数值分析

郑苏民 编

云南大学出版社

责任编辑：张世鸾
封面设计：丁群亚

数 值 分 析

郑苏民 著

*

云南大学出版社出版
（云南大学校内）
云南大学印刷厂印装

*

开本：787×1092／16 印张：18.75 字数:456千
1990年4月第1版 1990年4月第1次印刷
ISBN 7-81025-013-2／O·2 定价：7.50元

序

数值分析这门课程应用非常广泛。国防尖端的一些科研课题，例如核武器的研制，导弹的发射，国内、国际气象资料的汇总以求得天气图象，等等，都离不开数值分析。

本书作者郑苏民教授亲自在云南大学试教过这本教材，几经修改增删。就内容而言，符合大纲要求。在取材方面，着重基本概念、基础知识的阐述，内容丰富，重点突出。章节之间联系较紧密，前后呼应。著者选了大量例题，帮助读者理解和及时掌握课程内容及技巧。每章挑选难度适中的习题（共计两百多题），培养读者独立工作能力和学习积极性。在叙述方面，著者注意深入浅出，通俗易明。

本书主要内容有误差理论，各种插值方法，差分方法，迭代法，函数逼近，数值积分，线性方程，微分方程的数值解法，矩阵代数迭代技术。著作给出诸多实用方法，发人深省。

由于本书经过多次实践考验，它的问世（特别是在科技书籍比较难出版的现在）相信会受到广泛的欢迎。

朱德祥

1990.5.

前　言

本书是在作者多年讲授《数值分析》课程的讲义基础上，经过多次修改补充编写而成。其对象是非计算数学专业的广大工程与应用科学师生、工程技术人员及师范院校师生。所需的预备知识为高等数学、线性代数、算法语言等。为使读者易于学习使用，论述中不追求数学的严密性，而主要以分析方法来阐明问题。

数值分析是一门应用数学，在其发展中，许多名家曾为一些重要计算公式的确立作出过贡献，因而时常会遇到冠以他们名字的公式，例如牛顿-柯特斯公式、辛普生公式，欧拉公式等等，然而实质上许多这类公式不过是基本差分运算公式的一些特例。作者试图在部分章节中不从个别人名公式的推导入手，而从基本差分运算公式出发导出常用的数值积分、数值微分方程解法中的一些公式，从而使本书有别于其他同类教材，而具有更完美的系统性。

本书共分十章：第一章简要叙述数值计算中误差的产生、传播问题，引导读者对误差问题的重视；第二章为插值理论，是传统方法介绍，它是数值计算方法的基础；第三章介绍将差分符号作为"算子"而进行运算的基本方法，它能以简单明了的方式导出数值分析中许多常用的公式，还能导出一些未被人们所熟知的公式，是数值积分、微分、微分方程解法的基础；第四章总和的计算，包括有限项级数的求和与无穷级数的加速收敛方法，还对阶乘函数及其在总和计算中的作用作了初步介绍；第五章非线性方程求解方法，虽然传统上公认牛顿法是[redacted]用场合优于牛顿法；第六章讨论了函数[redacted]它们除了在函数逼近中有重要作用外，[redacted]运算为基础引人数值积分方法进而导出[redacted]高斯型求积方法的推广奠定了物质基础[redacted]基础引人数值求解微分方程的方法，[redacted]及；第九章线性代数方程组的直接解法[redacted]适用于稀疏矩阵计算的线性代数方程组[redacted]

本书给出许[redacted]乐于用自己熟悉的语言编制程序。限于[redacted]人本书。

云南师范大[redacted]鸾副教授曾对本书提出过一些中肯的意[redacted]

限于作者水[redacted]

作者谨识

1990 年 5 月

目　　录

第一章　误差理论的基本知识

人们在日常生活和生产实践中往往是不可能要求绝对准确的.由于测量手段的限制，对同一个物体进行多次同样的测量，往往有不同的结果. 测量精度高，各次所测数据相同的位数就多，而尾数往往受各种随机因素的影响而各不相同.于是，我们只能把每次得到的数据看成是被测物理量的一个近似值. 因为近似值与真值不可能完全相同，它们之间必然存在有一定差值，通常称为误差. 在任何实际测量和计算中，我们都不可能避免误差的产生. 一般，只要误差不超过给定的范围，也就满足了. 为此，我们必须考虑近似值本身的误差以及用这些近似值进行运算所得结果的误差情况，这在工程计算及实验测量中是很重要的. 这就是本章所要讨论的主要内容.

§1.1　数　学　准　备

本节我们将复习几条在微积分中已熟知的定理，这些定理将在本教材中用到，它们虽然简单，但在数值分析中却非常重要，我们仅列举出来而不作证明.

