

袁慰平 张令敏 黄新芹 闻震初

计算方法与实习

Computing Methods and Laboratory

东南大学出版社

Southeast University Press

计算方法与实习

袁慰平 张令敏 编
黄新芹 闻震初

何旭初 主审

东南大学出版社

内 容 提 要

全书分两篇。第一篇为计算方法，包括误差分析、方程求根、线性方程组求解、函数插值与曲线拟合、数值微积分、微分方程数值解法及矩阵特征值计算等八章，每章末均有复习思考题和习题，用于课堂教学；第二篇为计算实习，用于指导学生上机实习和供学生自学，其中共有七个实习，对主要算法都给出了框图、程序、例题及上机实习题。

该书取材适当，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。

责任编辑 徐步政

责任校对 张燕明

计算方法与实习

袁慰平 张令敏 编
黄新芹 闻震初

东南大学出版社出版

南京四牌楼2号

江苏省新华书店发行 东南大学印刷厂印刷

开本787×1092毫米 1/16印张 17.5字数 426千字

1989年1月第1版 1989年1月第1次印刷

印数：1—8000册

ISBN 7-81023-135-1

前 言

随着电子计算机的迅速发展和广泛应用，在众多的领域内，科学计算已逐渐超过和代替了实验方法。人们愈来愈认识到科学计算是科学研究的第三种方法，特别是工科大学的学生，应当具备这方面的知识与能力。现在，很多工科专业已经开设“计算方法”课，并列为大学生的必修课程。

本书是在多年讲授该课程的讲义基础上修订而成的。初稿自1982年以来在本校及兄弟院校中使用过多遍，反映良好。在内容取舍上，本书力求精简，并重视理论联系实际，在讲授计算方法的基本内容及计算机上的常用算法的同时，专门安排了计算实习的有关内容，包括主要算法的程序框图及用 BASIC 语言编写的程序与例题。我们认为，只要牢固地掌握了这些基本方法，便不难进一步学习计算方法中其它更为深入的内容。

全书分两篇，第一篇为计算方法，用于课堂讲授，一般需 40~45 学时；第二篇为计算实习，用于学生上机演算，应与第一篇并行使用。读者最好在读通例题后，转换为其它语言（如 FORTRAN 或 PASCAL 语言），再上机演算。

本书第一篇由袁慰平、张令敏编写，第二篇由黄新芹、闻震初编写。由于水平有限，缺点与错误在所难免，恳请读者批评指正。

南京大学数学系何旭初教授和王嘉松副教授仔细地审阅了全部书稿，提出了不少宝贵意见，我们作了修改和补充，最后又由何旭初教授审定。在本书使用、修改和出版过程中得到本教研组有关同志及东南大学出版社的支持与帮助，在此一并表示衷心感谢。

编 者

一九八八年六月

目 录

第一篇 计算方法

第一章 绪论

§ 1 计算方法的对象与特点	(1)
§ 2 误差的来源及误差的基本概念	(1)
2-1 误差的来源	(1)
2-2 绝对误差与绝对误差限	(2)
2-3 相对误差与相对误差限	(2)
2-4 有效数字	(3)
2-5 数据误差的影响	(4)
§ 3 机器数系	(5)
3-1 数的浮点表示	(5)
3-2 机器数系	(6)
3-3 机器数的相对误差限	(7)
§ 4 误差危害的防止	(7)
复习思考题	(13)
习题一	(13)

第二章 方程求根

§ 1 问题的提出	(15)
§ 2 二分法	(16)
§ 3 迭代法	(18)
3-1 迭代格式的构造及其收敛性	(18)
3-2 埃特金加速法	(23)
§ 4 牛顿迭代法	(26)
4-1 迭代格式的构造及其局部收敛性	(26)
4-2 简化牛顿法	(28)
4-3 拟牛顿法	(30)
4-4 牛顿下山法	(31)
§ 5 代数方程求根的劈因子法	(32)
复习思考题	(36)
习题二	(36)

第三章 线性方程组数值解法

§ 1 问题的提出	(38)
§ 2 消去法	(39)
2-1 高斯消去法	(39)
2-2 列主元消去法	(46)
§ 3 矩阵的直接分解及其在解方程组中的应用	(50)
3-1 矩阵分解的紧凑格式	(50)
3-2 改进平方根法	(54)
3-3 追赶法	(55)
§ 4 迭代法及其收敛性	(56)
4-1 雅可比迭代法	(58)
4-2 高斯-赛德尔迭代法	(61)
4-3 迭代法的收敛性	(62)
复习思考题	(67)
习题三	(68)

