

114

·研究生数学教材·

0241-43
SF8(2)

数值分析

(第2版)

孙志忠 袁慰平 闻震初 编著



A0966804

东南大学出版社

内 容 提 要

本书着重介绍适合于电子计算机上采用的数值计算方法及其理论。其内容有:误差分析,非线性方程求根,线性方程组数值解法,矩阵特征值与特征向量计算,函数插值与逼近,数值积分与数值微分,常微分方程数值解法,偏微分方程数值解法等。内容覆盖了国家教委工科研究生数学课程教学指导小组所制订的工科硕士生数值分析课程教学基本要求,同时还增加了一些工科专业所需要的内容,如机器数系、有理函数插值、FFT变换、振荡函数积分等。书中对各种计算方法的构造思想都作了较详细的阐述,对稳定性、收敛性、误差估计以及算法的优缺点等也作了适当的讨论。本书还挑选了部分东南大学工科研究生结合各自专业自选课题的计算实习,以此作为本书各章的应用实例。

本书可作为工科研究生的教材,也可供从事科学与工程计算的科技工作者阅读参考。

图书在版编目(CIP)数据

数值分析/孙志忠等编著. —2版. —南京:东南
大学出版社,2002.1

ISBN 7-81050-931-4

I. 数... II. 孙... III. 计算方法-高等学校-教
材 IV. 0241

中国版本图书馆 CIP 数据核字(2002)第 002020 号

东南大学出版社出版发行
(南京四牌楼2号 邮编210096)

出版人:宋增民

江苏省新华书店经销 华东有色地研所印刷厂印刷
开本 700mm×1000mm 1/16 印张 24.5 字数 494千字

2002年2月第2版 2002年2月第1次印刷

印数:5000 定价:29.8元

第 2 版说明

本书自 1992 年出版以来已有近 10 年,在此期间得到兄弟院校同行老师和读者的关心与支持,我们受到很大鼓舞,在此深表感谢。

在第 2 版的编写过程中汲取了广大读者的宝贵意见和我们在教学实践中的经验和体会,对第 1 版内容作了如下修改:

1. 各章中都有部分内容次序的调整以及例题的增减。
2. 第 1 章中增加了机器数的运算规则,以使学生对误差的传播及数值稳定性能理解得更深刻些。
3. 第 2 章中改写了二分法和劈因子法,给出了迭代发散的充分条件和 Aitken 加速算法收敛定理、Newton 迭代法大范围收敛定理的证明。
4. 第 3 章中增加了三对角方程组并行算法一节,使学生对并行计算机的数值方法有所了解。
5. 第 4 章中删去了:向后差分和中心差分的概念以及用其表示等距节点插值多项式的方法、样条插值中的三转角方程组、最佳一致逼近的 Remez 算法和和正交多项式这一节;充实了分段插值,增加了样条插值收敛性定理的证明;将 Chebyshev 多项式移至最佳一致逼近一节中;Legendre 多项式移至第 5 章 Gauss 求积公式一节中;改写和充实了最佳平方逼近。
6. 第 5 章中改写了复化求积公式和 Romberg 求积公式,特别强调了误差后验估计式在实际计算中的应用;改写了振荡函数的积分和重积分的近似计算;增加了插值型求导公式截断误差的表达式及其证明。
7. 第 6 章中统一地定义了显式公式、隐式公式以及单步公式、多步公式的局部截断误差和整体截断误差,减弱了单步法收敛定理中对增量函数所加的限制性条件,改写了单步法的自适应算法和加速算法,删去了自动变步长 Adams 预测校正算法,改写了线性多步法并增加了多步法的收敛性和稳定性。
8. 第 7 章中删去了矩阵广义特征值的计算一节。
9. 改写了第 8 章。对于差分方法强调如何建立差分格式及如何求解,给出了稳定性和收敛性的定义及有关结果;对于有限元方法,介绍了变分原理、Ritz - Galerkin 方法和有限元法的基本思想。
10. 各章调整和增删了部分习题,且每章均增列了一道上机作业题。

