21世纪高等学校教材

孙志忠 袁慰平 闻震初 编著

数值分析

A PRACTICAL NUMERICAL ANALYSIS

（第3版）

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

数值分析

（第3版）

孙志忠　袁慰平　闻震初　编著

东南大学出版社

内 容 提 要

本书着重介绍适合于电子计算机上采用的数值计算方法及其理论，内容包括误差分析、非线性方程求根、线性代数方程组数值解法、多项式插值与函数逼近、数值积分与数值微分、常微分方程数值解法、偏微分方程数值解法等。本书内容覆盖了教育部工科研究生数学课程教学指导小组所制订的工科硕士生数值分析课程教学基本要求，同时还增加了一些工科专业所需要的内容，如机器数系、有理函数插值、振荡函数积分等。书中对各种计算方法的构造思想都作了较详细的阐述，对稳定性、收敛性、误差估计以及算法的优缺点等也作了适当的讨论。本书还挑选了部分东南大学工科研究生结合各自专业自选课题的计算实习，以此作为本书各章的应用实例。

本书可作为各类工科专业研究生和数学系各专业本科生的教材或教学参考书，也可供从事科学与工程计算的科技工作者阅读参考。

图书在版编目(CIP)数据

数值分析/孙志忠，袁慰平，闻震初编著. —3版.
—南京：东南大学出版社，2010.12
ISBN 978-7-5641-2577-6

Ⅰ.①数… Ⅱ.①孙…②袁…③闻… Ⅲ.①数值计算 Ⅳ.O241

中国版本图书馆CIP数据核字(2010)第261686号

东南大学出版社出版发行
(南京四牌楼2号 邮编210096)
出版人：江建中
全国各地新华书店经销 南京玉河印刷厂印刷
开本：700mm×1000mm 1/16 印张：25.75 字数：525千字
2011年2月第3版 2011年2月第1次印刷
ISBN 978-7-5641-2577-6
定价：43.80元
(凡有印装质量问题，可直接与读者服务部联系。电话：025—83792328)

前　言

随着计算机的广泛使用与科学技术的迅速发展，使用计算机进行科学计算已成为科学研究、工程设计中越来越不可缺少的一个环节，它有时甚至代替或超过了实验所起的作用。因此，科学计算应该成为高级科技人员的一项基本功。为此，作为科学计算的核心——“数值分析”已被较多的工科专业列为硕士研究生的一门学位课程。

我校自 1981 年开设此课程以来，进行了专业需要的调查以及教学内容和教学方法的改革与研究，并在教学实践的基础上形成了本书的初稿。初稿在本校使用过多遍，得到研究生及其指导教师们的良好反应与支持。在这次出版时，再一次总结教学实践中的经验，作了修改和加工。

本书着重介绍了适合于电子计算机上采用的计算方法的构造及使用，对误差估计、方法的收敛性、稳定性、适用范围及优缺点等作了适当分析，对一些解法作了比较详细的推导并列举了较多的数值计算实例。其内容覆盖了国家教委工科研究生数学课程教学指导小组所制订的工科硕士研究生数值分析课程教学基本要求，同时还增加了一些工科专业所需要的内容，如误差分析中的机器数系；非线性方程求根的 Sturm 定理；插值与逼近中的重节点插值、有理函数插值及最佳一致逼近；数值积分中的振荡函数的积分、重积分的近似计算；常微分方程中的自适应算法及稳定性较好的单步隐格式；特征值计算中的广义特征值计算等等。增加的这些内容，若教学课时数较少则可以删去，不会影响教学的连续性；但若增加了这些内容，则在内容的深广度上更能适应工科专业发展的需要。因此，本书兼顾了多学时与少学时的要求。

本书每章末（除第 1、7 两章外）都有应用实例一节，取材于历届学生结合各自专业的自选课题。我们认为研究生学习“数值分析”课，一方面是提高数学素质，另一方面是为应用所学知识进行科学研究。于是我们要求学生在学完本课程后，必须结合所学内容及各自的专业自选研究课题做一次大型计算实习。这次，在讲义修改出版过程中，我们从历届学生做的大型计算实习中挑选了部分作为实例进入教材，有的实例尽管被我们作了简化和加工，但仍然可以从这些实例中初步看到本课程在各专业应用中的深广度。

本书的读者对象是工科研究生及从事数值计算的科技工作者。本书编写分工如下：第 1、3、5 章由袁慰平编写，第 2、6 章由黄新芹编写，第 4 章由闻震初编写，第 7、8 章由张令敏编写。

