，节点 $x_k = a + kh$，$h = \frac{b-a}{n}$，在区间$[x_k, x_{k+1}]$上梯形公式为

$$\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x \approx \frac{h}{2}[f(x_k) + f(x_{k+1})]$$

若节点加密一倍，区间长为$\frac{b-a}{2n}$，记$[x_k, x_{k+1}]$中点为 $x_{k+1/2}$在同一区间$[x_k, x_{k+1}]$上的复合梯形公式为

$$\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x \approx \frac{h}{4}[f(x_k) + 2f(x_{k+1/2}) + f(x_{k+1})]$$

于是

$$T_{2n} = \frac{h}{4}\sum_{k=0}^{n-1}[f(x_k) + f(x_{k+1})] + \frac{h}{2}\sum_{k=0}^{n-1} f(x_{k+1/2}) = \frac{1}{2}T_n + \frac{h}{2}\sum_{k=0}^{n-1} f(x_{k+1/2}) \qquad (6.3.1)$$

它表明 T_{2n}是在$\frac{1}{2}T_n$ 的基础上再加新节点 $x_{k+1/2}$的函数值之和乘新区间长$\frac{h}{2}$，而不必用(6.2.6)重新计算 T_{2n}，这时有误差估计式

$$I(f) - T_n = -\frac{b-a}{12}h^2 f''(\eta), \eta \in (a,b)$$

$$I(f) - T_{2n} = -\frac{b-a}{12}\left(\frac{h}{2}\right)^2 f''(\bar{\eta}), \bar{\eta} \in (a,b)$$

若 $f''(\eta) \approx f''(\bar{\eta})$，则得

$$\frac{I(f)-T_n}{I(f)-T_{2n}}\approx 4 \quad 或 \quad I(f)\approx T_{2n}+\frac{1}{3}(T_{2n}-T_n) \tag{6.3.2}$$

它表明用 $T_{2n}\approx I(f)$,其误差近似$\frac{1}{3}(T_{2n}-T_n)$.这也是在计算机上估计梯形公式误差的近似表达式.若$\frac{1}{3}|T_{2n}-T_n|\leqslant\varepsilon$(给定精度),则 $I(f)\approx T_{2n}$.

若在区间$[a,b]$中做 $2n$ 等分时,在$[x_k,x_{k+1}]$上用 Simpson 公式计算,则由(6.2.8)可知

$$\begin{aligned}S_n &= \frac{h}{6}\sum_{k=0}^{n-1}[f(x_k)+f(x_{k+1})]+\frac{4}{6}h\sum_{k=0}^{n-1}f(x_{k+1/2})\\ &= \frac{1}{3}T_n+\frac{2}{3}h\sum_{k=0}^{n-1}f(x_{k+1/2})=\frac{4}{3}T_{2n}-\frac{1}{3}T_n\end{aligned}$$

它恰好是(6.3.2)中 $I(f)$的近似值,即

$$I(f)\approx\frac{4}{3}T_{2n}-\frac{1}{3}T_n=S_n$$

它表明用(6.3.2)计算 $I(f)$,其精度已由 $I(f)-T_{2n}=O(h^2)$提高到 $I(f)-S_n=O(h^4)$.如果再将区间分半,使$[x_k,x_{k+1}]$分为 4 个小区间,长度为$\frac{h}{4}$,则可由(6.3.1)计算出 T_{4n}及 S_{2n},利用复合公式余项(6.2.9)得

$$I(f)-S_n=-\frac{b-a}{180}\left(\frac{h}{2}\right)^4 f^{(4)}(\eta),\eta\in(a,b)$$

$$I(f)-S_{2n}=-\frac{b-a}{180}\left(\frac{h}{4}\right)^4 f^{(4)}(\bar{\eta}),\bar{\eta}\in(a,b)$$

如果 $f^{(4)}(\eta)\approx f^{(4)}(\bar{\eta})$,则有

$$\frac{I(f)-S_n}{I(f)-S_{2n}}\approx 16 \quad 或 \quad I(f)\approx S_{2n}+\frac{1}{15}(S_{2n}-S_n) \tag{6.3.3}$$

从而有复合 Simpson 公式的误差估计

$$I(f)-S_{2n}\approx\frac{1}{15}(S_{2n}-S_n)$$

如果用(6.3.3)近似 $I(f)$,即

$$I(f)\approx\frac{16}{15}S_{2n}-\frac{1}{15}S_n \tag{6.3.4}$$

则精度可达到 $O(h^6)$.类似做法还可继续下去.这样对区间$[a,b]$逐次分半,利用公式(6.3.1)逐次递推.再由(6.3.2),(6.3.3)逐次构造出精度愈来愈高的计算积分 $I(f)$的公式,这就是 Romberg 求积的基本思想.

6.3.2 外推法与 Romberg 求积公式

仍从梯形公式出发,区间$[a,b]$中的节点如前所述.此时复合梯形公式可表示为

$$I(f)\approx\frac{h}{2}\sum_{k=0}^{n-1}[f(x_k)+f(x_{k+1})]=T_n=T(h),h=\frac{b-a}{n}$$

当$[a,b]$分为$2n$等分,区间长变为$\frac{h}{2}$时,记$T_{2n}=T\left(\frac{h}{2}\right)$,由于

$$T(h)=I(f)+\frac{b-a}{12}h^2f''(\eta),\eta\in(a,b)$$

$$\lim_{h\to 0}T(h)=T(0)=I(f)$$

此处将$T_n=T(h)$看作h的函数,将$T(h)$按h的幂展开,可得到以下结果.

定理 3.1 设f在$[a,b]$上的各阶导数存在,则复合梯形公式$T(h)$可展成

$$T(h)=I(f)+\alpha_1h^2+\alpha_2h^4+\cdots+\alpha_lh^{2l}+O(h^{2l+2}) \tag{6.3.5}$$

其中$\alpha_1,\alpha_2,\cdots,\alpha_l$为不依赖$h$的常数.

定理证明见[2].由(6.3.5),显然有$T(h)-I(f)=O(h^2)$.在(6.3.5)中若用$\frac{h}{2}$代替h则得

$$T\left(\frac{h}{2}\right)=I(f)+\alpha_1\left(\frac{h}{2}\right)^2+\alpha_2\left(\frac{h}{2}\right)^4+\cdots+\alpha_l\left(\frac{h}{2}\right)^{2l}+O(h^{2l+2}) \tag{6.3.6}$$

用4乘以(6.3.6)减去(6.3.5)除以3,则得

$$\frac{4T\left(\frac{h}{2}\right)-T(h)}{3}=I(f)+\beta_1\left(\frac{h}{2}\right)^4+\beta_2\left(\frac{h}{2}\right)^6+\cdots$$

若记

$$T_0(h)=T(h),T_1\left(\frac{h}{2}\right)=\frac{4T_0\left(\frac{h}{2}\right)-T_0(h)}{3} \tag{6.3.7}$$

显然

$$T_1(h)=I(f)+\beta_1h^4+\beta_2h^6+\cdots \tag{6.3.8}$$

$T_1\left(\frac{h}{2}\right)\approx I(f)$具有精度$O(h^4)$.实际上$T_1\left(\frac{h}{2}\right),T_1\left(\frac{h}{4}\right),\cdots$就是复合 Simpson 公式中的$S_n$,$S_{2n},\cdots$.为提高精度,可由$T_1(h)$及$T_1\left(\frac{h}{2}\right)$中消去$h^4$,得

$$T_2\left(\frac{h}{2}\right)=\frac{16T_1\left(\frac{h}{2}\right)-T_1(h)}{4^2-1}=I(f)+\gamma_1h^6+\gamma_2h^8+\cdots$$

用$T_2(h/2)\approx I(f)$精度为$O(h^6)$,如此逐次做下去,可得到

$$T_m\left(\frac{h}{2}\right)=\frac{4^mT_{m-1}\left(\frac{h}{2}\right)-T_{m-1}(h)}{4^m-1},m=1,2,\cdots \tag{6.3.9}$$

用$T_m\left(\frac{h}{2}\right)\approx I(f)$时,精度为$O(h^{2(m+1)})$,这种将步长$h$逐次减半,使$T_0(h),T_1(h),\cdots$逼近$I(f)$,以便精度逐次提高的方法称为外推法,它对于可展成$h$的幂级数的计算公式的加速收敛是很有效的,这里只将外推法用于计算积分$I(f)=\int_a^bf(x)\mathrm{d}x$.

下面若用 $T_0^{(k)}$ 表示将区间$[a,b]$二分 k 次得到的复合梯形公式,此时$[a,b]$分为 2^k 等分,步长 $h=\frac{b-a}{2^k}$,当 $k=0,1,\cdots$逐次得到 $T_0^{(0)},T_0^{(1)},T_0^{(2)},\cdots$即为 $n=2^k$ 等分的复合梯形公式,加速一次得序列$\{T_1^{(k)}\}$即为 Simpson 公式序列.加速 m 次则得$\{T_m^{(k)}\}$,由(6.3.9)可将它表示为

$$T_m^{(k)}=\frac{4^m T_{m-1}^{(k)}-T_{m-1}^{(k-1)}}{4^m-1},k=1,2,\cdots;m=1,2,\cdots,k \tag{6.3.10}$$

称为 Romberg 求积公式.计算从 $k=0$,即 $h=b-a$ 出发记 $T_0^{(k)}=T(h),h=\frac{b-a}{2^k}$,逐次二分得到 T 表(见表 6-2).

当 k 增加时,先由(6.3.1)根据 $T_0^{(k-1)}$ 算出 $T_0^{(k)}$,再由(6.3.10)对 $m=1,2,\cdots,k$ 计算 $T_m^{(k)}$.当 f 充分光滑时可证明

$$\lim_{k\to\infty}T_m^{(k)}=I(f),m=0,1,\cdots,k \quad (T\text{ 表任一列})$$

$$\lim_{k\to\infty}T_k^{(k)}=I(f) \qquad (T\text{ 表对角线})$$

计算到$|T_k^{(k)}-T_{k-1}^{(k-1)}|\leqslant\varepsilon$(精度要求)为止.

表 6-2 T 表

k	n	h	$T_0^{(k)}$	$T_1^{(k)}$	$T_2^{(k)}$	$T_3^{(k)}$	$T_4^{(k)}$
0	1	$b-a$	$T_0^{(0)}$				
1	2	$\frac{b-a}{2}$	$T_0^{(1)}$	$T_1^{(1)}$			
2	4	$\frac{b-a}{4}$	$T_0^{(2)}$	$T_1^{(2)}$	$T_2^{(2)}$		
3	8	$\frac{b-a}{8}$	$T_0^{(3)}$	$T_1^{(3)}$	$T_2^{(3)}$	$T_3^{(3)}$	
4	16	$\frac{b-a}{16}$	$T_0^{(4)}$	$T_1^{(4)}$	$T_2^{(4)}$	$T_3^{(4)}$	$T_4^{(4)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

例 6.4 用 Romberg 求积公式求 $I(f)=\int_0^1\frac{\sin x}{x}\mathrm{d}x$ 的近似值,使其具有 6 位有效数字.

解 本题直接用梯形递推公式(6.3.1)及 Romberg 求积公式(6.3.10),按 T 表依次计算

$$T_0^{(0)}=\frac{1}{2}[f(0)+f(1)]=0.920\,735\,5$$

$$T_0^{(1)}=\frac{1}{2}T_0^{(0)}+\frac{1}{2}f(0.5)=0.939\,793\,3$$

$$T_1^{(1)}=\frac{1}{3}(4T_0^{(1)}-T_0^{(0)})=0.946\,145\,9$$

其余计算结果见 T 表.

k	$T_0^{(k)}$	$T_1^{(k)}$	$T_2^{(k)}$	$T_3^{(k)}$
0	0.920 735 5			
1	0.939 793 3	0.946 145 9		
2	0.944 513 5	0.946 086 9	0.946 083 0	
3	0.945 690 9	0.946 083 3	0.946 083 1	0.946 083 1

由于 $|T_2^{(2)} - T_3^{(3)}| = 0.000\ 000\ 1 < \frac{1}{2} \times 10^{-6}$，故计算停止，$I \approx 0.946\ 083\ 1$ 即为所求.

例 6.5 证明等式

$$n \sin \frac{\pi}{n} = \pi - \frac{\pi^3}{3!\ n^2} + \frac{\pi^5}{5!\ n^4} - \cdots$$

试依据 $n\sin\frac{\pi}{n}(n = 3,6,12)$的值，用外推算法求 π 的近似值.

解 本题可利用 Taylor 展开式用外推原理求 π 的近似值. 可令 $f(n) = n\sin\frac{\pi}{n}$. 由 Taylor 公式展开得

$$\begin{aligned} n\sin\frac{\pi}{n} &= n\left[\frac{\pi}{n} - \frac{1}{3!}\left(\frac{\pi}{n}\right)^3 + \frac{1}{5!}\left(\frac{\pi}{n}\right)^5 - \frac{1}{7!}\left(\frac{\pi}{n}\right)^7 + \cdots\right] \\ &= \pi - \frac{\pi^3}{3!n^2} + \frac{\pi^5}{5!n^4} - \frac{\pi^7}{7!n^6} + \cdots \\ &= \pi\left[1 - \frac{1}{3!}\left(\frac{\pi}{n}\right)^2 + \frac{1}{5!}\left(\frac{\pi}{n}\right)^4 - \frac{1}{7!}\left(\frac{\pi}{n}\right)^6 + \cdots\right] \end{aligned}$$

若记 $T_n^{(0)} = n\sin\frac{\pi}{n} \approx \pi$，则其误差为 $O\left(\left(\frac{\pi}{n}\right)^2\right)$.

由外推法

$$T_n^{(1)} = \frac{1}{3}(4T_{2n}^{(0)} - T_n^{(0)}) \approx \pi, \text{其误差为 } O\left(\left(\frac{\pi}{n}\right)^4\right)$$

$$T_n^{(2)} = \frac{1}{15}(16T_{2n}^{(1)} - T_n^{(1)}) \approx \pi, \text{其误差为 } O\left(\left(\frac{\pi}{n}\right)^6\right)$$

根据以上公式计算结果如下表所示.

n	$T_n^{(0)} = n\sin\frac{\pi}{n}$	$T_n^{(1)}$	$T_n^{(2)}$
3	2.598 076		
6	3.000 000	3.133 975	
12	3.105 829	3.141 105	3.141 580

$\pi = 3.141\ 58$ 即为所求.

