

计算方法

主编 李福祥



哈尔滨工业大学出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

计 算 方 法

主 编 李福祥

副主编 曹作宝 毕 卉

哈爾濱工業大學出版社

内容简介

本书内容包括计算机上常用的各种数值计算方法,如插值法、最小二乘法、数值微积分、方程求根法、线性与非线性代数方程组解法、常微分方程初值问题的解法和矩阵特征值及特征向量的计算等。书中重点讨论了各种计算方法的构造原理和使用,对稳定性、收敛性、误差估计和优缺点等也作了适当的介绍。

本书可作高等工科院校非计算数学专业学生教材,也可供从事数值计算的科技工作者阅读参考。

图书在版编目(CIP)数据

计算方法/李福祥主编. —哈尔滨:

哈尔滨工业大学出版社, 2011. 2

ISBN 978-7-5603-2574-3

I. ①计… II. ①李… III. ①计算方法—高等学校—教材 IV. ①0241

中国版本图书馆 CIP 数据核字(2010)第 130390 号

责任编辑 赵文斌

封面设计 卞秉利

出版发行 哈尔滨工业大学出版社

社 址 哈尔滨市南岗区复华四道街 10 号 邮编 150006

传 真 0451-86414749

网 址 <http://hitpress.hit.edu.cn>

印 刷 肇东粮食印刷厂

开 本 787mm×960mm 1/16 印张 13.75 字数 284 千字

版 次 2011 年 2 月第 1 版 2011 年 2 月第 1 次印刷

书 号 ISBN 978-7-5603-2574-3

定 价 32.80 元

(如因印装质量问题影响阅读,我社负责调换)

前 言

随着科学技术的发展,电子计算机的应用日益广泛,科学与工程计算越来越显示出其重要性,与实验、理论三足鼎立,成为科学实践的三大手段之一,其应用范围渗透到所有的科学活动领域。由于这一原因,现在各大专院校非计算数学专业的研究生和高年级学生,已普遍学习“计算方法”课程。

本教材考虑到工科各专业对计算方法的实际需求,重点介绍了计算机上常用的基本计算方法的构造和使用;同时对计算方法的工作量、稳定性、收敛性、误差估计、适用范围及优缺点等也作了适当的分析。本书的叙述,采用了由简到繁、由个别到一般的方法,避免复杂化;例题与习题的选择,力求典型,数值计算简单,便于验证。

本书第一、五、八章由李福祥编写,第二、三、四章由曹作宝编写,第六、七、九章由毕卉编写。由于编者水平有限,缺点与疏漏在所难免,恳请读者批评指正。

本书获得了黑龙江省教育厅科技项目(11521045)、黑龙江省自然科学基金项目(A200811)、哈尔滨理工大学教育教学研究项目(P20090066)和哈尔滨理工大学青年科学研究基金项目(2009YF033)的支持,并在编写过程中得到了哈尔滨理工大学教务处、哈尔滨理工大学应用数学系及哈尔滨工业大学出版社等领导和老师们的关心和支持,在此表示感谢。

编 者

2010年6月

目 录

第 1 章 绪论	1
1.1 研究计算方法的必要性	1
1.2 机器数系	2
1.3 误差	3
1.4 设计算法的注意事项	12
思考题	15
习题	15
第 2 章 插值法	17
2.1 插值多项式的存在唯一性	17
2.2 拉格朗日插值多项式	19
2.3 牛顿插值多项式	24
2.4 埃尔米特插值	32
2.5 分段低次插值	35
2.6 三次样条插值函数	39
2.7 反插值	46
思考题	48
习题	48
第 3 章 数据拟合法	50
3.1 曲线拟合的最小二乘法	50
3.2 超定方程组的最小二乘解	57
3.3 一般最小二乘拟合	59
思考题	66
习题	66
第 4 章 数值积分与数值微分	67
4.1 数值积分的基本概念	67
4.2 牛顿-科兹公式	69
4.3 复合求积公式	73
4.4 龙贝格公式	77
4.5 高斯公式	82
4.6 数值微分	87
思考题	91