孙志忠 袁慰平 阎震初

2001 年 10 月

前 言

随着计算机的广泛使用与科学技术的迅速发展,使用计算机进行科学计算已成为科学研究、工程设计中越来越不可缺少的1个环节,它有时甚至代替或超过了实验所起的作用。因此,科学计算应该成为高级科技人员的1个基本功。为此,作为科学计算的核心——“数值分析”已被较多的工科专业列为硕士研究生的一门学位课程。

我校自1981年开设此课程以来,进行了专业需要的调查以及教学内容和教学方法的改革与研究,并在教学实践的基础上形成了本书的初稿。初稿在本校使用过多遍,得到研究生及其指导教师们的良好反应与支持。在这次出版时,再一次总结教学实践中的经验,作了修改和加工。

本书着重介绍了适合于电子计算机上采用的计算方法的构造及使用,对误差估计、方法的收敛性、稳定性、适用范围及优缺点等作了适当分析,对一些解法作了比较详细的推导并列举了较多的数值计算实例。其内容覆盖了国家教委工科研究生数学课程教学指导小组所制订的工科硕士研究生数值分析课程教学基本要求,同时还增加了一些工科专业所需要的内容,如误差分析中的机器数系;非线性方程求根的 Sturm 定理;插值与逼近中的重节点插值、有理函数插值及最佳一致逼近;数值积分中的振荡函数的积分、重积分的近似计算;常微分方程中的自适应算法及稳定性较好的单步隐格式;特征值计算中的广义特征值计算等等。增加的这些内容,若教学课时数较少则可以删去,不会影响教学的连续性;但若增加了这些内容,则在内容的深广度上更能适应工科专业发展的需要。因此,本书兼顾了多学时与少学时的要求。

本书每章末(除第1、7两章外)都有应用实例一节,取材于历届学生结合各自专业的自选课题。我们认为研究生学习“数值分析”课,一方面是提高数学素质,另一方面是为应用所学知识进行科学研究。于是我们要求学生在学习完本课程后,必须结合所学内容及各自的专业自选研究课题做1个大型计算实习。这次,在讲义修改出版过程中,我们将历届学生做的大型计算实习,挑选了部分作为实例进入教材,有的实例尽管被我们作了简化和加工,但仍然可以从这些实例中初步看到本课程在各专业应用中的深广度。

本书的读者对象是工科研究生及从事数值计算的科技工作者。本书编写分工如下:第1、3、5章由袁慰平编写,第2、6章由黄新芹编写,第4章由闻震初编写,第7、8章由张令敏编写。

南京大学苏煜城教授仔细审阅了全书,并提出了宝贵的意见,在苏先生的指点下,我们作了修改。在本书的形成、使用、修改和出版过程中,得到东南大学研究生院和东南大学出版社及有关工科研究生指导教师和本教研室同志的关心、支持和帮助,在此一并表示衷心感谢。

由于作者水平有限,缺点与错误在所难免,恳请读者批评指正。

编 者

1992年4月

目 录

1 绪论	(1)
1.1 数值分析的对象和特点	(1)
1.2 误差的基本概念	(1)
1.2.1 误差的来源	(1)
1.2.2 绝对误差	(2)
1.2.3 相对误差	(2)
1.2.4 有效数	(3)
1.2.5 数据误差对函数值的影响	(4)
1.3 机器数系	(6)
1.3.1 机器数系	(6)
1.3.2 机器数系的运算及误差估计	(8)
1.4 数值稳定问题	(11)
1.4.1 数值稳定性	(11)
1.4.2 病态问题	(14)
1.4.3 简化计算步骤,减少运算次数	(16)
习题 1	(17)
2 非线性方程的解法	(21)
2.1 概述	(21)
2.1.1 根的搜索	(21)
2.1.2 二分法	(22)
2.2 简单迭代法	(23)
2.2.1 迭代格式的构造	(24)
2.2.2 迭代法的收敛性	(25)
2.2.3 迭代法的收敛速度	(29)
2.2.4 Aitken 加速法	(31)
2.3 Newton 法	(34)
2.3.1 Newton 迭代格式及其几何意义	(34)
2.3.2 局部收敛	(35)
2.3.3 大范围收敛	(36)
2.3.4 求重根的修正 Newton 法	(40)

