普通高等教育“十一五”规划教材

21世纪研究生数学主干教材

丛书主编 陈 化

数值分析

陈晓江 黄樟灿 主编

科学出版社

北京

版权所有，侵权必究

举报电话：010-64030229；010-64034315；13501151303

内 容 简 介

本书是作者在多年为理工科硕士研究生讲授数值分析课程的基础上编写而成的. 全书共分 9 章，内容包括：绪论，插值、拟合与逼近，数值积分与数值微分，线性方程组的直接解法，线性方程组的迭代解法，矩阵特征值问题的数值解法，常微分方程的数值解法，非线性方程求根的数值方法，非线性方程求根的仿生方法. 本书从实用角度出发，介绍科学与工程计算中常用的数值计算方法和理论，介绍 Matlab 应用实例，配有大量的例题、习题和上机练习题供教师选用，每章有小结，书后有习题参考答案与提示.

本书可作为理工科大学非数学专业的研究生或数学专业高年级本科生的教材，也可作为科技工作者的参考书.

图书在版编目(CIP)数据

数值分析/陈晓江，黄樟灿主编. —北京：科学出版社，2010
普通高等教育"十一五"规划教材. 21 世纪研究生数学主干教材
ISBN 978-7-03-026265-3

Ⅰ. 数… Ⅱ. ①陈…②黄… Ⅲ. 数值计算—高等学校—教材 Ⅳ. O241

中国版本图书馆 CIP 数据核字(2009)第 231996 号

责任编辑：王雨舸 / 责任校对：董艳辉
责任印制：彭 超 / 封面设计：苏 波

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

武汉市新华印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2010 年 1 月第 一 版 开本：B5(720×1000)
2010 年 1 月第一次印刷 印张：19 1/2
印数：1—5 000 字数：375 000

定价：32.80 元

(如有印装质量问题，我社负责调换)

《21世纪研究生数学主干教材》丛书编委会

主　　编　陈　化

常务副主编　刘禄勤

副 主 编　吴传生　何　穗　刘安平

编　　委（按姓氏笔画为序）

王卫华　王展青　严国政　杨瑞琰　李　星

肖海军　罗文强　赵东方　黄樟灿　梅全雄

彭　放　彭斯俊　曾祥金　谢民育

《数值分析》编委会

主　　编　陈晓江　黄樟灿

副 主 编　王仲君　陈建业　尹　强

编　　委　陈晓江　黄樟灿　王仲君　陈建业　尹　强

《21世纪研究生数学主干教材》丛书序

《21世纪研究生数学主干教材》为高等学校研究生数学主干课程系列教材，大致划分为公共数学类、专业数学类两大块，经组编委员会审定，列选科学出版社普通高等教育“十一五”规划教材.

一、组编机构

《21世纪研究生数学主干教材》丛书由多所985和211大学联合组编：

丛书主编 陈 化

常务副主编 刘禄勤

副 主 编 吴传生 何 穗 刘安平

丛书编委(按姓氏笔画为序)

王卫华 王展青 严国政 杨瑞琰 李 星

肖海军 罗文强 赵东方 黄樟灿 梅全雄

彭 放 彭斯俊 曾祥金 谢民育

二、编写原则

质量. 质量是图书的生命，保持和发扬科学出版社“三高”、“三严”的传统特色，创造名牌大学和科学出版社双重品牌；适用性是教材的生命力所在，应明确读者对象，篇幅恰当.

系统. 研究生教材的系统性高于本科生教材，知识系统，体系完整，逻辑清晰，给学生留下选学和自学的内容和空间.

创新. 反映学科发展前沿、先进理念；在知识、内容等方面有所创新，有所贡献，体现教材的知识创新；紧跟和引领教学实践，在教学方法、教材结构、知识组织、详略把握、内容安排上有独到之处，体现教学实践创新.

三、指导思想

(1) 力求体系完整，结构严谨，内容精炼，循序渐进，推理简明，深入浅出，富有启发性，让学生打下坚实的理论基础.

(2) 恰当融入现代数学的新思想、新观点、新结果和新方法，使学生有较新的

学术视野.

(3) 为使学生巩固知识和提高应用能力,章末列出习题和思考题,并列出可进一步深入阅读的文献. 书末要给出索引.

