■新世纪大学数学系列教材

数值计算

SHU ZHI JI SUAN

主 编 王志军 王海红 孟红玲

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}dx\approx\frac{\pi}{n+1}\sum_{k=0}^{n}f\left(\cos\frac{2k+1}{2n+2}\pi\right)$$

河南大学出版社
HENAN UNIVERSITY PRESS

SHUZHI JISUAN

数值计算

主　编　王志军　王海红　孟红玲
副主编　徐　刚　郭　城　邓书显

河南大学出版社
·郑州·

图书在版编目(CIP)数据

数值计算/王志军,王海红,孟红玲主编. —郑州:河南大学出版社,2012.5
ISBN 978-7-5649-0733-4

Ⅰ.①数… Ⅱ.①王…②王…③孟… Ⅲ.①数值计算—高等学校—教材 Ⅳ.①0241

中国版本图书馆 CIP 数据核字(2012)第 090048 号

责任编辑 阮林要
责任校对 高丽燕
装帧设计 郭 灿

出版发行 河南大学出版社
地址:郑州市郑东新区商务外环中华大厦 2401 号
邮编:450046
电话:0371-86059750(职业教育出版分社)
0371-86059701(营销部)
网址:www.hupress.com
排 版 河南金河印务有限公司
印 刷 郑州海华印务有限公司
版 次 2012 年 8 月第 1 版
印 次 2012 年 8 月第 1 次印刷
开 本 787mm × 1092mm 1/16
印 张 14.5
字 数 344 千字
定 价 29.00 元

前　言

计算方法是利用数学模型解决现实生活中实际问题的一门学科，是数学领域中的一个重要分支. 随着计算机的广泛使用以及人们对科学计算的要求越来越高，计算方法的应用范围已拓展到所有的学科中，各领域的研究者都需要用科学、高效、准确的计算，解决本领域的数值计算问题.

本书根据理工科对科学计算的实际需求，详细地介绍了常用的基本计算方法的构造，对各种算法进行了归纳和推导，分析了各算法的稳定性、收敛性、误差估计等，说明了算法的适用范围并比较了优缺点. 本书选择了一些经典的例题及习题，阐述理论力求明确，数值计算力求简单、准确，并附了一些算法程序供读者参考.

本书由王志军、王海红、孟红玲担任主编，徐刚、郭城、邓书显担任副主编，由徐刚、孟红玲、邓书显统一定稿. 本书共分为八章，其中第 1、6 章和第 8 章的 8.2 节由王志军编写，第 2、3、7 章由王海红编写，第 5 章由郭城编写，第 4 章和第 8 章的 8.1 节由徐刚编写. 由于作者水平有限，书中难免出现疏漏及错误，恳请各位读者及同行批评指正.

本书在编写过程中得到了郑州师范学院数学系、河南财经政法大学数学与信息科学系、河南大学出版社领导和老师们的关心和支持，在此深表感谢.

编　者

2012 年 6 月

目　　录

第 1 章　数值计算概论 …… (1)
1.1　数值计算的对象与特点 …… (1)
1.2　误差与有效数字 …… (3)
1.3　误差估计与误差分析 …… (9)
1.4　误差的定性分析与运算原则 …… (13)

第 2 章　插值与曲线拟合 …… (18)
2.1　引言 …… (18)
2.2　Lagrange 插值 …… (20)
2.3　Newton 插值 …… (25)
2.4　Hermite 插值 …… (29)
2.5　分段低次插值 …… (32)
2.6　三次样条函数插值 …… (35)
2.7　最小二乘法的曲线拟合 …… (39)

第 3 章　数值积分与数值微分 …… (47)
3.1　数值求积公式 …… (47)
3.2　Newton-Cotes 求积公式 …… (50)
3.3　复化求积公式 …… (53)
3.4　Romberg 求积公式 …… (54)
3.5　高斯型求积公式 …… (56)
3.6　数值微分 …… (62)

第 4 章　线性方程组的数值解法 …… (68)
4.1　线性方程组概述及矩阵基础知识 …… (68)
4.2　线性方程组的直接解法 …… (71)
4.3　向量范数与矩阵范数 …… (88)
4.4　线性方程组的迭代解法 …… (107)

第 5 章 矩阵特征值与特征向量的计算 …… (120)
5.1 幂法 …… (120)
5.2 子空间迭代法 …… (126)
5.3 *QR* 算法 …… (129)
5.4 Jacobi 旋转法 …… (131)

第 6 章 非线性方程的数值解 …… (136)
6.1 引言 …… (136)
6.2 区间二分法 …… (137)
6.3 弦截法 …… (139)
6.4 切线法 …… (147)
6.5 一般迭代法 …… (151)

