西安交通大学研究生教育系列教材

数值分析

李乃成　梅立泉　编著

科学出版社

西安交通大学研究生教育系列教材

数值分析

李乃成　梅立泉　编著

科学出版社

北京

内 容 简 介

本书介绍了科学与工程计算中常用的数值计算方法及相关理论. 内容包括解线性方程组的直接法和迭代法、插值法、函数最优逼近、数值微积分、非线性方程 (组) 的迭代解法、矩阵特征值和特征向量的计算、常微分与偏微分方程数值解法等. 其中包含了一些在实际中有重要应用的新方法, 如求解超定方程组的最小二乘法、求解线性方程组的基于伽辽金原理的迭代法、奇异值分解、广义特征值问题的求解方法等. 同时, 对数值计算方法的计算效率、稳定性、收敛性、误差估计、适用范围及优缺点也进行了分析和介绍.

本书可作为高等院校数学系各专业本科生和各类工科专业研究生的教材或教学参考书, 也可供从事科学与工程计算的科研工作者阅读参考.

图书在版编目(CIP)数据

数值分析/李乃成, 梅立泉编著. —北京: 科学出版社, 2011
(西安交通大学研究生教育系列教材)
ISBN 978-7-03-032192-3

Ⅰ. ①数… Ⅱ. ①李… ②梅… Ⅲ. ①数值分析-研究生-教材 Ⅳ. ①O241

中国版本图书馆 CIP 数据核字(2011) 第 174452 号

责任编辑: 赵彦超 徐园园 房 阳 / 责任校对: 何艳萍
责任印制: 钱玉芬 / 封面设计: 王 浩

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

源海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

*

2011 年 9 月第 一 版 开本: B5(720 × 1000)
2011 年 9 月第一次印刷 印张: 21 1/2
印数: 1—2 500 字数: 421 000

定价: 65.00 元

(如有印装质量问题, 我社负责调换)

前　言

在现代科学研究与工程设计中, 计算机已成为不可或缺的有力工具. 科学计算与理论、实验三足鼎立, 成为科学实践的三大手段. 科教兴国, 要求现代理工科大学生掌握计算机常用的数值计算方法, 具有使用计算机解决实际问题的能力. “数值计算方法”已成为各高等学校本科生、研究生的必修课.

本书是为理工科大学各专业研究生 72 学时“数值计算方法”课程编写的教材. 以介绍常用、实用的数值计算方法为主, 在介绍经典计算方法的基础上, 本着推陈出新、与时俱进的精神, 增加了一些在实际中有着重要应用的、近年来发展的新方法. 如超定方程组的求解方法、求解线性方程组的基于伽辽金原理的迭代法、奇异值分解、广义特征值问题的求解方法等.

本书重点突出、层次分明、推导详细、清晰易懂, 内容由浅入深、循序渐进, 便于自学与教学, 可供从事科学与工程计算的科技工作者参考, 也适宜于相关专业人员自学.

承蒙邓建中教授在百忙中审阅本书, 并提出了不少宝贵的意见和建议. 本书的出版得到西安交通大学研究生院的支持、鼓励和资助. 科学出版社的有关同志对本书的出版给予了极大的帮助. 编者在此向他们表示衷心的感谢.

由于编者水平所限, 书中不妥之处在所难免, 恳请读者批评指正.

编　者

2010 年 10 月

目　录

第1章 绪 论

本章简要介绍数值分析研究的内容与特点、误差的来源与分类、数据误差的影响、计算机浮点数集和数值算法的稳定性等数值分析的基本概念.

1.1 数值分析研究的内容与特点

随着科学技术的发展和计算机的广泛应用, 解决科学研究和工程技术中的问题通常采用定量分析方法, 其过程分为两个阶段. 第一阶段, 利用有关科学知识和数学理论把实际问题提炼为数学问题 (即建立数学模型). 这一阶段属于应用数学范畴. 所提炼的数学问题一般比较复杂, 无法求出准确解, 只好采用数值计算方法求其近似解. 第二阶段, 针对数学问题提出用计算机求解的数值计算方法, 编写算法程序, 上机求解并分析计算结果. 这一阶段的工作属于计算数学范畴. 数值分析是计算数学的一个主要组成部分, 它研究用计算机求解各类数学问题的数值计算方法及其相关理论. 其内容包括线性方程组的求解、插值法、函数最优逼近法、数值积分与微分、非线性方程 (组) 的求解、矩阵特征值问题、常微分方程与偏微分方程的数值解法等.

可用于计算机求解的可行且有效的数值计算方法应具备以下特点:

(1) 可执行性. 设计算法要面向计算机, 计算机只能作加、减、乘、除和逻辑运算, 设计的算法中只能包含加、减、乘、除四则运算和逻辑运算.