(4) 在内容的取舍、叙述的方式、材料的组织安排等方面具有自己的特色.

(5) 公共数学教材注重强化学生的实验训练和实际动手能力,着力培养学生运用现代数学工具(软件)的能力;加强教学内容的应用性,注重案例分析,提高学生对数学知识、数学方法的应用能力及解决问题的能力.

四、主编职责

丛书组编委员会和出版社确定全套丛书的编写原则、指导思想和编写规范,在这一框架下,每本教材的主编对本书具有明确的责权利:

1. 拟定指导思想

按照丛书的编写原则和指导思想,拟出编写本书的指导思想和编写说明.

2. 明确特色和编写原则

教材的特色和闪光点;教改、课改动态,学科发展前沿、先进理念如何引入教材;知识和内容创新点及其编写方法;创新与继承的关系把握;教材系统性与教学实践性的关系处理和具体操作;严把教材质量关和适用性.

3. 掌握教材编写环节

(1) 把握教材编写人员水平,原则上要求博士、副教授以上,有多年课程教学经历,熟悉课程和学科领域的发展状况,有教材编写经验,有扎实的文字功底.

(2) 充分注意著作权问题,不侵犯他人著作权.

(3) 讨论、拟定教材提纲,并负责编写组的编写分工、协调与组织.

(4) 拟就内容简介、前言、目录、样章,统稿、定稿,确定交稿时间.

(5) 负责出版事宜,敦促编写组成员使用本教材,并优先选用本系列教材.

《21世纪研究生数学主干教材》组编委员会

2009年10月

试读结束:需要全本请在线购买: www.ertongbook.com

前　言

在科学与工程计算中，怎样选择与使用适当的数值计算方法，怎样估计计算结果的误差，怎样解释计算过程中的异常现象，已成为广大科技工作者迫切需要解决的问题. 由于这一原因，现在各院校对非数学专业的研究生和数学专业的高年级学生普遍开设“数值分析”课程. 本书就是作者在为理工科硕士研究生多年讲授数值分析课程的基础上编写而成的.

本书共分 9 章，内容包括：绪论，插值、拟合与逼近，数值积分与数值微分，线性方程组的直接解法，线性方程组的迭代解法，矩阵特征值问题的数值解法，常微分方程的数值解法，非线性方程求根的数值方法，非线性方程求根的仿生方法.

本书从实用的角度出发，通过实际问题引出基本概念，着重讲清原理，突出算法的构造和分析，并通过大量的例题帮助读者解决做题难的问题，每章最后一节介绍 *Matlab* 求解相关问题的应用实例，帮助读者提高解决实际问题的动手能力. 每章最后都有小结，并附有适当数量的习题和上机练习题，书后给出习题的参考答案与提示. 最后一章介绍了仿生方法，接触到最新的实用前沿，帮助读者用最新的方法解决实际问题.

本书的使用对象为理工科大学非数学专业的研究生或数学专业高年级本科生，也可作为科技工作者的参考书. 读者可根据不同的需要，选择适当的章节进行学习. 根据我们的教学实践，本书内容可在 72 学时内完成. 根据不同专业的需要，删去部分内容，可适用于 40～64 学时的教学需要.

本书由陈晓江、黄樟灿主编，王仲君、陈建业、尹强任副主编. 陈晓江编写第 1、2、3、7 章，王仲君编写第 4、5、6 章，陈建业编写第 8、9 章，尹强编写 *Matlab* 应用实例，陈晓江负责全书的统稿. 在本书的编写过程中，王卫华教授认真审阅了书稿，提出了修改意见，在此表示衷心感谢.

由于作者水平有限，本书的疏漏和不足在所难免，敬请读者批评指正.

编者

2009.10

目 录 CONTENTS

第1章 绪 论

随着科学技术的发展，科学与工程计算已推向科学活动的前沿. 科学与工程计算的范围扩大到了所有科学领域，并与科学实验、科学理论三足鼎立，相辅相成，成为人类科学活动的三大方法之一. 因此，熟练地运用计算机进行科学计算，已成为科技工作者的一项基本技能，这就要求人们去研究和掌握适用于计算机上使用的数值计算方法，而数值分析就是研究用计算机解决数学问题的数值计算方法及其有关理论.