6.4 Gauss 型求积公式

6.4.1 最高代数精确度求积公式

考虑带权积分

$$I(f)=\int_a^b \rho(x)f(x)\mathrm{d}x$$

其中 $\rho(x)$ 为权函数，它的求积公式为

$$I(f)=\int_a^b \rho(x)f(x)\mathrm{d}x\approx\sum_{k=0}^{n}A_kf(x_k)=I_n(f) \tag{6.4.1}$$

其中 A_k 为求积系数，不依赖于 f，x_k 为求积节点，在式(6.4.1)中 x_k，$A_k(k=0,1,\cdots,n)$ 作为待定参数，可选择使(6.4.1)对 $f(x)=x^m(m=0,1,\cdots,2n+1)$ 精确成立，从而得到关于 x_k，A_k 的 $(2n+2)$ 个参数的非线性方程组，由式(6.4.1)得

$$\sum_{k=0}^{n}A_kx_k^m=\int_a^b \rho(x)x^m\mathrm{d}x,\ m=0,1,\cdots,2n+1 \tag{6.4.2}$$

例如，当 $n=0$，$\rho(x)=1$ 时，求积公式为

$$\int_a^b f(x)\mathrm{d}x\approx A_0f(x_0)$$

当 $f(x)=1$，得 $A_0=b-a$；当 $f(x)=x$，得 $A_0x_0=\frac{1}{2}(b^2-a^2)$，于是 $x_0=\frac{a+b}{2}$，可得求积公式

$$\int_a^b f(x)\mathrm{d}x\approx(b-a)f\left(\frac{a+b}{2}\right) \tag{6.4.3}$$

称为中点求积公式，它的代数精确度为一次.

例 6.6 当 $n=1$ 时求积分公式 $\int_0^1\sqrt{x}f(x)\mathrm{d}x\approx A_0f(x_0)+A_1f(x_1)$ 的系数 A_0，A_1 及节点 x_0，x_1，使它具有最高代数精确度.

解 由代数精确度定义，公式对 $f(x)=1,x,x^2,x^3$ 精确成立，由(6.4.2)得

$$\begin{cases}A_0+A_1=\int_0^1\sqrt{x}\mathrm{d}x=\frac{2}{3}\\ A_0x_0+A_1x_1=\int_0^1x^{3/2}\mathrm{d}x=\frac{2}{5}\\ A_0x_0^2+A_1x_1^2=\int_0^1x^{5/2}\mathrm{d}x=\frac{2}{7}\\ A_0x_0^3+A_1x_1^3=\int_0^1x^{7/2}\mathrm{d}x=\frac{2}{9}\end{cases}$$

它是关于 x_0，x_1 与 A_0，A_1 的非线性方程组，由前两式有

$$\begin{bmatrix} A_0 \\ A_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ x_0 & x_1 \end{bmatrix}^{-1} \begin{bmatrix} \frac{2}{3} \\ \frac{2}{5} \end{bmatrix}$$

由后两式得

$$\begin{bmatrix} A_0 \\ A_1 \end{bmatrix} = \begin{bmatrix} x_0^2 & x_1^2 \\ x_0^3 & x_1^3 \end{bmatrix}^{-1} \begin{bmatrix} \frac{2}{7} \\ \frac{2}{9} \end{bmatrix}$$

故有

$$\begin{bmatrix} 1 & 1 \\ x_0 & x_1 \end{bmatrix}^{-1} \begin{bmatrix} \frac{2}{3} \\ \frac{2}{5} \end{bmatrix} = \begin{bmatrix} x_0^2 & x_1^2 \\ x_0^3 & x_1^3 \end{bmatrix}^{-1} \begin{bmatrix} \frac{2}{7} \\ \frac{2}{9} \end{bmatrix}$$

化简得

$$\begin{cases} -\frac{2}{3}(x_0 x_1) + \frac{2}{5}(x_0 + x_1) = \frac{2}{7} \\ -\frac{2}{3}(x_0 x_1)(x_0 + x_1) + \frac{2}{5}(x_0^2 + x_0 x_1 + x_1^2) = \frac{2}{9} \end{cases}$$

令 $u = x_0 x_1, v = x_0 + x_1$，上式可改写为

$$\begin{cases} -\frac{1}{3}u + \frac{1}{5}v = \frac{1}{7} \\ -\frac{1}{5}u + \frac{1}{7}v = \frac{1}{9} \end{cases}$$

解得 $u = x_0 x_1 = \frac{5}{21}, v = x_0 + x_1 = \frac{10}{9}$，从而求得

$$x_0 = 0.289\,949, x_1 = 0.821\,162$$

$$A_0 = 0.277\,556, A_1 = 0.389\,111$$

从例题看到直接解方程组(6.4.2)计算太复杂，$n \geqslant 2$ 时一般都不易求解. 但若先确定求积节点 $x_k (k = 0,1,\cdots,n)$，则由(6.4.2)求出系数 $A_k (k = 0,1,\cdots,n)$就容易了. 下面先证明求积公式(6.4.1)的代数精确度最高为$(2n+1)$次. 若令 $f(x) = \omega_{n+1}^2(x) = (x - x_0)^2 \cdots (x - x_n)^2$，则 $I_n(f) = \sum_{k=0}^{n} A_k f(x_k) = 0$，而

$$I(f) = \int_a^b \rho(x) f(x) \mathrm{d}x = \int_a^b \rho(x) \omega_{n+1}^2(x) \mathrm{d}x > 0$$

说明(6.4.1)对$(2n+2)$次多项式不精确成立，故它的最高代数精确度为$(2n+1)$次. 具有$(2n+1)$次代数精确度的求积公式称为 Gauss 型求积公式，相应的求积节点 $x_k (k = 0,1,\cdots,n)$称为 Gauss 点.

定理 4.1 插值求积公式(6.4.1)的节点 $a \leqslant x_0 < x_1 < \cdots < x_n \leqslant b$ 是 Gauss 点的充分必要条件是区间$[a,b]$上以这组节点为零点的多项式

$$\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$$

与任何次数不超过 n 的多项式 $P(x)\in H_n$ 带权 $\rho(x)$ 正交，即

$$\int_a^b \rho(x)P(x)\omega_{n+1}(x)\mathrm{d}x=0 \tag{6.4.4}$$

证明略，见[4].

根据此定理可知，Gauss 型求积公式的节点就是在 $[a,b]$ 上带权 $\rho(x)$ 正交多项式的零点，这就避免了根据代数精确度定义求解非线性方程组的困难. 在给出求积节点 $\{x_k\}_{k=0}^n$ 后，求积系数 $A_k(k=0,1,\cdots,n)$ 就可直接由解(6.4.2)的前 $(n+1)$ 个方程组得到. 而公式(6.4.1)的余项可通过 $f(x)$ 的 Hermite 插值多项式得到，设 $H_{2n+1}(x)\in H_{2n+1}$，满足插值条件

$$H_{2n+1}(x_i)=f(x_i),H'_{2n+1}(x_i)=f'(x_i),i=0,1,\cdots,n$$

于是有

$$f(x)=H_{2n+1}(x)+\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\omega_{n+1}^2(x),\ \xi\in(a,b)$$

两端乘权函数 $\rho(x)$，并从 a 到 b 积分，则得

$$I(f)=\int_a^b \rho(x)f(x)\mathrm{d}x=\sum_{k=0}^n A_kf(x_k)+R_n[f] \tag{6.4.5}$$

其中

$$R_n[f]=\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b \rho(x)\omega_{n+1}^2(x)\mathrm{d}x \tag{6.4.6}$$

(6.4.5)右端第一项是因为(6.4.1)对任何 $(2n+1)$ 次多项式精确成立，且 $H_{2n+1}(x_i)=f(x_i)(i=0,1,\cdots,n)$，则可得到，而 $R_n[f]$ 可利用积分中值定理得到的. 下面还可证明 Gauss 型求积公式(6.4.1)的稳定性和收敛性.

定理 4.2 Gauss 型求积公式(6.4.1)的系数 $A_k>0(k=0,1,\cdots,n)$.

证明 由于(6.4.1)对任何不大于 $(2n+1)$ 次的多项式精确成立，若取 $f(x)=l_k^2(x)$，其中 $l_k(x)$ 为 n 次插值基函数

$$l_k(x)=\frac{(x-x_0)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}\in H_n$$

于是 $f(x)=l_k^2(x)\in H_{2n}$，故有

$$0<\int_a^b \rho(x)l_k^2(x)\mathrm{d}x=\sum_{i=0}^n A_il_k^2(x_i)=A_k,k=0,1,\cdots,n$$

证毕.

再由定理 1.2 得如下推论.

推论 求积公式(6.4.1)是稳定的.

定理 4.3 设 $f\in C[a,b]$，则 Gauss 型求积公式(6.4.1)是收敛的，即

$$\lim_{n\to\infty} I_n(f)=\int_a^b \rho(x)f(x)\mathrm{d}x=I(f)$$

定理证明可见[4].

6.4.2 Gauss－Legendre 求积公式

若 $\rho(x)=1$，区间为$[-1,1]$的求积公式

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{k=0}^{n} A_k f(x_k) \tag{6.4.7}$$

其中节点$\{x_k\}_0^n$是 Legendre 多项式

$$P_{n+1}(x)=\frac{1}{2^{n+1}(n+1)!}\frac{\mathrm{d}^{(n+1)}}{\mathrm{d}x^{n+1}}[(x^2-1)^{n+1}]$$

的零点，则(6.4.7)称为 Gauss－Legendre 求积公式，简称 Gauss 求积公式．其系数

$$A_k=\int_{-1}^{1} l_k^2(x)\mathrm{d}x, l_k(x)=\frac{\tilde{P}_{n+1}(x)}{(x-x_k)\tilde{P}'_{n+1}(x_k)}$$

这里$\tilde{P}_{n+1}(x)$是最高项系数为 1 的 Legendre 多项式．余项可由(6.4.6)得到

$$\begin{aligned} R_n[f]&=\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_{-1}^{1}\tilde{P}_{n+1}^2(x)\mathrm{d}x \\ &=\frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3}f^{(2n+2)}(\eta),\ \eta\in(-1,1) \end{aligned} \tag{6.4.8}$$

例如当 $n=1$ 时，$P_2(x)=\frac{1}{2}(3x^2-1)=0$，则 $x_0=-\sqrt{\frac{1}{3}}, x_1=\sqrt{\frac{1}{3}}, A_0=A_1=1, R_1[f]=\frac{1}{135}f^{(4)}(\eta)$．它比 Simpson 公式(三点)的余项 $R_1[f]=-\frac{1}{90}f^{(4)}(\eta)$的系数绝对值小，Gauss 求积公式(6.4.7)的节点与系数可见表 6－3．

表 6－3

n	x_k	A_k	n	x_k	A_k
0	0	2	4	±0.906 179 845 9	0.235 926 885 1
				±0.538 469 310 1	0.478 628 670 5
1	±0.577 350 269 2	1		0	0.568 888 888 9
2	±0.774 596 669 2	5/9	5	±0.932 469 514 2	0.171 324 492 4
				±0.661 209 386 5	0.360 761 573 0
	0	8/9		±0.236 619 186 1	0.467 913 934 6
3	±0.861 136 311 6	0.347 854 845 1	6	±0.491 079 123 0	0.129 484 966 2
	±0.339 981 043 6	0.652 145 154 9		±0.741 531 185 6	0.279 705 391 5
				±0.405 845 151 4	0.381 830 050 5
				0	0.417 959 183 7

例 6.7 用四点($n=3$)的 Gauss 求积公式计算

$$I(f)=\int_0^{\frac{\pi}{2}} x^2\cos x\,\mathrm{d}x$$

解 先将区间$\left[0,\frac{\pi}{2}\right]$变换为$[-1,1]$,令 $x=\frac{\pi}{4}t+\frac{\pi}{4}$,

$$I(f)=\int_{-1}^{1}\left(\frac{\pi}{4}\right)^3(t+1)^2\cos\left[\frac{\pi}{4}(t+1)\right]\mathrm{d}t\approx\left(\frac{\pi}{4}\right)^3\sum_{k=0}^{3}A_k(t_k+1)^2\cos\left[\frac{\pi}{4}(t_k+1)\right]$$

其中 $-t_0=t_3=0.861\,136\,3$, $-t_1=t_2=0.339\,881\,0$

$A_0=A_3=0.347\,854\,8$, $A_1=A_2=0.652\,145\,2$

$I(f)\approx 0.467\,402$ (精确解 $I(f)=0.467\,401\cdots$)

这结果与用 $n=8$ 的 Romberg 求积相当.

6.4.3 Gauss-Chebyshev 求积公式

区间为$[-1,1]$,权函数 $\rho(x)=\frac{1}{\sqrt{1-x^2}}$的 Gauss 型求积公式,其节点 x_k 是 Chebyshev 多项式 $T_{n+1}(x)$的零点,即 $x_k=\cos\left[\frac{2k+1}{2(n+1)}\pi\right](k=0,1,\cdots,n)$,而 $A_k=\frac{\pi}{n+1}(k=0,1,\cdots,n)$,于是得到

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x\approx\frac{\pi}{n+1}\sum_{k=0}^{n}f\left[\cos\frac{2k+1}{2(n+1)}\pi\right]\tag{6.4.9}$$

称为 Gauss-Chebyshev 求积公式,公式的余项为

$$R_n(f)=\frac{2\pi}{2^{2(n+1)}(2n+2)!}f^{(2n+2)}(\eta),\eta\in(-1,1)\tag{6.4.10}$$

这种求积公式可用于计算奇异积分.

例 6.8 用三点和四点 Gauss-Chebyshev 求积公式计算积分 $I=\int_{-1}^{1}\frac{\mathrm{e}^x}{\sqrt{1-x^2}}\mathrm{d}x$,并估计误差.

