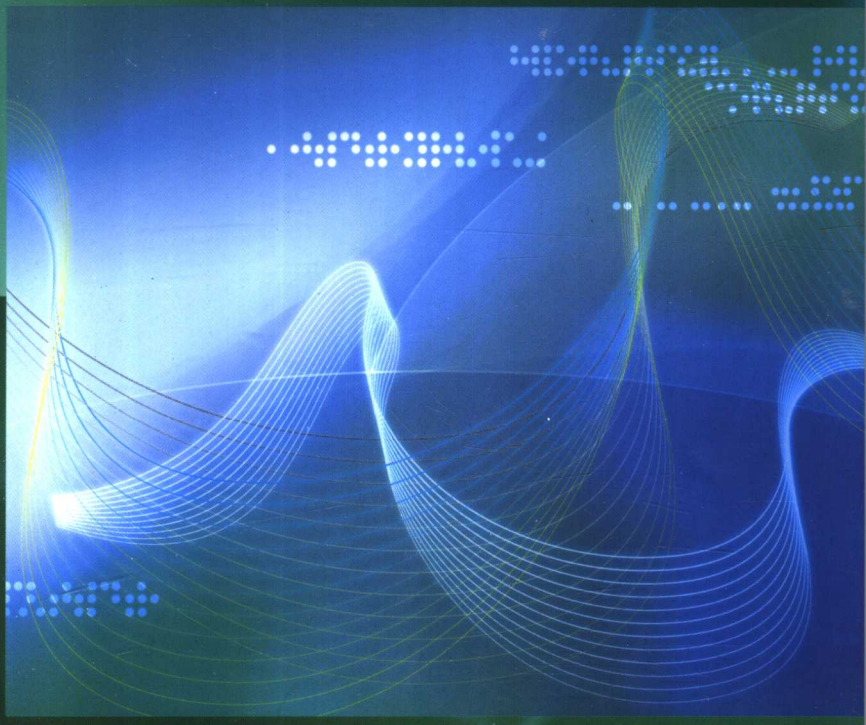
普通高等学校信息与计算科学专业系列丛书

普通高等教育“十一五”国家级规划教材

数值分析（上册）

主编 冯果忱 黄明游

编者 刘停战 刘 播 邹永魁 张树功



高等教育出版社
HIGHER EDUCATION PRESS

普通高等学校信息与计算科学专业系列丛书

普通高等教育“十一五”国家级规划教材

数 值 分 析

（上 册）

主编 冯果忱 黄明游
编者 刘停战 刘 播 邹永魁 张树功

高等教育出版社

内容提要

本书是为高等学校信息与计算科学本科教学而编写的, 强调数值计算的理论分析, 适用于较多学时的 "数值分析" 课程教学。全书共分上、下两册, 本书为上册, 主要介绍有关数值代数的内容、科学与工程计算中所出现的线性代数问题数值求解的算法设计原理、误差分析与收敛性估计等。

本书可作为高等学校信息与计算科学专业以及计算机类本科专业的教科书, 也可作为科学计算类课程的参考书, 供计算机学科、力学、物理学科各专业的本科生及相关人员阅读。

图书在版编目 (CIP) 数据

数值分析. 上册/ 冯果忱, 黄明游主编. —北京: 高等教育出版社, 2007.7

ISBN 978-7-04-021779-7

Ⅰ. 数… Ⅱ.①冯…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 (2007) 第 069496 号

策划编辑 李 蕊 **责任编辑** 李华英 **封面设计** 王凌波
责任绘图 尹文军 **版式设计** 马静如 **责任校对** 胡晓琪
责任印制 毛斯璐

出版发行	高等教育出版社	**购书热线**	010-58581118
社 址	北京市西城区德外大街 4 号	**免费咨询**	800-810-0598
邮政编码	100011	**网 址**	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经 销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印 刷	北京宏伟双华印刷有限公司	**畅想教育**	http://www.widedu.com
开 本	787×960 1/16	**版 次**	2007年 7 月第 1 版
印 张	12.75	**印 次**	2007年 7 月第 1 次印刷
字 数	230 000	**定 价**	16.40 元

