



工业和信息化部“十二五”规划教材

现代数值计算

(第2版)

Advanced Numerical Computing (2nd Edition)

同济大学计算数学教研室 编著

- 以必须够用为尺度，以实际应用为目的
- 以基本原理为基础，以基本技术为主线
- 以MATLAB为平台，加强学生的编程能力



名家系列

人民邮电出版社
POSTS & TELECOM PRESS



工业和信息化部“十二五”规划教材

现代数值计算

(第2版)

Advanced Numerical Computing (2nd Edition)

同济大学计算数学教研室 编著



名家系列

人民邮电出版社
北京

图书在版编目(CIP)数据

现代数值计算 / 同济大学计算数学教研室编著. --
2版. -- 北京: 人民邮电出版社, 2014. 9
ISBN 978-7-115-35993-3

I. ①现… II. ①同… III. ①数值计算—高等学校—
教材 IV. ①O241

中国版本图书馆CIP数据核字(2014)第152988号

内 容 提 要

本书是同济大学计算数学教研室几位老师集体智慧的结晶, 内容涉及数值计算的基本内容, 如函数插值与函数逼近、线性与非线性方程(组)的求解、数值积分与微分、矩阵的特征值与特征向量的计算、常微分方程的近似数值解, 还阐述了当今科学与工程研究中经常遇到的数值计算问题求解的新方法, 如快速傅里叶变换、蒙特卡罗随机方法(高维积分计算)、数值求导的稳定算法、大型线性方程组的分块迭代算法等; 在介绍一些重要的典型算法时, 附上了在工程中广泛使用的 MATLAB 程序. 书后附有丰富的习题和数值实验题, 并提供了配套的习题解答.

本书适合作为高等院校本科生和工科研究生“数值计算”课程的教材, 也适合相关科研人员参考.

-
- ◆ 编 著 同济大学计算数学教研室
责任编辑 武恩玉
责任印制 彭志环 杨林杰
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
大厂聚鑫印刷有限责任公司印刷
 - ◆ 开本: 787×1092 1/16
印张: 16.5 2014 年 9 月第 2 版
字数: 396 千字 2014 年 9 月河北第 1 次印刷
-

定价: 39.80 元

读者服务热线: (010)81055256 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京崇工商广字第 0021 号

第 2 版前言

本次修订版,作了如下的改动:

改正了第 1 版中的一些文字和公式错误,将第 1 版中第 6 章向量范数的定义放到第 1 章的误差定义中,使得可读性更强,增加了第 3 章关于混合插值的误差估计和第 5 章中关于辛普森求积公式的证明,增加了一些习题使得书的内容更加完整.

本次修订仍由同济大学计算数学教研室的有关教师集体完成.其中第 1 章和第 7 章由陈雄达完成,第 2 章和第 6 章由殷俊锋完成,第 3 章和第 4 章由陈素琴完成,第 5 章由徐承龙完成,第 8 章由关晓飞完成,第 9 章由王琤完成.徐承龙阅读了全书,提出了一些修改意见,最后由陈雄达排版.

本书的修订工作得到了同济大学研究生院和数学系领导的大力支持,同济大学计算数学教研室使用本书的全体教师提出了宝贵的意见,以及人民邮电出版社武恩玉编辑对本次修订工作的热忱关心和支持,我们深表感谢.

编 者

2014 年 1 月于同济大学

前 言

“数值分析”课程是科学计算方面的重要基础课程之一,承担着引导计算科学入门到介绍数值计算中各种基本算法的任务.随着科学技术的快速发展,对数值计算方面知识的要求也越来越高,这就迫切需要一本适合于目前学生学习的教材.本着这样的想法,同济大学计算数学教研室的有关教师在原有教材的基础上编写了本书.希望达到两个目的:一是在课程中介绍数值计算领域中的基本思想、基本理论与基本算法,如函数插值与函数逼近,线性与非线性方程(组)的求解,数值积分与微分,矩阵的特征值与特征向量的计算,微分方程的近似数值解;二是适当地介绍一些当今科学与工程研究中遇到的数值计算问题求解的新方法,如快速傅里叶变换,高维积分的蒙特卡罗方法,数值求导的稳定算法,大型线性方程组的分块迭代算法等.当然,由于课时的限制,我们只给出这些内容的一个初步介绍,有兴趣者可以参阅有关的参考书.本着实用的原则,同时也由于课时限制的原因,我们在介绍一些数学上比较深入的结论时,往往省略了相关的理论证明.在介绍一些重要的典型算法的同时,附上了在工程中广泛使用的 MATLAB 程序,以便于大家在修完此课程后能快速地上手做一些工程项目中的计算与编程问题.