定理 1.1　若$f(x)$在有限区间$[a,b]$内连续，且$f(a)$与$f(b)$符号相反，则在$[a,b]$中至少有一点ξ，使$f(\xi)=0$.

定理 1.2　若$f(x)$在$[a,b]$内连续，且若λ_1、λ_2为正常数，则在$[a,b]$中至少有一点ξ使

$$\lambda_1 f(a)+\lambda_2 f(b)=(\lambda_1+\lambda_2)f(\xi). \tag{1.1}$$

定理 1.3(罗尔定理)　若$f(x)$在$[a,b]$内连续，且在$[a,b]$内$f'(x)$存在. 若$f(a)=f(b)=0$，则在$[a,b]$内至少有一点ξ使

$$f'(\xi)=0. \tag{1.2}$$

定理 1.4(导数的中值定理)　若$f(x)$在$[a,b]$内连续，且在$[a,b]$内$f'(x)$存在，则在$[a,b]$内至少有一点ξ使

$$f(b)-f(a)=(b-a)f'(\xi). \tag{1.3}$$

定理 1.5　设$f(x)$在$[a,b]$内连续可积，且$|f(x)|\leqq M$，M为常数，则

$$\left|\int_a^b f(x)dx\right|\leqq\int_a^b|f(x)|dx\leqq M(b-a). \tag{1.4}$$

定理 1.6(积分的中值定理)　设$g(x)$在$[a,b]$内连续可积且不改变符号，又设$f(x)$在$[a,b]$内连续，则在$[a,b]$内至少有一点ξ使

$$\int_a^b f(x)g(x)dx=f(\xi)\int_a^b g(x)dx. \tag{1.5}$$

定理 1.7　若a为有限常数，u为x的可微函数，且$\dfrac{\partial F}{\partial x}$连续，则

$$\frac{d}{dx}\int_a^u F(x,s)ds=\int_a^u \frac{\partial F(x,s)}{\partial x}ds+F(x,u)\frac{du}{dx}. \tag{1.6}$$

定理 1.8 （柯西中值定理）

若 $g(x)$、$h(x)$ 是区间 $[a,b]$ 上的连续可微函数，则在区间内存在一数 ξ 使得

$$\frac{g(b)-g(a)}{h(b)-h(a)}=\frac{g'(\xi)}{h'(\xi)}. \tag{1.7}$$

在数值分析中有一个非常重要的工具是泰勒定理以及相应的泰勒级数．除了数值线性代数以外，在本教材的所有章节中都要用到．泰勒定理给出一个简单办法将 $f(x)$ 用多项式表出，因而有利于对 $f(x)$ 进行计算．

定理 1.9 （泰勒定理）

设 $f(x)$ 在 $[a,b]$ 内对某些 $n\geqq 0$ 值有 $n+1$ 阶连续导数，且 $x,x_0\in[a,b]$，则

$$f(x)=p_n(x)+R_{n+1}(x);$$

$$p_n(x)=f(x_0)+\frac{(x-x_0)}{1!}f'(x_0)+\cdots+\frac{(x-x_0)^n}{n!}f^{(n)}(x_0); \tag{1.8}$$

$$R_{n+1}(x)=\frac{1}{n!}\int_{x_0}^{x}(x-t)^n f^{(n+1)}(t)dt=\frac{(x-x_0)^{n+1}}{(n+1)!}f^{(n+1)}(\xi), \tag{1.9}$$

其中 ξ 是 x_0 与 x 间的某点．

应用泰勒定理，我们可得到下列标准公式

$$e^x=1+\frac{x}{1!}+\frac{x^2}{2!}+\cdots+\frac{x^n}{n!}+\frac{x^{n+1}}{(n+1)!}e^{\xi_x}; \tag{1.10}$$

$$\cos(x)=1-\frac{x^2}{2!}+\frac{x^4}{4!}\cdots+(-1)^n\frac{x^{2n}}{(2n)!}+(-1)^{n+1}\frac{x^{2n+2}}{(2n+2)!}\cos(\xi_x); \tag{1.11}$$

$$\sin(x)=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\cdots+(-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!}+(-1)^n\frac{x^{2n+1}}{(2n+1)!}\cos(\xi_x); \tag{1.12}$$

$$(1+x)^\alpha=1+C_1^\alpha x+C_2^\alpha x^2+\cdots+C_n^\alpha x^n+C_{n+1}^\alpha\frac{x^{n+1}}{(1+\xi_x)^{n+1-\alpha}}, \tag{1.13}$$

其中，α 为任意实数，未知数 ξ_x 位于 x 与 0 之间，

$$C_k^\alpha=\frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!},\quad k=1,2,3,\cdots.$$