本书如有缺页、倒页、脱页等质量问题, 请到所购图书销售部门联系调换。

物料号 21779-00

信息与计算科学专业系列教材编委会

序　言

由于计算机的发展与普及，借助计算机从事科学研究与工程设计已经成为重要手段。从而，借助计算机处理相应的数学模型的理论与方法日益活跃，这就产生了一门新兴的独立学科——计算数学。因为，在计算机上进行的计算当前仍以数值计算为主，因此这一学科也经常称为数值数学。计算数学的基础也称为数值分析。

线性代数问题是可以直接进行数值计算的，为使一个数学模型可以进行数值计算，通常需要借助数值逼近工具将其化为可计算的代数问题。因此数值分析通常包含两方面的内容：数值线性代数与数值逼近。参照现有的数值分析的惯例，本书还增加了一元非线性方程解法与常微分方程解法两方面的内容。此外，在本书的开头还介绍了数值计算的误差理论，这是数值计算应用的理论基础。

应当特别说明的是，本书着重讨论通用的计算方法的理论，为算法进一步发展做理论上的准备，而不一定涉及那些对某种特定问题运算快捷的算法。这是本书与流行的侧重于应用的计算方法方面书籍的差别，或者说本书不可作为计算方法教材。

本教材分上、下两册，上册讲述数值误差理论、线性代数计算方法以及一元方程求解，下册讲述一元逼近理论及常微分方程数值解法。

受教育部高等学校数学类专业教学指导分委员会的委托，吉林大学计算数学方面的教师们承担了本书的编写工作。这是计算数学方面的基础教材。在编写过程中我们充分利用了几十年来吉林大学计算数学方向的相关教材，并且本着与时俱进的精神，精心选择材料，尽可能征求任课教师的意见，力求完善。然而由于水平限制，一定有不尽如人意之处，我们欢迎来自各方面的意见和建议。

在完成本书的过程中，审稿人及出版社的同志们给予了很多帮助，特表衷心感谢。

冯果忱

2007 年春于长春

郑重声明

目　　录

绪　　论

实验方法与理论方法是推动科学技术发展的两大基本方法, 但它们也有局限性. 许多研究对象, 由于空间或时间的限制, 既不可能用理论精确描述, 也不可能用实验手段来实现.

计算机技术和计算技术的飞速发展, 为研究这些问题开辟了一条新的途径 —— 数值模拟或称为科学计算. 科学计算突破了实验和理论科学的局限, 在科技发展中起到越来越重要的作用. 可以认为, 科学计算已与实验、理论一起成为科学方法上不可或缺的三个主要手段.

计算数学的研究是科学计算的主要组成部分, 而数值分析则是计算数学的核心.

§1　数值分析的内容和特点

1.1　数值分析的内容

数值分析 (Numerical Analysis) 研究数值求解各类数学问题的方法和相应的数学理论. 研究的对象是数学问题, 所用的方法是数学方法, 因此也称为数值数学 (Numerical Mathematics).

数值分析研究的内容可划分为以下几个主要方面.

1. 数值代数. 主要包括线性代数方程组和非线性方程与方程组的数值解法、特征值与特征向量的数值计算等内容.

2. 数值逼近. 主要包括函数逼近 (特别是函数的插值逼近)、数值微分和数值积分等内容.

3. 常微分方程和动力系统的数值解法.

4. 偏微分方程的数值解法.

5. 最优化理论与方法. 主要研究在一定的约束条件下如何选取某些因素的值, 使某项或某几个指标达到最优.

6. 误差理论. 主要研究近似方法的误差 (即数值结果与原问题精确解之间的误差) 及影响误差的主要因素, 因为每一种数值方法严格说来都是近似方法. 这是数值分析中非常重要的一个问题.

本书主要介绍数值代数、数值逼近、常微分方程数值解法及相应的误差理论. 其他内容将在后续课程中介绍.

1.2 数值方法的特点

求解一个数学问题的数值方法是要给出该问题的一个近似的数值结果. 因此, 首先数值结果要能算得出来. 其次结果应有一定的精度, 满足实际问题的要求. 一般要求误差满足指定的值 ε. 第三, 计算时间应尽可能少.