(2) 理论可靠性. 算法要有可靠的理论分析, 即设计的算法要保证其收敛性、稳定性, 并给出其误差分析.

(3) 计算复杂性. 算法要有良好的计算复杂性, 计算复杂性包括时间复杂性 (算法的运算次数, 运算次数决定计算机的计算时间) 和空间复杂性 (占用计算机存储空间的大小).

最后还需通过数值试验验证算法的有效性.

由于数值算法的复杂性, 有些算法虽然在理论上不够严密, 但通过大量的数值试验和对比分析, 证明是行之有效的方法, 在实际中有着广泛的应用, 人们在应用中逐步完善它的理论基础.

鉴于数值计算方法的特点, 学习“数值分析”课程, 首先应掌握设计算法的思想和原理, 重视误差分析、收敛性、稳定性等理论基础; 其次要重视实践, 需要做

一些习题和在计算机上的数值试验, 掌握数值计算方法的计算过程, 加深对算法的理解.

1.2 误　差

1.2.1 误差的来源与分类

科学与工程计算中的数通常是近似数, 近似值与真正值之差称为误差. 误差的来源或分类有 4 种. 从实际问题提炼数学问题时, 由于实际问题比较复杂, 为使问题简化, 往往忽略了许多次要因素, 所提炼的数学问题是对实际问题的近似描述. 即使数学问题能求出准确解, 也与实际问题的真正解不同, 二者之差称为**模型误差**. 数学问题中含有若干参量, 如质量、温度、电压等, 它们的值是通过观测得到的, 难免存在着误差, 这种误差称为**观测误差**、**数据误差**或**参量误差**. 数学问题一般都比较复杂, 难以求其准确解. 通常用近似公式代替准确公式使之容易求解, 由此求出问题的近似解. 准确解与近似解之差称为**截断误差**或**方法误差**. 例如, 为计算函数值 $f(x)$, 取它的 n 次泰勒多项式

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

近似代替 $f(x)$, 即取 $f(x) \approx p_n(x)$, 则截断误差

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}x^n,$$

其中 ξ 介于 0 与 x 之间.

求解数学问题时, 由于计算机的数位数有限, 一般对数要进行舍入, 由此产生的误差称为**舍入误差**或**计算误差**.

"数值分析"课程只研究截断误差和舍入误差的估计、在计算过程中的传播及控制和对计算结果的影响.

1.2.2 绝对误差、相对误差与准确数字

定义 1.2.1　*设 x 为准确值, $\tilde{x}$ 是 x 的一个近似值, 则*

$$\Delta x = x - \tilde{x} \quad 或 \quad |\Delta x| = |x - \tilde{x}|$$

*称为近似值 $\tilde{x}$ 的*绝对误差.

通常, 准确值是不知道的, 因而绝对误差也是不知道的, 但可以估计出绝对误差的一个上界 ε, 即 $|\Delta x| \leqslant \varepsilon$. ε 称为**绝对误差界**或**绝对误差限**. 绝对误差或绝对误差界简称为误差.

虽然不知道准确值, 但由绝对误差界可知准确值所在的范围, 即 $\tilde{x}-\varepsilon \leqslant x \leqslant \tilde{x}+\varepsilon$, 并记

$$x=\tilde{x} \pm \varepsilon.$$

例如, 用千分尺测量工件的直径 d 时, 只能测得工件直径的近似值 $\tilde{d}$, 此时绝对误差

$$|\Delta d|=|d-\tilde{d}| \leqslant \frac{1}{2} \times 10^{-2} \text{ mm}.$$

所测工件的直径的范围为

$$\tilde{d}-\frac{1}{2} \times 10^{-2} \leqslant d \leqslant \tilde{d}+\frac{1}{2} \times 10^{-2}.$$

对于绝对值较大或较小的近似值, 绝对误差不足以刻画近似数的准确程度, 于是引进相对误差, 用相对误差的大小来衡量其准确程度.

定义 1.2.2　*设 x 为准确值, $\tilde{x}$ 是 x 的一个近似值, 则*

$$\delta x=\frac{x-\tilde{x}}{x} \quad \text{或} \quad |\delta x|$$

*称为近似值 $\tilde{x}$ 的*相对误差.