(1.13) 的重要特例为几何级数

$$\frac{1}{1-x}=1+x+x^2+\cdots+x^n+\frac{x^{n+1}}{1-x}+\cdots. \tag{1.14}$$

只要在 (1.13) 中令 $\alpha=-1$，且将 x 代以 $(-x)$，余项的形式较 (1.13) 简单，其正确性可分别在 (1.14) 两端乘以 $1-x$ 而得到验证．要将 (1.10)~(1.14) 构成无穷极数只要令 $n\to\infty$ 便可．(1.10)~(1.12) 对所有 x 收敛，(1.13)、(1.14) 仅对 $|x|<1$ 收敛．

用 (1.8) 可将任意函数 $f(x)$ 直接展开为泰勒级数，取用项数可根据需要确定．但有些函数由于微分的复杂性，有时可利用 (1.10)~(1.14) 间接展开为泰勒级数．下面给出几个例子，它们的余项形式均比直接用 (1.9) 来得简单．

例 1.1　令 $f(x)=e^{-x^2}$，在 (1.10) 中用 $-x^2$ 代 x，得

$$e^{-x^2}=1-x^2+\frac{x^4}{2!}-\cdots+(-1)^n\frac{x^{2n}}{n!}+(-1)^{n+1}\frac{x^{2n+2}}{(n+1)!}e^{\xi_x},$$

式中，$x^2\leqq\xi_x\leqq 0$.

例 1.2　令$f(x)=\mathrm{tg}^{-1}x$，而由(1.14)得

$$\frac{1}{1+u^2}=1-u^2+u^4-\cdots+(-1)^nu^{2n}+(-1)^{n+1}\frac{u^{2n+2}}{1+u^2}+\cdots,$$

在0到x之间积分得

$$\mathrm{tg}^{-1}x=x-\frac{x^3}{3}+\frac{x^5}{5}-\cdots(+)^n\frac{x^{2n+1}}{2n+1}+(-1)^{n+1}\int_0^x\frac{u^{2n+2}}{1+u^2}du+\cdots.$$

应用积分中值定理，可得

$$\int_0^x\frac{u^{2n+2}du}{1+u^2}=\frac{x^{2n+3}}{2n+3}\frac{1}{1+\xi_x^2},$$

ξ_x在0到x之间.

例 1.3　令$f(x)=\int_0^1\sin(xt)dt$，利用(1.12),

$$f(x)=\int_0^1\left[xt-\frac{x^3t^3}{3!}+\cdots+(-1)^{n-1}\frac{(xt)^{2n-1}}{(2n-1)!}+(-1)^n\frac{(xt)^{2n+1}}{(2n+1)!}\cos\xi_{xt}\right]dt$$

$$=\sum_{j=1}^{n}(-1)^{j-1}\frac{x^{2j-1}}{(2j-1)!(2j)}+(-1)^n\frac{x^{2n+1}}{(2n+1)!}\int_0^1t^{2n+1}\cos\xi_{xt}dt,$$

ξ_{xt}在0和xt之间.直接应用积分中值定理得

$$\int_0^1\sin xt\,dt=\sum_{j=1}^{n}(-1)^{j-1}\frac{x^{2j-1}}{(2j)!}+(-1)^n\frac{x^{2n+1}}{(2n+2)!}\cos\xi_x,$$

式中ξ_x在0与x之间.

泰勒级数可以容易地推广到二维以及更高维数．若$f(x,y)$为二独立之变数x, y的函数，则$f(x,y)$可以在给定点(x_0,y_0)处展开为级数．通常用$L(x_0,y_0,x_1,y_1)$表示连接(x_0,y_0)及(x_1,y_1)的直线段上点(x,y)的集合.

定理 1.10　令(x_0y_0)及$(x_0+\xi,y_0+\eta)$为给定点，假定$f(x,y)$为$L(x_0,y_0;x_0+\xi,y_0+\eta)$邻域对所有$(x,y)$有$n+1$阶连续可微函数，则

$$f(x_0+\xi,y_0+\eta)=f(x_0,y_0)+\sum_{j=1}^{n}\frac{1}{j!}\left(\xi\frac{\partial}{\partial x}+\eta\frac{\partial}{\partial x}\right)^jf(x,y)\Big|_{\substack{x=x_0\\y=y_0}}$$

$$+\frac{1}{(n+1)!}\left(\xi\frac{\partial}{\partial x}+\eta\frac{\partial}{\partial y}\right)^{n+1}f(x,y)\Big|_{\substack{x=x_0+\theta\xi\\y=y_0+\theta\eta}}$$

式中，$0\leqq\theta\leqq 1$，点$(x_0+\theta\xi,y_0+\theta\eta)$为直线$L(x_0,y_0;x_0+\xi,y_0+\eta)$上一个未知点.