求解一个数学问题一般会有多种数值方法. 在保证所需精度的条件下, 计算时间越少的数值方法越好. 对串行机而言, 这相当于计算量越少越好.

例 1.1 *考虑求解线性代数方程组*

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{1.1}$$

其中系数矩阵 $\boldsymbol{A}$ 是 $n \times n$ 的方阵, 其行列式 $D \equiv \det(\boldsymbol{A}) \neq 0$.

按克拉默 (Cramer) 法则, (1.1) 的解为

$$x_i = D_i/D, \quad i = 1, 2, 3, \cdots, n. \tag{1.2}$$

每个行列式按 Laplace 展开计算, 这就给出了一种求解 (1.1) 的数值方法. 下面分析该算法的计算量. (1.2) 共需计算 $n+1$ 个 n 阶行列式. 用 Laplace 展开计算 n 阶行列式, 约需 $n!$ 次乘法. 这样, 不计加法, 该算法共需 $(n+1)n! = (n+1)!$ 以上的乘法. 对于一个 20 阶的方程组, 就需要 $21! \approx 5.11 \times 10^{19}$ 以上的乘法. 设用每秒可做亿次乘法的计算机, 一年可做 $365 \times 24 \times 3\,600 \times 10^8 \approx 3.15 \times 10^{15}$ 次乘法. 所以, 在此计算机上用 Cramer 法则解 20 阶的线性代数方程组, 需要的时间在 $(5.11 \times 10^{19}) \div (3.15 \times 10^{15}) \approx 1.62 \times 10^4 = 1.62$ 万年以上.

而用后面将要介绍的高斯 (Gauss) 消元法求解, 乘、除法的运算次数不超过 3 074 次, 与 Cramer 法则比较运算量有天壤之别. 方程组的阶数增大, 运算量的差别还会更大. 应当指出, 数值方法的计算时间, 由计算机速度和数值方法的效率决定. 从某种意义上说, 对于减少计算时间, 提高数值方法的效率甚至比提高计算机速度更重要, 因为算法研究所需要的代价要小得多.

提高数值方法的计算效率, 减少计算时间, 对于某些需要 "实时" 计算的问题, 比如 "天气预报" 等, 显得尤为重要.

一个数值方法一般包括若干计算公式. 递推化是许多计算公式采取的形式, 其基本思想是将一个相对复杂的计算过程归结为简单过程的多次重复, 而这在计算机上可用 "循环" 来实现.

例 1.2 考虑对于给定的 x 计算多项式

$$P_n(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$$

的值.

将它改写成等价形式

$$\begin{aligned} P_n(x) &= (a_0x^{n-1} + a_1x^{n-2} + \cdots + a_{n-2}x + a_{n-1})x + a_n \\ &= ((a_0x^{n-2} + a_1x^{n-3} + \cdots + a_{n-2})x + a_{n-1})x + a_n \\ &= \cdots = (\cdots((a_0x + a_1)x + a_2)x + \cdots + a_{n-1})x + a_n, \end{aligned}$$

就可以得到递推公式

$$\begin{aligned} v_0 &= a_0, \\ v_k &= xv_{k-1} + a_k, \quad k = 1, 2, \cdots, n. \end{aligned}$$

递推结果 $v_n = P_n(x)$.

这一算法称为霍纳 (Horner) 算法, 我们也称它为秦九韶算法 (因为算法的基本思想是我国宋代数学家秦九韶最先提出的). 该算法逻辑结构简单, 并且乘、除法的计算量比直接的幂法节省一半.

例 1.3 计算积分

$$I_n = \int_0^1 x^n \mathrm{e}^{x-1} \mathrm{d}x.$$

解 利用分部积分公式, 可得递推关系

$$I_n = 1 - nI_{n-1}, \tag{1.3}$$

或

$$I_{n-1} = (1 - I_n)/n. \tag{1.4}$$

先计算出 $\widetilde{I}_7 = 0.112\ 4$, 利用 (1.4) 式, 递推可算出 $\widetilde{I}_6, \cdots, \widetilde{I}_0$. 计算结果见表 1, 表中 I_n 是精确值的舍入结果.