全书共分 9 章,由同济大学计算数学教研室有关任课教师集体编写.其中第 1 章“科学计算与 MATLAB”和第 7 章“非线性方程求根”由陈雄达编写,第 2 章“线性方程组的直接解法”和第 6 章“线性方程组的迭代解法”由殷俊锋编写,第 3 章“多项式插值与样条插值”和第 4 章“函数逼近”由陈素琴编写,第 5 章“数值积分与数值微分”由徐承龙编写,第 8 章“矩阵特征值与特征向量的计算”和第 9 章“常微分初边值问题数值解”由王琤编写.全书由徐承龙负责组织与协调,陈雄达与殷俊锋负责本书的排版和校对.本书适合作为高等院校本科生和工科研究生的教材与参考书,需要读者掌握高等数学、线性代数和初步的概率方面的知识.在编写过程中我们参考了同济大学数学教研室和国内外有关专家编写的相关教材,在此表示感谢.同济大学数学系和同济大学研究生院领导对本书的编写给予了大力支持,为此我们表示深切的谢意.

由于编写时间和水平的限制,本书不可避免出现错误,我们殷切希望广大读者提出批评与修改建议.

编 者

2009 年 7 月于同济大学

目 录

第 1 章 科学计算与 MATLAB	1	§3.1.3 插值基函数	58
§1.1 科学计算的意义	1	§3.2 拉格朗日插值	59
§1.2 误差基础知识	2	§3.2.1 拉格朗日插值基函数	59
§1.2.1 误差的来源	2	§3.2.2 拉格朗日插值多项式	59
§1.2.2 误差度量	2	§3.2.3 插值余项	61
§1.2.3 有效数字	3	§3.3 牛顿插值	62
§1.2.4 向量的误差	3	§3.3.1 差商	62
§1.2.5 计算机的浮点数系	4	§3.3.2 牛顿插值公式及其余项	65
§1.2.6 一个实例	4	§3.3.3 差分与等距节点的插值公式	66
§1.2.7 数值计算中应注意的几个问题	5	§3.4 埃尔米特插值	67
§1.3 MATLAB 软件	8	§3.4.1 两点三次埃尔米特插值	67
§1.3.1 简介	8	§3.4.2 埃尔米特插值多项式的余项	69
§1.3.2 向量和矩阵的基本运算	9	§3.4.3 $n+1$ 个点 $2n+1$ 次埃尔米特插值多项式 $H_{2n+1}(x)$ 及其余项 $R_{2n+1}(x)$	69
§1.3.3 流程控制	16	§3.5 三次样条插值	71
§1.3.4 脚本文件和函数文件	19	§3.5.1 样条插值概念的产生	71
§1.3.5 帮助系统	23	§3.5.2 三次样条函数	74
§1.3.6 画图功能	27	习题三	82
§1.3.7 数据操作	31	数值实验三	84
习题一	34	第 4 章 函数逼近	85
数值实验一	34	§4.1 内积与正交多项式	85
第 2 章 线性方程组的直接解法	36	§4.1.1 权函数和内积	85
§2.1 高斯消去法	36	§4.1.2 正交函数系	86
§2.2 矩阵的三角分解	40	§4.1.3 勒让德多项式	87
§2.2.1 LU 分解和 LDU 分解	40	§4.1.4 切比雪夫多项式	88
§2.2.2 乔列斯基分解	43	§4.1.5 其他正交多项式	90
§2.2.3 追赶法	45	§4.2 最佳一致逼近与切比雪夫展开	90
§2.2.4 分块三角分解	47	§4.2.1 最佳一致逼近多项式	90
§2.3 QR 分解和奇异值分解	48	§4.2.2 线性最佳一致逼近多项式的求法	92
§2.3.1 正交矩阵	48	§4.2.3 切比雪夫展开与近似最佳一致逼近多项式	93
§2.3.2 QR 分解	51	§4.3 最佳平方逼近	94
§2.3.3 奇异值分解	53	§4.3.1 预备知识	94
习题二	54	§4.3.2 最佳平方逼近	95
数值实验二	56	§4.4 曲线拟合的最小二乘法	99
第 3 章 多项式插值与样条插值	57	§4.4.1 最小二乘法	99
§3.1 多项式插值	57		
§3.1.1 多项式插值问题的定义	57		
§3.1.2 插值多项式的存在唯一性	58		