例 1.4　作为一个简单例子，考虑$f(x,y)=x/y$在点(1,1)处的展开，令$n=1$，于是

$$\frac{x}{y}=f(1,1)+(x-1)\frac{\partial f(x,y)}{\partial x}\Big|_{x=y=1}+(y-1)\frac{\partial f(x,y)}{\partial y}\Big|_{x=y=1}$$

$$+\frac{1}{2}\left[(x-1)^2\frac{\partial f(x,y)}{\partial x^2}\right]\Bigg|_{\substack{x=\delta\\y=\gamma}}$$

$$+2(x-1)(y-1)\frac{\partial^2 f(x,y)}{\partial x\partial y}\Bigg|_{\substack{x=\delta\\y=\gamma}}+(y-1)\frac{2\partial^2 f(x,y)}{\partial y^2}\Bigg|_{\substack{x=\delta\\y=\gamma}}$$

$$=1+(x-1)-(y-1)+\frac{1}{2}\left[(x-1)^2\cdot 0-2(x-1)(y-1)\frac{1}{\gamma^2}+(y-1)^2\frac{2\delta}{\gamma^2}\right],$$

δ、γ 为 $L(1,1;x,y)$ 线上的一点，当 x, y 很接近于 $(1,1)$ 时

$$\frac{x}{y}\approx 1+x-y.$$

§1.2 误 差 的 来 源

误差的来源是多种多样的，现就主要的几种叙述如下:

一、描述误差 为了描述客观现象及运动规律，往往需要将它们归结为数学问题，或叫建立数学模型.由于各种限制,数学模型不可能百分之百地描绘出客观现象，也就是说，会造成一定的误差.我们称之为"描述误差".

例 2.1 我们考虑一个从地面向空中发射的质量为 m 的炮弹，设炮弹经常保持在地面附近飞行.我们引用直角坐标 xy，其原点为发射点，y 轴垂直地指向天空.在时间 t 炮弹的位置可用一矢量表示 $\mathbf{r}(t)=x(t)\mathbf{i}+y(t)\mathbf{j}+z(t)\mathbf{k}$ 用牛顿第二定律便可写出该炮弹飞行的一种数学模型

$$m\frac{d^2\mathbf{r}(t)}{dt^2}=-mg\mathbf{k}-b\frac{d\mathbf{r}(t)}{dt}, \tag{1.16}$$

式中 $b>0$ 为一常数，g 为重力加速度.该方程式说明作用在炮弹上只有两种力，① 地球重力 ② 摩擦阻力，它的大小与速度 $|\mathbf{v}(t)|=\left|\frac{d\mathbf{r}(t)}{dt}\right|$ 成正比例而方向与速度方向相反.

从某种意义上说，这是一个很好的模型，在有些时候甚至不用考虑后一项摩擦阻力（例如在普通力学里面）.但这个模型没有考虑到与飞行平面相垂直的力，例如横向风力以及也没有考虑到由于地球转动造成的科利奥利斯力、此外摩擦阻力与速度的关系不一定是线性关系，而可能与 $|\mathbf{v}(t)|^\alpha$ $(\alpha\neq 1)$ 成比例.于是直接引用 (1.16) 为数学模型就必然会发生误差.

二、观测误差 由于所使用的工具不够精确或人眼观测时有出入而造成的误差.

三、截断误差 用有限项结果代替无限项，因忽略该有限项后面的高次小量而产生的误差为截断误差.例如将 $f(x)$ 在 0 附近 $|x|<1$ 范围内展开为幂级数

$$f(x)=\sum_{k=0}^{\infty}a_k x^k, \tag{1.17}$$

把和式中第 n 项以后的诸项丢掉，剩下的和式称为截断式

$$p_n(x)=\sum_{k=0}^{n-1}a_k x^k. \tag{1.18}$$

第 n 项之后各项之和将构成所谓截断误差，实际上就是展开式的余项

$$R_n(x)=|f(x)-p_n(x)|\leqslant\sum_{k=n}^{\infty}|a_k|. \tag{1.19}$$

对于交错级数，由于各项正负交替，只要每一项的绝对值小于前一项，并且当 n 趋于无限大时，第 n 项的极限为 0，则对于这样的级数估计误差很容易，只要用余项中的头一项就可以了，因为后面各项是正负相消的，余项的总和总是小于头一项的．例如