第 7 章 常微分方程初值问题数值解法 …… (160)
7.1 引言 …… (160)
7.2 几类简单的求解初值问题的数值方法 …… (160)
7.3 Runge-Kutta 方法 …… (164)
7.4 单步法的收敛性与稳定性 …… (167)
7.5 线性多步法 …… (169)
7.6 常微分方程组初值问题的数值解法 …… (174)

第 8 章 实训(基于 C 语言和 MATLAB) …… (179)
8.1 C 语言概述 …… (179)
8.2 MATLAB 简介 …… (201)

习题参考答案 …… (217)
参考文献 …… (222)

第1章 数值计算概论

1.1 数值计算的对象与特点

1.1.1 研究对象

数值计算方法是研究并解决数学问题的数值近似解的方法，简称计算方法，也叫数值分析等. 它的研究对象是利用计算机求解各种数学问题的数值方法及有关理论，其内容包括函数的数值逼近(代数插值与最佳逼近)、数值积分与数值微分、非线性(代数的与超越的)方程(组)的数值解法、数值线性代数(线性代数方程组的解法与矩阵特征值问题的计算)、常微分方程的数值解法及偏微分方程的数值解法等. 它之所以成为数学中的独立分支，一方面是数学本身发展为之提供了可能，另一方面是20世纪40年代电子计算机的问世使之成为必要. 现代计算机的出现为大规模的数值计算创造了条件，集中而系统地研究适用于计算机的数值方法立即变得十分迫切和必要. 数值分析并不仅是一些数值方法的简单积累，而且揭示包含在多种多样的数值方法之间的相同的结构和统一的原理. 它在大量的数值计算实践和理论分析工作的基础上迅速发展着，原有的方法有的使用至今，有的逐步被淘汰，而新的方法和新的理论不断地产生.

1.1.2 主要特征

数值计算有以下三个主要特点.

第一，面向计算机，提供实际可行的常用算法. 具体地讲，由于计算机能够进行加、减、乘、除四则运算，故需要把每个求解的数学问题用四则运算的有限形式的公式表达出来. 这种公式只能是多项式或有理分式的形式，通常称为算法，它是计算机能够直接处理的.

第二，能够任意逼近，有可靠的理论分析. 计算方法有两类算法，一类是精确的，另一类是近似的. 所谓精确算法是指在没有运算舍入误差的假设下，能在确定的运算次数内获得数学问题的精确解. 近似算法本身有方法误差，从而在任何有限的运算次数内能获得数学问题的近似解. 大多算法是近似算法. 由于计算机字长有限，每次运算都有舍入误差，因此无论精确算法还是近似算法都只能获得数学问题的近似解. 计算机需要人们用种种数学理论和方法来建立各个算法. 对近似算法要保证收敛性，即近似解能逼近精确解到任意的程度. 对每个算法要保证数值稳定性，这是指舍入误差对解的准确性影响不大.

第三，省时间省资源，有良好的计算复杂性. 一个算法的计算复杂性是指该算法包含的运算次数和所需的存储量. 近似算法的运算次数取决于算法的收敛速度. 求解一个数学

问题，是否选用或建立复杂性好的算法很重要，有时会影响到现有的计算机上能否真正实现.

想一想：数值计算课程的任务是什么？

从上述特点可以看出，数值计算的任务是：

1. 建立各种数学问题的数值计算算法的方法和理论. 通俗地讲，就是为各种实际问题提供有效的数值近似解方法.

2. 提供在计算机上实际可行的、理论可靠的、计算复杂性好的各种常用算法.

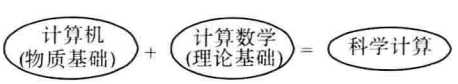
1.1.3 数值问题与数值方法

数值问题是指输入数据（即问题中自变量与原始数据）与输出数据之间函数关系的一个确定描述，输入输出数据可用有限维向量表示. 根据这种定义，**“数学问题”**不一定是**“数值问题”**，它往往要用数值逼近方法才能转化为数值问题. 函数的插值与逼近就是“数值分析”中最基本的问题之一，目的是提供各种简易的途径，将函数计算转化为数值问题，才能在计算机上处理. 对于一个给定的数值问题有许多不同的算法，称为数值方法，它们都给出问题的近似答案，但所需计算量和得到的精度可能相差很大，必须选择面向计算机计算复杂性好并有可靠理论分析的数值方法. 多数数学问题按建立数值方法的基本线索大体上可以归为以下两大类.

一类数学问题包含非有理函数或未知函数，如积分与微分计算、微分方程求解等. 这类问题建立数值方法的基本线索是：首先利用函数的数值逼近或离散化将原问题化为数值问题，然后去计算或求解数值问题以得到原问题的近似值或近似解. 例如，对利用各种方式得到的函数的逼近多项式进行求积分或求导数，可以引出各种形式的数值积分或数值微分公式；用一组离散点上特定值代替连续自变量的未知函数，通过各种逼近途径（包括插值、数值积分、数值微分以及泰勒（Taylor）展开等），将微分方程离散化，构成微分方程的各种数值解法.