南京大学苏煜城教授仔细审阅了全书，并提出了宝贵的意见，在苏先生的指点下，我们作了修改。在本书的形成、使用、修改和出版过程中，得到东南大学研究生院和东南大学出版社及有关工科研究生指导教师和本教研室同志的关心、支持和帮助，在此一并表示衷心感谢。

由于作者水平有限，缺点与错误在所难免，恳请读者批评指正。

编　者

1992 年 4 月

目　录

1 绪 论

1.1 数值分析的对象和特点

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。当今世界计算机已被广泛使用，因此数值分析所研究的应该是适合于计算机上使用的计算方法及其误差分析和收敛性、稳定性问题。

使用计算机通过计算方法或数值模拟的手段去解决科学或工程中的关键问题，简称为科学计算。它已成为科学研究、工程设计等越来越不可缺少的一个环节，有时甚至代替或超过了实验所起的作用。

最近半个世纪科学研究的实践使人们越来越清楚地认识到，当代科学研究方法论应该由实验、科学计算及理论三大环节所组成。也就是说，科学计算已成为一种新的科学研究方法。因此，作为科学计算的主体 —— 数值分析也就越来越被人们所重视。

1.2 误差的基本概念

1.2.1 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，它们之差称为误差。引起误差的原因是多方面的。

(1) 模型误差 将实际问题转化为数学问题即所谓的建立数学模型时，对被描述的实际问题进行了抽象和简化，因此数学模型只是客观现象的一种近似、粗糙的描述。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) 观测误差 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，而观测不可避免地会带有误差。这种误差称为观测误差。

(3) 截断误差 在计算中常常遇到只有通过无限过程才能得到最终结果，但实际计算时只能采用有限过程，如无穷级数求和，只能取前面有限项之和来近似代替，于是产生了有限过程代替无限过程的误差，称为截断误差。这是计算方法本身所带来的误差，所以也称为方法误差。

(4) 舍入误差 在计算中遇到的数据可能位数很多，也可能是无穷小数，如 $\sqrt{2}=1.414\,213\,56\cdots$，$\frac{1}{3}=0.333\,3\cdots$，但计算时只能对有限位进行运算，一般采用

四舍五入的办法。电子计算机计算时也有采用截尾的办法，如$\sqrt{2}$在8位字长的截断机里取成1.414 213 5。这类误差称为舍入误差，也称计算误差。

少量的舍入误差是微不足道的，但在电子计算机上作了成千上万次运算后，舍入误差的积累有时可能是十分惊人的。

由上述误差来源的分析可以得到如下结论：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。对于实际问题，在建立数学模型时本身已存在着模型误差和观测误差，因此既然描述问题的办法是近似的，那么要求解的绝对准确也就没有多大意义了。于是我们在计算方法里所讨论的求解都只要求出近似解，那种认为近似解不可靠、不准确的看法是错误的，应该认为求近似解是正常的，需要研究的问题是尽量设法减少误差，提高精度。从以上误差的四种来源的分析可以知道前两种误差是客观存在的，后两种误差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。

1.2.2 绝对误差

定义1.2.1 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**。

注意这样定义的误差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^*，也不可能算出误差的准确值，因此这个值虽然客观存在，但在实际计算中是很难得到的，而得到的往往是误差的某个范围，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

$$|e|=|x^*-x|\leqslant\varepsilon$$

称ε为近似值x的**绝对误差限**，简称**误差限**。有时也表示成$x^*=x\pm\varepsilon$。

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$x=20$ mm，由于直尺以毫米为刻度，所以其误差不超过0.5 mm，即

$$|x^*-20|\leqslant 0.5$$

从这个不等式我们不能得出准确值x^*，但却知道x^*的范围为

$$19.5\leqslant x^*\leqslant 20.5$$

对于给定的正数ε，若近似值x满足

$$|x^*-x|\leqslant\varepsilon$$

则在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。

1.2.3 相对误差

对于不同的物理量，绝对误差限的大小不能完全表示出近似值的精确程度。为

了更好地反映近似值的精确程度，必须考虑绝对误差与真值之比。

定义 1.2.2 设 x^* 为准确值，x 是 x^* 的一个近似值。称

$$e_r = \frac{x^* - x}{x^*}$$

为近似值 x 的**相对误差**。

在实际计算中，通常真值 x^* 总是难以求得的。人们常以

$$\bar{e}_r = \frac{x^* - x}{x}$$

作为相对误差。事实上，有

$$\bar{e}_r - e_r = \frac{\bar{e}_r^2}{1+\bar{e}_r} = \frac{e_r^2}{1-e_r}$$

因而当 $\bar{e}_r$ 和 e_r 有一为小量时，$\bar{e}_r - e_r$ 为该小量的 2 阶小量。

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

$$|e_r| \leqslant \varepsilon_r \quad 或 \quad |\bar{e}_r| \leqslant \varepsilon_r$$