习题	92
第 5 章 方程求根	93
5.1 增值寻根法与二分法	93
5.2 迭代法	96
5.3 迭代收敛的加速	101
5.4 牛顿法	103
5.5 割线法	106
思考题	108
习题	109
第 6 章 线性方程组的数值方法	110
6.1 高斯消元法	110
6.2 高斯主元素消元法	115
6.3 高斯—若当消元法	119
6.4 矩阵分解	123
6.5 向量和矩阵的范数	134
6.6 误差分析	141
思考题	146
习题	146
第 7 章 线性方程组的迭代法	148
7.1 迭代法及其收敛性	148
7.2 雅可比迭代法与高斯—塞德尔迭代法	152
7.3 超松弛迭代法	157
思考题	162
习题	162
第 8 章 常微分方程数值解法	164
8.1 欧拉法	164
8.2 龙格—库塔法	169
8.3 亚当斯方法	175
8.4 线性多步法	180
8.5 方程组与高阶方程的数值解法	182
8.6 边值问题的数值解法	185
思考题	188
习题	188
第 9 章 矩阵特征值与特征向量的计算	190
9.1 乘幂法与反幂法	190
9.2 子空间迭代法	201

9.3 对称矩阵的雅克比(Jacobi)旋转法	203
思考题	209
习题	209
参考文献	211

第 1 章 绪 论

1.1 研究计算方法的必要性

随着科学技术的发展,科学与工程计算已被推向科学活动的前沿。科学与工程计算在所有的科学领域得到应用,并与实验、理论三足鼎立,相辅相成,成为人类科学活动的三大方法之一。计算机作为科学与工程计算的主要工具越来越不可缺少,因而要求研究适合计算机使用的数值计算方法,解决数学问题的数值计算及有关理论。考察用计算机解决科学计算问题的一般过程,可以概括为:

实际问题→数学建模→计算方法→程序设计→验证及结果的可视化

计算方法以纯数学为基础,不仅是研究数学本身的理论,而且着重研究解决问题的数值方法及效果,包括方法的收敛性、稳定性以及误差分析,还要根据计算机的特点研究计算时间最短、需要计算机内存最少的计算方法。有些方法在理论上虽不够严密,但通过实际计算,对比分析等手段,证明是行之有效的方法,也应该采用。因此,计算方法既有纯数学的高度抽象性与严密科学性的特点,又具有应用的广泛性与实际试验的高度技术性特点,是一门与计算机使用密切结合的实用性很强的计算数学课程。

例如,考虑线性方程组的解,在“线性代数”中,只介绍解的存在唯一性及有关理论和精确解法,用这些理论和方法还不能直接在计算机上求解。众所周知,用克莱姆法则求解一个 n 阶线性方程组,要算 $n+1$ 个 n 阶行列式,总共需要 $(n-1)(n+1)n!$ 次乘法,当 n 充分大时,计算量是相当惊人的。如一个不算太大的 20 阶线性方程组大约要做 10^{20} 次乘法,这项计算即使用每秒百亿次的电子计算机去做,也要连续工作数千年才能完成,当然这是完全没有实际意义的。而如果用消元法,求解一个 n 阶线性方程组大约需要 $\frac{1}{3}n^3 + n^2$ 次乘法,对一个 20 阶的方程组即使用一台小型计算器也能很快解出来。这一简单的例子说明,能否正确地设计算法,是科学计算成败的关键。

1.2 机器数系

数值计算的工具是计算机,计算机的字长和运算方式对数值计算的结果有直接的影响。对给定的数值方法,一个注意到计算机有限字长和运算方式的程序员可以写出具有较高计算精度的程序;反之,就会得到十分粗糙的程序,得出完全失真的计算结果。因此,了解计算机数的表示和运算方式对使用计算机十分必要。

1.2.1 数的浮点表示

用于数值计算的计算机多采用浮点系统。因为用浮点方式表示的数,取值范围大,运算精度高,从而为编制程序提供了方便。

设 s 是 r 进制数, p 是 r 进制正负整数或零, r 进制数 x 可以用 s 和 r^p 的乘积表示为

$$x = s \times r^p \quad (1.1)$$

再设 s 的整数部分等于零,即 s 满足条件

$$-1 < s < 1 \quad (1.2)$$

则形如(1.1)而满足条件(1.2)的 r 进制数 x 称为 r 进制浮点数。 s 和 p 分别称为浮点数 x 的尾数和阶数。如果尾数的小数位数等于有限正整数 t ,则把 x 称为 t 位浮点数。