另一类数学问题主要是代数问题，包括线性代数方程组的求解和矩阵的特征值问题的计算，线性最小二乘法等不包含非有理函数，因其本身就是数值问题，可以直接去建立数值方法. 非线性（代数的或超越的）方程的求解也可归于此类，因为它们仅仅在求函数值时可能要借助于数值逼近. 这一类问题的数值方法大致可分为直接法、迭代法和变换约化法三种. 直接法是一种精确算法，一般仅用于解线性代数方程组. 迭代法的基本线索是针对确定类型的问题，寻求某种固定形式的递推公式，使得由公式产生的序列收敛于问题的解. 迭代法在计算方法中占有重要的地位，某些数值问题（如非线性方程等）应用迭代法求解. 目前，矩阵特征值问题的大多数重要的数值方法均利用相似变换将矩阵（近似地）约化为特殊形式的矩阵，从而使特征值和特征向量可以较方便地近似求得，这类方法是变换迭代法，可称为变换约化法.

想一想：1. 用自己的语言说一说计算方法是一门怎样的课程？



2. 怎样学好这门课程?

提几点意见供参考:

一、树立信心,克服"怕"的思想. 二、要先复习相关的数学基础知识. 三、要搞清每章要解决什么问题,如何解决. 搞清各种方法的思想及其数学原理,注重基本概念及基本方法不要死记硬背. 四、及时复习,在复习基础上做给定的习题. 习题要自己先做,不要一上来就看答案. 实在不会做再看解答,但必须自己搞清楚为什么这样做. 有条件的还可自己选做书本外的习题.

1.2 误差与有效数字

人们在工作或日常生活中遇到的数可分为两类,一类是精确地反映实际情况的,称为**精确数**,或称为**准确数**或**真值**,如教室里有 40 个人,数 40 是准确数;另一类数则不是这样,称为**近似数**,或称**某准确数的近似值**,如测量某桌子的长度为 110 厘米,则 110 是个近似数. 近似数与准确数之间存在一个差值,即**误差**. 在工作中出现误差的大小(或精度)往往是标志着一个工作人员的工作质量. 工作中若对近似数的问题处理不当,不是浪费时间,就是给工作带来损失,甚至严重损失. 因为近似数是大量存在的,所以在计算方法里首先讨论误差的概念就非常必要了. 误差的内容相当丰富,针对我们需要,在本章里主要讨论误差来源、近似数的误差及误差对计算结果的影响等.

1.2.1 误差的来源与分类

误差的来源,即产生误差的原因是多种多样的. 在数值计算中,为解决求方程近似值的问题,在实际问题中通常遇到以下几种误差.

1.2.1.1 模型误差(model error)

要将复杂的事物或现象做定量的分析,首先要抽象归结为数学模型,在这个过程中,总要抓住其中的主要矛盾,而忽略一些次要因素的影响,对问题做一些简化. 因此,数学模型和实际问题有一定的误差,这种误差称为**模型误差**. 数学模型中常包含某些参量,如质量、温度、电压等,此类量通过观察确定,产生的误差称为**观察误差**(observational error).

例如,质量为 m 的物体,在重力作用下自由下落,其下落距离 s 与时间 t 满足微分方程

$$m\frac{\mathrm{d}^2 s}{\mathrm{d}t^2} = mg, \tag{1-1}$$

其中 g 为重力加速度.

方程(1-1)就是自由落体的数学模型. 它忽略了空气阻力这个因素,从而由(1-1)求出的在某一时刻 t 的下落距离 s,也必然是近似的,是有误差的.

1.2.1.2 测量误差(measurement error)

在建模和具体运算过程中往往有若干参数或常数,它们大多是通过观察和测量得到的,由于精度的限制,这些数据一般是近似的,即有误差,这种误差称为**测量误差**.

例如(1-1)中的重力加速度 g,就是观测来的,观测值的精度,依赖于测量仪器的精密程度,还要依赖于人的操作标准度,等等. 其他还有阻力系数、比重等,都是观测来的,也都含有误差.

1.2.1.3 截断误差(truncation error)

由于实际运算只能完成有限项或有限步运算,因此要将有些需用极限或无穷过程进行的运算有限化,对无穷过程进行截断,这样产生的误差称为**截断误差**.

例如,用收敛的无穷级数的前 n 项和作为该级数和的近似值,它的余项就是截断误差,如以 $S_n(x)=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!}$ 代替 $e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!}+\cdots$ 其截断误差为 $R_n(x)=e^{\xi}\cdot\frac{x^{n+1}}{(n+1)!}$,其中 ξ 在 0 与 x 之间.