$$\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\frac{x^5}{5}+\cdots,$$

当 $x=0.01$ 时，在采用前三项作为函数的近似值时有

$$\ln(1.01)=0.01-\frac{(0.01)^2}{2}+\frac{(0.01)^3}{3}+R_4,$$

我们只要取第四项作为误差的估计就足够了，

$$R_4<\left|\frac{(0.01)^4}{4}\right|=0.0000000025.$$

对于一般的级数展开式的余项可用下公式计算

$$R_n(x)=\frac{1}{(n-1)!}\int_0^x f^{(n)}(x-t)t^{n-1}dt, \tag{1.20}$$

式中 $f^{(n)}$ 为 $f(x)$ 的 n 阶导数．

四、数据误差 通常在进行计算时所使用的数据都是近似的，因而总要带有一定的误差，即使是有理数，在分数情况下往往也只能表示成近似的小数．例如

$$\frac{1}{7}=0.1428571\cdots,$$

$$\frac{1}{4!}=\frac{1}{24}=0.0416666\cdots.$$

使用时也只能取有限的几位，仍然要产生截断误差或舍入误差．

五、舍入误差 对于许多永远除不尽的数，为了实际应用，总要截取有限的几位，而丢掉后面的数字，既然丢掉了就不可避免要产生误差．为了尽可能减少由于丢弃数字所带来的误差，通常采用所谓“四舍五入”的规则．这个规则规定如果保留 n 位有效数字，则凡是第 $n+1$ 位大于 5 的数就在第 n 位加 1，小于 5 则第 n 位不变．如果第 $n+1$ 位恰恰是 5，则在第 n 位为奇数时，第 n 位加 1，是偶数时第 n 位不变．有些人不注意这一点，结果凡遇到第 $n+1$ 位是 5 就往前进一，这往往造成一些附加的误差．这一点是很明显的，因为一般说来奇数与偶数出现机会相等，如果不管第 n 位是奇是偶，第 $n+1$ 位一律遇 5 进一，这就造成只遇到两个 5 就进 2 的可能，而实际上只能进一．

在实际计算时，舍入方法使用不当还会造成很大的误差，例如在计算两组三位有效数字相乘并求它们的差值．按理乘积只能取三位有效数字（多余的位数是无意义的），但在计算时就不能先舍入后求差而应先求差再舍入．例如下面两组三位有效数字的乘积之差

$$\begin{array}{lll} 0.438\times 0.563=0.246594 & & 0.247 \\ 0.436\times 0.563=\underline{0.245468} & & \underline{0.245} \\ \qquad\qquad\quad 0.001126 & & 0.002 \end{array}$$

显然应取 0.001．由于乘积只能取三位有效数字，乘积的差值 1.126×10^{-3} 只有第一位是

精确的，如果后面还有其它计算而我们都将上面第二、三、四位保留在计算中，那么这种不精确值便会向前传播，严重起来会使整个计算结果无意义．但我们也不能走向反面，过早进行舍入，在上述计算中我们用第三列数据表示，还未求差之前就进行了舍入计算，结果得到的是 2×10^{-3}，这与精确的 1×10^{-3} 相比误差达到 100%．

§ 1.3 绝对误差与相对误差

绝对误差是指某一数量的真值 x_T 与给定的近似值 x_A 之间的数量差，以 e_a 表示为

$$e_a = x_T - x_A. \tag{1.21}$$

但往往数量的真值我们是不知道的，我们只能求出近似值 x_A，因而绝对误差 e_a 也只能对其范围作一点估计．例如若求 $\sqrt{3} = 1.7320508\cdots$ 如果只取

$$\sqrt{3} = 1.732,$$

则 $$e_a = 0.0000508\cdots,$$

于是可认为 $$|e_a| < 0.0001,$$

或写为 $$|e_a| < \varepsilon_a \ ,\ \varepsilon_a \text{ 为误差上限}. \tag{1.22}$$

绝对误差通常不能表现出真正的准确度，例如测量高压交流电 32000V 能准确到 10~ 20V 就是很准确的了．但在研究放大器时，几个 mv（毫伏）都要认真对待．所以绝对误差不能表明度量工作的好坏，它只是两个数量之间的差值，并没有表明这些误差在整体中每一个单位所分配到的误差，而更能合理地衡量一个近似值的精确度的正是这种误差 — — 相对误差．相对误差定义为

$$e_r = \frac{|x_T - X_A|}{|x_T|}.$$

和绝对误差一样，我们还是不能定出 e_r 的准确值，只能估计它的范围，

$$|e_r| = \left|\frac{x_T - x_A}{x_T}\right| \leqq \varepsilon_r, \tag{1.23}$$

ε_r 为相对误差上限．

在电子计算机中常用的数制为二进、八进、十进或十六进数．若令 α 为基数，a_i 是基数为 α 的所有可能取的数，则一个浮点形式的数 x 可写为

$$x = \sigma\times (.a_1a_2\cdots a_na_{n+1}\cdots)_\alpha \times \alpha^e \tag{1.24}$$

