



21世纪高等学校教材

全国优秀畅销书

计算方法与实习

AN ELEMENTARY NUMERICAL ANALYSIS

袁慰平 孙志忠 吴宏伟 闻震初

(第 4 版)

东南大学出版社

SOUTHEAST UNIVERSITY PRESS

计算方法与实习

(第4版)

袁慰平 孙志忠 编
吴宏伟 闻震初

东南大学出版社

内 容 提 要

全书分两篇。第1篇为计算方法,包括误差分析、方程求根、线性方程组求解、函数插值、曲线拟合、数值微积分、常微分方程数值解法及矩阵特征值计算等8章,各章末有应用实例、内容小结、复习思考题和习题;第2篇为计算实习,供学生自学,用于指导学生上机实习,与第1篇各章相应共有8个实习,每一实习均给出了该实习的目的与要求、算法概要、用C++语言和Matlab编写并调试通过的程序、实例及上机实习题和答案。

本书取材适当,思路清晰,富有启发性,便于教学,可作为高等工科院校非数学专业学生的教材,也可作同等程度的自学教材,或科技人员的参考书。

图书在版编目(CIP)数据

计算方法与实习/袁慰平等编. —4版. —南京:东南大学出版社,2005.7

ISBN 7-5641-0199-7

I. 计... II. 袁... III. 电子计算机—计算方法—高等学校—教学参考资料 IV. TP301.6

中国版本图书馆CIP数据核字(2005)第047215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 姜堰市晨光印刷有限公司印刷
开本:700mm×1000mm 1/16 印张:18 字数:353千字

2005年12月第4版 2005年12月第17次印刷

印数:113001~118000 定价:23.00元

前 言

随着计算机的迅速发展和广泛应用,在众多的领域内,人们愈来愈认识到科学计算是科学研究的第3种方法,当今理工科大学学生,更应该具备这方面的知识与能力。“计算方法”是科学计算的一门主干课程,绝大多数理工科专业都已经开设了这门课程,并作为必修课程。计算数学是数学与计算机科学的交叉学科,它兼有这两门学科的基本特征,即既有数学的抽象性与严密性,又有计算机科学的实践性与技术性。基于以上认识,我们认为“计算方法”课程不但要讲授常用的计算方法,也要注重使学生能将所学内容在计算机上具体实现。在教材中,我们专门安排了计算实习的有关内容。

本书分两篇,第1篇为计算方法,用于课堂讲授,其中第2章、第4章至第7章均有应用实例一节,使学生知道计算方法在实际问题中的应用,这些内容可由学生自学;第2篇为计算实习,用于学生上机演算,应与第1篇平行使用。

1988年出了本书第1版,而当时计算方法与计算实习两者紧密结合的教材还不多见,故出版后深受同行的关注与支持。在吸取了广大读者的宝贵意见并结合自身在教学实践中的经验与体会,我们做了多次修改,不断完善,使之形成今天的第4版。第4版着重对第2篇计算实习作了修改,一是将原来的C语言程序改为C++程序;二是增加了Matlab语言编写的程序;三是给出了部分实习题答案。

学有余力的同学课外可同步阅读参考文献[1]。

本书第1版和第2版由袁慰平、张令敏、黄新芹、闻震初编写,从第3版开始由孙志忠、吴宏伟、袁慰平、闻震初编写。我们在编写中力求准确、精炼,但限于水平,可能还会有疏漏和不妥之处,恳请读者批评指正。