1.2.1.4 舍入误差(rounding error)

在实际计算中,无论用什么样的数字计算工具,只能用有限位数进行. 如有的电子计算机只能对十进制八位数进行运算,对超过八位的数,由于计算工具的限制,它会把它舍入成八位的数,再进行处理. 这样由于对数进行舍入产生的误差,称为**舍入误差**. 至于舍入的具体办法有大家熟悉的四舍五入法,另外还有去尾法(只舍不入)及收尾法(只入不舍).

例如,取 π 到四位小数的近似值:

用四舍五入法取,就是 $\pi\approx3.1416$;

用去尾法取,就是 $\pi\approx3.1415$;

用收尾法取,就是 $\pi\approx3.1416$.

例 1 假设 $L(t)$ 是金属棒在温度为 t 时的长度,其数学模型为

$$l(t)=l+\alpha t+\beta t^2,$$

其中 α,β 为参数. 有如下估计:

$$\alpha=0.001253\pm10^{6},\ \beta=0.000068\pm10^{-6},$$

则 $L(t)-l(t)$ 是模型误差,10^{-6} 是 α 与 β 的观测误差.

数学模型常常不能获得精确解,必须用数值方法求近似解,其误差称为**截断误差**或**方法误差**.

例 2 实际计算时,函数 f 用泰勒多项式 P_n 近似代替:

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

则数值方法的截断误差 $R(x)=f(x)-P_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1}$,其中 ξ 在 0 与 x 之间.

有了求解数学问题的算法后,由于计算机的字长有限,原始数据在计算机上表示会产生误差,每一次运算又可能产生新的误差,这种误差称为**舍入误差**或**计算误差**.

例 3 实际计算时,用 3.14159 近似代替 π,则舍入误差

$$R=\pi-3.14159=0.0000026\cdots$$

数值分析中,仅讨论截断误差和舍入误差.

想一想:我们课程重点关注的误差有哪几类?

我们主要关注两类误差：一是截断误差，即所有将数学问题离散化转化为数值计算的方法都会产生的误差，是误差估计的主要内容；另一种是舍入误差，是所有做数值计算都会产生的误差，计算中初始误差，原始数据产生的误差都归入此类，如何分析舍入误差是本书讨论的主要内容.

1.2.2 误差概念

1.2.2.1 绝对误差

定义 1.1 设 x 为准确值，a 为近似值，记

$$\Delta a = x - a$$

为 a 对 x 的**绝对误差**(absolute error)或**误差**.

由定义 1.1 可知，Δa 可正可负. 当 $\Delta a > 0$ 时，称 a 为 x 的**不足近似值**；当 $\Delta a < 0$ 时，称 a 为 x 的**过剩近似值**.

$|\Delta a|$ 的大小，标志着 a 的精度，一般地，**在同一量的不同近似值中，$|\Delta a|$ 越小，a 的精度越高**.

因为不总能求得准确值 x，所以必须引入绝对误差界的概念.

例如，若用厘米为最小刻度的直尺去量桌子的长，大约为 1.225 米，求 1.225 米的绝对误差.

此例中，桌子长的准确值 x 是未知的，因此 1.225 的绝对误差就不知道. 实际中这类问题很多，于是定义 1.1 就失去了实际意义.

所以，为解决这一问题，常根据测量工具的精度或计算情况的分析，用一个满足 $|\Delta a| \leqslant \delta a$ 的较小的正数 δa 表示绝对误差的上限，称数 δa 为近似数 a 的**绝对误差界**(或**绝对误差限**).

由 Δa 的定义可知

$$|\Delta a| = |x - a| \leqslant \delta a. \tag{1-2}$$

这样，就可知道准确值 x 所在的范围：

$$a - \delta a \leqslant x \leqslant a + \delta a. \tag{1-3}$$

所以，只能说估计误差，而不说计算误差. 在实用上，常用下述写法来刻画 a 的**精度**：$x = a \pm \delta a$.

例如，在真空中光速 $c = 299796 \pm 4$ 公里/秒，即

$$299792 \text{ 公里/秒} \leqslant c \leqslant 299800 \text{ 公里/秒}.$$

有了绝对误差概念，上述桌子长度问题就容易解决了. 设桌子长度的准确值为 x，则 x 必定在 1.220 米到 1.230 米之间，所以 $|\Delta a|$ 不会超过 0.5 厘米，即

$$|\Delta a| = |x - 1.225| \leqslant 0.5(\text{厘米}),$$

所以 $\delta a = 0.5$.

显然用去尾法或收尾法截取的近似数，其 $|\Delta a|$ 都不会超过近似数末位的一个单位，所以其绝对误差界可取为该近似数末位的一个单位.

例如，对圆周率 π，用去尾法或收尾法截取三位数，就得 3.14 和 3.15，它们的绝对误差界都是 0.01，$\delta 3.14 = 0.01$，$\delta 3.15 = 0.01$.