在实际计算中, 准确值 x 是不知道的, 但由于

$$\frac{x-\tilde{x}}{\tilde{x}}-\frac{x-\tilde{x}}{x}=\Delta x \frac{x-\tilde{x}}{x \tilde{x}}=\frac{(\Delta x)^2}{\tilde{x}(\tilde{x}+\Delta x)}=\frac{(\Delta x/\tilde{x})^2}{1+\Delta x/\tilde{x}},$$

两者之差是 $\Delta x/\tilde{x}$ 的平方级, 当 $|\Delta x|/|\tilde{x}|$ 较小时, 可以忽略不计. 这时, 相对误差可取为

$$|\delta x|=\frac{|x-\tilde{x}|}{|\tilde{x}|}.$$

若 $|\delta x| \leqslant \varepsilon_r$, 则 ε_r 称为**相对误差界或相对误差限**, 常用百分数表示, 并记为

$$x=\tilde{x}(1 \pm \varepsilon_r).$$

由上述定义可知, 绝对误差和绝对误差界是有量纲量, 而相对误差和相对误差界是无量纲量.

近似数 $\tilde{x}$ 的准确数字与误差界有关. 设

$$x=\pm x_1 x_2 \cdots x_m . x_{m+1} x_{m+2} \cdots x_{m+n} x_{m+n+1} \cdots,$$

若 x 的近似值 $\tilde{x}$ 取到小数后第 n 位, 则

$$\tilde{x}=\pm x_1 x_2 \cdots x_m . x_{m+1} \cdots \tilde{x}_{m+n},$$

其中最后一位

$$\tilde{x}_{m+n}=\begin{cases} x_{m+n}, & x_{m+n+1}\leqslant 4,\\ x_{m+n}+1, & x_{m+n+1}\geqslant 5.\end{cases}$$

这时, $|x-\tilde{x}|\leqslant 0.\underbrace{00\cdots0}_{n个}5=0.5\times10^{-n}=\dfrac{1}{2}\times10^{-n}$, 则称近似值 $\tilde{x}$ 准确到 n 位小数, 并将从该位起直到最左端的非零数字之间的一切数字称为近似值 $\tilde{x}$ 的**准确数字 (有效数字)**, 它的多少反映了 $\tilde{x}$ 近似 x 的准确程度.

有了准确数字的概念后, 数 1.23 和 1.230 是有区别的. 前者准确到两位小数, 后者准确到三位小数, 所以不能随意给数的末尾添零.

又如, $\pi=3.141\,592\,65\cdots$, 若取 π 的近似值分别为

$$\pi_1=3.14,\quad \pi_2=3.142,$$

则

$$|\pi-\pi_1|=0.001\,592\,65\cdots<0.005=0.5\times10^{-2}=\frac{1}{2}\times10^{-2},$$

$$|\pi-\pi_2|=0.000\,407\,346\cdots<0.000\,5=0.5\times10^{-3}=\frac{1}{2}\times10^{-3}.$$

由此可知近似值 π_1 准确到两位小数, 具有三位准确数字; 近似值 π_2 准确到三位小数, 具有四位准确数字.

若把 x 写成标浮点形式

$$x=\pm10^{m}\times0.x_1x_2\cdots,\quad x_1\neq0,$$

当 x 的近似值 $\tilde{x}$ 的绝对误差界为 $\dfrac{1}{2}\times10^{-n}$ 时, $\tilde{x}$ 准确到 n 位小数, 具有 $n+m$ 位准确数字.

若已知 $\tilde{x}$ 的相对误差界为 $\dfrac{1}{2}\times10^{-t}$, 即

$$\frac{|x-\tilde{x}|}{|\tilde{x}|}\leqslant\frac{1}{2}\times10^{-t},$$

则 $\tilde{x}$ 的绝对误差为

$$|x-\tilde{x}|\leqslant|\tilde{x}|\times\frac{1}{2}\times10^{-t}=10^{m}\times0.x_1x_2\cdots\times\frac{1}{2}\times10^{-t}\leqslant\frac{1}{2}\times10^{-(t-m)}.$$

由此可知, $\tilde{x}$ 至少具有 $t-m+m=t$ 位准确数字.

1.2.3　计算机中数的表示与舍入误差

计算机中的实数采用浮点表示法, 即将一个数分为指数和尾数两部分来表示. 设计算机采用 β 进制 (二进制、八进制、十六进制或十进制), 字长为 t 位. 按舍入原则, 将非零实数 x 在计算机上表示为

$$fl(x)=\tilde{x}=\pm\left\{\frac{x_1}{\beta}+\frac{x_2}{\beta^2}+\cdots+\frac{x_t}{\beta^t}\right\}\times\beta^l=\pm 0\,.\,x_1x_2\cdots x_t\times\beta^l,$$

其中, $x_1\in\{1,2,\cdots,\beta-1\}$, $x_i\in\{0,1,\cdots,\beta-1\}$ $(i=2,3,\cdots,t)$.