本书在编写过程中得到了东南大学教务处、数学系及东南大学出版社等领导 and 计算数学教研室的老师们的关心和支持,在此一并表示感谢。

编 者

2005年4月

目 录

第 1 篇 计算方法

1 绪论	(1)
1.1 计算方法的对象与特点	(1)
1.2 误差的来源及误差的基本概念	(1)
1.2.1 误差的来源	(1)
1.2.2 绝对误差与绝对误差限	(2)
1.2.3 相对误差与相对误差限	(2)
1.2.4 有效数字	(3)
1.2.5 数据误差的影响	(4)
1.3 机器数系	(6)
1.3.1 数的浮点表示	(6)
1.3.2 机器数系	(6)
1.3.3 机器数的相对误差限	(8)
1.4 误差危害的防止	(8)
1.4.1 使用数值稳定的计算公式	(8)
1.4.2 尽量避免两相近数相减	(11)
1.4.3 尽量避免用绝对值很大的数作乘数	(12)
1.4.4 防止大数“吃掉”小数	(12)
1.4.5 注意简化计算步骤,减少运算次数	(12)
小结	(14)
复习思考题	(14)
习题 1	(15)
2 方程求根	(17)
2.1 问题的提出	(17)
2.2 二分法	(18)
2.3 迭代法	(20)
2.3.1 迭代格式的构造及其敛散性条件	(20)

2.3.2	迭代法的局部收敛性	(26)
2.3.3	迭代法的收敛速度	(27)
2.3.4	埃特金加速法	(29)
2.4	牛顿法与割线法	(31)
2.4.1	牛顿迭代公式	(31)
2.4.2	局部收敛性	(32)
2.4.3	大范围收敛性	(34)
2.4.4	割线法	(36)
2.5	代数方程求根的劈因子法	(37)
2.6	应用实例:任一平面与螺旋线全部交点的计算	(40)
2.6.1	数学模型	(40)
2.6.2	关于交点个数的讨论	(41)
2.6.3	根的求法	(44)
2.6.4	根的个数趋于无穷时的“实时”求交点方法	(46)
	小结	(47)
	复习思考题	(47)
	习题2	(47)
3	线性方程组数值解法	(50)
3.1	问题的提出	(50)
3.2	消去法	(51)
3.2.1	三角方程组的解法	(51)
3.2.2	高斯消去法	(52)
3.2.3	追赶法	(57)
3.2.4	列主元高斯消去法	(58)
3.3	矩阵的直接分解及其在解方程组中的应用	(60)
3.3.1	矩阵分解的紧凑格式	(60)
3.3.2	改进平方根法	(64)
3.3.3	列主元三角分解法	(66)
3.4	向量范数和矩阵范数	(68)
3.4.1	向量范数	(68)
3.4.2	矩阵范数	(69)
3.5	迭代法	(71)

3.5.1	迭代法及其收敛性	(71)
3.5.2	雅可比迭代法	(75)
3.5.3	高斯-赛德尔迭代法	(77)
	小结	(79)
	复习思考题	(80)
	习题3	(81)
4	插值法	(84)
4.1	问题的提出	(84)
4.1.1	插值函数的概念	(84)
4.1.2	插值多项式的存在唯一性	(85)
4.2	拉格朗日插值多项式	(86)
4.2.1	基本插值多项式	(86)
4.2.2	拉格朗日插值多项式	(87)
4.2.3	插值余项	(88)
4.2.4	一类带导数插值条件的插值	(91)
4.3	差商、差分及牛顿插值多项式	(92)
4.3.1	差商及牛顿插值多项式	(93)
4.3.2	差分及等距节点插值公式	(97)
4.4	高次插值的缺点及分段插值	(100)
4.4.1	高次插值的误差分析	(100)
4.4.2	分段线性插值	(101)
4.4.3	分段二次插值	(103)
4.5	样条插值函数	(104)
4.5.1	三次样条插值函数	(105)
4.5.2	三次样条插值函数的求法	(105)
4.6	应用实例:丙烷导热系数的计算	(110)
	小结	(112)
	复习思考题	(112)
	习题4	(113)
5	曲线拟合	(115)
5.1	最小二乘原理	(115)
5.2	超定方程组的最小二乘解	(121)
5.3	应用实例:价格、广告与赢利	(123)