第四章 插值法

§ 1 问题的提出	(71)
1-1 插值函数的概念	(71)
1-2 插值多项式的存在唯一性	(72)
§ 2 拉格朗日插值多项式	(72)
2-1 线性插值和抛物插值	(73)
2-2 拉格朗日插值多项式	(75)
2-3 插值余项	(76)
§ 3 逐步线性插值	(79)
§ 4 差商、差分及牛顿插值公式	(81)
4-1 差商及牛顿插值公式	(82)
4-2 差分及等距节点插值公式	(86)
§ 5 高次插值的缺点及分段插值	(90)
5-1 高次插值的误差分析	(90)
5-2 分段低次插值	(92)
§ 6 埃尔米特插值	(93)
§ 7 样条插值	(96)
7-1 三次样条插值函数	(97)
7-2 三次样条插值函数的求法	(97)
复习思考题	(101)
习题四	(102)

第五章 曲线拟合法

§ 1 最小二乘原理	(104)
------------	-------

1-1	最小二乘问题	(104)
1-2	用最小二乘法求数据的拟合曲线	(105)
* § 2	正交多项式的曲线拟合	(109)
2-1	广义最小二乘拟合多项式	(109)
2-2	正交多项式的概念	(111)
2-3	勒让特多项式	(112)
2-4	用勒让特多项式作曲线拟合举例	(114)
2-5	等距节点上的正交多项式	(115)
	复习思考题	(120)
	习题五	(121)
第六章 数值积分与数值微分		
§ 1	数值积分问题的提出	(123)
(1-1	构造数值求积公式的基本思想	(123)
1-2	插值型求积公式	(124)
1-3	插值型求积公式的截断误差与代数精度的概念	(126)
§ 2	等距节点的求积公式	(128)
2-1	柯特斯系数	(128)
2-2	几种低阶牛顿-柯特斯公式的截断误差	(131)
2-3	复化求积公式与截断误差	(132)
§ 3	步长的自适应算法	(134)
§ 4	龙贝格求积公式	(135)
§ 5	数值微分	(138)
5-1	数值微分问题的提出	(138)
5-2	插值型的求导公式及截断误差	(139)
	复习思考题	(142)
	习题六	(142)
第七章 常微分方程数值解法		
§ 1	问题的提出	(144)
§ 2	欧拉方法	(145)
2-1	欧拉折线法	(145)
2-2	改进欧拉法及局部截断误差	(147)
§ 3	龙格-库塔方法	(150)
3-1	龙格-库塔方法的基本思想	(150)
3-2	二阶龙格-库塔公式	(150)
3-3	高阶龙格-库塔公式	(152)
3-4	步长的自适应问题	(154)
§ 4	线性多步法	(155)
4-1	阿当姆斯内插公式及误差	(155)

4-2 阿当姆斯外推公式及误差	(157)
§ 5 一阶方程组与高阶方程	(158)
5-1 一阶方程组	(158)
5-2 化高阶方程为一阶方程组	(159)
复习思考题	(161)
习题七	(161)
第八章 矩阵的特征值及特征向量的计算	
§ 1 问题的提出	(163)
§ 2 按模最大与最小特征值的求法	(163)
2-1 幂法	(164)
2-2 反幂法	(169)
§ 3 计算实对称矩阵特征值的雅可比法	(170)
* § 4 QR方法	(178)
4-1 矩阵A的QR分解	(178)
4-2 QR算法	(180)
复习思考题	(181)
习题八	(181)

第二篇 计算实习

实习一 方程求根

§ 1 二分法	(183)
§ 2 牛顿迭代法	(186)
实习题一	(189)

实习二 线性方程组的解法

§ 1 高斯消去法	(190)
§ 2 列主元消去法	(194)
§ 3 直接三角分解法	(204)
§ 4 改进平方根法	(210)
§ 5 追赶法	(212)
§ 6 迭代法	(214)
实习题二	(219)

实习三 插值法

§ 1 拉格朗日插值	(222)
§ 2 埃特金插值	(223)
§ 3 牛顿插值	(225)
实习题三	(228)