此外,如果还要求尾数 s 小数点后第一位数字不等于 0,也就是要求尾数 s 满足条件

$$r^{-1} \leq s < 1 \quad (1.3)$$

则形如(1.1)而满足条件(1.3)的浮点数称为 r 进制规格化浮点数。

例如,十进制数

$$0.003\ 012, 0.321\ 7, 283.4$$

的规格化浮点数分别为

$$0.301\ 2 \times 10^{-2}, 0.321\ 7 \times 10^0, 0.283\ 4 \times 10^3$$

二进制数

$$1001.101, 0.10101, 0.00101$$

的规格化浮点数分别为

$$0.1001101 \times 2^4, 0.10101 \times 2^0, 0.101 \times 2^{-2}$$

显然,只要数 $x \neq 0$,则 x 一定可以表示为规格化浮点数,这样一来,一个数的数量级就一目了然了。

1.2.2 机器数系

上面介绍的数的浮点表示方法为计算机所通用,是研究数值方法的基础,任一计算机只能用有限的位数来表示浮点的尾数和阶数。设进位制为 r ,阶数 p 满足条件

$$l \leq p \leq u \quad (1.4)$$

其中 l, u 为整数,它们主要由计算机用多少位数来表示阶数所确定。如果尾数的小数位数为 t ,则计算机数系由一切阶数满足(1.4)的 t 位 r 进制浮点数的集合 F 组成, F 中的浮点数具有以下形式

$$x = \pm \left(\frac{d_1}{r} + \frac{d_2}{r^2} + \cdots + \frac{d_t}{r^t} \right) \cdot r^p \triangleq \pm 0. d_1 d_2 \cdots d_t \times r^p \quad (1.5)$$

其中 $d_1, d_2, \cdots, d_t$ 为正整数,满足关系

$$0 \leq d_i \leq r-1, i=1, 2, \cdots, t \quad (1.6)$$

若对 $x \neq 0$,规定式(1.5)中 $d_1 \neq 0$,则 F 为规格化的浮点数系。不难证明, F 中共有

$$2(r-1)r^{t-1}(u-l+1)+1 \quad (1.7)$$

个浮点数。例如,若 $r=2, t=3, l=-1, u=2$,则相应的浮点数系 F 中共有33个浮点数。

当 $r=10, t=4, l=-99, u=99$ 时

$$-0.0001 \times 10^{-99}, 0.0001 \times 10^{-99}$$

是数系 F 中绝对值最小的非零数,而

$$-0.9999 \times 10^{99}, 0.9999 \times 10^{99}$$

是此数系中的最小数和最大数,若计算的中间结果超出了上述范围,则称为溢出。

由此可见,在计算机数系 F 中,数的个数有限,数系中的每一个数都是有理数。从整体看,数系中的数分布很不均匀;从局部看,阶数相同的数,又以相等的距离,分布在数轴的某一段上。所以计算机数系是由一些残缺不全,分布不均匀的数组成,如果运算结果超出了 F 的范围,则产生溢出。

1.3 误差

除了极个别的情况外,数值计算总是近似计算,实际计算结果与理论结果之间存在着误差。计算方法的任务之一是将误差控制在一定的容许范围内或者至少对误差有所估计。

1.3.1 误差的来源

误差一般有:模型误差、观测误差、截断误差和舍入误差。在计算方法中,主要讨论的是截断误差和舍入误差。

用计算机解决科学计算问题首先要建立数学模型,它是对被描述的实际问题进行抽象,简化而得到的,因而是近似的,把数学模型与实际问题之间出现的误差称为模型误差。

在数学模型中往往还有一些根据观测得到的物理量,如温度、长度、电压等,这些参量受测量工具及手段的影响,测量的结果不可能绝对准确,由此产生的误差称为观测误差。

在数学模型不能得到精确解时,通常要用数值方法求它的近似解,其近似解与精确解之间的误差称为截断误差或方法误差。例如,函数 $f(x)$ 用泰勒多项式

$$P_n(x) = f(0) + f'(0)x + \frac{1}{2!}f''(0)x^2 + \cdots + \frac{1}{n!}f^{(n)}(0)x^n \quad (1.8)$$