小结	(126)
复习思考题	(126)
习题 5	(126)
6 数值积分与数值微分	(128)
6.1 数值积分问题的提出	(128)
6.2 插值型求积公式	(129)
6.2.1 插值型求积公式	(129)
6.2.2 梯形公式、辛卜生公式和柯特斯公式	(130)
6.2.3 插值型求积公式的截断误差与代数精度	(131)
6.2.4 梯形公式、辛卜生公式和柯特斯公式的截断误差	(133)
6.3 复化求积公式	(135)
6.3.1 复化梯形公式	(135)
6.3.2 复化辛卜生公式	(136)
6.3.3 复化柯特斯公式	(137)
6.3.4 复化求积公式的阶	(139)
6.3.5 步长的自动选择	(139)
6.4 龙贝格求积公式	(140)
6.5 高斯求积公式简介	(145)
6.6 重积分的计算	(148)
6.7 数值微分	(151)
6.7.1 数值微分问题的提出	(151)
6.7.2 插值型求导公式及截断误差	(152)
6.8 应用实例:椭圆轨道长度的计算	(155)
小结	(157)
复习思考题	(157)
习题 6	(158)
7 常微分方程数值解法	(160)
7.1 问题的提出	(160)
7.2 欧拉方法	(160)
7.2.1 欧拉公式	(160)
7.2.2 梯形公式	(163)
7.2.3 改进欧拉公式	(164)
7.2.4 整体截断误差	(166)
7.3 龙格-库塔方法	(167)
7.3.1 龙格-库塔方法的基本思想	(167)

7.3.2	二阶龙格-库塔公式	(168)
7.3.3	高阶龙格-库塔公式	(169)
7.4	线性多步法	(173)
7.4.1	阿当姆斯内插公式	(174)
7.4.2	阿当姆斯外推公式	(175)
7.4.3	阿当姆斯预测校正公式	(176)
7.5	一阶方程组与高阶方程	(178)
7.5.1	一阶方程组	(178)
7.5.2	化高阶方程为一阶方程组	(179)
7.6	应用实例:摆球振动	(181)
	小结	(183)
	复习思考题	(184)
	习题 7	(184)
8	矩阵的特征值及特征向量的计算	(187)
8.1	问题的提出	(187)
8.2	按模最大与最小特征值的求法	(188)
8.2.1	幂法	(188)
8.2.2	反幂法	(194)
8.3	计算实对称矩阵特征值的雅可比法	(195)
8.4	QR 方法	(204)
8.4.1	矩阵 A 的 QR 分解	(204)
8.4.2	QR 算法	(207)
	小结	(208)
	复习思考题	(209)
	习题 8	(209)

第 2 篇 计算实习

1	舍入误差与数值稳定性	(211)
1.1	舍入误差与数值稳定性	(211)
	实习题 1	(215)
2	方程求根	(216)
2.1	二分法	(216)
2.2	牛顿迭代法	(219)
	实习题 2	(223)

3 线性方程组数值解法	(224)
3.1 列主元高斯消去法	(224)
3.2 矩阵直接三角分解法	(227)
3.3 迭代法	(230)
3.3.1 雅可比迭代法	(230)
3.3.2 高斯-赛德尔迭代法	(233)
实习题 3	(236)
4 插值法	(239)
4.1 拉格朗日插值多项式	(239)
4.2 牛顿插值多项式	(241)
实习题 4	(243)
5 曲线拟合	(244)
5.1 最小二乘法	(244)
实习题 5	(248)
6 数值积分	(249)
6.1 复化梯形公式与复化辛卜生公式的自适应算法	(249)
6.1.1 复化辛卜生公式	(249)
6.1.2 自适应梯形公式	(251)
6.2 龙贝格算法	(254)
实习题 6	(257)
7 常微分方程数值解法	(258)
7.1 改进欧拉方法	(258)
7.2 龙格-库塔方法	(261)
7.3 阿当姆斯方法	(263)
实习题 7	(267)
8 矩阵的特征值与特征向量的计算	(268)
8.1 幂法	(268)
实习题 8	(271)
实习题参考答案	(272)
参考文献	(276)