实习四 曲线拟合	
§ 1 最小二乘法.....	(229)
实习题四.....	(233)
实习五 数值积分	
§ 1 复化梯形法.....	(234)
§ 2 复化辛普生法.....	(235)
§ 3 自动变步长梯形法.....	(237)
§ 4 龙贝格公式.....	(238)
实习题五.....	(240)
实习六 常微分方程数值解法	
§ 1 欧拉方法.....	(241)
§ 2 龙格-库塔方法	(243)
§ 3 阿当姆斯方法.....	(246)
实习题六.....	(248)
实习七 矩阵的特征值与特征向量的计算	
§ 1 幂法.....	(249)
§ 2 雅可比法.....	(253)
实习题七.....	(260)
附录	
一 PC-8000 微型机上机简介.....	(262)
二 DPS-8 上机简介	(266)

第一篇 计算方法

第一章 绪论

§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。在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,即包括方法的收敛性、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,计算机内存最少的计算方法。有的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。因此计算方法除了有数学的抽象性与严格性的特点外,还有应用的广泛性与实际试验的技术性等特点,它是一门与计算机密切结合的实用性很强的课程。

§2 误差的来源及误差的基本概念

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,而观测难免不带误差,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。因此我们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，不准确的想法是错误的，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。在四种误差来源的分析中，前两种误差是客观存在的，后两种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。

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

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

ϵ 称为近似值 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 0.5mm ，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。

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 $x = 299796$ 公里/秒，误差限为 4 公里/秒，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 公里/秒，即 10 米/秒，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,当 e_r 很小时,也取 $e_r = e/x$ 作为 x 的相对误差。

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

$$|e_r| = \left| \frac{e}{x} \right| \leq \varepsilon_r$$

则称 ε_r 为 x 的相对误差限。

2-4 有效数字

在工程上对于测量得到的数经常表示成 $x \pm e$, 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:

$$x^* = \sqrt{3} = 1.732050808 \dots$$

取 3 位, $x_1 = 1.73, \quad \varepsilon_1 < 0.003$

取 5 位, $x_2 = 1.7321, \quad \varepsilon_2 < 0.00005$

它们的误差都不超过末位数字的半个单位, 即

$$|\sqrt{3} - 1.73| < \frac{1}{2} \times 10^{-2}, \quad |\sqrt{3} - 1.7321| < \frac{1}{2} \times 10^{-4}$$

定义 3 如果近似值的误差限是某一位上的半个单位, 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 有 n 位有效数字。

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; 取 $x_3 = 1.7320$, 则 x_3 只有 4 位有效数字, 因为它的误差已超过 $\frac{1}{2} \times 10^{-4}$ 。

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有效数字位应是 1.73 或 1.732 或 1.7321, 不能是 1.7320。同时, 在同一问题中, 参加运算的数, 都应该有相同位数的有效数。

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

$$236.48, \quad 0.0023471, \quad 9.0000, \quad 9000.0$$

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

例 2 指出下列各数是几位有效数字

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

按定义, 上述各数的有效位数分别是 5, 3, 4, 1, 1。

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，并取到一次项，则得 $e(y)$ 的近似表示式

$$\begin{aligned} e(y) &= y^* - y \approx \frac{\partial f(x_1, x_2)}{\partial x_1} (x_1^* - x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2} (x_2^* - x_2) \\ &= \frac{\partial f(x_1, x_2)}{\partial x_1} e_1 + \frac{\partial f(x_1, x_2)}{\partial x_2} e_2 \end{aligned} \quad (2.1)$$

其中 $e_1 = x_1^* - x_1, e_2 = x_2^* - x_2$ 。

y 的误差 $e(y)$ 实际上就是函数 $y=f(x_1, x_2)$ 在 (x_1, x_2) 处分别有增量 $\Delta x_1 = x_1^* - x_1 = e_1, \Delta x_2 = x_2^* - x_2 = e_2$ 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

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

函数值的相对误差

$$\begin{aligned} e_r(y) &= \frac{e(y)}{y} \approx \frac{\partial f(x_1, x_2)}{\partial x_1} \frac{x_1}{y} \frac{e_1}{x_1} + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{x_2}{y} \frac{e_2}{x_2} \\ &= \frac{\partial f(x_1, x_2)}{\partial x_1} \frac{x_1}{y} e_{1r} + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{x_2}{y} e_{2r} \end{aligned} \quad (2.2)$$