近似代替时,有误差

$$R_n(x) = f(x) - P_n(x) = \frac{1}{(n+1)!}f^{(n+1)}(\xi)x^{n+1} \quad (1.9)$$

其中 $\xi \in (0, x)$, 式(1.9)就是截断误差。

有了求解数学问题的计算公式以后,用计算机做数值计算时,由于计算机的字长有限,原始数据常常不属于计算机数系,而只能采用计算机数系中和它们比较接近的数来表示它们,由此产生的误差以及计算过程又可能产生新的误差,这些误差称为舍入误差。例如,用 3.141 59 近似代替 π , 产生的误差,即

$$R = \pi - 3.141\,59 = 0.000\,002\,6\cdots$$

就是舍入误差。

一般情况,每一步的舍入误差是微不足道的,但经过计算过程的传播和积累,舍入误差可能会对真值产生很大的影响。甚至在一些情况中,一次的舍入就会大大改变计算结果。

1.3.2 绝对误差和相对误差

设数 x (精确值)有一个近似值为 x^* , 记

$$e(x^*) \triangleq x^* - x \quad (1.10)$$

称 $e(x^*)$ 为近似值 x^* 的绝对误差,简称误差。注意这样定义的误差 $e(x^*)$ 可正可负。

准确值(也称精确值) x 一般是未知的,因而绝对误差 $e(x^*)$ 也是未知的,

但往往可以估计出绝对误差的一个上界,即可以找出一个正数 η ,使

$$|e(x^*)| \leq \eta \quad (1.11)$$

实践中用 $|e(x^*)|$ 尽可能小的上界 $\epsilon(x^*)$ 估计 x^* 的误差,称 $\epsilon(x^*)$ 为 x^* 的绝对误差限(或误差限)。

例如, $\pi = 3.141\,592\,653\,58\cdots$,若取 $\pi^* = 3.141\,59$,于是

$$|e(\pi^*)| \leq 0.000\,003$$

则 $\epsilon(x^*) = 0.000\,003$ 就可以作为用 π^* 近似表示 π 的绝对误差限。

显然,误差限 $\epsilon(x^*)$ 总是正数,且

$$|e(x^*)| \leq \epsilon(x^*) \quad (1.12)$$

即

$$x^* - \epsilon(x^*) \leq x \leq x^* + \epsilon(x^*) \quad (1.13)$$

这个不等式,在应用上常常采用如下写法

$$x = x^* \pm \epsilon(x^*) \quad (1.14)$$

例如,用毫米刻度的米尺测量一长度 x 时,如果该长度接近某一刻度 x^* ,则 x^* 作为 x 的近似值时

$$|e(x^*)| = |x^* - x| \leq \frac{1}{2} \text{ mm} = 0.5 \text{ mm}$$

它的误差限是 $\epsilon(x^*) = 0.5 \text{ mm}$ 。如果读出的长度为 $x^* = 765$,则 $|765 - x| \leq 0.5$,从这个不等式仍不能知道准确的 x 值,只知道 $764.5 \leq x \leq 765.5$ 。即 x 在区间 $[764.5, 765.5]$ 内。

绝对误差还不足以刻画近似数的精确程度,例如,有两个量 $x = 10 \pm 1$, $y = 1\,000 \pm 10$,虽然 x 的绝对误差限比 y 的绝对误差限小,但 $\frac{\epsilon(y^*)}{y^*} = \frac{10}{1\,000} = 1\%$

要比 $\frac{\epsilon(x^*)}{x^*} = \frac{1}{10} = 10\%$ 小,这说明 $y^* = 1\,000$ 作为 y 的近似值远比 $x^* = 10$ 作为 x 的近似值的近似程度要好得多。所以,除考虑误差的大小外,还应考虑准确值 x 本身的大小。我们把近似值的误差 $e(x^*)$ 与准确值 x 的比值,记作

$$e_r(x^*) \triangleq \frac{e(x^*)}{x} = \frac{x^* - x}{x} \quad (1.15)$$

称为近似值 x^* 的相对误差。

在实际计算中,由于真值 x 总是未知的,且由于

$$\frac{e(x^*)}{x} - \frac{e(x^*)}{x^*} = \frac{e(x^*)(x^* - x)}{xx^*} = \frac{[e_r(x^*)]^2}{1 + e_r(x^*)}$$

是 $e_r(x^*)$ 的平方项级,故当 $e_r(x^*)$ 较小时,常取

$$e_r(x^*) = \frac{e(x^*)}{x^*} = \frac{x^* - x}{x^*} \quad (1.16)$$

相对误差也可正可负,它的绝对值的上界称为该近似值的相对误差限,记作 $\epsilon_r(x^*)$,即

$$|e_r(x^*)| \leq \frac{\epsilon(x^*)}{|x^*|} \triangleq \epsilon_r(x^*) \quad (1.17)$$