1.1 数值分析的内容与特点

数值分析是计算数学的一个主要部分，计算数学是数学科学的一个分支，它研究用计算机求解各种数学问题的数值计算方法及其理论与软件实现. 一般地说，用计算机解决科学计算问题，首先需要针对实际问题提炼出相应的数学模型，然后为解决数学模型设计出数值计算方法，经过程序设计之后上机计算，求出数值结果，再由实验来检验. 概括为如图 1.1 所示. 其中根据数学模型提出求解的数值计算方法直到编出程序上机计算出近似结果，这一过程是计算数学的任务，也是数值分析研究的对象. 因此，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. 它以纯数学为基础，但却不完全像纯数学那样只研究数学本身的理论，而是着重研究数学问题求解的数值计算方法以及与此有关的理论，包括方法的收敛性、稳定性及误差分析；还要根据计算机的特点研究计算时间最省（或计算费用最省）的计算方法. 有的方法在理论上虽然还不够完善与严密，但通过对比分析、实际计算和实践检验等手段，被证明是行之有效的方法也可采用. 因此数值分析既有纯数学的高度抽象性与严密科学性的特点，又有应用数学的广泛性与实际试验的高度技术性的特点，是一门与计算机紧密结合的实用性很强的数学课程.

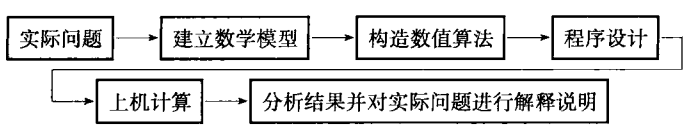

图 1.1

目前，计算机已成为数值计算的主要工具，数值分析的主要任务是研究适合计算机使用的、满足精度要求、节省计算时间的有效算法及其相关的理论. 在实现这

试读结束：需要全本请在线购买： www.ertongbook.com

些算法时往往还要根据计算机的容量、字长、速度等指标，研究具体的求解步骤和程序设计技巧. 数值分析的特点概括起来有四点：

第一，面向计算机，要根据计算机特点提供切实可行的有效算法. 即算法只能包括加、减、乘、除运算和逻辑运算，这些运算是计算机能直接处理的运算.

第二，有可靠的理论分析，能任意逼近并达到精度要求，对近似算法要保证收敛性和数值稳定性，还要对误差进行分析. 这些都建立在相应数学理论的基础上.

第三，要有好的计算复杂性，包括好的时间复杂性(计算时间少)和好的空间复杂性(占用存储单元少). 对很多数值问题使用不同算法，其计算复杂性将会大不一样，这也是建立算法要研究的问题，它关系到算法能否在计算机上实现.

第四，要有数值试验，即任何一个算法除了从理论上要满足上述三点外，还要通过数值试验证明是行之有效的.

例如，求解线性方程组 $Ax=b$，若 $\det(A)\neq 0$，则可用**克拉默(Cramer)法则**来求解. 设 A 为 20 阶矩阵，计算一个 20 阶行列式需要的乘法运算量为 $19\times 20!$，需要计算 21 个 20 阶的行列式，总的乘法运算量为

$$21\times 19\times 20!\approx 9.71\times 10^{20}$$

若用 1 万亿次/ 秒的计算机来运算，则一年可完成的乘法运算量为

$$10^{12}\times 365\times 24\times 3600\approx 3.15\times 10^{19}$$

求解 20 阶的线性方程组所需乘法运算的时间为

$$9.71\times 10^{20}\div(3.15\times 10^{19})\approx 30.83\ (\text{年})$$

即 30.83 年，显然这个运算时间在实际中是不可接受的. 而在实际问题中，如大型水利工程、天气预报等，需要求解的大型线性方程组的阶数一般都远远大于 20，若用上述方法显然无法解决. 这个例子说明求解线性方程组的克拉默法则在理论上虽然可行，但在实际应用中却不可行. 有人可能说，随着计算机的发展，运算速度提高、内存增大以及新结构计算机的出现，以前认为过于复杂而不能求解的问题将会得到解决. 但是，不论计算机如何发展，使用计算机的代价，即计算复杂性，都是需要考虑的.