§4.4.2 利用正交多项式做最小二乘拟合	102	§6.3 不定常迭代法	168
§4.4.3 非线性最小二乘问题	104	§6.3.1 最速下降法	169
§4.4.4 矛盾方程组	107	§6.3.2 共轭梯度法	172
§4.5 周期函数逼近与快速傅里叶变换	108	§6.3.3 广义极小残量法	175
§4.5.1 周期函数的最佳平方逼近	108	§6.3.4 预处理技术	180
§4.5.2 快速傅里叶变换 (FFT)	110	习题六	181
习题四	112	数值实验六	183
数值实验四	113	第 7 章 非线性方程求根	184
第 5 章 数值积分与数值微分	114	§7.1 非线性方程求根的基本问题	184
§5.1 几个常用积分公式及其复合积分公式	114	§7.2 二分法	187
§5.1.1 几个常用积分公式	114	§7.3 不动点迭代方法	188
§5.1.2 代数精度	116	§7.4 迭代加速	191
§5.1.3 积分公式的复合	118	§7.5 牛顿法	193
§5.2 变步长方法与外推加速技术	123	§7.6 割线法	199
§5.2.1 变步长梯形法	123	§7.7 非线性方程组简介	201
§5.2.2 外推加速技术与龙贝格求积方法	124	§7.8 非线性最小二乘问题	204
§5.3 牛顿-科茨公式	126	§7.9 大范围求解方法	206
§5.4 高斯公式	128	习题七	209
§5.4.1 高斯公式的定义及性质	128	数值实验七	210
§5.4.2 常用高斯型公式	132	第 8 章 矩阵特征值与特征向量的计算	211
§5.4.3 高斯型公式的应用	137	§8.1 前言	211
§5.5 多重积分的计算	140	§8.2 幂方法	213
§5.5.1 二重积分的计算	140	§8.2.1 乘幂法	213
§5.5.2 蒙特卡罗模拟求积法简介	143	§8.2.2 反幂法	217
§5.6 数值微分	146	§8.2.3 结合原点平移的反幂法	218
§5.6.1 基于拉格朗日插值多项式的求导方法	146	§8.3 QR 方法	219
§5.6.2 基于样条函数的求导方法	149	习题八	221
习题五	152	数值实验八	222
数值实验五	154	第 9 章 常微分方程初边值问题数值解	223
第 6 章 线性方程组的迭代解法	156	§9.1 欧拉公式及其改进	223
§6.1 范数和条件数	156	§9.1.1 欧拉公式	223
§6.1.1 矩阵范数	156	§9.1.2 数值积分与多步法	225
§6.1.2 扰动分析和条件数	157	§9.1.3 预估校正公式	228
§6.2 基本迭代法	159	§9.2 龙格-库塔公式	230
§6.2.1 雅可比迭代法	160	§9.3 收敛性与稳定性	235
§6.2.2 高斯-赛德尔迭代法	161	§9.3.1 显式单步法的收敛性	235
§6.2.3 超松弛 (SOR) 迭代法	162	§9.3.2 单步法的稳定性	238
§6.2.4 迭代的收敛性分析和误差估计	164	§9.4 微分方程组和刚性问题	240
		§9.5 有限差分法	244
		习题九	247
		数值实验九	248
		参考文献	249
		索引	250

表格目录

表 3-1	一般插值数据表	57	表 7-2	$x^3 - x - 1 = 0$ 的四个不同的 迭代方法	188
表 3-2	函数 $y = \ln x$ 数据表	60	表 7-3	方程 $x \ln x = 1$ 的不同迭代法	191
表 3-3	差商表	64	表 7-4	加速方法计算效果	192
表 3-4	函数 $y = \ln x$ 的差商表	66	表 7-5	艾特肯加速效果	193
表 3-5	两点埃尔米特插值数据	67	表 7-6	$\sqrt{2}$ 的不同近似值	196
表 3-6	两点埃尔米特插值数据	70	表 7-7	牛顿法和割线法的计算效果	200
表 3-7	三点埃尔米特插值数据	70	表 7-8	方程组的不同迭代效果比较	203
表 3-8	例 3.5.2 的数据表	80	表 8-1	计算结果	214
表 3-9	例 3.5.2 的差商表	80	表 8-2	乘幂法的计算结果	217
表 4-1	切比雪夫展开多项式系数	94	表 8-3	反幂法的计算结果	218
表 4-2	给定的数据表	100	表 8-4	带原点平移的反幂法的计算 结果	219
表 4-3	二次最小二乘拟合数据	101	表 9-1	欧拉法式的计算结果	225
表 4-4	多项式拟合数据表	102	表 9-2	梯形公式的计算结果	227
表 4-5	未知函数数据表	104	表 9-3	改进的欧拉公式的计算结果	230
表 4-6	函数 $y = a \sin bx$ 拟合数据	104	表 9-4	龙格-库塔公式阶数和次数的 关系	234
表 5-1	函数值表	114	表 9-5	标准四阶龙格-库塔方法的计算 结果	235
表 5-2	函数值表	120	表 9-6	三种不同方法的计算结果	237
表 5-3	三种不同方法的计算结果	121	表 9-7	两种方法的计算结果	240
表 5-4	区间折半法的计算结果	124	表 9-8	高阶常微分方程的数值解	242
表 5-5	龙贝格方法计算结果	126	表 9-9	刚性问题的计算结果	243
表 5-6	科茨系数	127	表 9-10	不同 N 情形下有限差分法近似 解的误差	246
表 5-7	勒让德高斯点及高斯系数	132			
表 5-8	部分拉盖尔高斯点及高斯系数	134			
表 5-9	埃尔米特高斯点及高斯系数	136			
表 5-10	函数 $f(x)$ 数据表	148			
表 7-1	二分法计算结果	187			

