计算数学丛书

舍入误差分析引论

唐珍编著

计算数学丛书

舍入误差分析引论

唐 珍 编著

上海科学技术出版社

责任编辑 唐林华

舍入误差分析引论

唐 珍 编著

上海科学技术出版社出版

（上海瑞金二路 450 号）

新华书店上海发行所发行　　上海市印刷三厂印刷

开本 787×1092 1/32　印张 9.5　字数 207,000

1987 年 11 月第 1 版　1987 年 11 月第 1 次印刷

印数：1—3,100

统一书号：13119·1436　定价：2.10 元

出 版 说 明

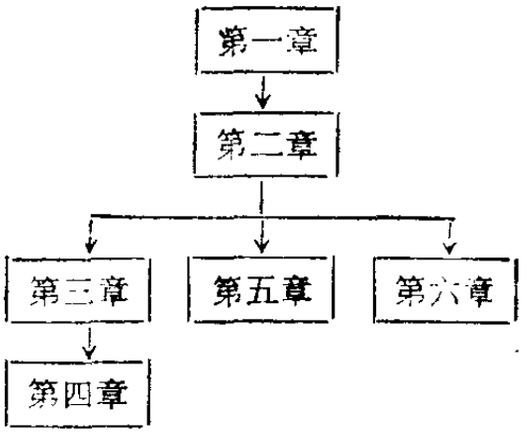
《计算数学丛书》是为了适应计算数学和计算机科学的发展，配合高等院校的计算数学教学的需要而组织的一套参考读物. 读者对象主要是高等院校数学系和计算机科学系的学生、研究生，亦可供高等院校数学系和计算机科学系的教师以及工矿企业、科研单位从事计算工作的技术人员参考.

本丛书向读者介绍近代计算方法的一些主要进展及其适用范围和实用效果. 每种书集中介绍一个专题，针对本专题的近代发展作综合性的介绍，内容简明扼要，重点突出，有分析，有评价，力图使读者对该专题的动向和发展趋势得到一个完整的了解.

本丛书已拟定的选题计有：《线性代数与多项式的快速算法》、《数论变换》、《数值有理逼近》、《矩阵特征值问题》、《索伯列夫空间引论》、《计算组合数学》、《样条与插值》、《不动点算法》、《广义逆矩阵的基本理论和计算方法》、《非线性方程的区间算法》、《奇异摄动中的边界层校正法》、《沃尔什函数理论与应用》、《多项式最佳逼近的实现》、《曲线曲面的数值表示和逼近》、《舍入误差分析引论》、《解边值问题的迦辽金法》、《非线性方程组迭代解法》、《外推法及其应用》、《蒙特卡罗方法》、《发展方程的有限元方法》、《数值解高维偏微分方程的分裂法》等二十余种，于一九八〇年初起陆续出版.

序　言

本书着重介绍舍入误差分析的一个主要方法——向后分析方法和误差分析在线代数方程组数值解法方面的某些进展. 第1、2章是基本的，第3、4章分析非稀疏矩阵的一些基本变换和非稀疏方程组的一些常用的直接解法，第5章分析稀疏方程组的消去法，第6章分析一些基本的迭代法. 各章逻辑关系如下：



本书可供高等院校数学系和计算机科学系有关专业的高年级大学生和研究生阅读，也可供科研设计单位、计算中心和工矿企业的有关科技人员参考.

胡诗秀同志认真阅读了原稿，指出了一些疏忽的地方. 此外，编写过程中得到甘肃省科学技术委员会领导的关怀和甘肃省计算中心不少同志的支持. 编者谨向这些同志表示深切的谢意.

由于编者水平和时间关系，缺点和错误在所难免，恳请读者批评指正.

唐　珍

于甘肃省计算中心

目　录

第 1 章

基本算术运算的误差分析

§1 引　　言

在本书中我们只考虑使用自动数字计算机的计算. 我们叙述的结果，对使用自动数字计算机的计算是非常重要的.