第 1 篇 计算方法

1 绪论

1.1 计算方法的对象与特点

计算方法是研究数学问题的数值解及其理论的一个数学分支,它涉及面很广,如:代数、微积分、微分方程等都有数值解的问题。自电子计算机成为数值计算的主要工具以来,计算方法主要研究适合于在计算机上使用的数值计算方法及与此相关的理论,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。某些在理论上虽然不够严格,但通过实际计算、对比分析等手段,被证明是行之有效的方法也可采用。因此计算方法除具有数学的抽象性与严格性外,还具有应用的广泛性与实际试验的技术性等,是一门与计算机密切结合的实用性很强的课程。

1.2 误差的来源及误差的基本概念

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,其差称为误差。引起误差的原因是多方面的。

1) 从实际问题转化为数学问题,即建立数学模型时,对被描述的实际问题进行了抽象和简化,忽略了一些次要因素,这样建立的数学模型虽然具有“精确”、“完美”的外衣,其实只是客观现象的一种近似。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) 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,而观测不可避免会带误差,这种误差称为**观测误差**。

3) 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,只能用有限过程来计算。如无穷级数求和,只能取前面有限项求和来近似代替,于是产生了有限过程代替无限过程的误差,称为**截断误差**。这是计算方法本身出现的误

差,所以也称为方法误差,这种误差是本课程中需要特别重视的。

4) 在计算中遇到的数据可能位数很多,也可能是无穷小数,如 $\sqrt{2}$ 、 $1/3$ 等,但计算时只能对有限位数进行运算,因而往往进行四舍五入,这样产生的误差称为舍入误差。少量舍入误差是微不足道的,但在电子计算机上完成了千百万次运算后,舍入误差的积累有时可能是十分惊人的。

由以上误差来源的分析可以看到: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。既然描述问题的方法都是近似的,那么要求解的绝对准确也就没有意义了。因此在计算方法里讨论的都是近似解,那种认为近似解是不可靠的、不准确的看法是错误的,应该认为求近似解是正常的,问题是怎样尽量设法减少误差,提高精度。在4种误差来源的分析中,前2种误差是客观存在的,后2种是由计算方法及计算过程所引起的。本课程是研究数学问题的数值解法,因此只涉及后2种误差。

1.2.2 绝对误差与绝对误差限

定义1 设 x^* 为准确值, x 是 x^* 的一个近似值,称 $e = x^* - x$ 为近似值 x 的绝对误差,简称误差。

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^* ,也不能算出误差 e 的准确值,因为这个值虽然客观存在,但实际计算中是得不到的,得到的只能是误差的某个范围,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,有时也可以表示成 $x^* = x \pm \epsilon$ 。

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 $x = 123 \text{ mm}$,由于直尺以毫米为刻度,所以其误差不超过 0.5 mm ,即

$$|x^* - 123| \leq 0.5$$

从这个不等式我们不能得出准确值 x^* ,但却知道 x^* 的范围

$$122.5 \leq x^* \leq 123.5$$

对于给定的正数 ϵ ,若近似值 x 满足

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

则在允许误差 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 相对误差与相对误差限

误差限的大小还不能完全表示出近似值的好坏。例如,测得光速的近似值为 $x = 299\,796 \text{ km/s}$,误差限为 4 km/s ,约为光速本身的十万分之一,显然测量是非

常准确的;如果测量运动员的跑速,误差限是 0.01 km/s,即 10 m/s,接近运动员的真正跑速,显然这是十分粗糙的测量。为了较好地反映近似值的精确程度,必须考虑误差与真值的比值,即相对误差。

定义 2 设 x^* 为准确值, x 是 x^* 的一个近似值,则称 $(x^* - x)/x^* = e/x^*$ 为近似值 x 的相对误差,记作 e_r 。

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, $\bar{e}_r - e_r$ 是该小量的二阶小量。