插图目录

图 1-1	河渠的图形	5	图 4-5	非线性最小二乘拟合	106
图 1-2	命令 plot 演示	27	图 5-1	三种不同方法的收敛速度的比较	122
图 1-3	命令 plot3 演示	28	图 5-2	被积函数的光滑性与高斯公式收敛速度的关系	138
图 1-4	命令 surf 演示	29	图 5-3	节点数与高斯公式的误差的关系	139
图 1-5	命令 surf 和 contour 用于二元函数	29	图 5-4	二重积分求积区域	141
图 1-6	图像的标注	30	图 5-5	误差 R_N 随节点 N 的变化规律	145
图 2-1	Givens 变换	49	图 7-1	函数 $f(x) = \frac{x \sin x}{x^2 + 1}$ 的图像	185
图 2-2	Householder 变换	51	图 7-2	非线性方程求根的不同困难程度	185
图 3-1	多项式插值的几何意义	58	图 7-3	不同的序列收敛速度	186
图 3-2	龙格函数及其等距节点上的 10 阶拉格朗日插值多项式与分段线性多项式	72	图 7-4	不动点迭代的局部收敛性	190
图 3-3	龙格函数及其切比雪夫节点上的 10 阶拉格朗日插值多项式	73	图 7-5	牛顿法的几何意义	194
图 3-4	例 3.5.2 的三次自然样条插值函数 $s(x)$	82	图 7-6	函数 $f(x) = x^2 + \sin 10x - 1$ 的图像	198
图 4-1	勒让德多项式的图形, $n = 0, 1, 2, 3$	88	图 7-7	割线法的几何意义	200
图 4-2	切比雪夫多项式的图形, $n = 0, 1, 2, 3, 4$	89	图 7-8	三个方程同伦算法的解曲线	207
图 4-3	最佳一次一致逼近多项式的几何意义	92	图 9-1	欧拉法的几何意义	224
图 4-4	例 4.4.3 数据点的分布情况	103	图 9-2	三种方法的收敛阶	238
			图 9-3	欧拉方法的稳定性	240
			图 9-4	高阶常微分方程的数值解	242

程序目录

函数 mysort.m	19	函数 gauss_lege.m	133
函数 mysort2.m	20	函数 gauss_laguerre.m	135
函数 mysort3.m	21	函数 gauss_her.m	136
函数 zhouchang.m	23	函数 herval.m	137
函数 fib.m	23	函数 jacobi.m	161
函数 fib.m (带注释)	25	函数 gs.m	161
函数 tridiagsolver.m	46	函数 sor.m	162
函数 lagrange.m	60	函数 cg.m	173
函数 chashang.m	64	函数 gmres.m	177
函数 nlfite.m	105	函数 newton.m	197
函数 nlfiteb.m	106	函数 homo.m	207
函数 fmid.m	118	函数 eigIPower.m	216
函数 ftrapz.m	119	函数 odeEuler.m	225
函数 fsimpson.m	119	函数 odeIEuler.m	229
函数 coeflege.m	133		

算法目录

算法 2.1.1	高斯消去法	37	算法 7.7.1	非线性方程组的 GS 方法	201
算法 2.2.1	杜利脱尔算法	42	算法 7.7.2	拟牛顿法	203
算法 2.2.2	克洛脱算法	42	算法 7.7.3	改进拟牛顿法	204
算法 2.2.3	乔列斯基算法	44	算法 7.8.1	高斯-牛顿法	205
算法 2.2.4	追赶法	45	算法 7.8.2	LM 算法	206
算法 5.2.1	区间折半法	123	算法 8.2.1	乘幂法	216
算法 6.2.1	雅可比迭代算法	160	算法 8.2.2	反幂法	217
算法 6.2.2	高斯-赛德尔迭代算法	161	算法 8.2.3	带原点平移的反幂法	218
算法 6.2.3	SOR 迭代算法	162	算法 8.3.1	QR 方法	219
算法 6.3.1	最速下降法	170	算法 9.1.1	欧拉公式	224
算法 6.3.2	共轭梯度法	173	算法 9.1.2	梯形公式	226
算法 6.3.3	广义极小残量法	177	算法 9.1.3	改进的欧拉公式	229
算法 6.3.4	左预处理共轭梯度法	180	算法 9.2.1	标准四阶四段龙格-库塔 公式	234
算法 7.2.1	二分法	187	算法 9.5.1	有限差分法	245
算法 7.5.1	牛顿下山法	197			

第 1 章 科学计算与 MATLAB

§1.1 科学计算的意义

数值计算是随着计算机的出现和大规模计算的需求而发展起来的一门新兴学科。数值计算主要考虑各种数学模型及其算法, 这些数学模型是为了解决各类应用领域, 特别是科学与工程计算领域的实际问题而提出的。为此, 数值计算有时也称为科学计算、工程计算或科学与工程计算。随着科学技术的发展, 计算机的性能和算法的效率, 即计算机的硬件和软件水平都有了飞速的提高, 需要求解的实际问题规模也成倍扩大, 其中的数学模型日趋复杂。通常, 这些数学模型是不能够精确地求解的, 这时需要简化模型并且提出相应的数值解法, 然后在计算机上编程实现, 求解这些问题并作实际检验。随着硬件性能的提高和软件上各种高效算法的出现, 人类的计算能力迅猛提高, 并同时期待能解决一些超大规模的具有挑战性的问题, 如基因测序、全球天气模拟等。对于同一个问题, 不同的算法在计算性能上可能相差百万倍甚至更多, 科学计算的主要任务就是设计高效可靠的数值算法。例如: 用一个每秒钟计算一亿次浮点运算的计算机求解一个 20 阶的线性代数方程组, 用克拉默 (Cramer) 法和行列式展开法计算至少需要 30 万年, 而用高斯消去法只不过用几秒钟而已。这个事实说明了两个问题: 一方面, 计算方法效率的提高速度往往比计算机性能的提高更快; 另一方面, 选择高效率的计算方法无疑是极其重要的。