一般情况下，用四舍五入法截取的近似数 a 的 $|\Delta a|$ 不超过 a 的末位的半个单位，所以 δa 可取该近似数末位的半个单位，设 10^s（s 为整数）为 a 的末位的计数单位，则

$$\delta a = \frac{1}{2} \times 10^s. \tag{1-4}$$

例如，$a = 3.1416$ 是对 π 用四舍五入法截取的近似值，6 所在的位的计数单位是 10^{-4}（$s = -4$），故有

$$\delta 3.1416 = 0.5 \times 10^{-4}.$$

特别地，当 $x = a$ 时，$\Delta a = 0$，$\delta a = 0$.

应当指出，在通常情况下，我们所说的绝对误差都是指绝对误差界而言的.

想一想：绝对误差能够比较不同近似数的精度吗？

先看一个例子，甲买了 1 吨煤有 10 公斤的误差，乙买了 100 公斤醋也有 10 公斤的误差，哪一个的情况更好些呢？显然，甲的情况好一些. 而乙的情况是不能允许的，原因是“误差太大”. 为什么太大呢？这是人们在头脑里已经把 10 公斤的误差与购买量做了比较的结果，即乙购买的醋每公斤就有 0.1 公斤的误差，而甲买的煤每公斤才有 0.01 公斤的误差.

绝对误差不能用来比较不同近似数的精度.

为此，我们引入表示近似值精度的另一尺度——相对误差.

1.2.2.2 相对误差

定义 1.2 在定义 1.1 的假设下，我们称比值

$$\Delta_r a = \frac{\Delta a}{x} = \frac{x - a}{x}$$

为 a 对 x 的**相对误差**（relative error）或近似数 a 的相对误差.

一般地，在同一量或不同量的几个近似值中，$|\Delta_r a|$ 小者，a 的精度高.

但在实际计算中，由于 x 未知，实际上总是将

$$\Delta_r^* a = \frac{\Delta a}{a} = \frac{x - a}{a} \tag{1-5}$$

作为 a 的相对误差. 记

$$|\Delta_r^* a| = \frac{|\Delta a|}{|a|} \leqslant \frac{\delta a}{|a|} \delta_r a, \quad |x| \neq 0, \tag{1-6}$$

知道绝对误差界 $\delta_r a$ 后便确定了 $\delta_r a$，显然，$\delta_r a$ 是 $|\Delta_r^* a|$ 的上界，称为 a 的**相对误差界**（或**相对误差限**）.

以后简称 $\delta_r a$ 为 a 关于 x 的相对误差. 有了这个公式就可以计算出上面量桌子长度的相对误差界为

$$\delta_r 1.225 = \frac{0.0005}{1.225} \leqslant 0.0004 = 0.04\%.$$

甲与乙买煤、买醋的相对误差界分别为

$$\delta_r 1000 = \frac{10}{1000} = 0.01 = 1\%,$$

$$\delta_r 100 = \frac{10}{100} = 0.1 = 10\%.$$

后者是前者的10倍,误差太大,工作质量不好.

由 $\delta_r a$ 的定义可知,相对误差表示在单位近似值中所含有的绝对误差,或者说,绝对误差在整个近似值中占有的比重,所以它是不名数,因而有广泛的适用性,它能更好地反映出误差的特性或近似数的精度.

1.2.3 有效数字

在实际工作中,人们总是愿意把近似数写成十进制的有限形式,因此就总结出不计算绝对误差和相对误差,而直接由组成近似数的数字个数来表示近似数精度的方法. 为此,需要建立有效数字的概念.

我们熟知,任何十进制数皆可写成10的幂级数的形式. 例如,

$$32.67 = 3 \times 10^1 + 2 \times 10^0 + 6 \times 10^{-1} + 7 \times 10^{-2},$$

$$0.0385 = 3 \times 10^{-2} + 8 \times 10^{-3} + 5 \times 10^{-4}.$$

一般地,可将准确数 x(设 $x>0$)的近似值 a 表示为

$$a = a_1 \times 10^{m-1} + a_2 \times 10^{m-2} + \cdots + a_p \times 10^{m-p} + \cdots + a_n \times 10^{m-n}. \tag{1-7}$$

其中,$a_1 \neq 0$,每个 a_i($i = 1,2,\cdots,n$)是0到9中的一个数字,p,n 为正整数,m(称为 a 的阶)为整数.

一般可表示为

$$a = \pm 10^m \times 0.a_1 a_2 \cdots a_n. \tag{1-8}$$

定义 1.3 对(1-7),如果

$$|\Delta a| = |x - a| \leqslant \frac{1}{2} \times 10^{m-p}, \tag{1-9}$$

则称 a **准确到** 10^{m-p} 位,或者说 a 具有 p 位**有效数字**(significant digits). 其中,$10^{m-1},\cdots,10^{m-p}$ 都是 a 的有效数位.