由定义可知,绝对误差与绝对误差限是有量纲的量,而相对误差和相对误差限是无量纲的量。

1.3.3 有效数字

如果近似值 x^* 的误差限是某一位的半个单位,该位到 x^* 的第一位非零数字共有 n 位,则称 x^* 有 n 位有效数字。

例如, $x = \pi = 3.141\,592\,653\,58\cdots$ 取 $x^* = 3.14$ 时

$$|x^* - x| \leq 0.002 < 0.005$$

所以, $x^* = 3.14$ 作为 π 的近似值时,就有 3 位有效数字;而取 $x^* = 3.141\,6$ 时,

$$|x^* - x| \leq 0.000\,008 < 0.000\,05$$

所以, $x^* = 3.141\,6$ 作为 π 的近似值时,就有 5 位有效数字。一般地,在 r 进制中,设近似值 x^* 可表示为

$$x^* = \pm(a_1 r^{-1} + a_2 r^{-2} + \cdots + a_n r^{-n}) \times r^m \quad (1.18)$$

$a_1 \neq 0$, 且

$$|x^* - x| \leq \frac{1}{2} r^{m-n} \quad (1.19)$$

则由定义可知, x^* 有 n 位有效数字。

当 $r=10$ 时,式(1.18)中表示十进制数,而当 $r=2$ 时,式(1.18)表示二进制规格化浮点数。

例 1 按四舍五入原则,写出下列各数具有 5 位有效数字的近似数

187.932 5, 0.037 855 51, 8.000 033, 2.718 281 8

解 按定义,上述各数具有 5 位有效数字的近似数分别是

187.93, 0.037 856, 8.000 0, 2.718 3

注意 $x=8.000\,033$ 的 5 位有效数字是 8.000 0,而不是 8,8 只有 1 位有效数字。

式(1.18)说明,有效位数与小数点的位置无关,而具有 n 位有效数字的近似数 x^* 其误差限为

$$\epsilon(x^*) = \frac{1}{2} r^{m-n} \quad (1.20)$$

在 m 相同的条件下,有效位数越多,则绝对误差限越小。而有效数字与相对误差限有下列关系。

定理 1.1 用式(1.18)表示的近似数 x^* ,若具有 n 位有效数字,则其相对

误差限为

$$|e_r(x^*)| \leq \frac{1}{2a_1} r^{-(n-1)} \quad (1.21)$$

证明 由式(1.18)知, $|x^*| \geq a_1 \cdot r^{m-1} > 0$, 故

$$|e_r(x^*)| = \frac{|x^* - x|}{|x^*|} \leq \frac{\frac{1}{2} r^{m-n}}{a_1 r^{m-1}} = \frac{1}{2a_1} r^{-(n-1)}$$

定理 1.2 由式(1.18)表示的近似数 x^* , 若满足

$$|e_r(x^*)| \leq \frac{1}{2(a_1 + 1)} r^{-(n-1)}$$

则 x^* 至少有 n 位有效数字。

证明 因为 $|x^* - x| = |x^*| \cdot |e_r(x^*)|$, 且 $|x^*| \leq (a_1 + 1) \cdot r^{m-1}$
故

$$|x^* - x| \leq (a_1 + 1) r^{m-1} \cdot \frac{1}{2(a_1 + 1)} r^{-(n-1)}$$

故 x^* 至少有 n 位有效数字。

定理 1.1, 1.2 说明, 近似数 x^* 的有效位数越多, 它的相对误差限越小; 反之, x^* 的相对误差越小, 它的有效位数越多。

例 2 要使 $\sqrt{20}$ 的近似值的相对误差限小于 0.1%, 要取几位有效数字。

解 由于 $4 < \sqrt{20} < 5$, 所以 $a_1 = 4$, 由定理 1.1 有

$$\frac{1}{2a_1} \times 10^{-(n-1)} \leq 0.1\%$$

即 $10^{n-4} \geq \frac{1}{8}$, 得 $n \geq 4$ 。故只要对 $\sqrt{20}$ 的近似数取 4 位有效数字, 其相对误差