表 1

n	7	6	5	4	3	2	1	0
$\widetilde{I}_n$	0.112 4	0.126 9	0.145 5	0.170 8	0.207 3	0.264 3	0.368 0	0.632 0
I_n	0.112 4	0.126 8	0.145 6	0.170 9	0.207 3	0.264 2	0.367 9	0.632 1

数值分析更为重要的特点 "误差估计" 与 "稳定性分析", 将在 §3 中专门介绍.

§2 数制与浮点运算

2.1 数制

一个数通常写成 $A = d_n d_{n-1} \cdots d_0 . d_{-1} d_{-2} \cdots d_{-m}$ 的形式, $0 \leqslant d_i \leqslant 9$, d_i 为整数, $d_n \neq 0$, 它也可写成级数形式

$$A = \sum_{j=-m}^{n} d_j 10^j. \tag{2.1}$$

这是十进制数的表示, 此时数是 10 幂次的级数.

计算机内部采用的数制, 不是十进制, 而是二进制、八进制等. 对照数的十进制表示 (2.1), 数的二进制表示就是它的 2 幂次级数

$$B = \sum_{j=-m}^{n} d_j 2^j,$$

其中 $d_j = 0$ 或 1, $d_n \neq 0$. 也可以将其写成二进制数的形式

$$B = (d_n d_{n-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots d_{-m})_2.$$

同样地, 当 $\beta \geqslant 2$ 为正整数时, 可以定义 β 进制的数

$$C = (d_n d_{n-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots d_{-m})_\beta = \sum_{j=-m}^{n} d_j \beta^j,$$

这里 $0 \leqslant d_j \leqslant \beta - 1, d_n \neq 0$.

一个数既可以用十进制表示, 即 $\beta = 10$, 也可以用 $\beta (\neq 10)$ 进制表示. 例如, π 的近似值 3.141 6 的二进制表示为

$$(11.001001000011111 \cdots)_2.$$

从此例看出, 有限位小数的十进制数用其他数制表示时, 可以是无限位小数.

β 进制数化成十进制数是简单的. 对 β 进制数

$$A = (h_n h_{n-1} \cdots h_1 h_0 . h_{-1} h_{-2} \cdots h_{-m})_\beta,$$

关于级数 $A = \sum\limits_{j=-m}^{n} h_j \beta^j$ 求和, 即得十进制数.

例如,

$$\begin{aligned}(11.001001)_2 &= 2^1 + 2^0 + 2^{-3} + 2^{-6} \\ &= 2 + 1 + \frac{1}{8} + \frac{1}{64} (= 2 + 1 + 0.125 + 0.015\ 625) \\ &= 3.140\ 625.\end{aligned}$$

将十进制数表成 β 进制数, 可将整数部分和小数部分分别转换.

整数部分的转换可以采用 Horner 多项式表达法. 将十进制整数

$$A = A_0 = d_n d_{n-1} \cdots d_0$$

除以 β, 得余数 r_0, 商数 A_1; 然后将 A_1 再除以 β, 又得余数 r_1, 商数 A_2; 依次类推, 直到 $A_k < \beta$, 并记 $r_k = A_k$. 这样,

$$\begin{aligned}(r_k r_{k-1} \cdots r_0)_\beta &= (\cdots((r_k\beta + r_{k-1})\beta + r_{k-2}) + \cdots + r_1)\beta + r_0 \\ &= \sum_{j=0}^{k} r_j \beta^j\end{aligned}$$

便是整数 A 的 β 进制表示.

例如, 用 Horner 法将 291 转换成二进制:

$291 \div 2$ 的商为 $145 = A_1$, 余数为 $1 = r_0$;

$145 \div 2$ 的商为 $72 = A_2$, 余数为 $1 = r_1$;

$72 \div 2$ 的商为 $36 = A_3$, 余数为 $0 = r_2$;

$36 \div 2$ 的商为 $18 = A_4$, 余数为 $0 = r_3$;

$18 \div 2$ 的商为 $9 = A_5$, 余数为 $0 = r_4$;

$9 \div 2$ 的商为 $4 = A_6$, 余数为 $1 = r_5$;

$4 \div 2$ 的商为 $2 = A_7$, 余数为 $0 = r_6$;

$2 \div 2$ 的商为 $1 = A_8$, 余数为 $0 = r_7$.