用于科学计算的大多数自动数字计算机是以二进制表示的数进行工作的，也有以三进制、八进制、十进制、十六进制等进行工作的. 所以，我们以 β 进制运算表示误差分析（β 叫做基），而一般地用我们习惯的十进制列举数值例子.

关于各种记号本书将采用以下的约定.

除指数、足标、向量维数、方阵阶数、某些习惯表示法及其它特别指出的记号外，小写希腊字母表示纯量；黑体小写拉丁字母表示列向量；斜体大写拉丁字母表示矩阵；草写体大写拉丁字母表示集合. 上标 T 表示转置. $\mathrm{diag}(\alpha_i)=\mathrm{diag}(\alpha_1, \alpha_2, \cdots, \alpha_n)$ 表示对角阵，其对角线元素为 $\alpha_1, \cdots, \alpha_n$. 一般，$a_i$ 表示矩阵 A 的第 i 列，α_{ij} 表示位于 (i, j) 位置的元素，而 e_i 表示单位阵 I 的第 i 列，且

$$e=\sum_i e_i.$$

ξ_i 表示 $\boldsymbol{x}$ 的元素，η_i 表示 $\boldsymbol{y}$ 的元素，$(\boldsymbol{x})_i$ 表示 $\boldsymbol{x}$ 的第 i 个分量.

当且仅当对每个 i 有 $\xi_i \leqslant \eta_i$ 时，

$$\boldsymbol{x} \leqslant \boldsymbol{y};$$

当且仅当对每个 i、j 有 $\alpha_{ij}\leqslant\beta_{ij}$ 时，

$$A\leqslant B.$$

$\delta(A)$ 表示 A 的行列式；但 δA，ΔA 或 $\delta[A]$ 表示 A 的摄动．$\lambda_i=\lambda_i(A)$ 表示 A 的特征值，并把它们按如下次序排列：

$$|\lambda_1(A)|\geqslant|\lambda_2(A)|\geqslant\cdots\geqslant|\lambda_n(A)|;$$

$\lambda=\lambda(A)$ 表示 A 的某个特征值．$\sigma_i(A)\geqslant 0$ 表示实矩阵 A 的奇异值，式中

$$\sigma_i^2(A)=\lambda_i(A^TA);$$

$\rho(A)$ 表示 A 的谱半径，即

$$\rho(A)=\max_i|\lambda_i(A)|.$$

我们预先假定读者对实数的近似值、绝对误差、相对误差、有效数字、固有误差、比例因子和向量空间、算法等概念以及求解线性代数方程组的一些常用算法是熟悉的．

本章主要讨论使用数字计算机时基本的和常用的定点及浮点算术运算的舍入误差分析．也一般地描述了向前舍入误差分析、向后舍入误差分析；算法的向后稳定性、向前稳定性、数值稳定性；问题的病态、良态、条件数等基本概念．

§2 定点算术运算和浮点算术运算

通常，在自动数字计算机上使用的运算方式主要有两种．第一种叫做定点算术运算（简称定点运算或定点）．用这种方式时，必须把计算加以组织，必要时引入比例因子，使得参加计算的每一个数 ξ 满足诸如

$$-1\leqslant\xi\leqslant 1 \qquad (2.1)$$

的某些不等式．在这样的不等式中，常常把数 $+1$ 排除在外，而不把数 -1 排除在外．但是，这使分析的细节在非本质之

点复杂化了. 所以，我们假定，数的允许区间是由(2.1)给出的区间.

一般说来，我们用固定个数的数字，即用 t 个 β 进数字，表示一个定点数. 换句话说，在一个定点数的 β 进小数点后有 t 个 β 进数字. 这时我们说计算机用 t 个 β 进数字构成的字进行工作. 定点数中数字的位数 t 叫做这个数的精度. 注意，数有符号(+号可以省略). 于是，任何一个 β 进定点数有形式

$$\pm 0.\xi_1\xi_2\cdots\xi_t, \tag{2.2}$$

式中数字 ξ_i 满足

$$0\leqslant\xi_i\leqslant\beta-1 \qquad (i=1, 2, \cdots, t). \tag{2.3}$$

在区间 $[-1, 1]$ 中，定点数形成等距点的有限集

$$\mathscr{F}=\mathscr{F}(\beta, t).$$

图 1.1 是 $\beta=2$, $t=3$ 时 15 点形成的 $\mathscr{F}$ 的图形表示.