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。

1.2.4 有效数字

在工程上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太麻烦,因此希望所写出的数本身就能表示它的准确程度,于是需要引进有效数字的概念。另外,当准确值 x^* 有很多位数时,常常按四舍五入原则得到 x^* 的前几位近似值 x 。例如:

$$x^* = \sqrt{3} = 1.732\ 050\ 808\cdots$$

取 3 位, $x_1 = 1.73$, $\epsilon_1 < 0.005$; 取 5 位, $x_2 = 1.732\ 1$, $\epsilon_2 < 0.000\ 05$ 。它们的误差都不超过末位的半个单位,即

$$|\sqrt{3} - 1.73| < \frac{1}{2} \times 10^{-2}, \quad |\sqrt{3} - 1.732\ 1| < \frac{1}{2} \times 10^{-4}$$

定义 3 如果近似值 x 的误差限是其某一位上的半个单位,且该位直到 x 的第 1 位非零数字一共有 n 位,则称近似值 x 有 n 位有效数字(见图 1-2-1)。

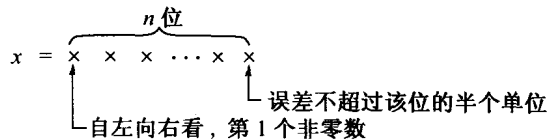


图 1-2-1 有效位数

如 $\sqrt{3}$ 的近似值取 $x_1 = 1.73$,则 x_1 有3位有效数字;取 $x_2 = 1.7321$,则 x_2 有5位有效数字;若取 $x_3 = 1.7320$,则 x_3 只有4位有效数字,因为它的误差限已超过 $\frac{1}{2} \times 10^{-4}$ 。

在讲了有效数字之后,我们规定今后所写出的数都应该是有效数字,如 $\sqrt{3}$ 的近似值根据所需要的不同位数的有效数字应是1.73或1.732或1.7321,而不能是1.7320。同时,在同一问题中,参加运算的数,都应尽可能有相同位数的有效数字。

例1 对下列各数写出具有5位有效数字的近似值:

$$236.478, \quad 0.00234711, \quad 9.000024, \quad 9.000034 \times 10^3$$

按定义,上述各数具有5位有效数字的近似值分别是

$$236.48, \quad 0.0023471, \quad 9.0000, \quad 9.0000 \times 10^3$$

注意 $x^* = 9.000024$ 的5位有效数字近似值是9.0000,而不是9,因为9只有1位有效数字。

例2 指出下列各数有几位有效数字:

$$2.0004, \quad -0.00200, \quad -9000, \quad 9 \times 10^3, \quad 2 \times 10^{-3}$$

按定义,上述各数的有效位数分别是5,3,4,1,1。

1.2.5 数据误差的影响

数值运算中由于所给数据的误差必然引起函数值的误差,这种数据误差的影响较为复杂,一般采用泰勒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 的绝对误差与相对误差。

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) 处作泰勒展开,并取一阶泰勒多项式,则得 $e(y)$ 的近似表示式为

$$\begin{aligned} e(y) &= y^* - y \approx \frac{\partial f(x_1, x_2)}{\partial x_1} (x_1^* - x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} (x_2^* - x_2) \\ &= \frac{\partial f(x_1, x_2)}{\partial x_1} e(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} e(x_2) \end{aligned} \quad (2.1)$$

式中, $e(x_1) = x_1^* - x_1$; $e(x_2) = x_2^* - x_2$ 。

式(2.1)的左端实际上就是函数 $y = f(x_1, x_2)$ 在 (x_1, x_2) 处分别有增量 $\Delta x_1 = x_1^* - x_1 = e(x_1)$, $\Delta x_2 = x_2^* - x_2 = e(x_2)$ 时,函数的全增量 Δy ,因此 $e(y)$

的近似表达式实质上就是 y 的全微分 dy , 即

$$e(y) = \Delta y \approx dy = \frac{\partial f(x_1, x_2)}{\partial x_1} dx_1 + \frac{\partial f(x_1, x_2)}{\partial x_2} dx_2$$