解 这里 $f(x)=\mathrm{e}^x$, $f^{(n)}(x)=\mathrm{e}^x$,由 Gauss-Chebyshev 求积公式(6.4.9)可得

$$I=\int_{-1}^{1}\frac{\mathrm{e}^x}{\sqrt{1-x^2}}\mathrm{d}x\approx\frac{\pi}{n+1}\sum_{k=0}^{n}\mathrm{e}^{x_k}$$

当 $n=2$ 时,$x_k=\cos\left(\frac{2k+1}{6}\pi\right)(k=0,1,2)$,求得

$$x_0=\cos\frac{\pi}{6}=\frac{\sqrt{3}}{2},x_1=\cos\frac{3}{6}\pi=0,x_2=\cos\frac{5}{6}\pi=-\frac{\sqrt{3}}{2}$$

代入上式得

$$I\approx\frac{\pi}{3}\left(\mathrm{e}^{\frac{\sqrt{3}}{2}}+\mathrm{e}^0+\mathrm{e}^{-\frac{\sqrt{3}}{2}}\right)=\frac{\pi}{3}(2.377\,44+1+0.420\,62)$$

$$\approx 3.977\,32$$

估计误差可用余项表达式(6.4.10),因 $f(x)=e^x$, $f^{(6)}(x)=e^x$,故

$$|R_2(f)|=\left|\frac{\pi}{2^5\cdot 6!}f^{(6)}(\eta)\right|\leqslant\frac{\pi}{2^5\cdot 6!}e\leqslant 3.706\ 5\times 10^{-4}$$

当 $n=3$ 时, $x_k=\cos\left(\frac{2k+1}{8}\pi\right)(k=0,1,2,3)$,求得

$$x_0=0.923\ 879\ 5, x_1=0.382\ 683\ 4,$$

$$x_2=-0.382\ 683\ 4, x_3=-0.923\ 879\ 5$$

$$I\approx\frac{\pi}{4}\sum_{k=0}^{3}e^{x_k}=\frac{\pi}{4}(2.519\ 044\ 1+1.466\ 213\ 8+$$

$$0.682\ 028\ 8+0.396\ 976\ 0)=3.977\ 462\ 62$$

误差

$$|R_3(f)|=\left|\frac{\pi}{2^7\cdot 8!}f^{(8)}(\eta)\right|\leqslant\frac{\pi}{2^7\cdot 8!}e\leqslant 1.655\times 10^{-6}.$$

习 题 六

1. 分别用复合梯形公式及复合 Simpson 公式计算下列积分.

(1) $\int_0^1\frac{x}{4+x^2}dx$, $n=8$　　(2) $\int_0^{\frac{\pi}{6}}\sqrt{4-\sin^2\varphi}d\varphi$, $n=6$

2. 用 Simpson 公式求积分 $\int_0^1 e^{-x}dx$,并估计误差.

3. 确定下列求积公式中的待定参数,使其代数精确度尽量高,并指明求积公式所具有的代数精确度.

(1) $\int_0^1 f(x)dx\approx Af(0)+Bf(x_1)+Cf(1)$

(2) $\int_{-2h}^{2h}f(x)dx\approx A_{-1}f(-h)+A_0f(0)+A_1f(h)$

(3) $\int_{-h}^{h}f(x)dx\approx Af(-h)+Bf(x_1)$

4. 计算积分 $I=\int_0^{\frac{\pi}{2}}\sin x dx$,若用复合 Simpson 公式要使误差不超过 $\frac{1}{2}\times 10^{-5}$,问区间 $\left[0,\frac{\pi}{2}\right]$ 要分为多少等分?若改用复合梯形公式达到同样精确度,区间 $\left[0,\frac{\pi}{2}\right]$ 应分为多少等分?

5. 用改变步长的复合梯形公式及复合 Simpson 公式计算积分 $I=\int_0^1\frac{x}{4+x^2}dx$,使事后误差估计分别达到 $\frac{1}{2}\times 10^{-3}$ 与 $\frac{1}{2}\times 10^{-5}$.

6. 用 Romberg 求积算法求积分 $\frac{2}{\sqrt{\pi}}\int_0^1 e^{-x^2}dx$,取 $k=3$.

7. 证明求积公式

$$\int_{x_0}^{x_1}f(x)dx\approx\frac{h}{2}[f(x_0)+f(x_1)]-\frac{h^2}{12}[f'(x_1)-f'(x_0)]$$

具有 3 次代数精确度，其中 $h = x_1 - x_0$.

8. 利用积分 $\int_2^8 \frac{1}{x}\mathrm{d}x = 2\ln 2$ 计算 ln2 时，若采用复合 Simpson 公式，问计算积分时应取多少节点才能使误差的绝对值不超过 $\frac{1}{2}\times 10^{-5}$?

9. 用三点 Gauss－Legendre 求积公式计算积分.

(1) $\int_{-1}^{1} \sqrt{x+1.5}\mathrm{d}x$　　(2) $\int_0^1 x^2 \mathrm{e}^{-x}\mathrm{d}x$

10. 用三点 Gauss－Chebyshev 求积公式计算积分 $I = \int_{-1}^{1} \frac{1}{\sqrt{1-x^4}}\mathrm{d}x$.

11. 建立 Gauss 型求积公式 $\int_0^1 \frac{1}{\sqrt{x}} f(x)\mathrm{d}x \approx A_0 f(x_0) + A_1 f(x_1)$.

12. 试确定常数 A, B, C，及 α，使求积公式

$$\int_{-2}^{2} f(x)\mathrm{d}x \approx Af(-\alpha) + Bf(0) + Cf(\alpha)$$

有尽可能高的代数精确度，并指出所得求积公式的代数精确度是多少．它是否为 Gauss 型的求积公式?

第七章　常微分方程数值解

7.1 引　　言

本章讨论常微分方程初值问题

$$\begin{cases} y' = f(x,y), & a \leqslant x \leqslant b \\ y(a) = y_0 \end{cases} \tag{7.1.1}$$

的数值解法，这也是科学与工程计算经常遇到的问题，由于只有很特殊的方程能用解析方法求解，而用计算机求解常微分方程的初值问题都要采用数值方法．通常我们假定(7.1.1)中 $f(x,y)$对 y 满足 Lipschitz 条件，即存在常数 $L>0$，使对 $\forall y_1, y_2 \in \mathbf{R}$，有

$$|f(x,y_1) - f(x,y_2)| \leqslant L|y_1 - y_2| \tag{7.1.2}$$

则初值问题(7.1.1)的解存在唯一．

假定(7.1.1)的精确解为 $y(x)$，求它的数值解就是要在区间$[a,b]$上的一组离散点 $a = x_0 < x_1 < \cdots < x_n < \cdots \leqslant b$ 上求$y(x)$的近似 $y_0, y_1, \cdots, y_n \cdots$．通常取 $x_n = a + nh\ (n = 0,1,\cdots)$，$h$ 称为步长，求(7.1.1)的数值解是按节点 $x_i\ (i = 1,2,\cdots)$的顺序逐步推进求得 $y_1, y_2, \cdots$．首先，要对方程做离散逼近，求出数值解的公式，再研究公式的局部截断误差，计算稳定性以及数值解的收敛性与整体误差等问题．

7.2 简单的单步法及基本概念

7.2.1 Euler 法、后退 Euler 法与梯形法

求初值问题(7.1.1)的一种最简单方法是将节点 x_n 的导数 $y'(x_n)$用差商 $\dfrac{y(x_n+h)-y(x_n)}{h}$代替，于是(7.1.1)的方程可近似写成

$$y(x_{n+1}) \approx y(x_n) + hf(x_n, y(x_n)), n = 0,1,\cdots \tag{7.2.1}$$

从 x_0 出发 $y(a) = y(x_0) = y_0$，由(7.2.1)求得 $y(x_1) \approx y_0 + hf(x_0,y_0) = y_1$ 再将 $y_1 \approx y(x_1)$代入(7.2.1)右端，得到 $y(x_2)$的近似 $y_2 = y_1 + hf(x_1,y_1)$，一般写成

$$y_{n+1}=y_n+hf(x_n,y_n),n=0,1,\cdots \tag{7.2.2}$$

称为解初值问题的 Euler 法.

Euler 法的几何意义如图 7－1 所示.初值问题(7.1.1)的解曲线 $y=y(x)$过点 $P_0(x_0,y_0)$,从 P_0 出发,以 $f(x_0,y_0)$为斜率作一段直线,与直线 $x=x_1$ 交点于 $P_1(x_1,y_1)$,显然有 $y_1=y_0+hf(x_0,y_0)$,再从 P_1 出发,以 $f(x_1,y_1)$为斜率作直线推进到 $x=x_2$ 上一点 P_2,其余类推,这样得到解曲线的一条近似曲线,它就是折线$\overline{P_0P_1P_2\cdots}$.

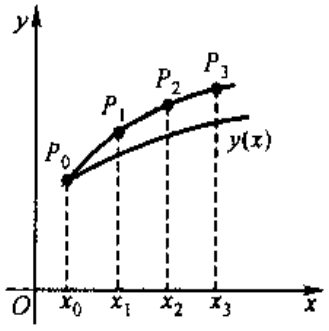

图 7－1

Euler 法也可利用 $y(x_{n+1})$的 Taylor 展开式得到,由

$$y(x_n+h)=y(x_n)+hy'(x_n)+\frac{h^2}{2}y''(\xi_n),\xi_n\in(x_n,x_{n+1}) \tag{7.2.3}$$

略去余项,以 $y_n\approx y(x_n)$,就得到近似计算公式(7.2.2).

另外,还可对(7.1.1)的方程两端由 x_n 到x_{n+1}积分得

$$y(x_{n+1})-y(x_n)=\int_{x_n}^{x_{n+1}}f(x,y(x))\mathrm{d}x \tag{7.2.4}$$

若右端积分用左矩形公式,用 $y_n\approx y(x_n),y_{n+1}\approx y(x_{n+1})$,则得(7.2.2).

如果在(7.2.4)的积分中用右矩形公式,则得

$$y_{n+1}=y_n+hf(x_{n+1},y_{n+1}),n=0,1,\cdots \tag{7.2.5}$$

称为后退(隐式)Euler 法.若在(7.2.4)的积分中用梯形公式,则得

$$y_{n+1}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},y_{n+1})],n=0,1,\cdots \tag{7.2.6}$$

称为梯形方法.

上述三个公式(7.2.2),(7.2.5)及(7.2.6)都是由 y_n 计算 y_{n+1},这种只用前一步即可算出 y_{n+1}的公式称为单步法,其中(7.2.2)可由 y_0 逐次求出 $y_1,y_2,\cdots$的值,称为显式方法,而(7.2.5)及(7.2.6)右端含有 $f(x_{n+1},y_{n+1})$当 f 对y 非线性时它不能直接求出 y_{n+1},此时应把它看作一个方程,求解 y_{n+1},这类方法称为稳式方法.此时可将(7.2.5)或(7.2.6)写成不动点形式的方程

$$y_{n+1}=h\beta f(x_{n+1},y_{n+1})+g$$

这里对式(7.2.5)有 $\beta=1,g=y_n$,对(7.2.6)则 $\beta=\frac{1}{2},g=y_n+\frac{h}{2}f(x_n,y_n)$,$g$ 与 y_{n+1}无关,可构造迭代法

$$y_{n+1}^{(s+1)}=h\beta f(x_{n+1},y_{n+1}^{(s)})+g,\quad s=0,1,\cdots \tag{7.2.7}$$

由于 $f(x,y)$ 对 y 满足条件(7.1.2),故有

$$|y_{n+1}^{(s+1)}-y_{n+1}|\leqslant h\beta|f(x_{n+1},y_{n+1}^{(s)})-f(x_{n+1},y_{n+1})|$$
$$\leqslant h\beta L|y_{n+1}^{(s)}-y_{n+1}|$$

当 $h\beta L<1$ 或 $h<\frac{1}{\beta L}$,迭代法(7.2.7)收敛到 y_{n+1},因此只要步长 h 足够小,就可保证迭代(7.2.7)收敛.对后退 Euler 法(7.2.5),当 $h<\frac{1}{L}$时迭代收敛,对梯形法(7.2.6),当 $h<\frac{2}{L}$时迭代序列收敛.

例 7.1 用 Euler 法、隐式 Euler 法、梯形法解

$$y'=-y+x+1,y(0)=1$$

取 $h=0.1$,计算到 $x=0.5$,并与精确解比较.

解 本题可直接用给出公式计算.由于 $f(x,y)=-y+x+1,h=0.1,x_0=0,y_0=1$,Euler 法的计算公式为

$$y_{n+1}=y_n+h(-y_n+x_n+1)$$
$$=(1-h)y_n+hx_n+h=0.9y_n+0.1x_n+0.1$$

$n=0$ 时,$y_1=0.9y_0+0.1x_0+0.1=1.000\ 000$.其余 $n=1,2,3,4$ 的计算结果见表7-1.

对隐式 Euler 法,计算公式为

$$y_{n+1}=y_n+h(-y_{n+1}+x_{n+1}+1)$$

解出

$$y_{n+1}=\frac{1}{1+h}(y_n+hx_{n+1}+h)=\frac{1}{1.1}(y_n+0.1x_n+0.11)$$

当 $n=0$ 时,$y_1=\frac{1}{1.1}(y_0+0.1x_0+0.11)=1.009\ 091$.其余 $n=1,2,3,4$ 的计算结果见表 7-1.

表 7-1 例 7.1 的三种方法及精确解的计算结果

x_n	Euler 法 y_n	隐式 Euler 法 y_n	梯形法 y_n	精确解 $y(x_n)$
0	1	1	1	1
0.1	1.000 000	1.009 091	1.004 762	1.004 837
0.2	1.010 000	1.026 446	1.018 594	1.018 731
0.3	1.029 000	1.051 315	1.040 633	1.040 818
0.4	1.056 100	1.083 014	1.070 097	1.070 320
0.5	1.090 490	1.120 922	1.106 278	1.106 531

对梯形法,计算公式为

$$y_{n+1}=y_n+\frac{h}{2}[(-y_n+x_n+1)+(-y_{n+1}+x_{n+1}+1)]$$

解得

$$y_{n+1}=\frac{1}{2+h}[(2-h)y_n+h(x_n+x_n+h)+2h]$$

$$=\frac{1}{2.1}(1.9y_n+0.2x_n+0.21)$$

当 $n=0$ 时，$y_1=\frac{1}{2.1}(1.9+0.21)=1.004\ 762$. 其余 $n=1,2,3,4$ 的计算结果见表 7－1.

本题的精确解为 $y(x)=x+e^{-x}$，表 7－1 列出三种方法及精确解的计算结果.

7.2.2 单步法的局部截断误差

解初值问题(7.1.1)的单步法可表示为

$$y_{n+1}=y_n+h\phi(x_n,y_n,y_{n+1},h),\quad n=0,1,\cdots \tag{7.2.8}$$

其中 ϕ 与 f 有关，称为增量函数，当 ϕ 含有 y_{n+1} 时，是隐式单步法，如(7.2.5)及(7.2.6)均为隐式单步法，而当 ϕ 不含 y_{n+1} 时，则为显式单步法，它表示为

$$y_{n+1}=y_n+h\phi(x_n,y_n,h),\quad n=0,1,\cdots \tag{7.2.9}$$

如 Euler 法(7.2.2)，$\phi(x,y,h)=f(x,y)$. 为讨论方便，我们只对显式单步法(7.2.9)给出局部截断误差概念.