则称 ε_r 为 x 的**相对误差限**。

绝对误差只能用来比较对同一个量所测得的不同近似值的准确程度，而相对误差却能用来刻画或比较任何近似值的准确程度。

1.2.4 有效数

工程技术中对于测量得到的数经常表示成 $x \pm \varepsilon$，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就太麻烦了。因此希望所表示的数本身就能显示出它的准确程度，于是需要引进有效数的概念。

定义 1.2.3 如果近似值 x 的误差限是其某一位上的半个单位，且该位直到 x 的第一个非零数字一共有 n 位（如图 1.2.1 所示），则称近似值 x 具有 n 位有效数字，用这 n 位有效数字表示的近似数称为有效数。

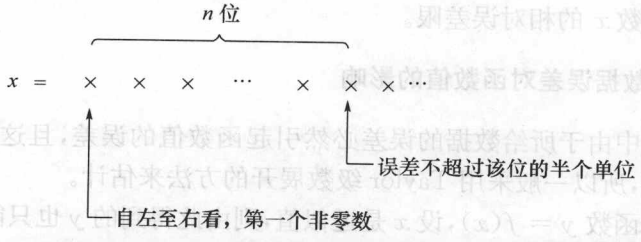

图 1.2.1 有效位数

如 π 的近似值取 $x_1 = 3.14$，则 x_1 有 3 位有效数字；取 $x_2 = 3.1416$，则 x_2 有 5 位有效数字；取 $x_3 = 3.1415$，则 x_3 只有 4 位有效数字。

在讲了有效数字之后，我们从此规定所写出的数都应该是有效数字，如 π 的近似值应是 3.14 或 3.142 或 3.1416，不能是 3.1415。

在科学记数法中通常将 n 位有效数字 x 表示成

$$x = \pm 0.\alpha_1\alpha_2\cdots\alpha_n \times 10^m$$

即

$$x = \pm(\alpha_1 \times 10^{-1} + \alpha_2 \times 10^{-2} + \cdots + \alpha_n \times 10^{-n}) \times 10^m \tag{1.2.1}$$

其中 m 为一整数，$\alpha_1, \alpha_2, \cdots, \alpha_n$ 都是 0 到 9 中的整数，且 $\alpha_1 \neq 0$。

按式(1.2.1)表示的有效数 x，其误差为

$$|x^* - x| \leqslant \frac{1}{2} \times 10^{m-n} \tag{1.2.2}$$

所以 x 的误差限为 $\varepsilon = \frac{1}{2} \times 10^{m-n}$，因此在 m 相同的情况下，n 越大则误差越小，亦即说明一个近似值的有效位数越多其误差限越小。由

$$\frac{|x - x^*|}{|x|} \leqslant \frac{\frac{1}{2} \times 10^{m-n}}{\alpha_1 \times 10^{-1} \times 10^m} = \frac{1}{2\alpha_1} \times 10^{-n+1}$$

知 x 的相对误差限为

$$\varepsilon_r = \frac{1}{2\alpha_1} \times 10^{-n+1} \tag{1.2.3}$$

式(1.2.3)表明一个近似值的有效位数越多，其相对误差限也越小。

由于 $\frac{1}{\alpha_1} \leqslant 1$，所以有时也简单地取

$$\varepsilon_r = \frac{1}{2} \times 10^{-n+1} \tag{1.2.4}$$

作为 n 位有效数 x 的相对误差限。

1.2.5 数据误差对函数值的影响

数值运算中由于所给数据的误差必然引起函数值的误差，且这种数据误差的影响较为复杂，所以一般采用 Taylor 级数展开的方法来估计。

对于一元函数 $y = f(x)$，设 x 是近似值，则由此得到的 y 也只能是近似值。我们来研究 y 的绝对误差和相对误差。设 x^* 是准确值，相应的函数 y 的准确值 $y^* = f(x^*)$，则函数值 y 的绝对误差为

$$e(y) = y^* - y = f(x^*) - f(x)$$

将 $f(x^*)$ 在 x 处作 Taylor 展开，并取 1 阶 Taylor 多项式，得 $e(y)$ 的近似表达式

$$e(y) \approx f'(x)(x^* - x) = f'(x)e(x)$$

式中 $e(x) = x^* - x$；函数值 y 的相对误差

$$e_r(y) = \frac{e(y)}{y} \approx \frac{f'(x)e(x)}{y} = \frac{xf'(x)}{f(x)}e_r(x)$$