函数的相对误差

$$\begin{aligned} e_r(y) &= \frac{e(y)}{y} \approx \frac{\partial f(x_1, x_2)}{\partial x_1} \frac{x_1}{y} \frac{e_1}{x_1} + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{x_2}{y} \frac{e_2}{x_2} \\ &= \frac{\partial f(x_1, x_2)}{\partial x_1} \frac{x_1}{y} e_r(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{x_2}{y} e_r(x_2) \end{aligned} \quad (2.2)$$

式中, $e_r(x_1), e_r(x_2)$ 分别是 x_1, x_2 的相对误差。

利用函数值的误差估计式(2.1)和(2.2), 可以得到两数和、差、积、商的误差估计:

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

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

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

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

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

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

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

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

例 3 已测得某物体行程 s^* 的近似值 $s = 800$ m, 所需时间 t^* 的近似值 $t = 35$ s. 若已知 $|t^* - t| \leq 0.05$ s, $|s^* - s| \leq 0.5$ m, 试求平均速度 v 的绝对误差限和相对误差限。

解 因为 $v = \frac{s}{t}$, 由商的误差估计式有

$$e(v) = e\left(\frac{s}{t}\right) \approx \frac{1}{t} e(s) - \frac{s}{t^2} e(t), \quad e_r(v) = e_r\left(\frac{s}{t}\right) \approx e_r(s) - e_r(t)$$

得

$$\begin{aligned} |e(v)| &\approx \left| \frac{1}{t} e(s) - \frac{s}{t^2} e(t) \right| \\ &\leq \frac{1}{t} |e(s)| + \frac{s}{t^2} |e(t)| \\ &\leq \frac{1}{35} \times 0.5 + \frac{800}{35^2} \times 0.05 \approx 0.0469 \leq 0.05 \end{aligned}$$

$$|e_r(v)| \approx |e_r(s) - e_r(t)|$$

$$\begin{aligned} &\leq |e_r(s)| + |e_r(t)| \\ &\leq \frac{0.5}{800} + \frac{0.05}{35} \approx 0.00205 \end{aligned}$$

所以平均速度 v 的绝对误差限和相对误差限分别为 0.05 和 0.00205。

应该指出,在由误差估计式得出绝对误差限和相对误差限的估计时,由于取了绝对值并用三角不等式放大,因此是按最坏情形得出的,所以由此得出的结果是很保守的。事实上,出现最坏情形的可能性是很小的。因此近年来出现了一系列关于误差的概率估计。一般说来为了保证运算结果的精确度,只要根据运算量的大小,比结果中所要求的有效数字的位数多取 1 位或 2 位进行计算就可以了。

1.3 机器数系

1.3.1 数的浮点表示

一个实数在科学计算中常常被表示成浮点形式。例如 456.789, -6.473, 0.00567, 0.321 等被分别表示成 0.456789×10^3 , -0.6473×10^1 , 0.567×10^{-2} , 0.321×10^0 , 其中 0.456789, -0.6473, 0.567, 0.321 等称为浮点表示的尾数部, $10^3, 10^1, 10^{-2}, 10^0$ 等称为浮点表示的定位部, 这种表示形式可以使得一个数的数量级一目了然, 更重要的是它可以扩大计算机表示数的范围。

一个基数为 β 的 t 位数字的浮点表示形式为

$$x = (\pm 0.a_1a_2 \cdots a_t) \beta^p \quad (3.1)$$

式中, $\beta \geq 2$ 是整数, 通常取 $\beta = 2, 8, 10, 16$; 每个 a_i 都是整数, 且 $0 \leq a_i \leq \beta - 1$; t 是计算机的字长; 带有符号的整数 p 称为指数, 也称为计算机的阶码, 它有固定的下限 L 和上限 U , 即 $L \leq p \leq U$, L, U 和 t 是由该计算机的硬件所决定的某些常数。尾数部

$$s = \pm 0.a_1a_2 \cdots a_t = \pm \left(\frac{a_1}{\beta} + \frac{a_2}{\beta^2} + \cdots + \frac{a_t}{\beta^t} \right) \quad (3.2)$$