2.3.5	Newton 法的变形	(41)
2.4	多项式方程的求根	(43)
2.4.1	实系数多项式零点的分布	(43)
2.4.2	劈因子法	(46)
2.5	应用实例:薄壳结构的静力计算	(49)
2.5.1	问题的背景	(49)
2.5.2	数学模型	(50)
2.5.3	计算方法与结果分析	(51)
	习题 2	(53)
3	线性代数方程组数值解法	(56)
3.1	引言	(56)
3.2	消去法	(57)
3.2.1	三角方程组的解法	(57)
3.2.2	Gauss 消去法	(58)
3.2.3	追赶法	(63)
3.2.4	列主元 Gauss 消去法	(64)
3.3	矩阵的直接分解法	(66)
3.3.1	矩阵的直接分解法	(66)
3.3.2	对称矩阵的直接分解法	(70)
3.3.3	列主元的三角分解法	(73)
3.4	方程组的性态与误差分析	(75)
3.4.1	向量范数	(76)
3.4.2	矩阵范数	(78)
3.4.3	方程组的性态及条件数	(85)
3.4.4	方程组近似解可靠性的判别	(88)
3.5	迭代法	(90)
3.5.1	迭代格式的一般形式	(90)
3.5.2	几个常用的迭代格式	(90)
3.5.3	迭代格式的收敛性	(94)
3.6	三对角方程组的并行算法	(99)
3.7	应用实例:纯电阻型立体电路分析	(103)
3.7.1	问题的背景	(103)
3.7.2	数学模型	(103)
3.7.3	计算方法与结果分析	(105)

习题 3	(107)
4 插值与逼近	(115)
4.1 Lagrange 插值	(115)
4.1.1 基本插值多项式	(115)
4.1.2 Lagrange 插值多项式	(116)
4.1.3 插值多项式的存在唯一性	(117)
4.1.4 插值余项	(117)
4.2 差商、差分和 Newton 插值	(120)
4.2.1 差商及 Newton 插值多项式	(121)
4.2.2 差分及等距节点 Newton 插值多项式	(125)
4.3 Hermite 插值	(127)
4.3.1 Lagrange 型 Hermite 插值多项式	(128)
4.3.2 Newton 型 Hermite 插值多项式	(131)
4.4 高次插值的缺点及分段插值	(134)
4.4.1 高次插值的误差分析	(134)
4.4.2 分段线性插值	(137)
4.4.3 分段 Hermite 插值	(138)
4.5 3 次样条插值	(139)
4.5.1 3 次样条插值函数	(139)
4.5.2 3 次样条插值函数的求法	(140)
4.5.3 3 次样条插值函数的收敛性	(144)
4.6 有理函数插值	(147)
4.7 最佳一致逼近	(153)
4.7.1 线性赋范空间	(153)
4.7.2 最佳一致逼近多项式	(154)
4.7.3 Chebyshev 多项式	(158)
4.7.4 近似最佳一致逼近多项式	(160)
4.8 最佳平方逼近	(162)
4.8.1 内积空间	(163)
4.8.2 离散数据的最佳平方逼近	(164)
4.8.3 超定方程组的最小二乘解	(167)
4.8.4 连续函数的最佳平方逼近	(169)
4.9 周期函数的逼近与快速 Fourier 变换	(171)
4.9.1 最佳平方三角逼近与三角插值	(171)