序列 $a_1\cdots a_n$ 常称为尾数，或有效位数，下标 α 指以 α 为基数的数制．e 为指数，$\sigma = \pm 1$ 为数 x 的符号，$+1$ 为正数，-1 为负数．

由于计算机的位数以及要求的位数所限（例如只存在 n 位）则 $n+1$ 位以后的数字便被舍去，于是产生了舍入误差．数量 x 在取 n 位尾数时可表为

$$fl(x) = \begin{cases} \sigma(.a_1a_2\cdots a_n)\alpha\times \alpha^e \ ,\ 0\leqq a_{n+1} < \dfrac{\alpha}{2}, \\ \sigma[(.a_1a_2\cdots a_n)\alpha + (00\cdots01)_\alpha]\times \alpha^e,\ \dfrac{\alpha}{2} \leqq a_{n+1} < \alpha. \end{cases} \tag{1.25}$$

式中 $(.00\cdots01)_\alpha$ 意味着 α^{-n}．(1.25) 表明了在任何数制下舍入的做法，如果第 $n+1$ 位

数 a_{n+1} 大于或等于$\frac{\alpha}{2}$，则在尾数加 1 或叫舍入，小于$\frac{\alpha}{2}$ 则舍去.

当取完 n 位尾数后，浮点数 $fl(x)$ 的误差将为

$$x-fl(x)=\sigma(.00\cdots 0a_{n+1}a_{n+2}\cdots)_\alpha\times\alpha^e=\sigma(.a_{n+1}a_{n+2}\cdots)_\alpha\times\alpha^{e-n},$$

其绝对舍入误差为

$$|e_a|=|x-fl(x)|\leqq\frac{1}{2}\alpha^{e-n}. \tag{1.26}$$

对于$\frac{\alpha}{2}\leqq a_{n+1}<\alpha$ 的数，(1.26) 也适用. 相应可用 (1.26) 算出其相对舍入误差

$$|e_r|=\frac{|x-fl(x)|}{|x|}\leqq\frac{\frac{1}{2}\alpha^{e-n}}{|x|}=\frac{\frac{1}{2}\alpha^{-n}}{|x|\alpha^{-e}}=\frac{\frac{1}{2}\alpha^{-n}}{|x|\frac{(.a_1a_2\cdots)_\alpha}{x}}\leqq\frac{1}{2}\frac{\alpha^{-n}}{(.10000)_\alpha}$$

$$=\frac{1}{2}\alpha^{-n+1}. \tag{1.27}$$

例如，当计算机使用二进制时，则 n 位尾数的相对误差为

$$|e_r|_2=\frac{|x-fl(x)|}{|x|}\leqq 2^{-n};$$

当采用十进制时，

$$|e_r|_{10}=\frac{|x-fl(x)|}{|x|}\leqq 5\times 10^{-n}.$$

有些计算机不采用舍入方法，而仅采用截断方法，即不管余项有多大，超过 n 位一律截去，这时数的表示仅采用 (1.25) 中的第一部份，同时绝对误差 (1.26) 应改为

$$|e_a|=|x-fl(x)|\leqq\alpha^{e-n}, \tag{1.28}$$

而相对误差为

$$|e_r|=\frac{|x-fl(x)|}{|x|}\leqq\alpha^{-n+1}. \tag{1.29}$$

§1.4 误 差 的 传 播

一、算术运算中误差的传播

我们将在本节研究由于数的计算而引起的误差效应. 我们来看基本的算术运算. 我们用符号 $\wedge$ 表示算术运算，如 $+$、$-$、$\times$、$/$ 号. 而用 $\wedge^*$ 表示同样运算的计算机操作，这意味着因而将产生舍入或截断. 设 x_A 和 y_A 为存在有误差的待运算数，其真值为

$$x_T=x_A+\varepsilon,\qquad y_T=y_A+\eta,$$

则在计算机中进行运算后 (进行了 $x_A\wedge^* y_A$ 运算) 其误差为

$$x_A\wedge y_T-x_A\wedge^* y_A=(x_T\wedge y_T-x_A\wedge y_A)+(x_A\wedge y_A-x_A\wedge^* y_A). \tag{1.30}$$

(1.30) 等号右边第一项称为传播误差，为二有误差的数通过运算后产生的附加误差，第二项称为舍入误差. 对第二项我们有

$$x_A \wedge^* y_A = fl(x_A \wedge y_A). \tag{1.31}$$

这表示首先对 $x_A \wedge y_A$ 作精确计算然后舍入（或截断）. 由 (1.27) 及 (1.31) 我们有

$$|x_A \wedge y_A - x_A \wedge^* y_A| \leqq \frac{\alpha}{2}|x_A \wedge y_A|\alpha^{-n} \tag{1.32}$$