而

$$x = s \times \beta^p \quad (3.3)$$

若规定 $a_1 \neq 0$, 则 $\beta^{-1} \leq |s| < 1$, 此时 x 称为规格化浮点数。今后除特别指出外, 都认为浮点数是规格化表示的。

1.3.2 机器数系

上述数的浮点表示, 几乎是当今所有计算机都采用的表示法。把计算机中浮点数所组成的集合加上“机器零”记为 F , 则 F 被以下 4 个参数所描述: 基数 β 、字长 t 、阶码范围 $[L, U]$, 我们称这个集合 F 为机器数系, 需要指出的是机器数系 F 是一个

离散分布不均匀的有限集。

例如, 设有一个二进制的 2 位字长的计算机, 即 $\beta = 2, t = 2$, 其指数 $p \in [-1, 1]$, 则它所能表示的数只有如下几个:

当 $p = -1$ 时, 有 $\pm(0.10 \times 2^{-1})_2 = \pm(0.25)_{10}$, $\pm(0.11 \times 2^{-1})_2 = \pm(0.375)_{10}$;

当 $p = 0$ 时, 有 $\pm(0.10 \times 2^0)_2 = \pm(0.5)_{10}$, $\pm(0.11 \times 2^0)_2 = \pm(0.75)_{10}$;

当 $p = 1$ 时, 有 $\pm(0.10 \times 2^1)_2 = \pm(1)_{10}$, $\pm(0.11 \times 2^1)_2 = \pm(1.5)_{10}$ 。

加上机器零, 共 13 个数, 它们在数轴上的表示见图 1-3-1。

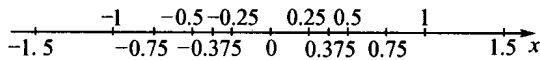


图 1-3-1 机器数的表示

不难证明集合 F 仅含有

$$1 + 2(\beta - 1)\beta^{r-1}(U - L + 1) \quad (3.4)$$

个数, 而且这些数不是等间隔地分布在数轴上。

当 $\beta = 10, t = 4, -L = U = 99$ 时, 此计算机的机器数系 F 仅含有 3 582 001 个数, -0.1000×10^{-99} 和 0.1000×10^{-99} 是该数系 F 中绝对值最小的非零数, 而 -0.9999×10^{99} 和 0.9999×10^{99} 分别是此数系 F 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。

由于机器数系有上述特性, 因此一个实数 x 进入计算机后, 成为计算机里的数, 称它为 x 的机器数, 用 $\text{fl}(x)$ 表示, 一般讲 $\text{fl}(x)$ 只是 x 的一个近似值。

例如, 对于数 $x = 0.6$, 在上述二进制 2 位字长计算机的机器数系 F 里找不到这个数, 通常取与 x 最靠近的数 0.5 作为 x 的近似值, 即它的机器数 $\text{fl}(0.6) = (0.10 \times 2^0)_2$ 。

目前的计算机分截断机和舍入机两种。对于截断机, $\text{fl}(x)$ 取 x 的前 t 位数字; 对于舍入机, $\text{fl}(x)$ 按四舍五入原则取 x 的前 t 位数字。

例 4 假设具有十进制、3 位字长、 $-L = U = 5$ 的 2 台计算机, 一台是截断机, 另一台是舍入机, 则它们对下述实数的规格化浮点数如表 1-3-1 所示。

表 1-3-1 实数的浮点表示

实数	截断机浮点数	舍入机浮点数
127 8	0.127×10^4	0.128×10^4
$-43 \frac{1}{3}$	-0.433×10^2	-0.433×10^2
0.005 669	0.566×10^{-2}	0.567×10^{-2}
123 456	溢出	溢出