1.2 计算机机器数系与浮点运算

微积分学的基础是实数系，而数值计算方法的理论则是建立在计算机机器数系的基础上. 为了设计高效、可靠的算法，这里简要介绍计算机机器数系的基本知识.

1.2.1 二进制数与计算机机器数系

在大多数计算机中，实数是以二进制形式表示的，并且在二进制实数系统中进

行运算. 这似乎与我们从屏幕上看到的不一样. 事实上，计算机首先将我们输入的十进制数转换为二进制数，然后在二进制实数系统中作运算，最后，再将结果转换为十进制数.

例 1 将 $x = 237$ 表示为二进制数.

解 将 x 展开成 2 的乘幂之和

$$\begin{aligned} x &= 237 \\ &= 1\times 2^7 + 1\times 2^6 + 1\times 2^5 + 0\times 2^4 + 1\times 2^3 + 1\times 2^2 + 0\times 2^1 + 1\times 2^0 \end{aligned}$$

即 x 的二进制表示为：$x = (11\,101\,101)_2$.

例 2 将 $x = 0.65\,625$ 与 $y = 0.7$ 分别表示为二进制数.

解 将 x 展开成 2 的负乘幂之和

$$x = 1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3} + 0\times 2^{-4} + 1\times 2^{-5}$$

即 x 的二进制表示为：$x = (0.10101)_2$. 用类似的方法可求得 $y = (0.1\overline{0110})_2$，这里，$\overline{0110}$ 表示 0110 的循环.

对于一般实数 x，将 x 展开成

$$\begin{aligned} x = \pm(&b_{J-1}\times 2^{J-1} + \cdots + b_1\times 2^1 + b_0\times 2^0 + b_{-1}\times 2^{-1} + b_{-2}\times 2^{-2} \\ &+ \cdots + b_{-n}\times 2^{-n} + \cdots) \end{aligned}$$

这样 x 的二进制表示为：$x = \pm(b_{J-1}\cdots b_1 b_0 . b_{-1} b_{-2}\cdots b_{-n}\cdots)_2$，$b_j$ 是 1 或 0. 例如

$$\begin{aligned} 18.25 &= 1\times 2^4 + 0\times 2^3 + 0\times 2^2 + 1\times 2^1 + 0\times 2^0 + 0\times 2^{-1} + 1\times 2^{-2} \\ &= (10\,010.01)_2 \end{aligned}$$

上述 x 的二进制表示可以写成与十进制类似的浮点形式

$$x = \pm 0.b_{J-1}\cdots b_1 b_0 b_{-1} b_{-2}\cdots b_{-n}\cdots \times 2^J$$

小数部分 $\pm 0.b_{J-1}\cdots b_1 b_0 b_{-1} b_{-2}\cdots b_{-n}\cdots$ 称为**尾数**，2 的指数 J 称为**阶码**是整数. 一般地，一个数可以有不同的浮点表示，例如

$$18.25 = 0.100\,1001\times 2^5 = 0.010\,010\,01\times 2^6$$

为了保证唯一性，通常规定非零数的尾数的第一位数字非零，即 $b_{J-1} = 1$. 在这种规定下的浮点表示，称为**规格化的二进制浮点数**.

在计算机中，一个非零数通常表示为如下二进制浮点形式

$$\pm 0.b_1 b_2\cdots b_t \times 2^m$$

其中 b_j ($j = 2, 3, \cdots, t$) 是 1 或 0，$b_1 = 1$；t 称为计算机的**字长**；阶码 m 有固定的上、下限，即 $L \leqslant m \leqslant U$，$L$，$U$ 和 t 随计算机而异. 上述形式的数称为**机器数**. 由于机器数的字长与阶码有限，因此计算机中的数是有限的. 事实上，计算机中共有 $2^t(U-L+1)+1$ 个机器数. 把计算机中的全体机器数组成的集合记为 F 或 $F(2, t, L, U)$，称为计算机机器数系. 机器数系 F 不是连续统，它是一个有限的、离散的、分布不均匀的集合. 不难验证，F 中任意非零数 y 满足