4.9.2	快速 Fourier 变换	(172)
4.10	应用实例:用样条函数设计公路平面曲线	(175)
4.10.1	问题的背景	(175)
4.10.2	数学模型	(175)
4.10.3	计算方法与结果分析	(176)
习题 4	(178)
5	数值积分与数值微分	(183)
5.1	数值积分的基本概念	(183)
5.2	插值型求积公式	(184)
5.2.1	插值型求积公式	(184)
5.2.2	代数精度	(187)
5.2.3	梯形公式、Simpson 公式和 Cotes 公式的截断误差	(190)
5.3	复化求积公式	(191)
5.3.1	复化梯形公式	(191)
5.3.2	复化 Simpson 公式	(195)
5.3.3	复化 Cotes 公式	(196)
5.3.4	复化求积公式的阶	(197)
5.4	Romberg 求积法	(198)
5.4.1	Romberg 求积公式	(198)
5.4.2	Romberg 求积法的一般公式	(201)
5.5	Gauss 求积公式	(202)
5.5.1	Gauss 求积公式	(203)
5.5.2	正交多项式	(206)
5.5.3	区间 $[-1, 1]$ 上的 Gauss 公式	(208)
5.5.4	区间 $[a, b]$ 上的 Gauss 公式	(209)
5.5.5	Gauss 公式的余项	(210)
5.5.6	Gauss 公式的稳定性和收敛性	(211)
5.5.7	带权积分	(213)
5.6	振荡函数的积分	(215)
5.7	重积分的近似计算	(220)
5.8	数值微分	(225)
5.8.1	数值微分问题的提出	(225)
5.8.2	插值型求导公式	(226)
5.8.3	样条求导	(229)

5.9	应用实例:混频器中变频损耗的数值计算	(230)
5.9.1	问题的背景	(230)
5.9.2	数学模型	(231)
5.9.3	计算方法与结果分析	(232)
	习题 5	(233)
6	常微分方程数值解法	(237)
6.1	微分方程数值解法概述	(237)
6.1.1	问题及基本假设	(237)
6.1.2	离散化方法	(238)
6.2	Euler 方法	(238)
6.2.1	Euler 公式	(238)
6.2.2	后退 Euler 公式	(241)
6.2.3	梯形公式	(242)
6.2.4	预测校正系统与改进 Euler 公式	(243)
6.2.5	整体截断误差	(245)
6.3	Runge - kutta 方法	(246)
6.3.1	Runge - kutta 方法的基本思想	(246)
6.3.2	2 阶 Runge - kutta 公式	(248)
6.3.3	高阶 Runge - kutta 公式	(250)
6.3.4	隐式 Runge - kutta 公式	(252)
6.4	单步方法的收敛性和稳定性	(253)
6.4.1	单步方法的收敛性	(253)
6.4.2	单步方法的稳定性	(256)
6.4.3	单步方法的自适应算法	(256)
6.4.4	单步方法的加速	(257)
6.5	线性多步法	(258)
6.5.1	基于数值积分的构造方法	(259)
6.5.1.1	Adams 显式公式	(259)
6.5.1.2	Adams 隐式公式	(261)
6.5.1.3	Adams 预测校正方法	(263)
6.5.1.4	Adams 公式的加速	(264)
6.5.2	基于 Taylor 展开的待定系数方法	(265)
6.5.3	多步法的收敛性和稳定性	(267)
6.5.4	绝对稳定性和绝对稳定域	(269)

6.6	1 阶微分方程组与高阶微分方程	(270)
6.6.1	1 阶微分方程组	(270)
6.6.2	高阶微分方程	(271)
6.6.3	刚性问题	(273)
6.7	边值问题的数值解法	(275)
6.7.1	试射法	(276)
6.7.2	差分法	(277)
6.8	应用实例:磁流体发电通道的数值计算	(280)
6.8.1	问题的背景	(280)
6.8.2	数学模型	(280)
6.8.3	计算方法与结果分析	(281)
	习题 6	(282)
7	矩阵特征值计算	(286)
7.1	引言	(286)
7.2	幂法及反幂法	(288)
7.2.1	求主特征值的幂法	(288)
7.2.2	幂法的加速技巧	(294)
7.2.3	反幂法	(297)
7.3	实对称矩阵的 Jacobi 法	(299)
7.3.1	Jacobi 法	(300)
7.3.2	Jacobi 法的变形	(305)
7.4	Givens 法和 Householder 法	(306)
7.4.1	把实对称矩阵约化为三对角阵	(306)
7.4.2	Sturm 序列与二分法	(311)
7.5	QR 算法	(313)
7.5.1	基本算法	(313)
7.5.2	具有位移的 QR 算法	(314)
	习题 7	(316)
8	偏微分方程的数值解法	(318)
8.1	抛物型方程的差分解法	(318)
8.1.1	网格剖分	(319)
8.1.2	古典显格式	(320)
8.1.3	古典隐格式	(322)
8.1.4	Crank - Nicolson 格式	(323)