因为 $A_8 < 2$, 故 $r_8 = A_8 = 1$. 由此即得 $291 = (100100011)_2 (= 2^8 + 2^5 + 2^1 + 2^0)$.

下面讨论将十进制纯小数化成 β 进制小数的方法. 将十进制纯小数

$$A = A_0 = 0.d_{-1}d_{-2}\cdots d_{-k}\cdots$$

写成 β 进制小数

$$A_0 = (0.g_{-1}g_{-2}\cdots g_{-l}\cdots)_\beta = \sum_{j=1}^{\infty} g_{-j}\beta^{-j}.$$

两边乘以 β 得

$$\beta A_0 = \sum_{j=1}^{\infty} g_{-(j+1)}\beta^{-j} + g_{-1}.$$

记 $A_1 = \beta A_0 - g_{-1} = \sum\limits_{j=1}^{\infty} g_{-(j+1)}\beta^{-j}$, 它是 βA_0 的小数部分, 而 g_{-1} 就是 βA_0 的整数部分. 同样地, 在 $A_1 = \beta A_0 - g_{-1}$ 的两端乘以 β, 就可以求出 g_{-2}, 它是

βA_1 的整数部分. 记 $A_2 = \beta A_1 - g_{-2}$, 它是 βA_1 的小数部分. 依次类推可以求出 $g_{-3}, g_{-4}, \cdots$.

例如, 求 $A = A_0 = 0.141\ 6$ 的二进制表示 $(0.g_{-1}g_{-2}\cdots)_2$.

$2 \times A_0 = 0.141\ 6 \times 2 = 0.283\ 2$, 整数部分 $g_{-1} = 0, A_1 = 0.283\ 2 - 0 = 0.283\ 2$;

$2 \times A_1 = 0.283\ 2 \times 2 = 0.566\ 4$, 整数部分 $g_{-2} = 0, A_2 = 0.566\ 4 - 0 = 0.566\ 4$;

$2 \times A_2 = 0.566\ 4 \times 2 = 1.132\ 8$, 整数部分 $g_{-3} = 1, A_3 = 1.132\ 8 - 1 = 0.132\ 8$;

$2 \times A_3 = 0.132\ 8 \times 2 = 0.265\ 6$, 整数部分 $g_{-4} = 0, A_4 = 0.265\ 6 - 0 = 0.265\ 6$;

$2 \times A_4 = 0.265\ 6 \times 2 = 0.531\ 2$, 整数部分 $g_{-5} = 0, A_5 = 0.531\ 2 - 0 = 0.531\ 2$;

$2 \times A_5 = 0.531\ 2 \times 2 = 1.062\ 4$, 整数部分 $g_{-6} = 1, A_6 = 1.062\ 4 - 1 = 0.062\ 4$;

.....

这样, $0.141\ 6 = (0.001\ 001\cdots)_2$. (因为是无穷小数, 故 $(0.001\ 001)_2 \neq 0.141\ 6$.)

2.2 浮点数

任意一个实数 x, 可以用整数部分加小数部分来表示, 也可以表示成

$$x = s \times 10^p, \tag{2.2}$$

这里 $s = \pm 0.d_1d_2\cdots d_k\cdots$ 是十进制数, 其中 $1 \leqslant d_1 \leqslant 9, 0 \leqslant d_i \leqslant 9, i = 2, 3, \cdots$, p 是整数, 称为**指数**, s 称为**尾数**. s 的小数点固定在 $d_1 \neq 0$ 之前, 因为有指数项 10^p 的作用, 将 x 写成整数部分和小数部分的表示时, 对不同的 x, 小数位置可以是浮动的, 因此称 (2.2) 式为 x 的 (十进制) 浮点数.