定义 2.1 设 $y(x)$ 是初值问题(7.1.1)的精确解，记

$$T_{n+1}=y(x_{n+1})-y(x_n)-h\phi(x_n,y(x_n),h) \tag{7.2.10}$$

称为显式单步法(7.2.9)在 x_{n+1} 的局部截断误差.

T_{n+1} 之所以称为局部截断误差，可理解为用公式(7.2.9)计算时，前面各步都没有误差，即 $y_n=y(x_n)$，只考虑由 x_n 计算到 x_{n+1} 这一步的误差，此时由(7.2.10)有

$$\begin{aligned}y(x_{n+1})-y_{n+1}&=y(x_{n+1})-[y_n+h\phi(x_n,y_n,h)]\\&=y(x_{n+1})-y(x_n)-h\phi(x_n,y(x_n),h)\\&=T_{n+1}\end{aligned}$$

局部截断误差(7.2.10)实际上是将精确解 $y(x)$ 代入(7.2.9)产生的公式误差，利用 Taylor 展开式可得到 $T_{n+1}=O(h^{p+1})$. 例如对 Euler 法(7.2.2)有 $\phi(x,y,h)=f(x,y)$，故

$$\begin{aligned}T_{n+1}&=y(x_{n+1})-y(x_n)-hf(x_n,y(x_n))\\&=y(x_n+h)-y(x_n)-hy'(x_n)=\frac{1}{2}h^2y''(x_n)+\frac{1}{6}h^3y'''(x_n)+\cdots\\&=O(h^2)\end{aligned}$$

它表明 Euler 法(7.2.2)的局部截断误差为 $T_{n+1}=\frac{h^2}{2}y''(x_n)+O(h^3)$，称 $\frac{h^2}{2}y''(x_n)$ 为局部截断误差主项.

定义 2.2 设 $y(x)$ 是初值问题(7.1.1)的精确解，若显式单步法(7.2.9)的局部截断误差 $T_{n+1}=O(h^{p+1})$，p 是展开式的最大整数，称 p 为单步法(7.2.9)的阶，含 h^{p+1} 的项称为局部截断误差主项.

根据定义,Euler法(7.2.2)中的 $p=1$ 故此方法为一阶方法.

对隐式单步法(7.2.8)也可类似求其局部截断误差和阶,如对后退 Euler 法(7.2.5)有局部截断误差

$$\begin{aligned}T_{n+1}&=y(x_{n+1})-y(x_n)-hf(x_{n+1},y(x_{n+1}))\\&=y(x_n+h)-y(x_n)-hy'(x_n+h)\\&=hy'(x_n)+\frac{h^2}{2}y''(x_n)+\frac{h^3}{3!}y'''(x_n)+\cdots-h[y'(x_n)+hy''(x_n)+\cdots]\\&=-\frac{h^2}{2}y''(x_n)+O(h^3)\end{aligned}$$

故此方法的局部截断误差主项为 $-\frac{h^2}{2}y''(x_n)$,$p=1$,也是一阶方法.对梯形法(7.2.6)同样有

$$\begin{aligned}T_{n+1}&=y(x_{n+1})-y(x_n)-\frac{h}{2}[f(x_n,y(x_n))+f(x_{n+1},y(x_{n+1}))]\\&=y(x_n+h)-y(x_n)-\frac{h}{2}[y'(x_n)+y'(x_n+h)]\\&=-\frac{h^3}{12}y'''(x_n)+O(h^4)\end{aligned}$$

它的局部误差主项为 $-\frac{h^3}{12}y'''(x_n)$,$p=2$,方法是二阶的.

7.2.3 改进 Euler 法

上述三种简单的单步法中,梯形法(7.2.6)为二阶方法,且局部截断误差最小,但方法是隐式的,计算要用迭代法.为避免迭代,可先用 Euler 法计算出 y_{n+1} 的近似 $\bar{y}_{n+1}$,将(7.2.6)改为

$$\begin{cases}\bar{y}_{n+1}=y_n+hf(x_n,y_n)\\y_{n+1}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},\bar{y}_{n+1})]\end{cases}\tag{7.2.11}$$

称为改进 Euler 法,它实际上是显式方法.即

$$y_{n+1}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},y_n+hf(x_n,y_n))]\tag{7.2.12}$$

右端已不含 y_{n+1}.可以证明 $T_{n+1}=O(h^3)$,$p=2$,故方法仍为二阶的,与梯形法一样,但用(7.2.11)计算 y_{n+1} 不用迭代.

例 7.2 用改进 Euler 法求例 7.1 的初值问题并与 Euler 法和梯形法比较误差的大小.

解 将改进 Euler 法用于例 7.1 的计算公式

$$\begin{aligned}y_{n+1}&=y_n+\frac{h}{2}[(-y_n+x_n+1)+(-y_n-h(-y_n+x_n+1)+x_{n+1}+1)]\\&=\left[1-\frac{h(2-h)}{2}\right]y_n+\frac{h(1-h)}{2}x_n+\frac{h}{2}(x_n+h)+\frac{h(2-h)}{2}\\&=0.905y_n+0.095x_n+0.1\end{aligned}$$

当 $n=0$ 时，$y_1=0.905y_0+0.095x_0+0.1=1.005\ 000$. 其余结果见表 7－2.

表 7－2　改进 Euler 法及三种方法的误差比较

x_n	改进 Euler 法 y_n	误差 $\lvert y_n-y(x_n)\rvert$	Euler 方法 $\lvert y_n-y(x_n)\rvert$	梯形法 $\lvert y_n-y(x_n)\rvert$
0.1	1.005 000	1.6×10^{-4}	4.8×10^{-3}	7.5×10^{-5}
0.2	1.019 025	2.9×10^{-4}	8.7×10^{-3}	1.4×10^{-4}
0.3	1.041 218	4.0×10^{-4}	1.2×10^{-2}	1.9×10^{-4}
0.4	1.070 802	4.8×10^{-4}	1.4×10^{-2}	2.2×10^{-4}
0.5	1.107 076	5.5×10^{-4}	1.6×10^{-2}	2.5×10^{-4}

从表 7－2 中看到改进 Euler 法的误差数量级与梯形法大致相同，而比 Euler 法小得多，它优于 Euler 法.

7.3　Runge－Kutta 方法

7.3.1　显式 Runge－Kutta 法的一般形式

上节已给出与初值问题(7.1.1)等价的积分形式

$$y(x_{n+1})-y(x_n)=\int_{x_n}^{x_{n+1}}f(x,y(x))\mathrm{d}x \tag{7.3.1}$$

只要对右端积分用不同的数值求积公式近似就可得到不同的求解初值问题(7.1.1)的数值方法，若用显式单步法

$$y_{n+1}=y_n+h\phi(x_n,y_n,h),\quad n=0,1,\cdots \tag{7.3.2}$$

当 $\phi(x_n,y_n,h)=f(x_n,y_n)$，即数值求积用左矩形公式，它就是 Euler 法(7.2.2)，方法只有一阶，若取

$$\phi(x_n,y_n,h)=\frac{1}{2}[f(x_n,y_n)+f(x_{n+1},y_n+hf(x_n,y_n))] \tag{7.3.3}$$

就是改进 Euler 法，这时数值求积公式是梯形公式的一种近似，计算时要用二个右端函数 f 的值，但方法是二阶的. 若要得到更高阶的公式，则求积分时必须用更多的 f 值，根据数值积分公式，可将(7.3.1)右端积分表示为

$$\int_{x_n}^{x_{n+1}}f(x,y(x))\mathrm{d}x=h\sum_{i=1}^{r}c_if(x_n+a_ih,y(x_n+a_ih))+O(h^{p+1})$$

注意，右端 f 中 $y(x_n+a_ih)$ 还不能直接得到，需要像改进 Euler 法(7.2.11)一样，用前面已算得的 f 值表示为(7.3.3)，一般情况可将(7.3.2)的 ϕ 表示为

$$\phi(x_n,y_n,h)=\sum_{i=1}^{r}c_ik_i \tag{7.3.4}$$

其中　　$k_1 = f(x_n, y_n)$

$$k_i = f\left(x_n + a_i h, y_n + h\sum_{j=1}^{i-1} b_{ij}k_j\right), \quad i = 2,3,\cdots,r$$

这里 $c_i, a_i, b_{ij}(i = 1,2,\cdots,r, j = 1,2,\cdots,i-1)$均为待定常数，公式(7.3.2)，(7.3.4)称为 r 级的显式 Runge－Kutta 法，简称 R－K 方法. 它每步计算 r 个 f 值(即 $k_1,\cdots,k_r$)，而 k_i 由前面 $(i-1)$个已算出的 $k_1, k_2, \cdots, k_{i-1}$表示，故公式是显式的. 例如当 $r = 2$ 时，公式可表示为

$$y_{n+1} = y_n + h(c_1 k_1 + c_2 k_2) \tag{7.3.5}$$

其中 $k_1 = f(x_n, y_n), k_2 = f(x_n + a_2 h, y_n + b_{21} h k_1)$. 改进 Euler 法(7.2.11)就是一个二级显式 R－K方法. 参数 c_1, c_2, a_2, b_{21}取不同的值，可得到不同公式.

7.3.2　二、三级显式 R－K 方法

对 $r = 2$ 的显式 R－K 方法(7.3.5)，要求选择参数 c_1, c_2, a_2, b_{21}，使公式的阶 p 尽量高，由局部截断误差定义

$$T_{n+1} = y(x_{n+1}) - y(x_n) - h[c_1 f(x_n, y(x_n)) + c_2 f(x_n + a_2 h, y_n + b_{21} h k_1)] \tag{7.3.6}$$

令 $y_n = y(x_n)$，对(7.3.6)式在(x_n, y_n)处按 Taylor 公式展开，由于

$$y(x_{n+1}) = y(x_n + h) = y_n + hy'_n + \frac{h^2}{2!}y''_n + \frac{h^3}{3!}y'''_n + O(h^4)$$

$$y'_n = f(x_n, y_n) = f_n$$

$$y''_n = \frac{\mathrm{d}}{\mathrm{d}x}f(x_n, y_n) = f'_x(x_n, y_n) + f'_y(x_n, y_n)f_n$$

$$f(x_n + a_2 h, y_n + b_{21} h k_1) = f(x_n, y_n) + f'_x(x_n, y_n)a_2 h + f'_y(x_n, y_n)(b_{21} h k_1) + O(h^2)$$

将上述结果代入(7.3.6)得

$$\begin{aligned} T_{n+1} &= hf_n + \frac{h^2}{2}[f'_x(x_n, y_n) + f'_y(x_n, y_n)f_n] + O(h^3) - \\ &\quad h[c_1 f_n + c_2(f_n + a_2 f'_x(x_n, y_n)h + b_{21} f'_y(x_n, y_n)f_n h + O(h^2))] \\ &= (1 - c_1 - c_2)hf_n + \left(\frac{1}{2} - c_2 a_2\right)h^2 f'_x(x_n, y_n) + \left(\frac{1}{2} - c_2 b_{21}\right)\cdot \\ &\quad h^2 f'_y(x_n, y_n)f_n + O(h^3) \end{aligned}$$

要使公式(7.3.5)具有的阶 $p = 2$，即 $T_{n+1} = O(h^3)$，必须

$$1 - c_1 - c_2 = 0, \quad \frac{1}{2} - c_2 a_2 = 0, \quad \frac{1}{2} - c_2 b_{21} = 0 \tag{7.3.7}$$

即

$$c_1 + c_2 = 1, \quad c_2 a_2 = \frac{1}{2}, \quad c_2 b_{21} = \frac{1}{2}$$

由此三式求 c_1, c_2, a_2, b_{21}的解不唯一. 因 $r = 2$，故 $c_2 \neq 0$，于是有解

$$c_1 = 1 - c_2, \quad a_2 = b_{21} = \frac{1}{2c_2} \tag{7.3.8}$$

它表明使(7.3.5)具有二阶的方法很多,只要 $c_2 \neq 0$ 都可得到二阶 R－K 方法.若取 $c_2 = \frac{1}{2}$,则 $c_1 = \frac{1}{2}, a_2 = b_{21} = 1$,则得改进 Euler 法(7.2.11),若取 $c_2 = 1$,则得 $c_1 = 0, a_2 = b_{21} = \frac{1}{2}$,此时(7.3.5)为

$$y_{n+1} = y_n + hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1\right) \tag{7.3.9}$$

其中 $$k_1 = f(x_n, y_n)$$

称为中点公式.后退 Euler 法(7.2.11)及中点公式(7.3.9)是两个常用的二级 R－K 方法,注意二级 R－K 方法只能达到二阶,而不可能达到三阶.因为 $r = 2$ 只有 4 个参数,要达到 $p = 3$ 则在(7.3.6)的展开式中要增加 3 项,即增加三个方程.加上(7.3.7)的三个方程求 4 个待定参数是无解的.当然 $r = 2, p = 2$ 的 R－K 方法(7.3.5)当 $c_2 \neq 0$ 取其他数时,也可得到其他公式,但系数较复杂,一般不再给出.