求解科学与工程计算领域中的问题一般要经历以下几个过程。首先根据实际问题构造相应的数学模型, 把它转换为可以计算的问题, 称为**数值问题**; 其次根据问题特点选择计算方法并编制程序; 最后在计算机上求解。科学计算的主要研究内容是提出数值问题, 设计高效的算法, 并探讨全过程中各种误差对近似解的影响。数值问题要求对有限个输入数据计算得出有限个输出数据, 这些输出数据通常称为数值解, 或者也可以理解为近似解。**数值算法**则是求解问题数值解的方法, 它是由有限个明确无歧义的操作组成的对输入数据的变换, 其中每一个操作都是计算机能够完成的, 例如, 仅包含加减乘除的运算。一个算法只有在保证可靠的前提下才有可能评价其性能的好坏, 通俗地讲, 可靠性方面包含诸如算法的收敛性、稳定性、误差估计等多方面的内容。评价一个算法的优劣应该考虑其时间复杂度 (即占用的计算机时间)、空间复杂度 (即占用的计算机存储空间) 以及逻辑复杂度 (即程序开发周期长短及维护的难易程度)。

由于各种科学计算问题最后通常都归结为求解一些基本的问题, 所以数值算法领域的许多工作者为这些基本问题设计了一些相对固定的高效算法, 并把它们设计成简单且容易调用的功能函数并形成软件包。但由于实际问题的复杂性及算法自身的适应性, 调用者必须自行选择适合自己问题的功能函数。现代数值计算领域流行的软件有 Maple、Mathematica、MATLAB 等, 但不仅限于这些软件, 更多软件可以在网上查询 (<http://www.netlib.org>), 其中, MATLAB 软件是在工程计算界广泛使用的深受计算工作者和工程师喜爱的软件之一。MATLAB 的官方网站是 <http://www.mathworks.com>。

鉴于实际问题的复杂性,通常将一个实际问题具体分解为一系列的子问题进行研究,数值工作者把这些子问题归纳总结为数学上不同的几类问题.本课程主要涉及以下几方面的问题:第 1 章余下部分为数值计算的基本知识及 MATLAB 软件简介,其后各章内容包括函数的插值与逼近、数值积分与数值微分、线性方程组的直接解法和迭代解法、非线性方程组的求解、矩阵特征值问题的求解以及常微分方程的数值解.

§1.2 误差基础知识

人们常用相对误差、绝对误差或有效数字来说明一个近似值的准确程度.这些概念在科学计算中被广泛应用,下面我们对有关概念作一介绍.

§1.2.1 误差的来源

我们把通过任何途径得到的数据或模型与真实情况之间的差异称为**误差**.误差的来源经常是多方面的.在建立数学模型的过程中,不可避免要忽略一些次要的因素,因而数学模型往往只是对实际问题的一种近似的表达,这两者之间的差异我们称为**模型误差**.同时数学模型中可能包含一些参数,它们可以通过仪器观测得到或通过经验得到,这种数据间的误差我们称为**观测误差**.数值分析通常假定数学模型真实地反映了客观实际,直接处理已经归纳总结出来的数值问题,因而这两类误差在数值分析中并不常见.数学模型问题通常要转化为数值问题才能被求解,经常使用的转化手段往往有离散化、有限展开等方式.我们称这种数值问题与数学模型之间的误差为**截断误差或方法误差**,通常引起方法误差的原因在于我们必须在有限的步骤内在计算机上得到结果.在用计算机实现数值方法的过程中,由于计算机表示的浮点数是固定的有限字长,因此计算机并不能精确地表示所有的数,这样,不仅原始输入数据有误差,中间计算的数据及最终输出结果也必然有误差.这种因为计算机有限字长引起的误差称为**舍入误差**,原始数据的误差导致最终结果也有误差的过程称为**误差传播**.

§1.2.2 误差度量

假设 x 是真值, \bar{x} 是它的近似值,则称 $\Delta x = x - \bar{x}$ 为该近似值的**绝对误差**,或简称**误差**.一般说来,真值通常是求不出来的,因此我们也不可能知道 Δx 的值,而只能有如下估计:

$$|\Delta x| = |x - \bar{x}| \leq \varepsilon, \quad (1.1)$$

数 ε 称为**绝对误差限或误差限**.于是有

$$\bar{x} - \varepsilon \leq x \leq \bar{x} + \varepsilon, \quad (1.2)$$

在工程上也记作 $x = \bar{x} \pm \varepsilon$.误差限给出了真值的范围,但不能很好地表示近似值的精确程度.例如测量珠峰高度为 8848m,误差不超过 1m;在测量运动员身高时就绝对不可以用这个误差限,否则结果是没有任何意义的.同样的误差限对于不同的数据,其反映近似真实的程度可以完全相反,因此必须同时考虑真值的大小.