对于二元函数 $y = f(x_1, x_2)$，设 x_1, x_2 是近似值，由此计算 y 得到的也只能是近似值。设 x_1^*, x_2^* 是准确值，其函数准确值为 $y^* = f(x_1^*, x_2^*)$。于是函数值 y 的绝对误差为

$$e(y) = y^* - y = f(x_1^*, x_2^*) - f(x_1, x_2)$$

将 $f(x_1^*, x_2^*)$ 在 (x_1, x_2) 处作 Taylor 展开，并取 1 阶 Taylor 多项式，得 $e(y)$ 的近似表达式

$$e(y) \approx \frac{\partial f(x_1, x_2)}{\partial x_1}(x_1^* - x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2}(x_2^* - x_2)$$

或

$$e(y) \approx \frac{\partial f(x_1, x_2)}{\partial x_1}e(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2}e(x_2) \tag{1.2.5}$$

式中

$$e(x_1) = x_1^* - x_1, \quad e(x_2) = x_2^* - x_2$$

由此可得函数值 y 的相对误差

$$e_r(y) = \frac{e(y)}{y} \approx \frac{\partial f(x_1, x_2)}{\partial x_1}\frac{1}{y}e(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2}\frac{1}{y}e(x_2)$$

或

$$e_r(y) \approx \frac{\partial f(x_1, x_2)}{\partial x_1}\frac{x_1}{f(x_1, x_2)}e_r(x_1) + \frac{\partial f(x_1, x_2)}{\partial x_2}\frac{x_2}{f(x_1, x_2)}e_r(x_2) \tag{1.2.6}$$

式中

$$e_r(x_1) = \frac{x_1^* - x_1}{x_1}, \quad e_r(x_2) = \frac{x_2^* - x_2}{x_2}$$

利用函数值的误差估计式(1.2.5) 和式(1.2.6) 可以得到两数和、差、积、商的误差估计：

$$e(x_1+x_2)\approx e(x_1)+e(x_2) \tag{1.2.7}$$

$$e(x_1-x_2)\approx e(x_1)-e(x_2) \tag{1.2.8}$$

$$e(x_1x_2)\approx x_2e(x_1)+x_1e(x_2) \tag{1.2.9}$$

$$e\left(\frac{x_1}{x_2}\right)\approx\frac{1}{x_2}e(x_1)-\frac{x_1}{x_2^2}e(x_2)\quad(x_2\neq 0) \tag{1.2.10}$$

$$e_r(x_1+x_2)\approx\frac{x_1}{x_1+x_2}e_r(x_1)+\frac{x_2}{x_1+x_2}e_r(x_2) \tag{1.2.11}$$

$$e_r(x_1-x_2)\approx\frac{x_1}{x_1-x_2}e_r(x_1)-\frac{x_2}{x_1-x_2}e_r(x_2) \tag{1.2.12}$$

$$e_r(x_1x_2)\approx e_r(x_1)+e_r(x_2) \tag{1.2.13}$$

$$e_r\left(\frac{x_1}{x_2}\right)\approx e_r(x_1)-e_r(x_2)\quad(x_2\neq 0) \tag{1.2.14}$$

例 1.2.1 计算$\sqrt{2\,001}-\sqrt{1\,999}$,并分析计算结果具有几位有效数字。

解 记$x_1^*=\sqrt{2\,001}$,$x_2^*=\sqrt{1\,999}$,则它们的6位有效数为$x_1=44.732\,5$,$x_2=44.710\,2$。

第一种方法:$x_1^*-x_2^*\approx x_1-x_2=44.732\,5-44.710\,2=0.022\,3$

第二种方法:$x_1^*-x_2^*=\dfrac{2}{x_1^*+x_2^*}\approx\dfrac{2}{x_1+x_2}=\dfrac{2}{44.732\,5+44.710\,2}$