就可小于 0.1%。因此, 可取 $\sqrt{20} \approx 4.472$ 。

1.3.4 误差传播

1. 误差分析的重要性

在数值计算方法中, 除了研究数学问题的算法外, 还要研究计算结果的误差是否满足精度要求, 这就是误差估计问题。下面举例说明误差分析的重要性。

例 3 计算 $I_n = \int_0^1 \frac{x^n}{x+10} dx$, 并估计误差。

解 因为

$$I_n + 10I_{n-1} = \int_0^1 \frac{x^n + 10x^{n-1}}{x+10} dx = \int_0^1 x^{n-1} dx = \frac{1}{n}$$

可得递推关系

$$I_n = \frac{1}{n} - 10I_{n-1}, n=1, 2, \dots$$

其中

$$I_0 = \int_0^1 \frac{1}{x+10} dx = \ln 11 - \ln 10 = \ln 1.1$$

如果取 $\tilde{I}_0 = 0.095\,310$ 作为 I_0 的近似, 则其误差为

$$|e(\tilde{I}_0)| = |\tilde{I}_0 - I_0| \leq 0.000\,000\,2$$

并由递推公式

$$(A) \quad \begin{cases} \tilde{I}_0 = 0.095\,310 \\ \tilde{I}_n = \frac{1}{n} - 10\tilde{I}_{n-1}, n=1, 2, \dots \end{cases}$$

计算结果见表 1.1 的 $\tilde{I}_n$ 列。

从表中看出, $\tilde{I}_5 > \tilde{I}_4$, 且 $\tilde{I}_6$ 出现负值, 这与一切 $\tilde{I}_{n-1} > \tilde{I}_n > 0$ 相矛盾。因此, 当 n 较大时, 用 $\tilde{I}_n$ 近似 I_n 显然是不正确的。这里计算公式与每步计算都是正确的, 那么什么原因使计算结果出现错误呢? 主要就是初值 $\tilde{I}_0$ 有误差 $e(\tilde{I}_0)$, 由此引起以后各步计算的误差 $e(\tilde{I}_n)$, 它满足关系

$$e(\tilde{I}_n) = -10e(\tilde{I}_{n-1}), n=1, 2, \dots$$

从而

$$e(\tilde{I}_n) = (-10)^n e(\tilde{I}_0)$$

这说明 $\tilde{I}_0$ 有误差 $e(\tilde{I}_0)$, 则 $\tilde{I}_n$ 就有 $e(\tilde{I}_0)$ 的 $(-10)^n$ 倍误差。

下面换一种计算方法。当 $0 < x < 1$ 时

$$\frac{1}{11}x^n \leq \frac{x^n}{10+x} \leq \frac{1}{10}x^n$$

所以

$$\frac{1}{11(n+1)} \leq I_n \leq \frac{1}{10(n+1)}$$

粗略地取

$$I_7^* = \frac{1}{2} \left(\frac{1}{11 \times 8} + \frac{1}{10 \times 8} \right) = 0.011\,932$$

然后将递推公式倒过来使用, 即由公式

$$(B) \quad \begin{cases} I_7^* = 0.011\,932 \\ I_{n-1}^* = \frac{1}{10} \left(\frac{1}{n} - I_n^* \right), n=7, 6, \dots, 1 \end{cases}$$

计算结果见表 1.1 的 I_n^* 列。尽管 I_7^* 是粗略地取的,有很大误差 $e(I_7^*)$,但因误差随传播逐步缩小, $e(I_0^*)$ 比 $e(I_n^*)$ 缩小了 $(-10)^n$ 倍。故计算的数值可靠,可用 I_n^* 近似 I_n 。

表 1.1

n	$\tilde{I}_n$	I_n^*
0	0.095 310	0.095 310
1	0.046 900	0.046 898
2	0.031 000	0.031 018
3	0.023 333	0.023 514
4	0.016 667	0.018 454
5	0.033 333	0.015 357
6	-0.166 667	0.013 093
7	1.809 524	0.011 932

此例说明,在数值计算中如不注意误差分析,用了类似(A)的计算公式,就会出现“差之毫厘,失之千里”的错误结果。尽管数值计算中估计误差比较困难,仍应重视计算过程中的误差分析。