若 x 是不为零的真值, \bar{x} 是它的近似值,则称 $\delta x = \Delta x/x = (x - \bar{x})/x$ 为该近似值的**相对误差**(真值为零的情况没有定义).若可求得某数 ε_r , 满足

$$|\delta x| = \frac{|x - \bar{x}|}{|x|} \leq \varepsilon_r, \quad (1.3)$$

则称 ε_r 为相对误差限. 由于真值难以求出, 假如 \bar{x} 也非零, 通常也使用 $\delta x = \Delta x / \bar{x}$.

§1.2.3 有效数字

当 x 有很多位数字, 为规定其近似数的表示法, 使得用它表示的近似数自身就指明了相对误差的大小, 我们引入有效数字的概念.

设十进制数有如下的标准形式:

$$x = \pm 10^m \times 0.x_1x_2 \cdots x_nx_{n+1} \cdots, \quad (1.4)$$

其中 m 为整数, $\{x_i\} \subseteq \{0, 1, 2, \cdots, 9\}$ 且 $x_1 \neq 0$. 对 x 四舍五入保留 n 位数字, 得到近似值 \bar{x} :

$$\bar{x} = \begin{cases} \pm 10^m \times 0.x_1x_2 \cdots x_n, & x_{n+1} \leq 4, \\ \pm 10^m \times 0.x_1x_2 \cdots (x_n + 1), & x_{n+1} \geq 5. \end{cases} \quad (1.5)$$

容易证明, 四舍五入近似数的误差限满足

$$|x - \bar{x}| \leq 10^m \times \left(\frac{1}{2} \times 10^{-n} \right) = \frac{1}{2} \times 10^{m-n}. \quad (1.6)$$

设 x 的近似值 \bar{x} 有如下标准形式

$$\bar{x} = \pm 10^m \times 0.x_1x_2 \cdots x_n \cdots x_p, \quad (1.7)$$

其中 m 为整数, $\{x_i\} \subseteq \{0, 1, 2, \cdots, 9\}$ 且 $x_1 \neq 0, p \geq n$. 如果有

$$|x - \bar{x}| \leq \frac{1}{2} \times 10^{m-n}, \quad (1.8)$$

则称 \bar{x} 为 x 的具有 n 位有效数字的近似数, 其中 x_1, x_2, \cdots, x_n 分别称为第 1 位到第 n 位有效数字. 当 $p = n$ 时, 称 \bar{x} 为有效数, 即全由有效数字组成的数是有效数. 有效数的误差限是末位数单位的一半, 其本身就体现了误差界, 因此有效数末尾是不可以随便添加零的.

§1.2.4 向量的误差

某些问题的解可能是一个 n 维向量 \mathbf{x} , $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$. 为了度量向量的误差, 我们引入向量的范数.

定义 1.2.1 对任意 n 维向量 \mathbf{x} , 若对应非负实数 $\|\mathbf{x}\|$, 且满足

- (1) $\|\mathbf{x}\| \geq 0$, 当且仅当 $\mathbf{x} = \mathbf{0}$ 时等号成立;
- (2) 对任意实数 α , $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$;
- (3) 对任意的 n 维向量 \mathbf{x} 和 \mathbf{y} , $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$,

则称 $\|\mathbf{x}\|$ 为向量 \mathbf{x} 的范数.

设 $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$, 定义

$$\begin{aligned} \|\mathbf{x}\|_1 &= |x_1| + |x_2| + \cdots + |x_n|, \\ \|\mathbf{x}\|_2 &= \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}, \\ \|\mathbf{x}\|_\infty &= \max_{1 \leq i \leq n} |x_i|, \end{aligned}$$

为向量的 1 范数、2 范数和无穷范数.

易证, 这三个范数都满足上述范数的三条性质.

譬如, 对于向量的 2 范数 $\|\mathbf{x}\|_2$, 前两条性质显然成立.

对于性质 3, 使用柯西不等式, 有

$$\begin{aligned}\|\mathbf{x} + \mathbf{y}\|_2^2 &= \sum_{i=1}^n |x_i + y_i|^2 \leq \sum_{i=1}^n |x_i|^2 + 2 \sum_{i=1}^n |x_i| \cdot |y_i| + \sum_{i=1}^n |y_i|^2 \\ &\leq \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 \\ &= (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2.\end{aligned}$$

即性质 3 成立, 所以 $\|\mathbf{x}\|_2$ 是一个向量范数.

若向量 \mathbf{x} 有近似值 $\bar{\mathbf{x}}$, 则定义该近似向量误差为 $\|\mathbf{x} - \bar{\mathbf{x}}\|$, 这里范数可根据实际需要选取.