对 $r = 3$ 的情形,要计算三个 k 值,即

$$\phi(x_n, y_n, h) = c_1k_1 + c_2k_2 + c_3k_3$$

其中 $$k_1 = f(x_n, y_n), \quad k_2 = f(x_n + a_2h, y_n + b_{21}hk_1)$$

$$k_3 = f(x_n + a_3h, y_n + b_{31}hk_1 + b_{32}hk_2)$$

将 k_2, k_3 按二元函数在 (x_n, y_n) 处按 Taylor 公式展开,然后代入局部截断误差表达式,可得

$$T_{n+1} = y(x_n + h) - y(x_n) - h\phi(x_n, y_n, h) = O(h^4)$$

可得三阶方法,其系数应满足方程

$$\begin{cases} c_1 + c_2 + c_3 = 1 \\ a_2 = b_{21},\ a_3 = b_{31} + b_{32} \\ c_2a_2 + c_3a_3 = \frac{1}{2},\ c_2a_2^2 + c_3a_3^2 = \frac{1}{3} \\ c_3a_2b_{32} = \frac{1}{6} \end{cases} \tag{7.3.10}$$

这是 8 个未知数 6 个方程的方程组,解也是不唯一的,通常 $c_3 \neq 0$.一种常见的三级三阶 R－K 方法是下面的 Kutta 三阶方法:

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 4k_2 + k_3) \tag{7.3.11}$$

$$k_1 = f(x_n,\ y_n)$$

$$k_2 = f\left(x_n + \frac{1}{2}h,\ y_n + \frac{h}{2}k_1\right)$$

$$k_3 = f(x_n + h,\ y_n - hk_1 + 2hk_2)$$

7.3.3 四阶 R－K 方法及步长的自动选择

利用二元函数 Taylor 展开式可以确定(7.3.4)中 $r = 4, p = 4$ 的 R－K 方法,经典的四阶 R

－K方法是：

$$y_{n+1}=y_n+\frac{h}{6}(k_1+2k_2+2k_3+k_4) \tag{7.3.12}$$

$$k_1=f(x_n,\ y_n)$$

$$k_2=f\left(x_n+\frac{1}{2}h,\ y_n+\frac{1}{2}hk_1\right)$$

$$k_3=f\left(x_n+\frac{1}{2}h,\ y_n+\frac{1}{2}hk_2\right)$$

$$k_4=f(x_n+h,\ y_n+hk_3)$$

它的局部截断误差 $T_{n+1}=O(h^5)$，故 $p=4$，这是最常用的四阶R－K方法，数学库中都有用此方法求解初值问题的软件.这种方法的优点是精度较高，缺点是每步要算4个右端函数值，计算量较大.

例7.3 用经典四阶R－K方法解例7.1的初值问题 $y'=-y+x+1, y(0)=1$，仍取 $h=0.1$，计算到 $x_5=0.5$，并与改进Euler法、梯形法在 $x_5=0.5$ 处比较其误差大小.

解 用四阶R－K方法公式(7.3.12)，此处 $f(x,y)=-y+x+1, x_0=0, y_0=1, h=0.1$，于是当 $n=0$ 时

$$k_1=f(x_0,y_0)=-y_0+x_0+1=0$$

$$k_2=f\left(x_0+\frac{1}{2}h,y_0+\frac{h}{2}k_1\right)$$

$$=\left(-1+\frac{1}{2}h\right)y_0+\left(1-\frac{h}{2}\right)x_0+1=0.05$$

$$k_3=f\left(x_0+\frac{1}{2}h,y_0+\frac{h}{2}k_2\right)$$

$$=\left(-1+\frac{h}{2}-\frac{h^2}{4}\right)y_0+\left(1-\frac{h}{2}+\frac{h^2}{4}\right)x_0+1=0.047\ 5$$

$$k_4=f(x_0+h,y_0+hk_3)$$

$$=\left(-1+h-\frac{h^2}{2}+\frac{h^3}{4}\right)y_0+\left(1-h+\frac{h^2}{2}-\frac{h^3}{4}\right)x_0+1$$

$$=0.095\ 25$$

于是 $y_1=y_0+\frac{h}{6}(k_1+2k_2+2k_3+k_4)=1+\frac{0.1}{6}\times 0.290\ 25=1.004\ 837\ 50$，按公式(7.3.12)可算出

$$y_2=1.018\ 730\ 90,\quad y_3=1.040\ 818\ 42$$

$$y_4=1.070\ 320\ 29,\quad y_5=1.106\ 530\ 94$$

此方法误差：$|y_5-y(x_5)|=2.8\times 10^{-7}$

改进Euler法误差：$|y_5-y(x_5)|=5.5\times 10^{-4}$

梯形法误差：$|y_5-y(x_5)|=2.5\times 10^{-4}$

可见四阶 R－K 方法的精度比二阶方法高得多.

用四阶 R－K 方法求解初值问题(7.1.1)精度较高,但要从理论上给出误差 $|y_n - y(x_n)|$ 的估计式则比较困难.那么应如何判断计算结果的精度以及如何选择合适的步长 h? 通常是通过不同步长在计算机上的计算结果近似估计.设 $y(x_n)$ 在 x_n 处的值 $y_n = y(x_n)$,当 $x_{n+1} = x_n + h$ 时,$y(x_{n+1})$ 的近似为 $y_{n+1}^{(h)}$,于是由四阶 R－K 方法有

$$T_{n+1} = y(x_{n+1}) - y_{n+1}^{(h)} \approx c_n h^5$$

若以 $\frac{h}{2}$ 为步长,计算两步到 x_{n+1},则有

$$T_{n+1} = y(x_{n+1}) - y_{n+1}^{\left(\frac{h}{2}\right)} \approx 2c_n \left(\frac{h}{2}\right)^5$$

于是得

$$\frac{y(x_{n+1}) - y_{n+1}^{(h)}}{y(x_{n+1}) - y_{n+1}^{\left(\frac{h}{2}\right)}} \approx 2^4$$

即

$$y(x_{n+1}) \approx y_{n+1}^{\left(\frac{h}{2}\right)} + \frac{1}{15}\left(y_{n+1}^{\left(\frac{h}{2}\right)} - y_{n+1}^{(h)}\right)$$

或

$$\left|y(x_{n+1}) - y_{n+1}^{\left(\frac{h}{2}\right)}\right| \approx \frac{1}{15}\left|y_{n+1}^{\left(\frac{h}{2}\right)} - y_{n+1}^{(h)}\right| \tag{7.3.13}$$

它给出了误差的近似估计.如果 $\frac{1}{15}\left|y_{n+1}^{\left(\frac{h}{2}\right)} - y_{n+1}^{(h)}\right| \leqslant \varepsilon$($\varepsilon$ 为给定精度),则认为以 $\frac{h}{2}$ 为步长的计算结果 $y_{n+1}^{\left(\frac{h}{2}\right)}$ 满足精度要求,若 $\frac{1}{15}\left|y_{n+1}^{\left(\frac{h}{2}\right)} - y_{n+1}^{(h)}\right| \ll \varepsilon$,则还可放大步长.因此(7.3.13)提供了自动选择步长的方法.

7.4 单步法的收敛性与绝对稳定性

7.4.1 单步法的收敛性

定义 4.1 设 $y(x)$ 是初值问题(7.1.1)的精确解,y_n 是单步法(7.3.2)在 $x_n = x_0 + nh$ 处产生的近似解,若

$$\lim_{h \to 0} y_n = y(x_n)$$

则称方法(7.3.2)产生的数值解 $\{y_n\}$ 收敛于 $y(x_n)$.

实际上,定义中 x_n 是一固定点,当 $h \to 0$ 时 $n \to \infty$,n 不是固定的.因 $h = \frac{x_n - x_0}{n}$ 显然方法收敛,则在固定点 $x = x_n$ 处的整体误差 $e_n = y(x_n) - y_n = O(h^p)$,当 $p \geqslant 1$ 时 $\lim\limits_{h \to 0} e_n = 0$.

下面定理给出方法(7.3.2)收敛的条件.

定理 4.1 设初值问题(7.1.1)的单步法(7.3.2)是 p 阶方法($p \geqslant 1$),且函数 ϕ 对 y 满足

Lipschitz 条件，即存在常数 $L>0$，使对 $\forall y,z\in\mathbf{R}$，均有

$$|\phi(x,y,h)-\phi(x,z,h)|\leqslant L|y-z|$$

则方法(7.3.2)收敛，且 $e_n=y(x_n)-y_n=O(h^p)$.

定理证明略. 可见[3].

7.4.2 绝对稳定性

用单步法(7.3.2)求数值解 $y_0,y_1,\cdots,y_n,\cdots$，由于原始数据及计算过程舍入误差影响，实际得到的不是 y_n 而是 $\bar{y}_n=y_n+\rho_n$，其中 ρ_n 是误差，再计算下一步得到

$$\bar{y}_{n+1}=\bar{y}_n+h\phi(x_n,\bar{y}_n,h)$$

以 Euler 法为例，若令 $\rho_n=\bar{y}_n-y_n$，则

$$\rho_{n+1}=\rho_n+h[f(x_n,\bar{y}_n)-f(x_n,y_n)]=[1+hf'_y(x_n,\eta)]\rho_n \tag{7.4.1}$$

如果 $|1+hf'_y|\leqslant 1$，则从 y_n 计算到 y_{n+1} 误差不增长，它是稳定的. 但如果条件不满足就不稳定.

例 7.4 $y'=-100y,y(0)=1$，精确解为 $y(x)=e^{-100x}$，用 Euler 法求解得

$$y_{n+1}=y_n+hf(x_n,y_n)=(1-100h)y_n,n=0,1,\cdots$$

若取 $h=0.025$，则 $y_{n+1}=-1.5y_n$，当 $n\to\infty$，$\lim\limits_{n\to\infty}|y_n|=\infty$，而 $\lim\limits_{x\to\infty}y(x)=\lim\limits_{x\to\infty}e^{-100x}=0$，显然计算是不稳定的.

如果用后退 Euler 法(7.2.5)解此例，仍取 $h=0.025$，则

$$y_{n+1}=y_n+hf_{n+1}=y_n-2.5y_{n+1},\text{即 } y_{n+1}=\frac{1}{3.5}y_n$$

显然当 $\lim\limits_{n\to\infty}y_n=0$，计算是稳定的.

由此看到稳定性与方法有关，也与 f'_y 有关，在此例中 $f'_y=-100$. 在研究方法的稳定性时，通常不必对一般的 $f(x,y)$ 进行讨论，而只针对模型方程

$$y'=\lambda y,\mathrm{Re}(\lambda)<0 \tag{7.4.2}$$

这里 λ 可能为复数. 规定 $\mathrm{Re}(\lambda)<0$ 是因为 $\mathrm{Re}(\lambda)>0$ 时微分方程(7.4.2)本身是不稳定的，而讨论数值方法(7.3.2)的稳定性，必须在微分方程本身稳定的前提下进行. 另一方面，对初值问题(7.1.1)，若将 $f(x,y)$ 在 (x_n,y_n) 处线性展开，可得

$$f(x,y)\approx f(x_n,y_n)+f'_x(x_n,y_n)(x-x_n)+f'_y(x_n,y_n)(y-y_n),$$

于是方程(7.1.1)可近似表示为

$$y'=\lambda y+g(x_n,y_n),\lambda=f'_y(x_n,y_n)$$

它表明用模型方程(7.4.2)是合理的，至于模型方程(7.4.2)中所以用复数 λ 是因为初值问题(7.1.1)如果是方程组，即 $y\in\mathbf{R}^m$，$f\in\mathbf{R}^m$，则 f'_y 是 $(m\times m)$ 阶矩阵，其特征值可能是复数. 当然对单个方程，λ 就是实数，此时只要规定 $\lambda<0$ 即可.

用单步法(7.3.2)解模型方程(7.4.2)可得到

$$y_{n+1}=E(h\lambda)y_n \tag{7.4.3}$$

其中 $E(h\lambda)$ 依赖所选方法，如用 Euler 法则

$$y_{n+1}=y_n+h\lambda y_n=(1+h\lambda)y_n,E(h\lambda)=1+h\lambda \tag{7.4.4}$$

此时由(7.4.1)看到误差方程也为 $\rho_{n+1}=E(h\lambda)\rho_n$,与(7.4.4)是一样的.因此对一般单步法(7.3.2)误差方程也与(7.4.3)一致.下面再考虑二阶 R－K 方法有

$$\begin{aligned}y_{n+1}&=y_n+\frac{h}{2}[\lambda y_n+\lambda(y_n+h\lambda y_n)]\\&=[1+h\lambda+\frac{1}{2}(h\lambda)^2]y_n,E(h\lambda)=1+h\lambda+\frac{1}{2}(h\lambda)^2\end{aligned}$$

对四阶 R－K 方法,可得

$$E(h\lambda)=1+h\lambda+\frac{1}{2}(h\lambda)^2+\frac{1}{3!}(h\lambda)^3+\frac{1}{4!}(h\lambda)^4$$

定义 4.2 将单步法(7.3.2)用于解模型方程(7.4.2),若得到(7.4.3)中的 $|E(h\lambda)|<1$ 则称方法是绝对稳定的.在复平面上复变量 $h\lambda$ 满足 $|E(h\lambda)|<1$ 的区域,称为方法(7.3.2)的绝对稳定域,它与实轴的交点称为绝对稳定区间.

例如对 Euler 法,$|E(h\lambda)|=|1+h\lambda|<1$ 在复平面 $h\lambda$ 上是以(－1,0)为圆心,以 1 为半径的单位圆域内部,当 λ 为实数时,则得绝对稳定区间为 $-2<h\lambda<0$,因 $\lambda<0$,故有 $0<h<\frac{2}{-\lambda}$.在例 7.4 中 $\lambda=-100$,$h<\frac{2}{100}=0.02$ 时方法稳定,而例中取 $h=0.025$ 故不稳定.

对后退 Euler 法(7.2.5),$y_{n+1}=y_n+h\lambda y_{n+1}$

$$y_{n+1}=\frac{1}{1-h\lambda}y_n,|E(h\lambda)|=\left|\frac{1}{1-h\lambda}\right|<1$$

因 $\lambda<0$,故 $|1-h\lambda|>1$,其绝对稳定域是以(1,0)为圆心的单位圆外部,绝对稳定区间为 $-\infty<h\lambda<0$,即对任何 $h>0$ 方法都是绝对稳定的.

二阶 R－K 方法的绝对稳定区间为 $-2<h\lambda<0$.

三阶 R－K 方法的绝对稳定区间为 $-2.51<h\lambda<0$.

四阶 R－K 方法的绝对稳定区间为 $-2.785<h\lambda<0$.

例 7.5 用经典四阶 R－K 方法计算初值问题

$$y'=-20y(0\leqslant x\leqslant 1),y(0)=1$$

步长取 $h=0.1$ 及 0.2,给出计算误差并分析其稳定性.

解 本题直接按 R－K 方法(7.3.12)的公式计算.因精确解为 $y(x)=e^{-20x}$,其计算误差 $|y_n-y(x_n)|$ 如表所示.

x_n	0.2	0.4	0.6	0.8	1.0
$h=0.1$	0.092 795	0.012 010	0.001 366	0.000 152	0.000 017
$h=0.2$	4.98	25.0	125.0	625.0	3 125.0

从计算结果看到,$h=0.2$ 时误差很大,这是由于在 $\lambda=-20$,$h=0.2$ 时 $\lambda h=-4$,面四阶 R－K

方法的绝对稳定区间为[-2.785,0],故 $h=0.2$ 时计算不稳定,误差很大.而 $h=0.1$ 时 $\lambda h=-2$,其值在绝对稳定区间[-2.785,0]内,计算稳定,故结果是可靠的.