我们先研究几个特殊情形的传播误差：

(*a*) 乘法 $\wedge = \times$ 或 $\cdot$ 或不写任何符号，运算 $x_A y_A$ 的传播误差为

$$x_T y_T - x_A y_A = x_T y_T - (x_T - \varepsilon)(y_T - \eta) = x_T \eta + y_T \varepsilon - \varepsilon\eta.$$

$$e_r|_{x_A y_A} = \frac{x_T y_T - x_A y_A}{x_T y_T} = \frac{\eta}{y_T} + \frac{\varepsilon}{x_T} - \frac{\varepsilon}{x_T}\frac{\eta}{y_T} = e_r|_{x_A} + e_r|_{y_A} - e_r|_{x_A} e_r|_{y_A}. \tag{1.33}$$

$|e_r|_{x_A}|e_r|_{y_A} \ll 1$ 时，我们有

$$e_r|_{x_A y_A} \approx e_r|_{x_A} + e_r|_{y_A}. \tag{1.34}$$

(*b*) 除法 $\wedge = /$ 根据类似方式有

$$e_r|_{x_A / y_A} = \frac{e_r|_{x_A} - e_r|_{y_A}}{1 - e_r|_{y_A}}. \tag{1.35}$$

当 $e_r|_{y_A} \ll 1$ 时

$$e_r|_{x_A / y_A} \approx e_r|_{x_A} - e_r|_{y_A}. \tag{1.36}$$

对于乘法和除法，误差的传播并不很快. 它们的相对误差等于各个因子相对误差之和或差

(*c*) 加法和减法 $\wedge = +$ 或 $-$，我们有如下关系

$$(x_T \pm y_T) - (x_A \pm y_A) = (x_T - x_A) \pm (y_T - Y_A), \tag{1.37}$$

$$|(x_T \pm y_T) - (x_A \pm y_A)| \leqq |x_T - x_A| + |y_T - y_A|. \tag{1.38}$$

就是说近似数的和与差的绝对误差等于它们绝对误差的和与差. 对于相对误差，应分别进行讨论.

对和的相对误差，可写出

$$\frac{(x_T + y_T) - (x_A + y_A)}{x_T + y_T} = \frac{x_T - x_A}{x_T}\frac{x_T}{x_T + y_T} + \frac{y_T - y_A}{y_T}\frac{y_T}{x_T + y_T}.$$

设 x_A 具有较大的误差

$$\left|\frac{(x_T + y_T) - (x_A + y_A)}{x_T + y_T}\right| < \left|\frac{x_T - x_A}{x_T}\right|\left|\frac{x_T}{x_T + y_T} + \frac{y_T}{x_T + y_T}\right| = e_r|_{x_A}. \tag{1.39}$$

这就是说和的相对误差限不超过相对各项中最不准确的一项的相对误差限，对于多个数之和，这些讨论也是正确的.

差的相对误差为

$$\frac{(x_T - y_T) - (x_A - y_A)}{x_T - y_T} = \frac{x_T - x_A}{x_T}\frac{x_T}{x_T - y_T} - \frac{y_T - y_A}{y_T}\frac{y_T}{x_T - y_T}.$$

若 $x_A \gg y_A$，则 $\dfrac{y_T}{x_T - y_T}$ 将很小，所以

$$\left|\frac{(x_T-y_T)-(x_A-y_A)}{x_T-y_T}\right|\approx\left|\frac{x_T-x_A}{x_T}\right|=|e_r|_{x_A}. \tag{1.40}$$

就是说二数相差很大时，相对误差取决于大的那个数的相对误差．问题比较严重的发生在两个几乎相等的数字之差上，这时位于前面的大多数相同的数字因相消而变为零，剩下的差值往往丧失了规定的有效位数，从而造成很大的误差．

例 4.1　求解 $x^2-26x+1=0$.

我们可得到这个方程的两个解

$$x_T^{(1)}=13+\sqrt{168},\qquad x_T^{(2)}=13-\sqrt{168}.$$

当我们采用五位数平方根时，$\sqrt{168}=12.961$，于是

$$|\sqrt{168}-12.961|\leqq 0.0005,$$

我们求得

$$x_A^{(1)}=25.961,\qquad x_A^{(2)}=0.039,$$

由此可求得各误差如下：

$$e_a|_{x_A^{(1)}}=e_a|_{x_A^{(2)}}\leqq 0.0005,$$

$$e_r|_{x_A^{(1)}}\leqq\frac{0.0005}{29.9605}\approx 1.9\times 10^{-5},$$

$$e_r|_{x_A^{(1)}}\leqq\frac{0.0005}{x_T^{(2)}}\leqq\frac{0.0005}{0.0385}\approx 1.3\times 10^{-2}.$$