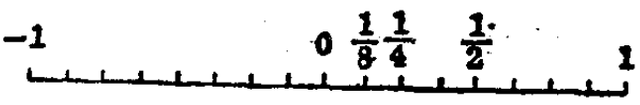


图 1.1

任何一个想在计算机中表示的位于 $[-1, 1]$ 中的实数 ξ 都可以用定点数近似地表示. 第一种表示方法是"四舍五入"法，即用最接近它的定点数 ξ_R(ξ 的"四舍五入")来逼近. 即我们取数 $|\xi|+\frac{1}{2}\beta^{-t}$ 在基 β 中的无穷表达式，并丢弃从小数点往右第 t 位后的所有数字. 这样，ξ_R 就由 ξ 的符号和刚才得到的 t 位的数构成. 于是，我们有下列定理.

定理 2.1 若实数 $\xi\in[-1, 1]$，则

$$\xi_R=\xi+\delta \quad \text{其中} \quad |\delta|\leqslant\frac{1}{2}\beta^{-t}. \tag{2.4}$$

这个定理在用"四舍五入"法的定点舍入误差理论中是最

基本的. 它提供用 ξ_R 替代 ξ 时引进的绝对误差的一种方便的表达式.

例 2.1 令 $\beta=10$, $t=2$ 和 $\xi=c$(Euler 常数), c 的十进小数表示为 $0.577216\cdots$, 于是 $\xi_R=0.58$. 因此

$$\delta=\xi_R-\xi=0.003784\cdots.$$

因为 $\frac{1}{2}\beta^{-t}=\frac{1}{2}(10^{-2})=0.005$, 显然, δ 满足不等式 $|\delta|\leqslant\frac{1}{2}\beta^{-t}$.

现在, 我们证明: $\frac{1}{2}\beta^{-t}$ 是 $|\delta|$ 的**精确界**, 即, 它是可以达到的. 为此目的, 考虑 $\xi=\beta^{-1}+\frac{1}{2}\beta^{-t}$, 那时 $\xi_R=\beta^{-1}+\beta^{-t}$. 于是

$$\delta=\xi_R-\xi=\frac{1}{2}\beta^{-t}.$$

用定点数近似地表示区间 $[-1, 1]$ 中的实数的第二种方法是**切断法**. 即, 在 $[-1, 1]$ 中给定某实数 ξ, 我们取 ξ_0(ξ 的切断)为 $\mathscr{F}$ 中的最接近 ξ 且满足条件 $|\xi_0|\leqslant|\xi|$ 的唯一数. ξ_0 可以用下述方法得到: 取数 ξ 在基 β 下的无穷表达式(如为有限表示, 则最后位以后全为零), 并且丢弃小数点后前 t 位数字后的所有数字. 于是, 我们有下列定理.

定理 2.2 若实数 $\xi\in[-1, 1]$, 则

$$\xi_0=\xi+\delta,\quad \text{其中}\quad |\delta|<\beta^{-t}. \tag{2.5}$$

用定点数表示实数的其它方法, 我们就不在这里叙述了.

为方便起见, 我们定义以下的简化记号.

令 ν 代表用定点数表示实数 $\xi\in[-1, 1]$ 时的绝对**舍入误差界单位**, 则

$$\nu=\begin{cases}\frac{1}{2}\beta^{-t} & (\text{“四舍五入”方法}),\\ \beta^{-t} & (\text{切断方法}).\end{cases}$$

若必须以高于 t 的精度进行工作，则我们可以使用以 t 个 β 进数位的倍数表示的数，并且我们把它叫做多精度算术运算，以区别于精度为 t 的单精度算术运算. 今后，我们把单精度定点数叫做标准定点数. 通常，多精度算术运算需要使用专门的子程序，并且每一个算术运算比对应的单精度算术运算要耗费长几倍的时间.