7.5 线性多步法

7.5.1 线性多步法的一般公式

前面给出了求解初值问题(7.1.1)的单步法,其特点是计算 y_{n+1} 时只用到 y_n 的值,此时 y_0, $y_1,\cdots,y_{n-1},y_n$ 的值均已算出.如果在计算 y_{n+1} 时除用 y_n 的值外,还用到 $y_{n-1},y_{n-2},\cdots,y_{n-k+1}$ 的值,这就是多步法.若记 $x_k=x_0+kh$,h 为步长,$y_k\approx y(x_k)$,$f_k=f(x_k,y_k)(k=0,1,\cdots,n)$,则线性多步法可表示为

$$y_{n+1}=\sum_{i=0}^{k-1}(\alpha_i y_{n-i})+h\sum_{i=-1}^{k-1}(\beta_i f_{n-i}),n=k-1,k,\cdots \tag{7.5.1}$$

其中 α_i,β_i 为常数,若 $\alpha_{k-1}^2+\beta_{k-1}^2\neq 0$,称(7.5.1)为线性 k 步法.计算时用到前面已算出的 k 个值 $y_n,y_{n-1},\cdots,y_{n-k+1}$.当 $\beta_{-1}=0$ 时,(7.5.1)为显式方法,当 $\beta_{-1}\neq 0$ 则称(7.5.1)为隐式多步法.隐式方法与梯形方法一样,计算时要用迭代法求 y_{n+1}.多步法(7.5.1)的局部截断误差定义也与单步法类似.

定义 5.1 设 $y(x)$是初值问题(7.1.1)的精确解,线性多步法(7.5.1)在 x_{n+1} 处的局部截断误差定义为

$$T_{n+1}=y(x_{n+1})-\sum_{i=0}^{k-1}\alpha_i y(x_{n-i})-h\sum_{i=-1}^{k-1}\beta_i y'(x_{n-i}) \tag{7.5.2}$$

若 $T_{n+1}=O(h^{p+1})$,则称线性多步法(7.5.1)是 p 阶的.

如果我们希望得到的多步法是 p 阶的,则可利用 Taylor 公式展开,将 T_{n+1} 在 x_n 处展开到 h^{p+1} 阶,它可表示为

$$T_{n+1}=C_0y(x_n)+C_1hy'(x_n)+\cdots+C_ph^py^{(p)}(x_n)+C_{p+1}h^{p+1}y^{(p+1)}(x_n)+O(h^{p+2}) \tag{7.5.3}$$

注意,(7.5.2)式按 Taylor 展开可得

$$\begin{aligned}T_{n+1}&\approx y(x_n+h)-\sum_{i=0}^{k-1}\alpha_i y(x_n-ih)-h\sum_{i=-1}^{k-1}\beta_i y'(x_n-ih)\\&=y(x_n)+hy'(x_n)+\cdots+\frac{h^p}{p!}y^{(p)}(x_n)+\frac{h^{p+1}}{(p+1)!}y^{(p+1)}(x_n)-\\&\quad\sum_{i=0}^{k-1}\alpha_i\left[y(x_n)-ihy'(x_n)+\frac{(ih)^2}{2!}y''(x_n)+\cdots\right]-\\&\quad h\sum_{i=-1}^{k-1}\beta_i\left[y'(x_n)-ihy''(x_n)+\frac{(ih)^2}{2!}y'''(x_n)+\cdots\right]\end{aligned}$$

经整理比较系数可得

$$\begin{cases} C_0 = 1 - \sum_{i=0}^{k-1} \alpha_i \\ C_1 = 1 - \sum_{i=1}^{k-1} (-i)\alpha_i - \sum_{i=-1}^{k-1} \beta_i \\ C_j = \frac{1}{j!}\left[1 - \sum_{i=1}^{k-1} (-i)^j \alpha_i\right] - \frac{1}{(j-1)!} \sum_{i=-1}^{k-1} (-i)^{j-1} \beta_i, j = 2,3,\cdots,p+1 \end{cases} \tag{7.5.4}$$

若线性多步法(7.5.1)为 p 阶,则可令

$$C_0 = C_1 = \cdots = C_p = 0, \quad C_{p+1} \neq 0$$

于是得局部截断误差

$$T_{n+1} = C_{p+1} h^{p+1} y^{(p+1)}(x_n) + O(h^{p+2}) \tag{7.5.5}$$

右端第一项称为局部截断误差主项. C_{p+1} 称为误差常数. 要使多步法(7.5.1)逼近初值问题(7.1.1),方法的阶 $p \geqslant 1$,当 $p=1$ 时,则 $C_0 = C_1 = 0$,由(7.5.4)得

$$\alpha_0 + \alpha_1 + \cdots + \alpha_{k-1} = 1, \sum_{i=1}^{k-1} i\alpha_i - \sum_{i=-1}^{k-1} \beta_i = -1 \tag{7.5.6}$$

称为相容性条件.

公式(7.5.1)当 $k=1$ 时即为单步法,若 $\beta_{-1} = 0$,由(7.5.6)则得

$$\alpha_0 = 1, \beta_0 = 1$$

式(7.5.1)就是 $y_{n+1} = y_n + hf_n$,即为 Euler 法. 此时 $C_2 = \frac{1}{2} \neq 0$,方法为 $p=1$ 阶. 若 $\beta_{-1} \neq 0$,由 $C_0 = 0$ 得 $\alpha_0 = 1$,为确定 β_{-1} 及 β_0,必须令 $C_1 = C_2 = 0$,由(7.5.4)得

$$\beta_{-1} + \beta_0 = 1 \text{ 及 } \beta_{-1} = \frac{1}{2}$$

此时(7.5.1)就是 $y_{n+1} = y_n + \frac{h}{2}(f_n + f_{n+1})$,即为梯形法.

由 $$C_3 = \frac{1}{3!}[1-0] - \frac{1}{2!}\beta_{-1} = \frac{1}{6} - \frac{1}{4} = -\frac{1}{12}, T_{n+1} = -\frac{1}{12}h^3 y'''(x_n) + O(h^4)$$

故 $p=2$,方法是二阶的,与 7.1 节中给出的结果相同.

实际上,当 k 给定后,则可利用(7.5.4)求出公式(7.5.1)中的系数 α_i 及 β_i,并求得 T_{n+1} 的表达式(7.5.5).

7.5.2 Adams 显式与隐式方法

形如

$$y_{n+1} = y_n + h \sum_{i=-1}^{k-1} \beta_i f_{n-i}, n = k-1, k, \cdots \tag{7.5.7}$$

的 k 步法称为 Adams 方法,当 $\beta_{-1} = 0$ 时为 Adams 显式方法,当 $\beta_{-1} \neq 0$ 时,称为 Adams 隐式方法.

对初值问题(7.1.1)的方程两端从 x_n 到 x_{n+1} 积分得

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

显然只要对右端的积分用插值求积公式，求积节点取为 $x_{n-k+1},\cdots,x_n,x_{n+1}$ 即可推出形如(7.5.7)的多步法，但这里我们仍采用 Taylor 展开的方法直接确定(7.5.7)的系数 $\beta_i(i=-1,0,\cdots,k-1)$. 对比(7.5.1)可知，此时 $\alpha_0=1,\alpha_1=\alpha_2=\cdots=\alpha_{k-1}=0$，只要确定 $\beta_{-1},\beta_0,\cdots,\beta_{k-1}$ 即可. 现在若 $k=4$ 且 $\beta_{-1}=0$，即为 4 步的 Adams 显式方法

$$y_{n+1}=y_n+h(\beta_0 f_n+\beta_1 f_{n-1}+\beta_2 f_{n-2}+\beta_3 f_{n-3})$$

其中 $\beta_0,\beta_1,\beta_2,\beta_3$ 为待定参数，若直接用(7.5.4)，可知此时 $C_0=1-\alpha_0=1-1=0$ 自然成立，再令 $C_1=C_2=C_3=C_4=0$ 可得

$$\begin{cases}\beta_0+\beta_1+\beta_2+\beta_3=1\\ \beta_1+2\beta_2+3\beta_3=-\dfrac{1}{2}\\ \beta_1+4\beta_2+9\beta_3=\dfrac{1}{3}\\ \beta_1+8\beta_2+27\beta_3=-\dfrac{1}{4}\end{cases}$$

解此方程组得 $\beta_0=\dfrac{55}{24},\beta_1=-\dfrac{59}{24},\beta_2=\dfrac{37}{24},\beta_3=-\dfrac{9}{24}$.

由此得到

$$C_5=\frac{1}{5!}-\frac{1}{4!}\sum_{i=0}^{3}(-i)^4\beta_i=\frac{251}{720}$$

于是得到四阶 Adams 显式方法及其余项为

$$y_{n+1}=y_n+\frac{h}{24}(55f_n-59f_{n-1}+37f_{n-2}-9f_{n-3}) \tag{7.5.8}$$

$$T_{n+1}=\frac{251}{720}h^5y^{(5)}(x_n)+O(h^6) \tag{7.5.9}$$

若 $\beta_{-1}\neq 0$，则可得到 $p=4$ 的 Adams 隐式公式，则 $k=3$ 并令 $C_1=C_2=C_3=C_4=0$，由(7.5.4)可得

$$\begin{cases}\beta_{-1}+\beta_0+\beta_1+\beta_2=1\\ \beta_{-1}-\beta_1-2\beta_2=\dfrac{1}{2}\\ \beta_{-1}+\beta_1+4\beta_2=\dfrac{1}{3}\\ \beta_{-1}-\beta_1-8\beta_2=\dfrac{1}{4}\end{cases}$$

解得 $\beta_{-1}=\dfrac{9}{24},\beta_0=\dfrac{19}{24},\beta_1=-\dfrac{5}{24},\beta_2=\dfrac{1}{24}$，而 $C_5=\dfrac{1}{5!}-\dfrac{1}{4!}\sum_{i=-1}^{2}(-i)^4\beta_i=-\dfrac{19}{720}$，于是得到四阶 Adams 隐式方法及余项为

$$y_{n+1}=y_n+\frac{h}{24}(9f_{n+1}+19f_n-5f_{n-1}+f_{n-2}) \tag{7.5.10}$$

$$T_{n+1}=-\frac{19}{720}h^5y^{(5)}(x_n)+O(h^6) \tag{7.5.11}$$

一般情形，k 步 Adams 显式方法是 k 阶的，$k=1$ 即为 Euler 法，$k=2$ 为

$$y_{n+1}=y_n+\frac{h}{2}(3f_n-f_{n-1})$$

$k=3$ 时，$y_{n+1}=y_n+\frac{1}{12}(23f_n-16f_{n-1}+5f_{n-2})$.

k 步隐式方法是$(k+1)$阶公式，$k=1$ 为梯形法，$k=2$ 为三阶隐式 Adams 公式

$$y_{n+1}=y_n+\frac{h}{12}(5f_{n+1}+8f_n-f_{n-1})$$

k 步的 Adams 方法计算时必须先用其他方法求出前面 k 个初值$(y_0,y_1,\cdots,y_{k-1})$才能按给定公式算出后面各点的值，它每步只需计算一个新的 f 值，计算量少，但改变步长时前面的 $y_{n-1},y_{n-2},\cdots,y_{n-k+1}$也要跟着重算，不如单步法简便.

例 7.6 用四阶显式 Adams 方法及四阶隐式 Adams 方法解初值问题

$$y'=-y+x+1,0\leqslant x\leqslant 1,y(0)=1,\text{步长 } h=0.1$$

用到的初始值由精确解 $y(x)=e^{-x}+x$ 计算得到.

解 本题直接由公式(7.5.8)及(7.5.10)计算得到. 对于显式方法，将 $f(x)=-y+x+1$ 直接代入式(7.5.8)得到

$$y_{n+1}=y_n+\frac{0.1}{24}(55f_n-59f_{n-1}+37f_{n-2}-9f_{n-3}),n=3,4,\cdots,9$$

其中 $f_i=-y_i+x_i+1,x_i=ih,h=0.1$.

对于隐式方法，由式(7.5.10)可得到

$$y_{n+1}=y_n+\frac{0.1}{24}\times[9(-y_{n+1}+x_{n+1}+1)+19f_n-5f_{n-1}+f_{n-2}]$$

直接求出 y_{n+1}，而不用迭代，得到

$$y_{n+1}=\frac{8}{8.3}y_n+\frac{1}{249}\times(9x_{n+1}+9+19f_n-5f_{n-1}+f_{n-2}),$$
$$n=2,3,\cdots,9$$

计算结果如表所示.

x_n	精确解 $y(x_n)$ $=e^{-x_n}+x_n$	Adams 显式方法		Adams 隐式方法	
		y_n	$\lvert y(x_n)-y_n\rvert$	y_n	$\lvert y(x_n)-y_n\rvert$
0.3	1.040 818 22			1.040 818 01	2.1×10^{-7}
0.4	1.070 320 05	1.070 322 92	2.87×10^{-6}	1.070 319 66	3.9×10^{-7}
0.5	1.106 530 66	1.106 535 48	4.82×10^{-6}	1.106 530 14	5.2×10^{-7}
0.6	1.148 811 64	1.148 818 41	6.77×10^{-6}	1.148 811 01	6.3×10^{-7}
0.7	1.196 585 30	1.196 593 40	8.10×10^{-6}	1.196 584 59	7.1×10^{-7}
0.8	1.249 328 96	1.249 338 16	9.20×10^{-6}	1.249 328 19	7.7×10^{-7}
0.9	1.306 569 66	1.306 579 62	9.96×10^{-6}	1.306 568 84	8.2×10^{-7}
1.0	1.367 879 44	1.367 889 96	1.05×10^{-5}	1.367 878 59	8.5×10^{-7}

7.5.3 Adams 预测－校正方法

上述给出的 Adams 显式方法计算简单，但精度比隐式方法差，而隐式方法由于每步要做迭代，计算不方便．为了避免迭代，通常可将同阶的显式 Adams 方法与隐式 Adams 方法结合，组成预测－校正方法．以四阶方法为例，可用显式方法(7.5.8)计算初始近似 $y_{n+1}^{(0)}$，这个步骤称为预测(Predictor)，以 P 表示，接着计算 f 值(Evaluation)，$f_{n+1}^{(0)}=f(x_{n+1},y_{n+1}^{(0)})$，这个步骤用 E 表示，然后用隐式公式(7.5.10)计算 y_{n+1}，称为校正(Corrector)，以 C 表示，最后再计算 $f_{n+1}=f(x_{n+1},y_{n+1})$，为下一步计算做准备．整个算法如下：

$$\begin{cases}\text{预测 P:}\ y_{n+1}^{P}=y_n+\dfrac{h}{24}(55f_n-59f_{n-1}+37f_{n-2}-9f_{n-3})\\ \text{求值:E:}\ f_{n+1}^{P}=f(x_{n+1},y_{n+1}^{P})\\ \text{校正 C:}\ y_{n+1}=y_n+\dfrac{h}{24}(9f_{n+1}^{P}+19f_n-5f_{n-1}+f_{n-2})\\ \text{求值 E:}\ f_{n+1}=f(x_{n+1},y_{n+1})\end{cases}\tag{7.5.12}$$