$$2^{L-1} \leqslant |y| \leqslant 2^U(1-2^{-t})$$

机器数有单精度与双精度之分，字长 t 的值规定了机器数的精度. 一般地，单精度数 $t=23$，约为十进制的 7 位有效数字；双精度数 $t=52$，约为十进制的 15 位有效数字. 字长越大，机器数的精度越高. 阶码 m 的值规定了机器数的绝对值范围，单精度数阶码 m 的范围为 $-127 \leqslant m \leqslant 128$，其绝对值范围为 $2^{-128} \sim 2^{128}$，即 $10^{-38} \sim 10^{38}$；双精度阶码 m 的值为 $-1023 \leqslant m \leqslant 1024$，其绝对值范围为 $2^{-1024} \sim 2^{1024}$，即 $10^{-308} \sim 10^{308}$.

在计算中，当数据的绝对值不在上述范围之内时，称为溢出，小于机器数下限时，称为下溢出，此时，对应机器数取为零；大于机器数上限时称为上溢出，此时，对应机器数被取作无穷大，程序停止执行.

1.2.2 数据的表示与浮点运算

无论怎样的计算机，其机器数系 $F(2, t, L, U)$ 都是一个有限的集合，它所表示的实数只是实数系的一小部分. 绝大多数实数输入计算机时，要转换为有限字长的二进制机器数，总要经“舍”或“入”而由一个与之相近的机器数代替. 实数 x 对应的机器数记为 $fl(x)$.

一般地，设 $x = \pm 0.b_1b_2\cdots b_t\cdots \times 2^m$，且 $2^{L-1} \leqslant |x| \leqslant 2^U(1-2^{-t})$，则

$$fl(x) = \operatorname{sgn}(x)\,\bar{a} \times 2^m \tag{1.2.1}$$

其中

$$\bar{a} = \begin{cases} 0.b_1b_2\cdots b_t, & \text{若 } b_{t+1} = 0 \\ 0.b_1b_2\cdots b_t + 2^{-t}, & \text{若 } b_{t+1} = 1 \end{cases} \tag{1.2.2}$$

这种获取机器数的方法称为**舍入法**；另一种获取机器数的方法称为**截断法**. 此时，对应上述 x 的机器数为

$$fl(x) = \operatorname{sgn}(x)0.b_1b_2\cdots b_t \times 2^m$$

例 3 将实数 $x = 2.656\,25$ 与 $y = 0.1$ 分别表示为 $F(2, 8, -19, 19)$ 中的机器数.

解 因为 $x = 2.656\,25 = 0.101\,010\,1 \times 2^2 \in F$，所以

$$fl(x) = x = 0.101\,010\,1 \times 2^2$$

而 $y = 0.1 = (0.0\,\overline{0011})_2 = 0.\overline{1100} \times 2^{-3} \notin F$，但 $2^{-20} \leqslant |y| < 2^{20}$. 按舍入法，则

$$fl(y) = 0.110\,011\,01 \times 2^{-3} = 0.100\,097\,656$$

按截断法，则

$$fl(y) = 0.110\,011\,00 \times 2^{-3} = 0.099\,609\,375$$

以上介绍了二进制机器数系，机器数系不仅可以是二进制，还可以是 β 进制，例如八进制、十六进制、十进制等. β 进制机器数系可记为 $F(\beta, t, L, U)$，F 中任意非零数 y 可表示为

$$y = \pm 0.b_1 b_2 \cdots b_t \times \beta^m$$

其中 $0 \leqslant b_j \leqslant \beta - 1$ $(j = 2, 3, \cdots, t)$，$b_1 \neq 0$；t 称为机器数的字长；阶码 m 满足 $L \leqslant m \leqslant U$. 特别地，十进制机器数系为 $F(10, t, L, U)$. 在下面的讨论中，为适应人们的习惯，采用十进制机器数系. 类似于二进制机器数系，十进制机器数系 $F(10, t, L, U)$ 也按两种方法获取机器数：舍入法或截断法. 前者是按四舍五入原则截取 x 尾数的前 t 位数，后者是直接截取 x 尾数的前 t 位数作为机器数的尾数.

例 4 将实数 π 表示为 $F(10, 5, -19, 19)$ 中的机器数.