8.1.5 Richardson 格式	(326)
8.2 差分格式的稳定性 and 收敛性	(331)
8.2.1 差分格式的稳定性	(331)
8.2.2 差分格式的收敛性	(338)
8.3 双曲型方程的差分解法	(339)
8.3.1 显格式	(340)
8.3.2 隐格式	(343)
8.4 椭圆型方程的差分解法	(346)
8.4.1 差分方程的建立	(346)
8.4.2 差分格式解的先验估计及其收敛性	(349)
8.5 变分原理	(352)
8.5.1 2 次函数的极值	(352)
8.5.2 边值问题的变分原理	(354)
8.6 Ritz - Galerkin 方法	(357)
8.7 有限元法	(360)
8.7.1 三角形剖分	(360)
8.7.2 试探函数空间的构造	(361)
8.7.3 试探函数空间的基函数	(362)
8.7.4 有限元方程	(363)
8.8 应用实例:水污染方程的有限差分解法	(365)
8.8.1 问题的背景	(365)
8.8.2 数学模型	(366)
8.8.3 计算方法与结果分析	(366)
习题 8	(367)
参考文献	(371)

1 绪 论

1.1 数值分析的对象和特点

数值分析是寻求数学问题近似解的方法、过程及其理论的一个数学分支。当今世界计算机已被广泛使用,因此数值分析所研究的应该是适合于计算机上使用的计算方法及其误差分析和收敛性、稳定性问题。

使用计算机通过计算方法或数值模拟的手段去解决科学或工程中的关键问题,简称为科学计算。它已成为科学研究、工程设计等越来越不可缺少的1个环节,有时甚至代替或超过了实验所起的作用。

最近半个世纪科学研究的实践使人们越来越清楚地认识到,当代科学研究方法论应该由实验、科学计算及理论三大环节所组成。也就是说,科学计算已成为一种新的科学研究方法。因此,作为科学计算的主体——数值分析也就越来越被人们所重视。

1.2 误差的基本概念

1.2.1 误差的来源

一个物理量的真实值和我们算出的值往往不相等,它们之差称为误差。引起误差的原因是多方面的。

(1) **模型误差** 将实际问题转化为数学问题,就是所谓的建立数学模型时,对被描述的实际问题进行了抽象和简化,因此数学模型只是客观现象的一种近似、粗糙的描述。这种数学模型与实际问题之间出现的误差称为模型误差。

(2) **观测误差** 在给出的数学模型中往往涉及一些根据观测得到的物理量,如电压、电流、温度、长度等,而观测不可避免地会带有误差。这种误差称为观测误差。

(3) **截断误差** 在计算中常常遇到只有通过无限过程才能得到的结果,但实际计算时,只能采用有限过程,如无穷级数求和,只能取前面有限项之和来近似代替,于是产生了有限过程代替无限过程的误差,称为截断误差。这是计算方法本身所带来的误差,所以也称为方法误差。

(4) **舍入误差** 在计算中遇到的数据可能位数很多,也可能是无穷小数,如 $\sqrt{2} = 1.414\ 213\ 56\dots$, $\frac{1}{3} = 0.333\ 3\dots$,但计算时只能对有限位进行运算,一般采用四舍

五入的办法。电子计算机计算时也有采用截尾的办法,如 $\sqrt{2}$ 在8位字长的截断机里取成1.414 213 5。这类误差称为舍入误差,也称计算误差。

少量的舍入误差是微不足道的,但在电子计算机上作了成千上万次运算后,舍入误差的积累有时可能是十分惊人的。

由上述误差来源的分析可以得到如下结论:误差是不可避免的,要求绝对准确、绝对严格实际上是办不到的。对于实际问题,在建立数学模型时,本身已存在着模型误差和观测误差,因此既然描述问题的方法是近似的,那么要求解的绝对准确也就没有多大意义了。于是我们在计算方法里所讨论的求解都只要求出近似解,那种认为近似解不可靠、不准确的想法是错误的,应该认为求近似解是正常的,需要研究的问题是尽量设法减少误差,提高精度。从以上误差的四种来源的分析可以知道前两种误差是客观存的,后两种误差是由计算方法所引起的。本课程是研究数学问题的数值解法,因此只涉及后两种误差。