一般地, 向量的相对误差 $\|\mathbf{x} - \bar{\mathbf{x}}\|/\|\mathbf{x}\|$ 并不能直接解读出各分量的有效位. 如令 $\mathbf{x} = (1.000, 0.001)^T$, $\bar{\mathbf{x}} = (0.999, 0.002)^T$, 则采用无穷范数时 $\|\mathbf{x} - \bar{\mathbf{x}}\|_\infty/\|\mathbf{x}\|_\infty = 0.001$. 但是 \mathbf{x} 第二个分量误差达到了 100%.

§1.2.5 计算机的浮点数系

计算机内部通常使用浮点数进行实数的运算. 计算机的浮点数是仅有有限字长的二进制数, 大部分实数存入计算机时需要做四舍五入, 由此引起的误差称为舍入误差. 一个浮点数的表示由正负号、小数形式的尾数以及为确定小数点位置的阶三部分组成. 例如单精度实数用 32 位的二进制表示, 其中符号占 1 位, 尾数占 23 位, 阶数占 8 位. 这样一个规范化的计算机单精度数 (零除外) 可以写成如下形式:

$$\pm 2^p \times (0.\alpha_1\alpha_2\alpha_3 \cdots \alpha_{23})_2, \quad |p| \leq 2^7 - 1, \quad p \in Z, \alpha_i \in \{0, 1\}. \quad (1.9)$$

上面记号中, Z 表示整数集. 二进制的非零数字只有 1, 所以 $\alpha_1 = 1$. 阶数的 8 位中须有 1 位表示阶数的符号, 所以阶数的值占 7 位. 凡是能够写成上述形式的数称为机器数. 设机器数 a 有上述形式, 则与之相邻的机器数为 $b = a + 2^{p-23}$ 和 $c = a - 2^{p-23}$. 这样, 区间 (c, a) 和 (a, b) 中的数无法准确表示, 计算机通常按规定用与之最近的机器数表示.

设实数 x 在机器中的浮点 (float) 表示为 $fl(x)$, 我们把 $x - fl(x)$ 称为舍入误差. 如当 $x \in \left[\frac{c+a}{2}, \frac{a+b}{2}\right) = [a - 2^{p-1-23}, a + 2^{p-1-23})$ 时, 用 a 表示 x , 记为 $fl(x) = a$. 其相对误差满足

$$|\varepsilon_r| = \left| \frac{x - fl(x)}{fl(x)} \right| \leq \frac{2^{p-1-23}}{2^{p-1}} = 2^{-23} \approx 10^{-6.9}. \quad (1.10)$$

上式表明单精度实数有 6~7 位有效数字.

二进制阶数最高为 $2^7 - 1$, 相应于十进制的阶数 38, 即 $(2^7 - 1)\lg 2$. 因此单精度实数 (零除外) 的数量级不大于 10^{38} 且不小于 10^{-38} . 当输入数据、输出数据或中间数据太大而无法表示时, 计算过程将会非正常停止, 此现象称为上溢 (overflow); 当数据太小而只能用零表示时, 计算机将此数置零, 精度损失, 此现象称为下溢 (underflow). 下溢并不总是有害的, 在做浮点运算时, 我们需要考虑数据运算可能产生的上溢及有害的下溢.

§1.2.6 一个实例

下面我们通过一个简单的例子来说明, 一个实际问题从提出到解决过程中出现的各种误差.

例 1.2.1 有一艘驳船, 宽度为 5m, 欲驶过一个河渠. 该河渠有一个直角弯道, 形状和尺寸如图 1-1 所示. 试问, 要驶过这个河渠, 驳船的长度不能超过多少米?

解: 易知, 驳船的长度有如下关系

$$l = l_1 + l_2 = \frac{10 - 5 \cos \theta}{\sin \theta} + \frac{12 - 5 \sin \theta}{\cos \theta} = f(\theta), \quad (1.11)$$

式中, l_1, l_2 分别为直角拐角处到船两头的距离. 驳船如若能通过河渠, 则其最大长度应是上式右端函数的最小值. 因此, 该问题就转化为求解极小化问题

$$\min f(\theta) = \frac{10 - 5 \cos \theta}{\sin \theta} + \frac{12 - 5 \sin \theta}{\cos \theta}, \quad (1.12)$$

或者, 求解非线性函数零点的问题

$$f'(\theta) = \frac{5 - 10 \cos \theta}{\sin^2 \theta} + \frac{12 \sin \theta - 5}{\cos^2 \theta} = 0. \quad (1.13)$$

可以证明, 对于任意 $\theta \in (0, \frac{\pi}{2})$, $f''(\theta) > 0$. 因此 (1.12) 式的极小点即是 (1.13) 式的零点, 这两者完全等价. 通过本教材将要介绍的近似计算方法, 我们可以知道上述两个问题的解为

$$\theta^* = 0.73, \quad f(\theta^*) = 21. \quad (1.14)$$

因此, 驳船的长度不能超过 21m.