$0\,.\,x_1x_2\cdots x_t$ 称为尾数部分, β^l 称为指数部分. 指数 l 是整数, 也称为阶码. 阶码 l 的取值范围为 $L\leqslant l\leqslant U(L<0,U>0)$. L,U 分别称为指数的下、上界. $fl(x)$ 称为**规格化浮点数**. 计算机所能表示的全部浮点数的集合称为计算机的**浮点数集**, 记为

$$F(\beta,t,L,U)=\{0\}\bigcup\{fl(x)=\pm 0\,.\,x_1x_2\cdots x_t\times\beta^l\}.$$

显然, 浮点数集 $F(\beta,t,L,U)$ 中共有 $2(\beta-1)\beta^{t-1}(U-L+1)+1$ 个数, 它所能表示的数的范围为

$$fl_{\min}(x)=\beta^{L-1}\leqslant|\,fl(x)\,|\leqslant\beta^U\times(1-\beta^{-t})=fl_{\max}(x).$$

由此可知, 当 $|x|>fl_{\max}(x)$ 时, 计算机无法表示 (这种情况称为**"上溢"**), 计算机中断运行; 当 $x\neq 0$ 但 $|x|<fl_{\min}(x)$ 时, 计算机令 $fl(x)=0$ (这种情况称为**"下溢"**), 这时计算机虽然继续运算, 但可能会得出难以预料的结果. 因此, **在计算过程中应尽量调整计算顺序, 避免出现"上溢"、"下溢"的情况**. 由此还可知, 计算机字长有限, 一个实数 x 一般只能用最接近的浮点数 $fl(x)$ 代替它, 如

$$x=\pm 0\,.\,x_1x_2\cdots x_tx_{t+1}\cdots\times\beta^l,\quad fl(x)=\pm 0\,.\,x_1x_2\cdots\tilde{x_t}\times\beta^l,$$

$fl(x)$ 尾数部分最末位可能有半个单位的误差, 由此将产生舍入误差. 浮点数 $fl(x)$ 的绝对误差与相对误差分别为

$$|x-fl(x)|\leqslant\frac{1}{2}\beta^{-t}\times\beta^l=\frac{1}{2}\beta^{l-t},$$

$$\frac{|x-fl(x)|}{|\,x\,|}\leqslant\frac{\frac{1}{2}\beta^{l-t}}{\beta^{l-1}}=\frac{1}{2}\beta^{-(t-1)}.$$

相对误差限 $\frac{1}{2}\beta^{-(t-1)}$ 只与计算机数的进制和字长有关, 称为**计算机的相对精度**.

在计算机的每步运算中, 得到的数都可能有舍入误差, 如两个数加、减运算, 计算机先对阶, 使得阶小的数与阶大的数的阶码相同, 然后尾数相加减, 最后对计算结果规格化. 例如, 在浮点数集 $F(10,4,-10,10)$ 中, 两个数 $a=0.123\,4\times10^3$ 与 $b=0.456\,7\times10^{-1}$ 相加,

$$\begin{aligned}a+b&=0.123\,4\times10^3+0.456\,7\times10^{-1}=0.123\,4\times10^3+0.000\,045\,67\times10^3\\&=0.123\,4\times10^3+0.000\,0\times10^3=0.123\,4\times10^3=a,\end{aligned}$$

得到了 $a+b\ (b\neq 0)=a$ 的结果, 即大数 a“吃掉”了小数 b.

一般地, 为避免大数“吃掉”小数, **若干个数相加时, 宜采取绝对值较小的数先加的原则**.

为了减少舍入误差和节省计算时间, **在计算中应尽量简化计算步骤, 减少运算次数**. 下面以两个例子予以说明.

例如, 计算多项式 $p_n(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n$ 的值, 若先计算各项 a_kx^k, 再逐项相加, 则需进行 $1+2+3+\cdots+n=\dfrac{n(n+1)}{2}$ 次乘法和 n 次加法; 若用秦九韶算法①, 将 $p_n(x)$ 改写为

$$p_n(x)=(\cdots((a_nx+a_{n-1})x+a_{n-2})x+\cdots+a_1)x+a_0,$$

则只需 n 次乘法和 n 次加法.

又如, 计算 ABx, 其中, A,B 为 n 阶矩阵, x 为 n 维向量. 若按 $(AB)x$ 的次序计算, 则算出 AB 的 n^2 个数各需 n 次乘法和 $n-1$ 次加法; 再算出 $(AB)x$ 的 n 个数各需 n 次乘法和 $n-1$ 次加法, 因此, 共需 $n^2\cdot n+n\cdot n=n^3+n^2$ 次乘法和 $n^2\cdot(n-1)+n\cdot(n-1)=n^3-n$ 次加法. 同理, 若按 $A(Bx)$ 的次序计算, 则只需 $2n^2$ 次乘法和 $2n^2-2n$ 次加法.