1.2.2 绝对误差

定义 1.1 设 x^* 为准确值, x 是 x^* 的 1 个近似值。称 $e = x^* - x$ 为近似值 x 的绝对误差,简称误差。

注意这样定义的误差 e 可正可负,所以绝对误差不是误差绝对值。通常我们不能算出准确值 x^* ,也不可能算出误差的准确值,因此这个值虽然客观存在,但在实际计算中是很难得到的,而得到的往往是误差的某个范围,即根据测量工具或计算情况估计出误差的绝对值不超过某正数 ϵ ,即

$$|e| = |x^* - x| \leq \epsilon$$

称 ϵ 为近似值 x 的绝对误差限,简称误差限。有时也表示成 $x^* = x \pm \epsilon$ 。

例如用毫米刻度的直尺测量一长度为 x^* 的物体,测得其长度的近似值为 $x = 20$ mm,由于直尺以毫米为刻度,所以其误差不超过 0.5 mm,即

$$|x^* - 20| \leq 0.5$$

从这个不等式我们不能得出准确值 x^* ,但却知道 x^* 的范围为

$$19.5 \leq x^* \leq 20.5$$

对于给定的正数 ϵ ,若近似值 x 满足

$$|x^* - x| \leq \epsilon$$

则在 ϵ 范围内认为 x 就是 x^* ,也即近似值 x 和真值 x^* 关于允许误差 ϵ 可以看成是“重合”的,或者说值 x 关于允许误差 ϵ 是准确的。

1.2.3 相对误差

对于不同的物理量绝对误差限的大小不能完全表示出近似值的精确程度。为了更好地反映近似值的精确程度,必须考虑绝对误差与真值之比。

定义 1.2 设 x^* 为准确值, x 是 x^* 的 1 个近似值。称

$$e_r = \frac{x^* - x}{x^*}$$

为近似值 x 的相对误差。

在实际计算中,通常真值 x^* 总是难以求得的。人们常以

$$\bar{e}_r = \frac{x^* - x}{x}$$

作为相对误差。事实上,

$$\bar{e}_r - e_r = \frac{\bar{e}_r^2}{1 + \bar{e}_r} = \frac{e_r^2}{1 - e_r}$$

因而当 \bar{e}_r 和 e_r 有一为小量时, $\bar{e}_r - e_r$ 为该小量的 2 阶小量。

计算相对误差与计算绝对误差具有相同的困难,因此通常也只能考虑相对误差限,即如果有正数 ϵ_r ,使

$$|e_r| \leq \epsilon_r \quad \text{或} \quad |\bar{e}_r| \leq \epsilon_r$$

则称 ϵ_r 为 x 的相对误差限。

绝对误差只能用来比较对同一量所测得的不同近似值的准确程度,而相对误差却能用来刻画或比较任何近似值的准确程度。

1.2.4 有效数

工程技术中对于测量得到的数经常表示成 $x \pm \epsilon$,它虽然表示了近似值 x 的准确程度,但用这个量进行数值计算就太麻烦了。因此希望所表示的数本身就能显示出它的准确程度,于是需要引进有效数的概念。

定义 1.3 如果近似值 x 的误差限是其某 1 位上的半个单位,且该位直到 x 的第一位非零数字一共有 n 位,则称近似值 x 具有 n 位有效数字,用这 n 位有效数字表示的近似数称为有效数(如图 1.1 所示)。

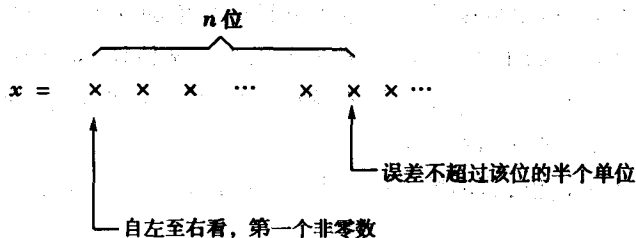


图 1.1 有效位数

如 π 的近似值取 $x_1 = 3.14$, 则 x_1 有 3 位有效数字; 取 $x_2 = 3.1416$, 则 x_2 有 5 位有效数字; 取 $x_3 = 3.1415$, 则 x_3 只有 4 位有效数字。