第二种运算方式叫做**浮点算术运算**(简称浮点算术，或浮点). 采用这种方式时，把每一个数 ξ 用一个有序数对 α, γ 表示成

$$\xi=\beta^{\gamma}(\alpha), \tag{2.6}$$

式中 γ 是整数，它有一定的范围；α 是满足

$$-\frac{1}{\beta}\geqslant\alpha\geqslant-1 \text{ 或 } \frac{1}{\beta}\leqslant\alpha\leqslant 1 \tag{2.7}$$

的数. 通常，把数 γ 叫做**指数或阶码**，把数 α 叫做**小数部分或数值部分或尾数**. 在区间(2.7)中，可能把端点之一排除在外. 但是，和定点算术运算一样，我们不管这种可能性. 这样一来，任何非零浮点数(+号可以省略)有形式

$$\beta^{\gamma}(\pm 0.\alpha_1\alpha_2\cdots\alpha_t), \tag{2.8}$$

式中整数 γ 满足不等式

$$-\mu'\leqslant\gamma\leqslant\mu'', \tag{2.9}$$

整数 $\alpha_1, \alpha_2, \cdots, \alpha_t$ 满足不等式

$$1\leqslant\alpha_1\leqslant\beta-1, \tag{2.10}$$

$$0\leqslant\alpha_i\leqslant\beta-1 \quad (i=2, 3, \cdots, t). \tag{2.11}$$

浮点数小数部分中数字的位数 t 叫做**浮点数的精度**.

关于数零的浮点表示，各个计算机是不相同的. 在一些计算机上用取小数部分的值为零来表示，例如，$\alpha=0$，在另一些计算机上用取指数 γ 为某一很大负数的办法来表示，例如，

$\gamma=-\mu'$. 决定用哪一种表示，不是很重要的. 我们假定用取小数部分和指数的值均为零 ($\alpha=\gamma=0$) 的办法来表示浮点数零.

在实数轴上，浮点数形成不等距点的有限集 $\mathscr{G}=\mathscr{G}(\beta, t, \mu', \mu'')$. 图 1.2 是 $\beta=2$, $t=3$, $\mu'=1$ 和 $\mu''=2$ 时 33 点形成的 $\mathscr{G}$ 的图形表示.

图 1.2

任何一个想在计算机中表示的在一定范围内的实数 ξ 都可以用刚才描述的集合 $\mathscr{G}$ 中的浮点数来逼近. 第一种逼近方法是"四舍五入"法，即，给定 ξ，我们用 $\mathscr{G}$ 中最接近的数 ξ_R 来表示 ξ；在表示不唯一情况下，我们选择 $\mathscr{G}$ 中两个最接近 ξ 的数中具有较大小数部分者为 ξ_R. 如果 $\xi=0$，则取 $\xi_R=0$. 如果 $\xi\neq 0$，则选择 α' 和 γ 使得

$$|\xi|=\beta^{\gamma}(\alpha'), \tag{2.12}$$

其中

$$\beta^{-1}\left(1-\frac{1}{2}\beta^{-t}\right)\leqslant\alpha'<1-\frac{1}{2}\beta^{-t}. \tag{2.13}$$

如果 γ 在允许区间 $-\mu'\leqslant\gamma\leqslant\mu''$ 之外，则不可能用适当的指数来表示 ξ，我们就说 ξ 在浮点数范围之外，这种情况我们将不讨论. 如果 γ 在允许区间内，则取数 $\alpha'+\frac{1}{2}\beta^{-t}$ 在基 β 中的无穷表达式(如果是有限表达式，则最后位以后全为零)，并且丢弃从小数点往右第 t 位后的所有数字. 这样，ξ_R 就由 ξ 的符号、刚才得到的 t 位的数和在(2.12)及(2.13)中确定的 β^{γ} 构成.

现在我们证明下述定理.