其中 e_{1r}, e_{2r} 分别是 x_1, x_2 的相对误差。

利用函数值的误差估计，可以得到两数和、积、商的误差估计：

$$e(x_1 + x_2) \approx e_1 + e_2$$

$$e(x_1 \cdot x_2) \approx x_2 e_1 + x_1 e_2$$

$$e\left(\frac{x_1}{x_2}\right) \approx \frac{1}{x_2} e_1 - \frac{x_1}{x_2^2} e_2 \quad (x_2 \neq 0)$$

$$e_r(x_1 + x_2) \approx \frac{x_1}{x_1 + x_2} e_{1r} + \frac{x_2}{x_1 + x_2} e_{2r}$$

$$e_r(x_1 x_2) \approx e_{1r} + e_{2r}$$

$$e_r\left(\frac{x_1}{x_2}\right) \approx e_{1r} - e_{2r}$$

例 3 已测得某物体行程 s^* 的近似值为 $s=800$ 米，所需时间 t^* 的近似值为 $t=35$ 秒，

若已知 $|t^* - t| \leq 0.05$ 秒, $|s^* - s| \leq 0.5$ 米, 试求平均速度 v 的绝对误差限和相对误差限。

解 因为 $v = \frac{s}{t}$

所以绝对误差

$$e(v) \approx \frac{\partial v}{\partial s} e(s) + \frac{\partial v}{\partial t} e(t)$$

$$\frac{\partial v}{\partial s} = \frac{1}{t}, \quad \frac{\partial v}{\partial t} = -\frac{s}{t^2}$$

因此在近似值 $s=800$ 米, $t=35$ 秒时

$$\begin{aligned} |e(v)| &\leq \left| \frac{\partial v}{\partial s} \right| |e(s)| + \left| \frac{\partial v}{\partial t} \right| |e(t)| \\ &= \frac{1}{35} \times 0.5 + \frac{800}{35^2} \times 0.05 \approx 0.046939 < 0.05 \end{aligned}$$

所以绝对误差限 $e(v) = 0.05$ 米/秒, 相对误差限为

$$\varepsilon_r(v) = \frac{e(v)}{v} = \frac{0.05}{\frac{800}{35}} < 0.0022 = 0.22\%$$

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, 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, 这样有可能引起在计算中保留过多的数字, 从而增加了不必要的计算量。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一位或两位进行计算就可以了。

§ 3 机器数系

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

$$x = \pm (0.a_1 a_2 \dots a_t) \beta^p \quad (3.1)$$

这里 $\beta \geq 2$ 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 是精度, 为计算机的字长; 带有符号的数 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

$$s = \pm 0.a_1 a_2 \dots a_t$$

$$= \pm \left(\frac{a_1}{\beta} + \frac{a_2}{\beta^2} + \dots + \frac{a_t}{\beta^t} \right) \quad (3.2)$$

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

若规定 $a_1 \neq 0$ ，则 $\beta^{-1} < |s| < 1$ ，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。

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记为 F ，则 F 被以下四个参数所描述：基数 β 、精度 t 、阶码范围 $[L, U]$ ，我们称这个集合为机器数系，需要指出的是机器数系 F 是一个离散的有限集。

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：

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

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

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

及机器零，共 13 个数，它们在数轴上的表示如图 1-1

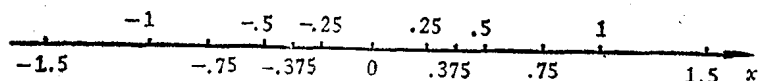


图 1-1

不难证明集合 F 仅含有

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

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 3582001 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 $fl(x)$ 表示，一般讲 $fl(x)$ 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 $fl(0.6)=0.5$ 。

目前的计算机分截断机和舍入机两种。对于截断机， $fl(x)$ 取 x 的前 t 位数字；对于舍入机， $fl(x)$ 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， $-L=U=5$ 的两台计算机，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表示如下：

实数	截断机浮点数	舍入机浮点数
1728	0.127×10^4	0.128×10^4
$-43\frac{1}{3}$	-0.433×10^2	-0.433×10^2
0.005669	0.566×10^{-2}	0.567×10^{-2}
123456	溢出	溢出

这两台计算机能表示的最大数与最小数是 $\pm 0.999 \times 10^5$ ，因此数 123456 超出它所能表示的范围。

3-3 机器数的相对误差限

对于舍入机， $fI(x)$ 与 x 相差至多为小数点右边第 $(t+1)$ 位数字的 $\beta/2$ ，这样，如果 $fI(x)$ 的尾数是 s ，指数是 p ，那么 $x = (s+\eta)\beta^p$ ，其中 $|\eta| \leq \beta^{-t}/2$ ，因此有

$$x - fI(x) = (s+\eta)\beta^p - s\beta^p = \eta\beta^p$$

因而 $fI(x)$ 的相对误差

$$\frac{|x - fI(x)|}{|x|} = \frac{|\eta|}{|s+\eta|} \leq \frac{\frac{1}{2}\beta^{-t}}{|s+\eta|}$$