不管 $x_A^{(2)}$ 本身数据如何精确，$e_r|_{x_A^{(2)}}$ 却有一非常大的相对误差，因为在计算$\sqrt{168}$时采用了 12.961，就使得计算 $x_A^{(2)}$ 时将有效数字丧失了．

为了减少因有效数字丧失而引起的误差，我们一般先变直接相减为和的倒数，再进行计算．如上例中的

$$x_T^{(2)}=13-\sqrt{168}=\frac{1}{13+\sqrt{168}}\approx\frac{1}{25.961}\approx 0.03851932\equiv x_A^{(2)},$$

从而有

$$\begin{aligned}|x_T^{(2)}-x_A^{(2)}|&\leqq\left|x_T^{(2)}-\frac{1}{25.961}\right|+\left|\frac{1}{25.961}-0.03851932\right|\\&\leqq 7\times 10^{-7}+5\times 10^{-9}\approx 7\times 10^{-7},\end{aligned}$$

$$e_r|_{x_A^{(2)}}\leqq\frac{7\times 10^{-7}}{x_T^{(2)}}+\frac{5\times 10^{-7}}{x_T^{(2)}}\leqq 1.9\times 10^{-5}+1.3\times 10^{-7}=1.9\times 10^{-5}.$$

用此方法能使相对误差减得很小．

至于相对舍入误差，由(1.32)可得

$$e_s=\left|\frac{x_A\wedge y_A-x_A\wedge^* y_A}{x_T\wedge y_T}\right|\leqq\frac{\alpha^{-n+1}}{2}|1-e_r|, \tag{1.41}$$

其中 e_r 表相对传播误差．于是总相对误差 e_{rT} 为

$$e_{rT}=e_r+e_s=\frac{\alpha^{-n+1}}{2}+\left(1-\frac{\alpha^{-n+1}}{2}\right)e_r. \tag{1.41}$$

把各种具体运算法产生的相对误差 e_r 代入，易于各到诸如乘、除、加、减法的总相对误差．由 (1.42) 可见如果有效位数 n 很大，主要误差为传播误差．当 n 不大时，舍入误差的比重就加大．特别是在利用递推公式作递推运算时，舍入误差的存在可能导致完全错误的结果．我们看如下的例子：

例 4.2　数列 $1, \frac{1}{3}, \frac{1}{9}, \frac{1}{27}, \frac{1}{81}, \cdots$，设 $p_0 = 1, p_1 = \frac{1}{3}$，则其余各项可由如下递推公式产生：

$$p_n = \frac{10}{3} p_{n-1} - p_{n-2},$$

试用有舍入的五位数字进行计算．表 1.1　给出计算出的 p_n 以及正确的 p_n 之值．

表 1.1

n	计算值 p_n	正确值 p_n
0	1.00000	1.00000
1	.33333	.33333
2	.11110	.11111
3	$.37000\times 10^{-1}$	$.37037\times 10^{-1}$
4	$.12230\times 10^{-1}$	$.12346\times 10^{-1}$
5	$.37660\times 10^{-2}$	$.41152\times 10^{-2}$
6	$.32300\times 10^{-3}$	$.13717\times {}^{-2}$
7	$-.26893\times 10^{-2}$	$.45725\times 0^{-3}$
8	$-.92872\times 10^{-2}$	$.15242\times 10^{-3}$

我们看到由于仅使用 5 位数字并采取有舍入的计算方法，误差累计到一定程度就使计算值变得面目全非，甚至使计算值成为负值，就完全失去意义了．

为减少舍入误差造成的影响，常采取的措施是增大有效位，例如在通常的计算机的位数基础上采用双倍或三倍精度，即使 (1.42) 中的 n 充分大．当然采用双精度的缺点是耗费机时太多，同时也不能完全杜绝舍入误差的恶性增长，仅只延缓其产生影响的时间．

二、函数计算中的误差处理问题

(*a*)　将直接相减变为平方相减

例如　$h(x) = f(x_T) - f(x_A)$ 可写为

$$h(x) = [f(x_T) - f(x_A)]\frac{[f(x_T) + f((x_A)]}{[f(x_T) + f(x_A)]} = \frac{f^2(x_T) - f^2(x_A)}{f(x_T) + f(x_A)}. \tag{1.43}$$

这种办法只有在 $f^2(x+\varepsilon) - f^2(x)$ 所丧失的数字少于 $f(x+\varepsilon) - f(x)$ 时才有价值，然而往往还是可以挽救一到两位有效数字．

(*b*)　先对差值的函数作变换，再作数值计算

例如　求 $1 - \cos x$ 在 x 很小时的值时可写为

$$1 - \cos x = 2\sin^2 \frac{x}{2},$$