$=0.022\,360\,684\,5\cdots\approx 0.022\,360\,7$

现在来分析上述两种方法所得结果各具有几位有效数字。由

$$\begin{aligned}|e(x_1-x_2)|&\approx|e(x_1)-e(x_2)|\leqslant|e(x_1)|+|e(x_2)|\\&\leqslant\frac{1}{2}\times10^{-4}+\frac{1}{2}\times10^{-4}=10^{-4}<\frac{1}{2}\times10^{-3}\end{aligned}$$

知按第一种方法所得结果(至少)具有2位有效数字。

由

$$\begin{aligned}\left|e\left(\frac{2}{x_1+x_2}\right)\right|&\approx\left|-\frac{2}{(x_1+x_2)^2}e(x_1+x_2)\right|\\&\approx\left|-\frac{2}{(x_1+x_2)^2}[e(x_1)+e(x_2)]\right|\\&\leqslant\frac{2}{(x_1+x_2)^2}(|e(x_1)|+|e(x_2)|)\\&\leqslant\frac{2}{(44.732\,5+44.710\,2)^2}\left(\frac{1}{2}\times10^{-4}+\frac{1}{2}\times10^{-4}\right)\\&=0.25\times10^{-7}<\frac{1}{2}\times10^{-7}\end{aligned}$$

知按第二种方法计算所得结果(至少)具有6位有效数字。由此也不难看出第一种算法确实只具有2位有效数字。

通过本例可以看出当两个相近的数相减时会造成有效位数的减少。事实上根据式(1.2.12)可知当 x_1 和 x_2 相近时，$\frac{x_1}{x_1-x_2}$ 和 $\frac{x_2}{x_1-x_2}$ 的绝对值会很大，这时 $|e_r(x_1-x_2)|$ 可能比 $|e_r(x_1)|+|e_r(x_2)|$ 大得多。因此，在实际计算中应当尽可能避免两相近数相减，否则会使计算精度大大降低。

1.3 机器数系

1.3.1 机器数系

电子计算机中数的表示大都采用浮点表示的形式，并以该形式存贮和运算，这种形式与科学记数法非常相似。

设一台计算机有 n 位字长，采用 β 进制，阶码为 p，$L\leqslant p\leqslant U$(这里 L，U 和 n 都是由该计算机的硬件所决定的某些常数)，则在此计算机中数的浮点表示为

$$x=\pm(0.\alpha_1\alpha_2\cdots\alpha_n)\beta^p \tag{1.3.1}$$

亦即

$$x=\pm\left(\frac{\alpha_1}{\beta}+\frac{\alpha_2}{\beta^2}+\cdots+\frac{\alpha_i}{\beta^i}+\cdots+\frac{\alpha_n}{\beta^n}\right)\beta^p$$

其中 α_i 为满足

$$0\leqslant\alpha_i\leqslant\beta-1\quad(i=1,2,\cdots,n)$$

的整数，$\alpha=\pm 0.\alpha_1\alpha_2\cdots\alpha_n$ 称为尾数，β 称为浮点数的基。

若规定 $\alpha_1\neq 0$，则称此数为规格化的浮点数。

我们把形如式(1.3.1)的所有规格化浮点数的全体及机器零组成的集合称为**机器数系**，记作 $F(\beta,n,L,U)$，即

$$F(\beta,n,L,U)=\{0\}\cup\{x\mid x=\pm(0.\alpha_1\alpha_2\cdots\alpha_n)\beta^p\}$$

式中，$1\leqslant\alpha_1\leqslant\beta-1$；$0\leqslant\alpha_i\leqslant\beta-1$，$i=2,3,\cdots,n$；$L\leqslant p\leqslant U$。

由此可见，机器数系是一个离散的由有限个有理数所组成的集合。$F(\beta,n,L,U)$ 中共有

$$1+2(\beta-1)\beta^{n-1}(U-L+1) \tag{1.3.2}$$

个数，其中绝对值最大的数为 $\pm\left(\frac{\beta-1}{\beta}+\frac{\beta-1}{\beta^2}+\cdots+\frac{\beta-1}{\beta^n}\right)\beta^U=\pm(1-\beta^{-n})\beta^U$，绝对值最小的非零数为 $\pm\left(\frac{1}{\beta}+\frac{0}{\beta^2}+\cdots+\frac{0}{\beta^n}\right)\beta^L=\pm\beta^{-1+L}$。