同样地, 可定义 β 进制的浮点数

$$x = \pm(0.d_1d_2\cdots d_k\cdots)_\beta \times \beta^p,$$

其中 $1 \leqslant d_1 < \beta, 0 \leqslant d_i < \beta, i = 2, 3, \cdots$.

每种计算机是按固定位数的有限位浮点数进行运算的, 这样, 尾数的位数是固定的, 设为 t, 同时限制指数有上界 M 和下界 $-m$. 于是计算机上的数 (简称机器数) 有如下表示形式:

$$x = \pm(0.d_1d_2\cdots d_t)_\beta \times \beta^p,$$

其中 $1 \leqslant d_1 < \beta, 0 \leqslant d_i < \beta, i = 2, 3, \cdots, t, -m \leqslant p \leqslant M, t$ 称为字长, t, p, m, M 均为正整数, β 是数制, 多数计算机取 $\beta = 2$ (也有取 $\beta = 8, 16$ 的). 记 $s = \pm 0.d_1d_2\cdots d_t$. 数字 0 的表示一般约定 $s = 0, p = -m$.

浮点数可以有多种表示方法 (目前大部分计算机系统的二进制浮点数采用的是 IEEE 标准), 但是它们的基本原理是一样的.

对于固定的计算机, β, t, m, M 是固定的, 因此计算机所能表示的数的范围是固定的, 将其记为 $F(\beta,t,m,M)$, 相应的计算机系统记为 (β,t,m,M).

易见, 对于任意的 $x \in F(\beta,t,m,M), x \neq 0$, 有

$$\beta^{-m-1} \leqslant |x| < \beta^M.$$

运算中若出现绝对值小于 β^{-m-1} 的数, 计算机就下溢, 并用 0 代替此数 (如它作除数, 会出现什么现象?). 若出现绝对值大于等于 β^M 的数, 计算机就上溢 (通常会停机).

属于

$$G = \{x \in \mathbb{R} : \beta^{-m-1} \leqslant |x| < \beta^M\}$$

的实数 x, 不一定属于 $F(\beta,t,m,M)$. 此时可按某种规则确定 x 的一个近似值 $\tilde{x} \in F(\beta,t,m,M)$ 来表示 x, 于是, 对于任意的 $x \in G$, 均可用机器数 $fl(x)$ 来表示. 这就定义了从 G 到 $F(\beta,t,m,M)$ 的函数 fl.

在现有计算机中 $fl(x)$ 可按舍入和切断两种规则之一确定. 设 $x \in G$,

$$x = \pm(0.d_1d_2\cdots d_td_{t+1}\cdots)_\beta \times \beta^p.$$

1) 舍入: $fl(x)$ 取成与 x 最接近的机器数:

$$|x - fl(x)| = \min_{f\in F} |x - f|.$$

此时, $fl(x)$ 实际是这样得到的: 若 $d_{t+1} < \frac{1}{2}\beta$, 则将 d_{t+1} 及其以后各数舍去; 若 $d_{t+1} \geqslant \frac{1}{2}\beta$, 则将 d_{t+1} 及其以后各数舍去, 并将 d_t 换成 $d_t + 1$ (这还意味着如果 $d_t + 1 = \beta$, 则进位). 这实际上是通常的舍入规则.

2) 切断: $fl(x)$ 取成满足 $|fl(x)| \leqslant |x|$ 并与 x 最接近的机器数:

$$|x - fl(x)| = \min_{f\in F, |f|\leqslant|x|} |x - f|.$$

这相当于在 x 的表示中截去 d_{t+1} 及其以后的所有数.

定理 令

$$\delta = \frac{fl(x) - x}{x}, \quad \mu = \begin{cases} \frac{1}{2}\beta^{-t+1}, & \text{舍入情形}, \\ \beta^{-t+1}, & \text{切断情形}. \end{cases} \tag{2.3}$$

则有

$$|x - fl(x)| \leqslant \varepsilon \equiv \begin{cases} \frac{1}{2}\beta^{p-t}, & \text{舍入情形}, \\ \beta^{p-t}, & \text{切断情形}, \end{cases} \tag{2.4}$$

$$fl(x) = x(1+\delta), \quad |\delta| \leqslant \mu.$$