解 $fl(\pi) = 0.314\,16 \times 10$ （舍入式） $fl(\pi) = 0.314\,15 \times 10$ （截断式）

下面讨论计算机中浮点数的运算. 如前所述，计算机只能进行加、减、乘、除四则运算，而且机器数系对四则运算并不封闭，就是说 F 中任意两数的和、差、积、商不一定都在 F 中. 此时，计算机自动将计算结果用 F 中机器数表示出来.

设 x 和 y 都是机器数，即 $x, y \in F(10, t, L, U)$，它们的算术运算符合下述规则：

(1) 加减法. 先对阶(靠高阶)，后运算，再舍入.

(2) 乘除法. 先运算，再舍入.

在运算中，不妨假定计算机具有双精度累加寄存器，即在运算时先保留 $2t$ 位，最后再把第 $t+1$ 位的数进行四舍五入. 下面举例说明：

例 5 设 $x = 0.505\ 561\ 28 \times 10^{-3}$，$y = 0.231\ 627\ 43 \times 10^{2}$，$z = -0.231\ 621\ 32 \times 10^{2}$，在 $F(10, 8, -29, 29)$ 中，按舍入式，分别计算 $x+y+z$ 与 xy.

解 按两种方法求和：

$$\begin{aligned}
(1)\ fl(x+y+z) &= fl[(x+y)+z] \\
&= fl[fl(0.000\,005\,055\,612\,8 \times 10^{2} + 0.231\,627\,43 \times 10^{2}) \\
&\quad - 0.231\,621\,32 \times 10^{2}] \qquad \text{(对阶，靠高阶)} \\
&= fl(0.231\,632\,49 \times 10^{2} - 0.231\,621\,32 \times 10^{2}) \\
&= 0.111\,700\,00 \times 10^{-2} \\
(2)\ fl(x+y+z) &= fl[x+(y+z)] \\
&= fl[0.505\,561\,28 \times 10^{-3} + fl(0.231\,627\,43 \times 10^{2} \\
&\quad - 0.231\,621\,32 \times 10^{2})] \\
&= fl(0.505\,561\,28 \times 10^{-3} + 0.611\,000\,00 \times 10^{-3}) \\
&= 0.111\,656\,13 \times 10^{-2}
\end{aligned}$$

精确结果为 $x+y+z = 0.111\,656\,128 \times 10^{-2}$，显然，方法(2)的结果较准确.

$$\begin{aligned}
fl(xy) &= fl(0.505\,561\,28 \times 10^{-3} \times 0.231\,627\,43 \times 10^{2}) \\
&= 0.117\,101\,86 \times 10^{-1}
\end{aligned}$$

由上例可以看出，在计算机机器数系中，人们所熟悉的加减法的交换律与结合律是不成立的，特别在某些加法运算中，运算顺序对计算结果有很大影响. 关于这个问题，本书将在后面的误差分析中还会谈到.

1.3 数值计算的误差

用数值计算方法来解决实际问题，不可避免的会产生误差. 数值分析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 误差的来源与分类

(1) 模型误差. 数学模型与实际问题之间的误差称为**模型误差**.

一般来说，生产和科研中遇到的实际问题是比较复杂的，要用数学模型来描述，需要进行必要的简化，忽略一些次要的因素，这样建立起来的数学模型与实际问题之间一定有误差. 它们之间的误差就是模型误差.

(2) 观测误差. 实验或观测得到的数据与实际数据之间的误差称为**观测误差或数据误差**.

数学模型中通常包含一些由观测(实验)得到的数据，例如用 $s(t)=\frac{1}{2}gt^2$ 来描述初始速度为 0 的自由落体下落时距离和时间的关系，其中重力加速度 $g\approx 9.8\ \mathrm{m/s^2}$ 是由实验得到的，它和实际重力加速度之间是有出入的. 其间的误差就是观测误差.

(3) 截断误差. 数学模型的精确解与数值方法得到的数值解之间的误差称为**方法误差或截断误差**.