特别地,当 a 准确到末位(即 $p=n$)时,称 a 为 n 位**有效数字**,其中,$a_1,a_2,\cdots,a_n$ 分别称为 a 的第1,第2,…,第 n 位**有效数字**.

当 x 是准确值有很多位数时,常常需要按字长限制取 x 的位数来确定近似值 a,这个 a 按四舍五入规则来选取,能保证绝对误差最小. 例如,

$$x = \pi = 3.1415926\cdots$$

取三位,$a=3.14$,在所有三位数中3.14与 π 的误差最小;取五位,$a=3.1416$,在所有五位数中3.1416与 π 误差最小. 它们的误差均不超过末位的半个单位,即

$$|\pi - 3.14| \leqslant \frac{1}{2} \times 10^{-2}, \quad |\pi - 3.1416| \leqslant \frac{1}{2} \times 10^{-4}.$$

实际上,有效数字的概念无非是说:按四舍五入规则得到的近似值,其每一位数都是有效数字. 如果是收尾法或去尾法选取的近似值,其每一位数就不一定是有效数字,需先根据其绝对误差决定最末一位.

如3.10625(± 0.07),因为其十分位的半个单位是0.05,绝对误差0.07超过了十分位的半个单位,所以个位才是它的最末一位,其有效数字只是一位.

再如,以22/7的近似值3.142857作为 π 的近似值,因为

$$|\pi - 3.142857| = 0.00126\cdots < 0.002,$$

所以取 $\delta 3.142857 < 0.002 < 0.5 \times 10^{-2}$，$m = 1$，$m - p = -2$，故 $p = 3$，即 3.142857 是 π 的具有三位有效数字的近似值.

显然，有效数字位数与小数点的位置无关. 有效位数越多，绝对误差和相对误差就都越小.

有了有效数字概念后，以下写法是有区别的：

$$34.01,\ 34.0100,$$

前者表示四位有效数字，后者则表示六位有效数字.

我们约定：**原始数据都要用有效数字写**. 凡是不表明绝对（或相对）误差界的近似数，都被认为是有效数.

必须指出，定义 1.3 中的 $|\Delta a| \leqslant \frac{1}{2} \times 10^{m-p}$ 应理解为

$$10^{m-p-1} \leqslant |\Delta a| \leqslant \frac{1}{2} \times 10^{m-p}. \tag{1-10}$$

在定义 1.3 中表明了近似数的有效数字位数与绝对误差界的关系，下面指出有效数字（位）与相对误差界的关系.

定理 1.1 如果形如（1-7）的近似数 a 有 p 位有效数字，则

$$10^{-(p+1)} < \delta_r a \leqslant \frac{1}{2a_1} \times 10^{-(p-1)}. \tag{1-11}$$

证明 由（1-7）知

$$a_1 \times 10^{m-1} \leqslant a \leqslant (a_1 + 1) \times 10^{m-1}, \tag{1-12}$$

所以由（1-7）及定义 1.3 得

$$\delta_r a = \frac{\Delta a}{a} \leqslant \frac{0.5 \times 10^{m-p}}{a_1 \times 10^{m-1}} = \frac{1}{2a_1} \times 10^{-(p-1)}. \tag{1-13}$$

另一方面，由（1-10）及（1-12）得

$$\delta_r a = \frac{\Delta a}{a} \geqslant \frac{10^{m-p-1}}{(a_1 + 1) \times 10^{m-1}} = \frac{1}{a_1 + 1} \times 10^{-p} \geqslant 10^{-(p+1)}.$$

证毕.

例如，取 3.14 为 π 的近似值，其相对误差界由（1-11）可得

$$\delta_r 3.14 = \frac{\Delta 3.14}{3.14} \leqslant \frac{1}{2 \times 3} \times 10^{-(3-1)} = \frac{1}{6} \times 10^{-2} \approx 0.17\%.$$

若用（1-6）计算，则 $\delta_r 3.14 = \frac{\Delta 3.14}{3.14} < \frac{0.0016}{3.14} = 0.051\%$，可见用（1-11）估计 $\delta_r a$ 虽然简便，但结果比较粗糙.

推论 如果 $\delta_r a > \frac{1}{2} \times 10^{-p}$，则 a 含有的有效数字（位）的位数不超过 p.

证明 用反证法. 假设 a 有 $p+1$ 位有效数字（位），则由定理 1.1 得

$$\delta_r a < \frac{1}{2a_1} \times 10^{-p} \leqslant \frac{1}{2} \times 10^{-p},$$

这与题设矛盾. 证毕.

例 4　求$\sqrt{6}$的近似值 a，使 $\delta_r a \leqslant \frac{1}{2} \times 10^{-3}$.