2. 四则运算的误差传播

设 x_1, x_2 的近似值分别为 x_1^*, x_2^* , 有误差

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

如果以 $x_1^* + x_2^*$, $x_1^* - x_2^*$ 分别作为 $x_1 + x_2$, $x_1 - x_2$ 的近似值,则有

$$e(x_1^* \pm x_2^*) = e(x_1^*) \pm e(x_2^*) \quad (1.22)$$

即和的误差是误差之和,差的误差是误差之差,进一步有

$$|e(x_1^* \pm x_2^*)| \leq |e(x_1^*)| + |e(x_2^*)|$$

即

$$\epsilon(x_1^* \pm x_2^*) = \epsilon(x_1^*) + \epsilon(x_2^*) \quad (1.23)$$

所以误差限之和是和或差的误差限。以上的结果适用于任意多个近似数的和或差。而相对误差有

$$e_r(x_1^* + x_2^*) = \frac{x_1}{x_1 + x_2} e_r(x_1^*) + \frac{x_2}{x_1 + x_2} e_r(x_2^*) \quad (1.24)$$

即和的相对误差等于各项相对误差的加权平均。

若 x_1 与 x_2 同号,则式(1.24)右端 $e_r(x_1^*)$ 与 $e_r(x_2^*)$ 的系数满足

$$0 < \frac{x_1}{x_1 + x_2}, \frac{x_2}{x_1 + x_2} < 1$$

且
$$\frac{x_1}{x_1 + x_2} + \frac{x_2}{x_1 + x_2} = 1 \quad (1.25)$$

此时,由式(1.24)可得

$$|e_r(x_1^* + x_2^*)| \leq \max\{|e_r(x_1^*)|, |e_r(x_2^*)|\}$$

即

$$\epsilon_r(x_1^* + x_2^*) \leq \max\{\epsilon_r(x_1^*), \epsilon_r(x_2^*)\} \quad (1.26)$$

和的相对误差限不超过各项相对误差限中的最大者。

若 x_1 与 x_2 异号,则式(1.24)中两个系数的绝对值至少有一个大于1,如果这时 x_1 与 $-x_2$ 相当接近,则式(1.24)中的两个系数的绝对值都可能很大,从而使 $e_r(x_1^* + x_2^*)$ 很大,在这种情况下,原始数据的误差会对计算结果产生相当大的影响。

如果以 $x_1^* \cdot x_2^*$ 与 $\frac{x_1^*}{x_2^*}$ 分别作为 $x_1 \cdot x_2$ 与 $\frac{x_1}{x_2}$ 的近似值,则有

$$e(x_1^* \cdot x_2^*) \approx x_2^* e(x_1^*) + x_1^* e(x_2^*) \quad (1.27)$$

$$e\left(\frac{x_1^*}{x_2^*}\right) \approx \frac{x_2^* e(x_1^*) - x_1^* e(x_2^*)}{(x_2^*)^2} \quad (1.28)$$

于是

$$\epsilon(x_1^* \cdot x_2^*) \approx |x_2^*| \epsilon(x_1^*) + |x_1^*| \epsilon(x_2^*) \quad (1.29)$$

$$\epsilon\left(\frac{x_1^*}{x_2^*}\right) \approx \frac{|x_2^*| \epsilon(x_1^*) + |x_1^*| \epsilon(x_2^*)}{|x_2^*|^2} \quad (1.30)$$

例4 求解二次方程 $x^2 - 26x + 1 = 0$,并估计误差。

解 利用二次方程的求根公式得

$$x_1 = 13 - \sqrt{168}, x_2 = 13 + \sqrt{168}$$

取 $\sqrt{168} = 12.961$, 有 $|\sqrt{168} - 12.961| \leq 0.0005$, 于是

$$x_1^* = 13 + 12.961 = 25.961$$

$$x_2^* = 13 - 12.961 = 0.039$$

$$|e(x_1^*)| = |e(x_2^*)| \leq 0.0005$$

而相对误差

$$|e_r(x_1^*)| \leq \frac{0.0005}{25.961} \approx 1.9 \times 10^{-5}$$

$$|e_r(x_2^*)| \leq \frac{0.0005}{0.039} \approx 1.3 \times 10^{-2}$$

尽管 x_2^* 的绝对误差比较小,但相对误差却很大。原因是计算 x_2^* 时有效数字的丢失比较多。所以,如果把计算 x_2^* 的公式改为

$$x_2 = 13 - \sqrt{168} = \frac{1}{13 + \sqrt{168}}$$