例如，由泰勒(Taylor)公式得

$$e^x=1+x+\frac{x^2}{2!}+\cdots+\frac{x^n}{n!}+R_n(x)$$

用 $p_n(x)=1+x+\frac{x^2}{2!}+\cdots+\frac{x^n}{n!}$ 近似代替 e^x，这时的截断误差为

$$R_n(x)=\frac{e^{\xi}}{(n+1)!}x^{n+1},\quad \xi\text{介于 }0\text{ 与 }x\text{ 之间}$$

(4) 舍入误差. 计算中遇到的数据可能位数很多或是无穷小数，如 $\sqrt{2}=1.414\ 213\ 56\cdots$，受机器字长的限制，无穷小数和位数很多的数必须舍入成一定的位数(机器字长). 舍入方法有：① 舍入法，如将 $1.414\ 213\ 56\cdots$ 四舍五入为 $1.414\ 213\ 6$；② 截断法，如 $\sqrt{2}$ 在八位字长的截断机里取成 $1.414\ 213\ 5$. 这样产生的误差称为**舍入误差**. 少量的舍入误差是微不足道的，但是在计算机作了成千上万

次运算后，舍入误差的累积有时可能是十分惊人的. 它取决于算法的稳定性. 如果算法能够累积大量的误差，这种算法是不稳定的，反之称为稳定算法.

研究计算结果的误差是否满足精度要求就是误差估计问题，本书主要讨论算法的截断误差与舍入误差，而截断误差将结合具体算法讨论. 为分析数值运算的舍入误差，先要对误差基本概念做简单介绍.

1.3.2 绝对误差与相对误差

定义 1 设 x 为准确值，x^* 为 x 的一个近似值，称 $e^* = x^* - x$ 为近似值的**绝对误差**，简称**误差**.

注意，这样定义的误差 e^* 可正可负，当绝对误差为正时近似值偏大，称为强近似值；当绝对误差为负时近似值偏小，称为弱近似值.

通常我们不能算出准确值 x，当然也不能算出误差 e^* 的准确值，只能根据测量工具或计算情况估计出误差的绝对值不超过某个正数 ε^*，也就是误差绝对值的一个上界. ε^* 称为近似值的**绝对误差**，简称为**误差限**，它总是正数.

一般情形 $|x^* - x| \leqslant \varepsilon^*$，即 $x^* - \varepsilon^* \leqslant x \leqslant x^* + \varepsilon^*$. 这个不等式有时也表示为 $x = x^* \pm \varepsilon^*$. 例如，用卡尺测量一个圆杆的直径为 $x^* = 350$ mm，它是圆杆直径 x 的近似值，由卡尺的精度知这个近似值的误差不会超过 0.5 mm，则有

$$|x^* - x| = |350 - x| \leqslant 0.5(\text{mm})$$

于是该圆杆的直径为 $x = 350 \pm 0.5(\text{mm})$.

用 $x = x^* \pm \varepsilon^*$ 表示准确值可以反映它的准确程度，但不能说明近似值的好坏. 例如，测量一根 10 cm 长的圆钢时发生了 0.5 cm 的误差，和测量一根 10 m 长的圆钢时发生了 0.5 cm 的误差，其绝对误差都是 0.5 cm，但是，后者的测量结果显然比前者要准确得多. 这说明决定一个量的近似值的好坏，除了要考虑绝对误差的大小，还要考虑准确值本身的大小，这就需要引入相对误差的概念.

定义 2 设 x 为准确值，x^* 为 x 的一个近似值，近似值 x^* 的误差 e^* 与准确值 x 的比值 $\dfrac{e^*}{x} = \dfrac{x^* - x}{x}$ 称为近似值 x^* 的**相对误差**，记为 e_r^*.

在实际计算中，由于准确值 x 是不知道的，通常取 $e_r^* = \dfrac{e^*}{x^*} = \dfrac{x^* - x}{x^*}$ 作为 x^* 的相对误差，条件是 $e_r^* = \dfrac{e^*}{x^*}$ 较小，此时

$$\frac{e^*}{x} - \frac{e^*}{x^*} = \frac{e^*(x^* - x)}{x^* x} = \frac{(e^*)^2}{x^*(x^* - e^*)} = \frac{(e^*/x^*)^2}{1 - (e^*/x^*)}$$

是 e_r^* 的平方项级，故可忽略不计.

相对误差也可正可负，它的绝对值上界称为**相对误差限**，记为 ε_r^*，即

试读结束：需要全本请在线购买： www.ertongbook.com