解　因为$\sqrt{6} = 2.4494\cdots$，$a_1 = 2$，设 a 有 n 位有效数字，由定理 1.1 有

$$\delta_r a \leqslant \frac{1}{4} \times 10^{-(n-1)}.$$

令

$$\frac{1}{4} \times 10^{-(n-1)} \leqslant \frac{1}{2} \times 10^{-3},$$

求满足此不等式的最小正整数 n，可得 $n = 4$. 故取 $a = 2.449$，就合乎要求，因为

$$\delta_r 2.449 \leqslant \frac{1}{4} \times 10^{-3}.$$

反之，如果已知近似数 a 的相对误差界及 a_1，则由下述定理可估计出 a 的有效数字（位）的位数.

定理 1.2　如果已知近似数 a 的相对误差界

$$\delta_r a \leqslant \frac{1}{2(a_1 + 1)} \times 10^{-(p-1)}, \tag{1-14}$$

则 a 至少有 p 位有效数字（位）.

证明　由（1－6）、（1－12）及（1－14）知

$$\delta a = a \times \delta_r a \leqslant (a_1 + 1) \times 10^{m-1} \times \frac{1}{2(a_1 + 1)} \times 10^{-(p-1)} = \frac{1}{2} \times 10^{m-p},$$

由定义 1.3 知 a 至少有 p 位有效数字. 证毕.

1.3　误差估计与误差分析

1.3.1　算术运算的误差界

定理 1.3　设 x 与 y 是精确值，a 与 b 是相对的近似值，绝对误差界分别为 δa 与 δb，则

$$\delta(a \pm b) = \delta a \pm \delta b,$$

$$\delta(ab) \approx |a| \delta b + |b| \delta a,$$

$$\delta\left(\frac{a}{b}\right) \approx \frac{|a| \delta b + |b| \delta a}{|b|^2} \quad (b \neq 0).$$

例 1　$a = 1.21 \times 3.65 + 9.81$，其中每个数据的绝对误差界为 0.005，$a$ 的绝对误差界

$$\begin{aligned} \delta(a) &= \delta(1.21 \times 3.65) + \delta(9.81) \\ &\approx 1.21 \times 0.005 + 3.65 \times 0.005 + 0.005 \\ &= 0.0293 \leqslant 0.03. \end{aligned}$$

定理 1.4　设 a 与 b 是近似值，相对误差界分别为 $\delta_r a$ 与 $\delta_r b$，则

$$\delta_r(a + b) = \max\{\delta_r a, \delta_r b\} \quad (a \text{ 与 } b \text{ 同号}),$$

$$\delta_r(a - b) = \frac{|a| \delta_r a + |b| \delta_r b}{|a - b|} \quad (a \text{ 与 } b \text{ 同号}),$$

$$\delta_r(ab)\approx\delta_r a+\delta_r b,$$

$$\delta_r\left(\frac{a}{b}\right)\approx\delta_r a+\delta_r b\ (b\neq 0).$$

例 2 例 1 中 a 的相对误差界

$$\begin{aligned}\delta_r(a)&=\max\{\delta_r(1.21\times 3.65),\delta(9.81)\}\\&\approx\max\{\delta_r(1.21)+\delta_r(3.65),\delta_r(9.81)\}\\&=\max\left\{\frac{\delta(1.21)}{1.21}+\frac{\delta(3.65)}{3.65},\frac{\delta(9.81)}{9.81}\right\}\\&=\max\left\{\frac{0.005}{1.21}+\frac{0.005}{3.65},\frac{0.005}{9.81}\right\}\\&\approx\max\{0.0055,0.0005\}\\&=0.0055.\end{aligned}$$

1.3.2 函数求值的误差估计

在计算一元或多元函数的值时,由于自变量数据不精确会产生误差.

设 f 是一元函数,要计算在点 x 的函数值,但又知 x 的近似值为 a,以 $f(a)$ 近似 $f(x)$,$f(a)$ 的绝对误差界 $\delta(f(a))$ 可用泰勒公式估计. 假定 f 在包含 x 与 a 的一个开区间上存在足够阶的导数,则有

$$\Delta(f(a))=f(x)-f(a)=f'(a)(x-a)+\frac{f''(\xi)}{2!}(x-a)^2,$$

其中 ξ 在 x 与 a 之间,取绝对值,得

$$\begin{aligned}|\Delta(f(a))|&=|f(x)-f(a)|\\&\leqslant|f'(a)|\delta a+\frac{|f''(\xi)|}{2!}(\delta a)^2.\end{aligned}$$

假定 $|f''(a)|$ 与 $|f'(a)|$ 相比不太大,则可以忽略高阶项,得

$$\delta(f(a))\approx|f'(a)|\delta a. \tag{1-15}$$

但是,如果 $|f'(a)|$ 是零或值很小,则要考虑后面的项. 特别地,若

$$f'(a)=f''(a)=\cdots=f^{(k-1)}(a)=0,\quad f^{(k)}(a)\neq 0,$$

且 $|f^{(k+1)}(\xi)|$(ξ 在 x 与 a 之间任意变动)不很大,则

$$\delta(f(a))\approx\frac{|f^{(k)}(a)|}{k!}(\delta a)^k. \tag{1-16}$$

例 3 设 $f(x)=\sin x,\alpha=45°,\beta=90°,\delta\alpha=0.1°,\delta\beta=0.2°$. 因为

$$f'(a)=\cos\alpha=\frac{\sqrt{2}}{2},\quad f'(\beta)=0,\quad f''(\beta)=-\sin\beta=-1,$$