由于 $fI(x)$ 是规格化的，因此 $|s+\eta| \geq \beta^{-1}$ ，所以相对误差以 $\beta^{t-1}/2$ 为界。

对于截断机， $fI(x)$ 在小数点右边第 $(t+1)$ 位数字上至多相差 β ，因此相对误差限为 β^{1-t}

综上所述，我们有如下结论：在浮点数字范围内，每个非零数 x ，其机器数 $fI(x)$ 的相对误差限为

$$\frac{|x - fI(x)|}{|x|} \leq \begin{cases} \frac{1}{2}\beta^{1-t}, & \text{舍入机} \\ \beta^{1-t}, & \text{截断机} \end{cases} \quad (3.5)$$

所以当使用的计算机确定后，一个机器数的相对误差限也就确定了，此相对误差限通常称为计算机的精度。如通常用的 8 位字长的十进制计算机，其机器数的相对误差限为

$$\frac{1}{2} \times 10^{1-8} = \frac{1}{2} \times 10^{-7}, \quad \text{舍入机}$$

$$10^{1-8} = 10^{-7}, \quad \text{截断机}$$

§ 4 误差危害的防止

误差分析在数值运算中是一个又重要而又复杂的问题，因为每步运算都有可能产生误差，而一个工程或科学计算问题往往要算千万次，如果每步运算都分析误差，这是不可能的，也是不必要的。这里提出的若干原则，就是为了鉴别计算结果的可靠性，和防止误差危

害现象的产生。

(1) 使用数值稳定的计算公式

什么是数值稳定的计算公式呢？我们先来研究一个例子。

例5 建立积分 $I_n = \int_0^1 \frac{x^n}{x+5} dx$ $n=0, 1, \dots, 20$

的递推关系式，并研究它的误差传递。

解 由

$$I_n + 5I_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5} dx = \int_0^1 x^{n-1} dx = \frac{1}{n}$$

和

$$I_0 = \int_0^1 \frac{dx}{x+5} = \ln 6 - \ln 5$$

可建立下列递推关系

$$\begin{cases} I_n = -5I_{n-1} + \frac{1}{n}, & n=1, 2, \dots, 20 \\ I_0 = \ln 6 - \ln 5 \end{cases} \quad (4.1)$$

计算出 I_0 后，由递推关系式可逐次求出 I_1, I_2, \dots, I_{20} 的值。但在计算 I_0 时有舍入误差，设为 e_0 ，并设求得的 I_0 的近似值为 \bar{I}_0 ，即 $e_0 = I_0 - \bar{I}_0$ ， I_1 的近似值为 $\bar{I}_1 = -5\bar{I}_0 + 1$ ，于是 $I_1 - \bar{I}_1 = -5(I_0 - \bar{I}_0)$ 即 $e_1 = -5e_0$ ，因此在使用递推公式中，实际算得的都是近似值 $\bar{I}_n (n=1, 2, \dots, 20)$ 。现在来研究误差 e_0 是怎么传递的。

由 $I_1 - \bar{I}_1 = -5(I_0 - \bar{I}_0) = (-1)^2 5^2 e_0$ 可推得

$$I_n - \bar{I}_n = (-1)^n 5^n e_0.$$

由此看出误差 e_0 对第 n 步的影响是扩大 5^n 倍，当 n 较大时，误差将淹没真值，因此用 \bar{I}_n 近似 I_n 显然是不正确的，这种递推公式不宜采用。但是，我们若由

$$I_n = -5I_{n-1} + \frac{1}{n}$$

解出

$$I_{n-1} = -\frac{1}{5}I_n + \frac{1}{5n} \quad (4.2)$$

如能先求出 I_{20} ，则由递推式 (4.2) 亦可依次算出 $I_{19}, I_{18}, \dots, I_1, I_0$ 。

由

$$I_n = \int_0^1 \frac{x^n}{x+5} dx$$

使用广义积分中值定理 (见注)

$$I_n = \frac{1}{\xi+5} \int_0^1 x^n dx = \frac{1}{\xi+5} \cdot \frac{1}{n+1} \quad 0 \leq \xi \leq 1$$

注：若函数 $\phi(x)$ 在区间 $[a, b]$ 上不变号，则 $\int_a^b f(x)\phi(x) dx = f(\xi) \int_a^b \phi(x) dx$, $\xi \in [a, b]$ 。