1.2.4　数据误差影响的估计

前面已述数学问题中的参数 $x_1,x_2,\cdots,x_n$ 存在着数据误差, 设其近似值为 $\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n$. 数学问题的解与这些参数有关, 记为

$$y=\phi(x_1,x_2,\cdots,x_n).$$

设其近似解为

$$\tilde{y}=\phi(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n).$$

由泰勒展开式知, 近似解 $\tilde{y}$ 的绝对误差和相对误差分别为

$$\begin{aligned}\Delta y=&y-\tilde{y}=\phi(x_1,x_2,\cdots,x_n)-\phi(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n)\\ \approx&\sum_{i=1}^{n}\frac{\partial\phi(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n)}{\partial x_i}(x_i-\tilde{x}_i)=\sum_{i=1}^{n}\frac{\partial\phi(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n)}{\partial x_i}\Delta x_i,\end{aligned}$$

$$\delta y=\frac{\Delta y}{y}\approx\sum_{i=1}^{n}\frac{\partial\phi(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n)}{\partial x_i}\frac{x_i}{y}\frac{\Delta x_i}{x_i}=\sum_{i=1}^{n}\frac{\partial\phi(\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n)}{\partial x_i}\frac{x_i}{y}\delta x_i.$$

① 秦九韶算法也称为霍纳 (Horner) 算法, 秦九韶于 1247 年提出此法, 比霍纳 1819 年提出此算法早 500 多年.

以上两式中系数的绝对值 $\left|\dfrac{\partial\phi}{\partial x_i}\right|$ 和 $\left|\dfrac{\partial\phi}{\partial x_i}\dfrac{x_i}{y}\right|$ 是数学问题的近似解 $\tilde{y}$ 的绝对误差和相对误差关于参数 x_i 绝对误差和相对误差的放大或缩小倍数. 若其值较大, 则 $\tilde{x}_i$ 的小误差就会导致近似解 $\tilde{y}$ 的大误差. $\left|\dfrac{\partial\phi}{\partial x_i}\right|$ 和 $\left|\dfrac{\partial\phi}{\partial x_i}\dfrac{x_i}{y}\right|$ 称为数学问题的**条件数**, 其值很大的问题称为**病态问题**. 换句话说, **病态问题就是数据发生微小的变化将引起其解发生剧烈变化的问题**. 这是数学问题本身的固有特征. 对病态问题必须采取相应的处理措施和特殊的计算方法, 以减少误差影响.

注意到当 $y=\phi(x_1,x_2,\cdots,x_n)\approx 0$ 时, 相对误差的条件数 $\left|\dfrac{\partial\phi}{\partial x_i}\dfrac{x_i}{y}\right|$ 很大. 因此, **凡是计算结果接近于零的问题往往是病态问题**. 例如, 相近的两个数相减, 因差接近于零, 就是病态问题. 因此, 在设计算法时, 应尽量避免两个相近的数相减. 常用的做法是变换算式. 这是因为**数学上等价的公式在数值计算时并不总是等效的.** 例如, 当 $|x|$ 的绝对值充分大时,

$$\sqrt{x+1}-\sqrt{x}=\frac{1}{\sqrt{x+1}+\sqrt{x}},\quad \frac{1}{x}-\frac{1}{x+1}=\frac{1}{x(x+1)},$$

$$\ln\left(x+\frac{1}{x}\right)-\ln\left(x-\frac{1}{x}\right)=\ln\left(\frac{x^2+1}{x^2-1}\right).$$

当 $|x|$ 的绝对值充分小时,

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

$$\sin x-x=-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots.$$

又如, 二次方程 $ax^2+bx+c=0$ 的求根公式为

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a},$$

当 $b^2\gg 4\mid ac\mid$ 时, $b\approx\sqrt{b^2-4ac}$, 在求根公式中总有一式存在相近数相减的情况, 使得计算结果不可靠. 为避免两个相近数相减, 在计算机上求解宜取

$$x_1=\frac{-b-\operatorname{sgn}(b)\sqrt{b^2-4ac}}{2a},$$

其中 $\operatorname{sgn}(x)$ 为符号函数, 即

$$\operatorname{sgn}(x)=\begin{cases}1, & x\geqslant 0,\\ -1, & x<0.\end{cases}$$

由韦达定理取 $x_2=\dfrac{c}{ax_1}$.