所以,在把 $\delta\alpha$ 与 $\delta\beta$ 化为弧度后有

$$\delta(\sin\alpha)\approx|\cos\alpha|\delta\alpha=\frac{\sqrt{2}}{2}\times 0.1\times\frac{\pi}{180}\approx 1.2\times 10^{-3},$$

$$\delta(\sin\beta)\approx\frac{|\sin\beta|}{2!}(\delta\beta)^2=\frac{1}{2}\times\left(0.2\times\frac{\pi}{180}\right)^2\approx 6.1\times 10^{-6}.$$

多元函数值的误差界可用多元函数的泰勒公式得到.

设 n 元函数 $y=f(x_1,x_2,\cdots,x_n)$ 在点 $X=(x_1,x_2,\cdots,x_n)$ 附近可微,$A(a_1,a_2,\cdots,a_n)$ 为 X 附近的点,a_i 为 x_i 的近似值($i=1,2,\cdots,n$),e_i 是 a_i 的绝对误差,我们讨论用 $f(A)$ 代替 $f(x)$ 时的误差.

由数学分析知道,函数的改变量,即误差为

$$\Delta f=f(x_1,x_2,\cdots,x_n)-f(a_1,a_2,\cdots,a_n)=\mathrm{d}f+o(\rho).$$

其中,$o(\rho)=\sqrt{\sum_{i=1}^{n}e_i^2}$,当 $\rho\to 0$ 时,$o(\rho)$ 为比 ρ 高阶的无穷小量,$\mathrm{d}f=\sum_{i=1}^{n}\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}e_i$ 是 f 在点 A 处的全微分. 我们就取 $\mathrm{d}f$ 为 f 在点 A 处的绝对误差,记为 $e_f=\mathrm{d}f$,则

$$|e_f|\leqslant\delta(f(a_1,a_2,\cdots,a_n))\approx\sum_{i=1}^{n}\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\right|\cdot\delta a_i. \tag{1-17}$$

显然,$\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\right|$ 越大,初始数据的误差 e_i 对计算结果的影响也越大,称 $\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\right|$ 为在绝对误差意义下 f 在点 A 处的条件数,记为 $\mathrm{cond}\ (f)_e$. 当 $\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\right|$ 很大时,称 f 在点 A 处是坏条件的;否则,是好条件的.

但对很多问题来说,计算结果的精度是用相对误差来描述的,于是设 $a_i\neq 0$,a_i 的相对误差记为 $\varepsilon_{a_i}=\frac{e_i}{a_i}$;设 $f(A)\neq 0$,f 的相对误差记为 $\varepsilon_f=e_f/f(A)$,则由(1-17)得

$$\varepsilon_f=\sum_{i=1}^{n}\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\cdot\frac{a_i}{f(A)}\cdot\varepsilon_{a_i},$$

$$|\varepsilon_f|\leqslant\delta f=\sum_{i=1}^{n}\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\cdot\frac{a_i}{f(A)}\right|\cdot\delta a_i. \tag{1-18}$$

显然,$\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\cdot\frac{a_i}{f(A)}\right|$ 反映误差 ε_{a_i} 对计算结果的影响,其越大,初始数据的相对误差 ε_{a_i} 对结果的影响也越大,称 $\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\cdot\frac{a_i}{f(A)}\right|$ 为在相对误差意义下 f 在点 A 处的条件数,记为 $\mathrm{cond}(f)$. 当 $\left|\frac{\partial f(a_1,a_2,\cdots,a_n)}{\partial x_i}\cdot\frac{a_i}{f(A)}\right|$ 很大时,称 f 在点 A 处是坏条件的;否则,是好条件的.(注意,在不提在什么意义下时,是指在相对误差意义下)在坏条件下,用 $f(A)$ 代替 $f(x)$ 的误差会很大,这时的代替是不适宜的.

例如,设 $f(x)=x^2+x-10100$. 当 $x=100$ 时,$f(100)=0$;当 $x\doteq a=99$ 时,$f(99)=-200$,则

$$\mathrm{cond}\ (f)_e=|f'(a)|=199,\quad \mathrm{cond}(f)=\left|f'(a)\cdot\frac{a}{f(a)}\right|=\left|199\times\frac{99}{-200}\right|\doteq 99.$$

可见,不论在什么意义下,$f(x)$ 在 a 处是坏条件的,用 $f(99)$ 代替 $f(100)$ 作为其近似值是