在讲了有效数字之后,我们从此规定所写出的数都应该是有效数字,如 π 的近似值应是 3.14 或 3.142 或 3.1416, 不能是 3.1415。

在科学记数法中通常将 n 位有效数字 x 表示成

$$x = \pm 0.a_1 a_2 \cdots a_n \times 10^m$$

即

$$x = \pm (a_1 \times 10^{-1} + a_2 \times 10^{-2} + \cdots + a_n \times 10^{-n}) \times 10^m \quad (1.1)$$

其中 m 为一整数, a_1, a_2, \cdots, a_n 都是 0 到 9 中的整数, 且 $a_1 \neq 0$ 。

按式(1.1)表示的有效数 x , 其误差为

$$|x^* - x| \leq \frac{1}{2} \times 10^{m-n} \quad (1.2)$$

所以 x 的误差限为 $\epsilon = \frac{1}{2} \times 10^{m-n}$, 因此在 m 相同的情况下, n 越大则误差越小, 亦即说明 1 个近似值的有效位数越多其误差限越小。由

$$\frac{|x - x^*|}{|x|} \leq \frac{\frac{1}{2} \times 10^{m-n}}{a_1 \times 10^{-1} \times 10^m} = \frac{1}{2a_1} \times 10^{-n+1}$$

知 x 的相对误差限为

$$\epsilon_r = \frac{1}{2a_1} \times 10^{-n+1} \quad (1.3)$$

式(1.3)表明 1 个近似值的有效位数越多, 其相对误差限也越小。

由于 $\frac{1}{a_1} \leq 1$, 所以有时也简单地取

$$\epsilon_r = \frac{1}{2} \times 10^{-n+1} \quad (1.4)$$

作为 n 位有效数字 x 的相对误差限。

1.2.5 数据误差对函数值的影响

数值运算中由于所给数据的误差必然引起函数值的误差, 且这种数据误差的影响较为复杂, 所以一般采用 Taylor 级数展开的方法来估计, 如计算 $y = f(x_1, x_2)$ 的值。设给定值(即数据) x_1, x_2 是近似值, 则由此计算得到的 y 也只能是近似值。现在来研究 y 的绝对误差和相对误差。

设 x_1^*, x_2^* 为准确值, 其函数准确值为 $y^* = f(x_1^*, x_2^*)$ 。于是函数值 y 的误差是

$$e(y) = y^* - y = f(x_1^*, x_2^*) - f(x_1, x_2)$$

将 $f(x_1^*, x_2^*)$ 在 (x_1, x_2) 处作 Taylor 展开, 并取 1 阶 Taylor 多项式, 得 $e(y)$ 的近似表达式

$$e(y) \approx \frac{\partial f(x_1, x_2)}{\partial x_1} (x_1^* - x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} (x_2^* - x_2)$$

或

$$e(y) \approx \frac{\partial f(x_1, x_2)}{\partial x_1} e(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} e(x_2) \quad (1.5)$$

式中

$$e(x_1) = x_1^* - x_1, \quad e(x_2) = x_2^* - x_2$$

由此可得函数值 y 的相对误差

$$e_r(y) = \frac{e(y)}{y} \approx \frac{\partial f(x_1, x_2)}{\partial x_1} \frac{1}{y} e(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{1}{y} e(x_2)$$

或

$$e_r(y) \approx \frac{\partial f(x_1, x_2)}{\partial x_1} \frac{x_1}{f(x_1, x_2)} e_r(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{x_2}{f(x_1, x_2)} e_r(x_2) \quad (1.6)$$