定理 2.3 如果 ξ 是在浮点数范围内的实数，则

$$\xi_R=\xi(1+\varepsilon),\quad 其中\quad |\varepsilon|\leqslant\frac{1}{2}\beta^{1-t}. \tag{2.14}$$

证明 我们假定 $\xi>0$，因为 $\xi<0$ 时可以类似地处理，而 $\xi=\xi_R=0$ 时是明显的. 令 γ 是满足

$$\beta^{\gamma-1}\leqslant\xi<\beta^{\gamma}$$

的唯一整数. 在区间 $[\beta^{\gamma-1},\ \beta^{\gamma}]$ 中，浮点数以间隔 $\beta^{\gamma-t}$ 均匀分布. 最接近 ξ 的是 ξ_R，而且它应该位于与 ξ 距离不大于 $\frac{1}{2}\beta^{\gamma-t}$ 的范围内，即

$$|\xi_R-\xi|\leqslant\frac{1}{2}\beta^{\gamma-t}.$$

因为 $\beta^{\gamma-1}\leqslant\xi$，所以，我们有

$$\frac{|\xi_R-\xi|}{|\xi|}\leqslant\frac{\frac{1}{2}\beta^{\gamma-t}}{\beta^{\gamma-1}}=\frac{1}{2}\beta^{1-t}.$$

最后，由于

$$\varepsilon=(\xi_R-\xi)/\xi,$$

所以，定理得到证明. 证毕.

这个定理在用“四舍五入”法的浮点舍入误差理论中是最基本的. 它提供用 ξ_R 替代 ξ 时引进的相对误差的一种方便的表达式.

例 2.2 令 $\beta=10$，$t=5$ 和 $\xi=e$，e 的十进小数表示为

$$2.718282\cdots=10^1\times0.2718282\cdots,$$

于是 $\gamma=1$，$\xi_R=10^1(0.27183)$，

$$\varepsilon=(\xi_R-\xi)/\xi=0.000006651\cdots.$$

因为 $\frac{1}{2}\beta^{1-t}=\frac{1}{2}\times10^{-4}=0.00005$，显然，$\varepsilon$ 满足不等式 $|\varepsilon|<\frac{1}{2}\beta^{1-t}$.

现在我们证明 $\frac{1}{2}\beta^{1-t}$ 差不多是 $|\varepsilon|$ 的精确界. 为此，我们考虑

$$\xi=\beta^{\gamma}\left(1+\frac{1}{2}\beta^{1-t}\right).$$

那时
$$\frac{|\xi_R-\xi|}{|\xi|}=\frac{\frac{1}{2}\beta^{1-t}\times\beta^{\gamma}}{\beta^{\gamma}\left(1+\frac{1}{2}\beta^{1-t}\right)}=\frac{\frac{1}{2}\beta^{1-t}}{1+\frac{1}{2}\beta^{1-t}}.$$

对实际的 t 值，等式十分接近于 $\frac{1}{2}\beta^{1-t}$.

用浮点数表示一定范围内实数的第二种方法是**切断法**. 即，在浮点数范围内给定某实数，我们取 ξ_0 为 $\mathscr{G}$ 中最接近 ξ 且满足条件 $|\xi_0|\leqslant|\xi|$ 的唯一数. ξ_0 可以用下述方法得到：取数 ξ 在基 β 下的无穷表达式（在有限表示时最后位以后全为零），并且丢弃前 t 个有效数字后的所有数字. 于是，我们有下列定理.

定理 2.4 如果 ξ 是浮点数范围内的实数，则

$$\xi_c=\xi(1+\varepsilon),\quad \text{其中}\quad |\varepsilon|<\beta^{1-t}.$$

用浮点数表示实数的其它方法我们就不在这里叙述了.

为方便起见，我们定义以下的简化记号.

令 μ 表示**相对舍入误差界单位**，则

$$\mu=\begin{cases}\frac{1}{2}\beta^{1-t} & (\text{“四舍五入”方法}),\\ \beta^{1-t} & (\text{切断方法}).\end{cases}$$

在不用机器指令代码直接提供浮点运算的计算机上，必须用子程序实现浮点运算. 为了使子程序运算速度较快，常常给浮点数分配两个字，一个给 α，一个给 γ.

我们将用同一符号 t 来表示定点数及浮点数的精度，并