公式(7.5.12)称为四阶 Adams 预测－校正方法(PECE)．

利用(7.5.8)和(7.5.10)的局部截断误差(7.5.9)和(7.5.11)可对预测－校正方法(7.5.12)进行修改，在(7.5.12)中的步骤 P 有

$$y(x_{n+1})-y_{n+1}^{P}\approx\frac{251}{720}h^5y^{(5)}(x_n)$$

对于步骤 C 有

$$y(x_{n+1})-y_{n+1}\approx-\frac{19}{720}h^5y^{(5)}(x_n)$$

两式相减可得 $h^5y^{(5)}(x_n)\approx-(y_{n+1}^{P}-y_{n+1})$

于是有

$$y(x_{n+1})-y_{n+1}^{P}\approx-\frac{251}{720}(y_{n+1}^{P}-y_{n+1})$$

$$y(x_{n+1})-y_{n+1}\approx\frac{19}{720}(y_{n+1}^{P}-y_{n+1})$$

若用 y_{n+1}^{C} 代替上式 y_{n+1}，并令

$$y_{n+1}^{PM}=y_{n+1}^{P}+\frac{251}{720}(y_{n+1}^{C}-y_{n+1}^{P})$$

$$y_{n+1}=y_{n+1}^{C}-\frac{19}{720}(y_{n+1}^{C}-y_{n+1}^{P})$$

显然 y_{n+1}^{PM}，y_{n+1} 比 y_{n+1}^{P}，y_{n+1}^{C} 更好，但注意到 y_{n+1}^{PM} 的表达式中 y_{n+1}^{C} 是未知的，因此改为

$$y_{n+1}^{PM}=y_{n+1}^{P}+\frac{251}{720}(y_n^{C}-y_n^{P})$$

下面给出修正的预测－校正格式(PMECME)．

$$\begin{cases} \mathrm{P}: y_{n+1}^{\mathrm{P}} = y_n + \dfrac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \\ \mathrm{M}: y_{n+1}^{\mathrm{PM}} = y_{n+1}^{\mathrm{P}} + \dfrac{251}{720}(y_n^{\mathrm{C}} - y_n^{\mathrm{P}}) \\ \mathrm{E}: f_{n+1}^{\mathrm{PM}} = f(x_{n+1}, y_{n+1}^{\mathrm{PM}}) \\ \mathrm{C}: y_{n+1}^{\mathrm{C}} = y_n + \dfrac{h}{24}(9f_{n+1}^{\mathrm{PM}} + 19f_n - 5f_{n-1} + f_{n-2}) \\ \mathrm{M}: y_{n+1} = y_{n+1}^{\mathrm{C}} - \dfrac{19}{720}(y_{n+1}^{\mathrm{C}} - y_{n+1}^{\mathrm{P}}) \\ \mathrm{E}: f_{n+1} = f(x_{n+1}, y_{n+1}) \end{cases} \tag{7.5.13}$$

经过修正后的 PMECME 格式比原来 PECE 格式提高一阶.

7.5.4 Milne 方法与 Hamming 方法

与 Adams 显式方法不同的另一类四阶显式方法的计算公式形如

$$y_{n+1} = y_{n-3} + h(\beta_0 f_n + \beta_1 f_{n-1} + \beta_2 f_{n-2}) \tag{7.5.14}$$

这里 $\beta_0, \beta_1, \beta_2$ 为待定常数,此公式也是 $k=4$ 步方法,即计算 y_{n+1} 时要用到 $y_n, y_{n-1}, y_{n-2}, y_{n-3}$ 4 个值.为了确定 $\beta_0, \beta_1, \beta_2$,当然可以利用公式(7.5.4)直接算出,但下面我们直接利用 Taylor 展开式确定 $\beta_0, \beta_1, \beta_2$,使它的阶尽量高.方法(7.5.14)的局部截断误差为

$$T_{n+1} = y(x_n + h) - y(x_n - 3h) - h[\beta_0 y'(x_n) + \beta_1 y'(x_n - h) + \beta_2 y'(x_n - 2h)]$$

将它在 x_n 点展成 Taylor 级数,得

$$\begin{aligned} T_{n+1} = {} & y(x_n) + hy'(x_n) + \frac{h^2}{2!}y''(x_n) + \frac{h^3}{3!}y'''(x_n) + \frac{h^4}{4!}y^{(4)}(x_n) + \frac{h^5}{5!}y^{(5)}(x_n) + \cdots - \\ & \left[y(x_n) - 3hy'(x_n) + \frac{(3h)^2}{2}y''_n - \frac{(3h)^3}{3!}y'''(x_n) + \frac{(3h)^4}{4!}y^{(4)}(x_n) - \frac{(3h)^5}{5!}y^{(5)}(x_n) + \cdots\right] - \\ & h\left\{\beta_0 y'(x_n) + \beta_1\left[y'(x_n) - hy''(x_n) + \frac{h^2}{2}y'''(x_n) - \frac{h^3}{3!}y^{(4)}(x_n) + \frac{h^4}{4!}y^{(5)}(x_n) - \cdots\right] + \right. \\ & \left.\beta_2\left[y'(x_n) - 2hy''(x_n) + \frac{(2h)^2}{2!}y'''(x_n) - \frac{(2h)^3}{3!}y^{(4)}(x_n) + \frac{(2h)^4}{4!}y^{(5)}(x_n) - \cdots\right]\right\} \\ = {} & [1 + 3 - (\beta_0 + \beta_1 + \beta_2)]hy'(x_n) + \left(\frac{1}{2} - \frac{9}{2} + \beta_1 + 2\beta_2\right)h^2 y''(x_n) + \\ & \left(\frac{1}{6} + \frac{27}{6} - \frac{1}{2}\beta_1 - \frac{4}{2}\beta_2\right)h^3 y'''(x_n) + \left(\frac{1}{24} - \frac{81}{24} + \frac{1}{6}\beta_1 + \frac{8}{6}\beta_2\right)h^4 y^{(4)}(x_n) + \\ & \left(\frac{1}{5!} + \frac{3^5}{5!} - \frac{1}{4!}\beta_1 - \frac{2^4}{4!}\beta_2\right)h^5 y^{(5)}(x_n) + O(h^6) \end{aligned}$$

要使公式的阶尽量高,要令前 3 项系数为 0.即

$$4 - (\beta_0 + \beta_1 + \beta_2) = 0, \quad -4 + \beta_1 + 2\beta_2 = 0, \quad \beta_1 + 4\beta_2 = \frac{28}{3}$$

解得 $\beta_2 = \dfrac{8}{3}$, $\beta_1 = -\dfrac{4}{3}$, $\beta_0 = \dfrac{8}{3}$,代入公式,$h^4 y^{(4)}(x_n)$ 的系数为 0,故

$$T_{n+1}=\frac{14}{45}h^5y^{(5)}(x_n)+O(h^6) \tag{7.5.15}$$

于是得四阶方法

$$y_{n+1}=y_{n-3}+\frac{4}{3}h(2f_n-f_{n-1}+2f_{n-2}) \tag{7.5.16}$$

称为 Milne 公式,它的局部截断误差为(7.5.15).

与(7.5.16)配对的隐式方法为 $k=3$ 的多步法,它的一般形式可表示为

$$y_{n+1}-\alpha_0y_n+\alpha_1y_{n-1}+\alpha_2y_{n-2}+h(\beta_{-1}f_{n+1}+\beta_0f_n+\beta_1f_{n-1})$$

要求公式的阶 $p=4$,可直接用(7.5.4),并令 $C_0=C_1=C_2=C_3=C_4=0$,可得

$$\begin{cases}\alpha_0+\alpha_1+\alpha_2=1\\ -\alpha_1-2\alpha_2+\beta_{-1}+\beta_0+\beta_1=1\\ \alpha_1+4\alpha_2+2\beta_{-1}-2\beta_1=1\\ -\alpha_1-8\alpha_2+3\beta_{-1}+3\beta_1=1\\ \alpha_1+16\alpha_2+4\beta_{-1}-4\beta_1=1\end{cases} \tag{7.5.17}$$

若令 $\alpha_2=0$,可解出 $\alpha_0=0,\alpha_1=1,\beta_{-1}=\beta_1=\frac{1}{3},\beta_0=\frac{4}{3}$,于是得到下列四阶方法

$$y_{n+1}=y_{n-1}+\frac{h}{3}(f_{n+1}+4f_n+f_{n-1}) \tag{7.5.18}$$

称为 Simpson 公式,它的局部截断误差为

$$T_{n+1}=-\frac{1}{90}h^5y^{(5)}(x_n)+O(h^6) \tag{7.5.19}$$

用 Simpson 公式与 Milne 公式(7.5.16)相匹配,用(7.5.16)做预测,(7.5.18)做校正,由于(7.5.18)的稳定性较差,因此通常较少使用.为了改善稳定性,可重新选择四阶的隐式公式,Hamming 通过试验,发现在(7.5.17)中若令 $\alpha_1=0$,得到的公式稳定性较好,此时(7.5.17)的解为 $\alpha_0=\frac{9}{8},\alpha_2=-\frac{1}{8},\beta_{-1}=\frac{3}{8},\beta_0=\frac{6}{8},\beta_1=-\frac{3}{8}$,于是得四阶多步法

$$y_{n+1}=\frac{1}{8}(9y_n-y_{n-2})+\frac{3}{8}h(f_{n+1}+2f_n-f_{n-1}) \tag{7.5.20}$$

称为 Hamming 公式,它的局部截断误差为

$$T_{n+1}=-\frac{1}{40}h^5y^{(5)}(x_n)+O(h^6) \tag{7.5.21}$$

用 Milne 公式(7.5.16)与 Hamming 公式(7.5.20)相匹配,并利用截断误差公式(7.5.15)与(7.5.21)改进计算结果.类似 Adams 预测-校正格式(7.5.13),可得以下的预测-校正格式(PMECME):

$$\begin{cases} \mathrm{P}: y_{n+1}^{\mathrm{P}} = y_{n-3} + \dfrac{4}{3}h(2f_n - f_{n-1} + 2f_{n-2}) \\ \mathrm{M}: y_{n+1}^{\mathrm{PM}} = y_{n+1}^{\mathrm{P}} + \dfrac{112}{121}(y_n^{\mathrm{C}} - y_n^{\mathrm{P}}) \\ \mathrm{E}: f_{n+1}^{\mathrm{PM}} = f(x_{n+1}, y_{n+1}^{\mathrm{PM}}) \\ \mathrm{C}: y_{n+1}^{\mathrm{C}} = \dfrac{1}{8}(9y_n - y_{n-2}) + \dfrac{3}{8}h(f_{n+1}^{\mathrm{PM}} + 2f_n - f_{n-1}) \\ \mathrm{M}: y_{n+1} = y_{n+1}^{\mathrm{C}} - \dfrac{9}{121}(y_{n+1}^{\mathrm{C}} - y_{n+1}^{\mathrm{P}}) \\ \mathrm{E}: f_{n+1} = f(x_{n+1}, y_{n+1}) \end{cases} \tag{7.5.22}$$

例 7.7 用四步四阶显式 Milne 公式及三步四阶隐式 Hamming 公式解初值问题

$$y' = -y + x + 1, 0 \leqslant x \leqslant 1, y(0) = 1, \text{步长 } h = 0.1$$

初值 y_0, y_1, y_2, y_3 仍由精确解 $y(x) = \mathrm{e}^{-x} + x$ 给出，要求计算到 $x_5 = 0.5$ 为止，给出计算结果及误差，并与例 7.6 结果比较.

解 直接用公式(7.5.16)及(7.5.20)计算. 用 Milne 法计算公式为

$$y_{n+1} = y_{n-3} + \frac{0.4}{3} \times (2f_n - f_{n-1} + 2f_{n-2}), n = 3, 4$$

其中 $f_i = -y_i + x_i + 1$

$$f_0 = -y_0 + x_0 + 1 = 0$$

$$f_1 = -y_1 + x_1 + 1 = -1.004\ 837\ 42 + 1.1 = 0.095\ 162\ 58$$

$$f_2 = -y_2 + x_2 + 1 = -1.018\ 730\ 75 + 1.2 = 0.181\ 269\ 25$$

$$f_3 = -y_3 + x_3 + 1 = -1.040\ 818\ 22 + 1.3 = 0.259\ 181\ 78$$

$$y_4 = y_0 + \frac{0.4}{3} \times (2f_3 - f_2 + 2f_1) = 1.070\ 322\ 60$$

$$f_4 = -y_4 + x_4 + 1 = 0.329\ 677\ 40$$

$$y_5 = y_1 + \frac{0.4}{3} \times (2f_4 - f_3 + 2f_2) = 1.106\ 532\ 29$$

误差

$$|y(x_4) - y_4| = 2.55 \times 10^{-6}, |y(x_5) - y_5| = 1.63 \times 10^{-6}$$

用 Hamming 方法(7.5.20)计算公式为

$$y_{n+1} = \frac{1}{8} \times (9y_n - y_{n-2}) + \frac{0.3}{8} \times (f_{n+1} + 2f_n - f_{n-1})$$

$$= \frac{1}{8} \times (9y_n - y_{n-2}) + \frac{0.3}{8} \times (-y_{n+1} + x_{n+1} + 1 + 2f_n - f_{n-1})$$

可解得

$$y_{n+1} = \frac{1}{8.3} \times (9y_n - y_{n-2}) + \frac{0.3}{8.3} \times (x_{n+1} + 2f_n - f_{n-1}),$$

$$n = 2, 3, 4$$

$$y_3 = \frac{1}{8.3} \times (9y_2 - y_0) + \frac{0.3}{8.3} \times (x_3 + 1 + 2f_2 - f_1)$$

$$= \frac{1}{8.3} \times 8.168\ 576\ 75 + \frac{0.3}{8.3} \times 1.567\ 375\ 92 = 1.040\ 818\ 02$$

$$f_3 = -y_3 + x_3 + 1 = 0.259\ 181\ 98$$

$$y_4 = \frac{1}{8.3} \times (9y_3 - y_1) + \frac{0.3}{8.3} \times (x_4 + 1 + 2f_3 - f_2) = 1.070\ 319\ 66$$

$$f_4 = -y_4 + x_4 + 1 = 0.329\ 680\ 34$$

$$y_5 = \frac{1}{8.3} \times (9y_4 - y_2) + \frac{0.3}{8.3} \times (x_5 + 1 + 2f_4 - f_3) = 1.106\ 530\ 10$$

误差

$$|y(x_3) - y_3| = 2.0 \times 10^{-7}, |y(x_4) - y_4| = 3.9 \times 10^{-7} |y(x_5) - y_5| = 5.6 \times 10^{-7}$$

从所得结果可见 Milne 方法误差比显式 Adams 方法误差略小，而 Hamming 方法与隐式 Adams 方法误差相当.