式中

$$e_r(x_1) = \frac{x_1^* - x_1}{x_1}, \quad e_r(x_2) = \frac{x_2^* - x_2}{x_2}$$

利用函数值的误差估计式(1.5)和式(1.6)可以得到两数和、差、积、商的误差估计:

$$e(x_1 + x_2) \approx e(x_1) + e(x_2) \quad (1.7)$$

$$e(x_1 - x_2) \approx e(x_1) - e(x_2) \quad (1.8)$$

$$e(x_1 x_2) \approx x_2 e(x_1) + x_1 e(x_2) \quad (1.9)$$

$$e\left(\frac{x_1}{x_2}\right) \approx \frac{1}{x_2} e(x_1) - \frac{x_1}{x_2^2} e(x_2) \quad (x_2 \neq 0) \quad (1.10)$$

$$e_r(x_1 + x_2) \approx \frac{x_1}{x_1 + x_2} e_r(x_1) + \frac{x_2}{x_1 + x_2} e_r(x_2) \quad (1.11)$$

$$e_r(x_1 - x_2) \approx \frac{x_1}{x_1 - x_2} e_r(x_1) - \frac{x_2}{x_1 - x_2} e_r(x_2) \quad (1.12)$$

$$e_r(x_1 x_2) \approx e_r(x_1) + e_r(x_2) \quad (1.13)$$

$$e_r\left(\frac{x_1}{x_2}\right) \approx e_r(x_1) - e_r(x_2) \quad (x_2 \neq 0) \quad (1.14)$$

例 1.1 计算 $100 - \sqrt{9995}$ 。

解 记 $x_1^* = 100, x_2^* = \sqrt{9995}, x_1 = 100, x_2 = 99.974997$, 则 $e(x_1) = 0$, $|e(x_2)| \leq \frac{1}{2} \times 10^{-6}$, 即 x_1 是精确的, x_2 具有 8 位有效数。易知

$$x_1^* - x_2^* = \frac{5}{x_1^* + x_2^*}$$

第 1 种方法

$$x_1^* - x_2^* \approx x_1 - x_2 = 100 - 99.974997 = 0.025003$$

第 2 种方法

$$\frac{5}{x_1^* + x_2^*} \approx \frac{5}{x_1 + x_2} = \frac{5}{100 + 99.974997} = 0.0250031257\dots$$

现在来分析上述 2 种方法的精度。

由

$$\begin{aligned} |e(x_1 - x_2)| &\approx |e(x_1) - e(x_2)| \leq |e(x_1)| + |e(x_2)| \\ &\leq 0 + \frac{1}{2} \times 10^{-6} = \frac{1}{2} \times 10^{-6} \end{aligned}$$

知第 1 种方法所得结果具有 5 位有效数字。

由

$$\begin{aligned} \left| e\left(\frac{5}{x_1 + x_2}\right) \right| &\approx \left| -\frac{5}{(x_1 + x_2)^2} e(x_1 + x_2) \right| \\ &\approx \left| -\frac{5}{(x_1 + x_2)^2} [e(x_1) + e(x_2)] \right| \\ &\leq \frac{5}{(x_1 + x_2)^2} (|e(x_1)| + |e(x_2)|) \\ &\leq \frac{5}{(100 + 99.974997)^2} \left(0 + \frac{1}{2} \times 10^{-6} \right) \\ &= 0.62515629 \times 10^{-10} \\ &< 10^{-10} < \frac{1}{2} \times 10^{-9} \end{aligned}$$

知第 2 种方法所得结果具有 8 位有效数字,其 8 位有效数为 0.025003126。

由本例我们看出当两个相近的数相减时会造成有效位数的减少。事实上由式 (1.12) 可知当 x_1 和 x_2 相近时, $\frac{x_1}{x_1 - x_2}$ 和 $\frac{x_2}{x_1 - x_2}$ 的绝对值会很大。这时 $|e(x_1 - x_2)|$ 可能比 $|e(x_1)| + |e(x_2)|$ 大得多。因此,在实际计算中应当尽可能避免两相近数相减,否则会使计算精度大大降低。

1.3 机器数系

1.3.1 机器数系

电子计算机中数的表示大都采用浮点表示的形式,并以这种形式存贮和运算,这种形式与科学记数法非常相似。

设有一台计算机,有 n 位字长,采用 β 进制,阶码为 p , $L \leq p \leq U$, 这里 L, U 和 n 都是由该计算机的硬件所决定的某些常数,则在此计算机中数的浮点表示为

$$x = \pm (0.a_1 a_2 \dots a_n) \beta^p \quad (1.15)$$

亦即