从实际的角度出发, 我们知道, 一艘驳船能不能通过河渠应该是一个复杂的问题: 驳船不完全是长方形的, 而且可能和水深也有关系. 我们把它简化为长方形的, 并且只在二维平面内考虑该问题, 这就造成了模型误差. 我们无法精确求解极小化问题 (1.12) 式或求零点问题 (1.13) 式, 用近似求解的方法代替精确求解的方法, 造成了方法误差. 测量的数据, 如 5m、10m、12m 等, 都带有误差, 称为观测误差. 因此, 结论中的数据也只给出了两位有效数字, 是一个近似的答案. 由于初始数据的误差导致最终答案的误差, 这个过程就是误差的传播过程.

§1.2.7 数值计算中应注意的几个问题

舍入误差在实际计算中几乎是不可避免的, 定量地分析舍入误差的积累过程往往都是非常繁杂的. 一个可行的方法是研究舍入误差是否能够得到有效的控制, 不会影响到计算结果的实际效用. 一个算法, 如果在一定的条件下, 其舍入误差在整个运算过程中能够得到有效控制或者舍入误差的增长不影响产生可靠的结果, 则称该算法是数值稳定的, 否则称为数值不稳定的.

例 1.2.2 计算 $S_n = \int_0^1 \frac{x^n}{x+5} dx$, 其中 $n = 0, 1, 2, \dots, 8$.

解: 由于

$$S_n + 5S_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5} dx = \frac{1}{n}, \quad (1.15)$$

取 $S_0 = \ln 6 - \ln 5 = 0.182$, 利用公式 $S_n = \frac{1}{n} - 5S_{n-1}$, 可以逐步得到如下的数值, 计算过程

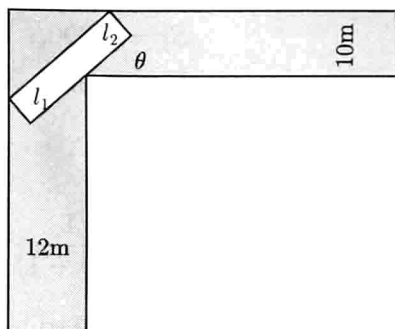


图 1-1 河渠的图形

中所有的数精确到小数点后 3 位:

$$\begin{aligned} S_1 &= 0.090, & S_2 &= 0.050, & S_3 &= 0.083, & S_4 &= -0.165, \\ S_5 &= 1.025, & S_6 &= -4.958, & S_7 &= 24.933, & S_8 &= -124.540. \end{aligned} \quad (1.16)$$

通过简单的积分估计, 有

$$\frac{1}{6(n+1)} = \int_0^1 \frac{x^n}{6} dx \leq S_n \leq \int_0^1 \frac{x^n}{5} dx = \frac{1}{5(n+1)}. \quad (1.17)$$

所以上述的 8 个计算结果中, 那些负数或者大于 1 的结果都是不可接受的. 当然, 其他结果也可能有比较大的误差.

下面我们分析造成这种现象的原因. 假设 S_n 的真值为 S_n^* , 误差为 ε_n , 即 $\varepsilon_n = S_n^* - S_n$. 对于真值, 我们也有关系式 $S_n^* + 5S_{n-1}^* = \frac{1}{n}$. 综合两个递推等式, 有

$$\varepsilon_n = -5\varepsilon_{n-1}. \quad (1.18)$$

这就意味着哪怕开始只有一点点误差, 就算整个过程都保留很长的小数位, 只要 n 足够大, 按照这种每计算一步误差增长 5 倍的方式, 所得的结果总是不可信的. 因此整个算法是数值不稳定的.

换一种方式, 若我们把计算方式改为

$$S_{n-1} = \frac{1}{5n} - \frac{1}{5}S_n, \quad n = 8, 7, \dots, 1. \quad (1.19)$$

则误差就会以每计算一步缩小到 $1/5$ 的方式进行. 用这样的方式计算, 可以先用上面的估计式计算出 S_8 :

$$S_8 = \frac{1}{2} \left(\frac{1}{6 \times 9} + \frac{1}{5 \times 9} \right) \approx 0.020. \quad (1.20)$$

逐步计算有

$$\begin{aligned} S_7 &= 0.021, & S_6 &= 0.024, & S_5 &= 0.028, & S_4 &= 0.034, \\ S_3 &= 0.043, & S_2 &= 0.058, & S_1 &= 0.088, & S_0 &= 0.182. \end{aligned} \quad (1.21)$$

这样的计算结果和实际是很相近的. 对 S_8 不同的估计方式, 最后得到的结果也相似: 只需对递推计算公式进行同样的误差分析就可以得到这个结论.

误差的传播在一些实际的问题中经常是很复杂的, 不像上面的例子那样可以得到一个误差传播的具体的公式. 事实上, 这个误差传播的具体公式需要假定 $1/n$ 的计算是完全精确的.

通过对误差传播规律的简单分析, 下面我们指出在数值计算中应该注意的基本问题.

§1.2.7.1 避免相近的数相减

在数值计算中, 两个相近的数相减时有效数字会损失. 例如计算

$$y = \sqrt{x+1} - \sqrt{x}, \quad (1.22)$$

其中 x 是比较大的数, 例如 $x = 1000$. 取 4 位有效数字计算, 有

$$y = \sqrt{1001} - \sqrt{1000} = 31.64 - 31.62 = 0.02. \quad (1.23)$$