例 7.8 将例 7.7 的初值问题用修正的 Milne－Hamming 预测－校正公式计算 y_5 及 y_6，初值 y_0, y_1, y_2, y_3 仍用已算出的精确解，即 $y_0 = 1, y_1 = 1.004\ 837\ 42, y_2 = 1.018\ 730\ 75, y_3 = 1.040\ 818\ 22$，给出计算结果及误差.

解 根据修正的 Milne－Hamming 预测－校正公式(7.5.22)得

$$y_5^{P} = 1.106\ 532\ 99$$

$$y_5^{PM} = y_5^{P} + \frac{112}{121} \times (y_4^{C} - y_4^{P}) = 1.106\ 530\ 364$$

$$f(x_5, y_5^{PM}) = -y_5^{PM} + x_5 + 1 = 0.393\ 469\ 636$$

$$y_5^{C} = \frac{1}{8} \times (9y_4 - y_2) + \frac{0.3}{8} \times [f(x_5, y_5^{PM}) + 2f_4 - f_3]$$

$$= 1.106\ 530\ 419$$

$$y_5 = y_5^{C} - \frac{9}{121} \times (y_5^{C} - y_5^{P}) = 1.106\ 530\ 61$$

$$f_5 = -y_5 + x_5 + 1 = 0.393\ 469\ 39$$

$$y_6^{P} = y_2 + \frac{0.4}{3} \times (2f_5 - f_4 + 2f_3) = 1.148\ 813\ 73$$

$$y_6^{PM} = y_6^{P} + \frac{112}{121} \times (y_5^{C} - y_5^{P}) = 1.148\ 811\ 35$$

$$f_6^{PM} = -y_6^{PM} + x_6 + 1 = 0.451\ 188\ 65$$

$$y_6^{C} = \frac{1}{8} \times (9y_5 - y_3) + \frac{0.3}{8} \times (f_6^{PM} + 2f_5 - f_4) = 1.148\ 811\ 44$$

$$y_6 = y_6^{C} - \frac{9}{121} \times (y_6^{C} - y_6^{P}) = 1.148\ 811\ 61$$

误差

$$|y(x_5)-y_5|=4.97\times10^{-8},\ |y(x_6)-y_6|=2.61\times10^{-8}$$

从结果看,此方法误差比四阶 Adams 隐式法和四阶 Hamming 方法小,这与理论分析一致.

7.6 一阶方程组与高阶方程数值方法

考虑一阶常微分方程组的初值问题

$$\begin{cases}\dfrac{dy_i}{dx}=f_i(x,y_1,\cdots,y_N) & x\in[x_0,b]\\ y(x_0)=y_{i0} & i=1,2,\cdots,N\end{cases}\tag{7.6.1}$$

若用向量形式表示,可记为 $\boldsymbol{y}=(y_1,y_2,\cdots,y_N)^{\mathrm{T}}$,$\boldsymbol{f}=(f_1,f_2,\cdots,f_N)^{\mathrm{T}}$,初始条件 $\boldsymbol{y}(x_0)=\boldsymbol{y}_0=(y_{10},y_{20},\cdots,y_{N0})^{\mathrm{T}}$,于是(7.6.1)可写成

$$\begin{cases}\dfrac{d\boldsymbol{y}}{dx}=\boldsymbol{f}(x,\boldsymbol{y}), & x\in[x_0,b],\boldsymbol{y}\in\mathbf{R}^N\\ \boldsymbol{y}(x_0)=\boldsymbol{y}_0\end{cases}\tag{7.6.2}$$

(7.6.2)形式上同初值问题(7.1.1)类似,只要看成向量方程即可.因此前面关于单个方程的初值问题数值方法均适用于方程组(7.6.2),相应理论也可类似地得到.

下面仅对如下两个数值方法作说明.

梯形法

$$\boldsymbol{y}_{n+1}=\boldsymbol{y}_n+\frac{1}{2}h\left[\boldsymbol{f}(x_n,\boldsymbol{y}_n)+\boldsymbol{f}(x_{n+1},\boldsymbol{y}_{n+1})\right]\tag{7.6.3}$$

四阶 R－K 方法

$$\boldsymbol{y}_{n+1}=\boldsymbol{y}_n+\frac{h}{6}(\boldsymbol{k}_1+2\boldsymbol{k}_2+2\boldsymbol{k}_3+\boldsymbol{k}_4)\tag{7.6.4}$$

其中

$$\boldsymbol{k}_1=\boldsymbol{f}(x_n,\boldsymbol{y}_n)$$

$$\boldsymbol{k}_2=\boldsymbol{f}\left(x_n+\frac{h}{2},\boldsymbol{y}_n+\frac{h}{2}\boldsymbol{k}_1\right)$$

$$\boldsymbol{k}_3=\boldsymbol{f}\left(x_n+\frac{h}{2},\boldsymbol{y}_n+\frac{h}{2}\boldsymbol{k}_2\right)$$

$$\boldsymbol{k}_4=\boldsymbol{f}(x_n+h,\boldsymbol{y}_n+h\boldsymbol{k}_3)$$

这些公式形式上与单个方程初值问题的数值方法完全一样,但注意这里 $\boldsymbol{y}$ 及 $\boldsymbol{f}$ 均为 N 维向量,在计算机上求解时可直接从数学软件库中选择所要方法,它们都是按方程组初值问题编写的,当 $N=1$ 时就是前面讨论的单个方程情形.

对于高阶微分方程初值问题,原则上总可归结为一阶方程组,例如下列 m 阶微分方程

$$y^{(m)}=f(x,y,y',\cdots,y^{(m-1)})\tag{7.6.5}$$

初始条件为

$$y(x_0)=y_0,y'(x_0)=y_0',\cdots,y^{(m-1)}(x_0)=y_0^{(m-1)}\tag{7.6.6}$$

只要引进新变量

$$y_1 = y, y_2 = y', \cdots, y_m = y^{(m-1)}$$

则可将 m 阶方程(7.6.5)化为如下一阶方程组

$$\begin{cases} y_1' = y_2 \\ y_2' = y_3 \\ \quad \vdots \\ y_{m-1}' = y_m \\ y_m' = f(x, y_1, y_2, \cdots, y_m) \end{cases} \tag{7.6.7}$$

初始条件(7.6.6)则相应化为

$$y_1(x_0) = y_0, y_2(x_0) = y_0', \cdots, y_m(x_0) = y_0^{(m-1)} \tag{7.6.8}$$

习 题 七

1. 用 Euler 法解初值问题

$$y' = x^2 + 100y^2, y(0) = 0$$

取步长 $h = 0.1$,计算到 $x = 0.3$(保留到小数点后 4 位).

2. 用改进 Euler 法和梯形法解初值问题

$$y' = x^2 + x - y, y(0) = 0$$

取步长 $h = 0.1$,计算到 $x = 0.5$,并与准确解 $y = -e^{-x} + x^2 - x + 1$ 相比较.

3. 用改进 Euler 法计算积分

$$y = \int_0^x e^{-t^2} dt$$

在 $x = 0.5, 0.75, 1$ 时的近似值(保留到小数点后 6 位).

4. 试证明梯形法中的迭代公式(7.2.7)的收敛性.设条件(7.1.2)及 $\beta hL < 1$ 成立.

5. 对模型方程 $y' = \lambda y (\lambda < 0)$,证明隐式 Euler 法(7.2.5)对任何步长 $h > 0$ 绝对稳定.

6. 证明中点公式(7.3.9)

$$y_{n+1} = y_n + hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1\right), k_1 = f(x_n, y_n)$$

是二阶的,并求其局部截断误差主项.

7. 用四阶 R－K 方法求解初值问题

$$y' = 3y/(1+x), 0 \leqslant x \leqslant 1, y(0) = 1$$

取步长 $h = 0.2$.

8. 对于初值问题

$$y' = -100(y - x^2) + 2x, y(0) = 1$$

(1) 用 Euler 法求解,步长 h 应取在什么范围内计算才稳定?

(2) 若用梯形法求解,对步长 h 有无限制?

(3) 若用四阶 R－K 方法求解,步长 h 如何选取?

9. 用四步四阶的 Adams 显式方法

$$y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})$$

求解初值问题

$$y' = 3x - 2y, 0 \leqslant x \leqslant 0.5, y(0) = 1$$

取 $h = 0.1$.

10. 证明线性二步法

$$y_{n+1} + (b-1)y_n - by_{n-1} = \frac{1}{4}h[(b+3)f_{n+1} + (3b+1)f_{n-1}]$$

当 $b \neq -1$ 时方法为二阶，当 $b = -1$ 时方法为三阶.

11. 用形如

$$y_{n+1} = \alpha(y_n + y_{n-1}) + h(\beta_0 f_n + \beta_1 f_{n-1})$$

的线性二步法解

$$y' = f(x, y), y(x_0) = y_0$$

试确定参数 α, β_0, β_1，使方法具有尽可能高的阶数，并求出局部截断误差主项.

计算实验题

要求

1. 用 Matlab 语言或你熟悉的其他算法语言编程序，使之尽量具有通用性.

2. 上机前充分准备，复习有关算法，写出计算步骤，反复查对程序，列出上机步骤.

3. 完成计算后写出计算实验报告，内容包括：计算机型号和所用机时，算法步骤描述，变量说明，程序清单，输出计算结果，结果分析和小结等.

4. 根据教师要求选做下列习题中的 2～3 道题.

(一) 求下列方程的实根，准确到 10^{-6}.

(1) $x^2-3x+2-e^x=0$

(2) $x^3+2x^2+10x-20=0$

实验要求

(1) 用自己设计的一种线性收敛迭代法计算，然后再用 Steffensen 加速迭代法计算.

(2) 用 Newton 法计算，要求输出迭代初值、各次迭代值及迭代次数. 比较各方法优缺点.

(二) 给定方程组

(1) $$\begin{bmatrix} 3.01 & 6.03 & 1.99 \\ 1.27 & 4.16 & -1.23 \\ 0.987 & -4.81 & 9.34 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

(2) 将(1)中系数矩阵中系数 3.01 改为 3.00，0.987 改为 0.990，其他元素不变.

(3) $$\begin{bmatrix} 10 & -7 & 0 & 1 \\ -3 & 2.099\,999 & 6 & 2 \\ 5 & -1 & 5 & -1 \\ 2 & 1 & 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}=\begin{bmatrix} 8 \\ 5.900\,001 \\ 5 \\ 1 \end{bmatrix}$$

实验要求

(1) 用 $\boldsymbol{LU}$ 分解和列主元 Gauss 消去法分别求解上述三个方程组.

(2) 输出 $\boldsymbol{Ax}=\boldsymbol{b}$ 中矩阵 $\boldsymbol{A}$ 及向量 $\boldsymbol{b}$，$\boldsymbol{A}=\boldsymbol{LU}$ 分解的 $\boldsymbol{L}$ 及 $\boldsymbol{U}$，$\det \boldsymbol{A}$ 及解向量 $\boldsymbol{x}$.

(3) 输出列主元法行交换次序及解向量 $\boldsymbol{x}$ 和 $\det \boldsymbol{A}$，并与(2)的结果比较.

(三) 给定列表函数如下

x	0.2	0.4	0.6	0.8	1.0
$f(x)$	0.979 865 2	0.917 771 0	0.808 034 8	0.638 609 3	0.384 373 5

试求 $f(x)$的三次样条插值函数 $s(x)$,分别满足:

(1) 自然边界条件. (2) $f'(0,2)=0.202\ 71, f'(1,0)=1.557\ 41$.

实验要求

(1) 用追赶法求三弯矩方程的解向量$(M_0, M_1, \cdots, M_4)$.

(2) 求 $s(0.2+0.1i)$的值,$i=0,1,\cdots,8$.

(四) 给出积分

(1) $\int_0^2 x^2 e^{-x^2} dx$ (2) $\int_{\frac{\pi}{2}}^{\frac{3}{4}\pi} \cot x dx$

实验要求

(1) 用 Romberg 算法计算上面积分,到$\left| T_{k-1}^{(k-1)} - T_k^{(k)} \right| < 10^{-6}$时结束,要求输出 T 表.

(2) 用 5 点 Gauss 求积公式计算,输出计算结果.

(3) 分析比较计算结果.

(五) 给定初值问题

(1) $\begin{cases} y' = \dfrac{1}{x^2} - \dfrac{y}{x}, & 1 \leqslant x \leqslant 2 \\ y(1) = 1 \end{cases}$

(2) $\begin{cases} y' = -50y + 50x^2 + 2x, & 0 \leqslant x \leqslant 1 \\ y(0) = \dfrac{1}{3} \end{cases}$

实验要求

(1) 用改进 Euler 法($h=0.05$)及四阶经典 Runge-Kutta 法($h=0.1$),求(1)的数值解,并打印 $x=1+0.1i(i=0,1,\cdots,10)$的值.

(2) 用经典四阶 R-K 方法解(2).步长分别取为 $h=0.1, h=0.025, h=0.01$,计算并打印 $x=0.1i(i=0.1,0.2,\cdots,1)$各点的数值解及准确解,并分析结果.(初值问题(2)的准确解为 $y(x)=\dfrac{1}{3}e^{-50x}+x^2$)

参考文献

[1] 李庆扬.数值分析复习与考试指导.北京:高等教育出版社,2000
[2] 李庆扬,王能超,易大义.数值分析(第四版).北京:清华大学出版社,2001
[3] 关治,陆金甫.数值分析基础.北京:高等教育出版社,1998
[4] 李庆扬,易大义,王能超.现代数值分析.北京:高等教育出版社,1995