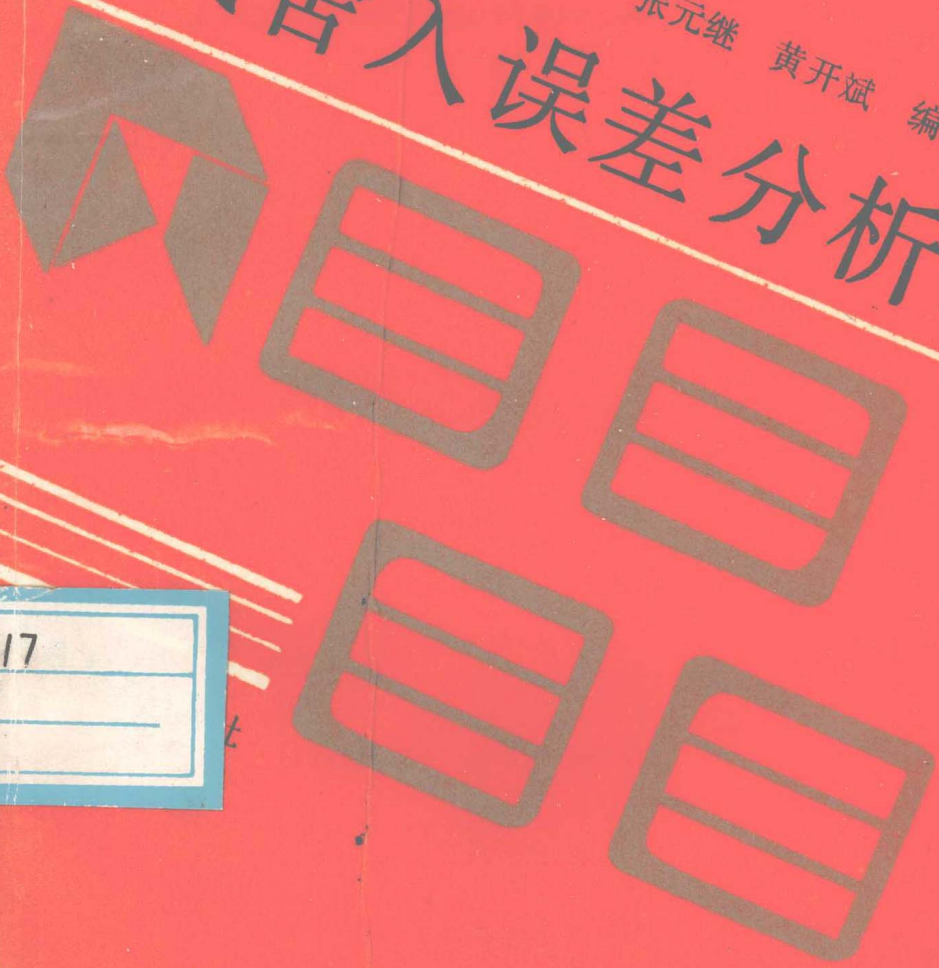
现代舍入误差分析

张元继　黄开斌　编著



现代舍入误差分析

张元继　黄开斌　编著

南京大学出版社

1990·南京

内 容 简 介

全书共6章.第1章介绍现代舍入误差分析的基础知识；第2—5章讨论解线性方程组、计算矩阵特征值和求多项式零点这三类问题的性态，并分析所述算法的数值稳定性和数值结果的精度；第6章介绍F. Stummel研究的向前误差分析方法.每章附有一定数量的习题和主要参考文献.

本书叙述深入浅出，并较多地介绍一些近代的研究成果，使读者能较系统地掌握现代舍入误差分析的理论和方法.可作为计算数学专业本科生和研究生教材，也可供从事工程与科学计算的科技工作者参考.

现代舍入误差分析

张元继　黄开斌　编著

何旭初　校

*

南京大学出版社出版

（南京大学校内）

江苏省新华书店发行　　丹徒县印刷厂印刷

*

开本787×1092　1/32　　印张8.625　字数194千

1990年5月第1版　　1990年5月第1次印刷

印数1—1500

ISBN 7-305-00223-2/O·18

定　价　1.90元

序　言

随着电子计算机的问世和发展，使得执行含有大量算术运算的计算成为可能。这不仅推动计算数学的飞速发展，而且进一步激励人们去研究舍入误差的累积对计算结果的影响，终使现代误差分析成为计算数学的一个重要分支。

在电子计算机上算题，为了得到符合精度要求的数值结果，必须研究该计算问题的性态和所用算法的数值稳定性。现代舍入误差分析研究的这两大课题是算法研究的理论基础之一，也是构造新的有效算法的一个重要依据。

现代舍入误差分析的研究始于J. Von Neumann和H. H. Goldstine的经典论文：Numerical Inverting of Matrices of High Order。至60年代中期，J.H.Wilkinson对其作出了卓越的贡献，形成了向后误差分析的方法。他对数值代数中的主要算法作了误差分析，并鉴定了它们的优劣。其专著“Rounding Errors in Algebraic Processes”（1963）和“The Algebraic Eigenvalue Problem”（1965）被公认为是这方面的经典著作。70年代以来，人们开始用向后误差分析方法探讨计算数学其他领域计算问题的性态和一些算法（包括并行算法）的数值稳定性。同时将舍入误差分析列入理工科有关专业研究生基础课和大学生选修课内容。80年代起，F·Stummel 在研究向前误差分析上做出了一系列的成果，为人们研究现代舍入误差分析又提供

了一个新的方法。

本书是根据我们近8年来在南京大学、南京师范大学两校数学系计算数学专业开设研究生基础课、高年级学生选修课所编写的讲义《误差分析》和编著者与合作者近几年的研究工作基础上编写而成。初稿还在一些兄弟院校交流过。本书假定读者已具备一定的线性代数与数值分析的基础知识。全书共六章：第一章介绍现代舍入误差分析的基础知识，它是学习以后各章的基础；第二章到第五章是讨论解线性代数方程组、矩阵特征值计算和求多项式零点这三类计算问题的性态，并分析所述算法的数值稳定性和数值结果的精度；第六章介绍F·Stummel研究的向前误差分析方法。每章附有一定数量的习题和主要参考文献目录。

在编写过程中，我们力求做到叙述深入浅出，尽可能多地介绍一些近代研究成果，使读者能较系统地掌握现代舍入误差分析的理论和方法，也衷心希望对从事科学与工程计算的同志们在他们的研究工作中能有所裨益。

何旭初教授在本书的编写过程中给予了直接指导，徐利治教授、徐家福教授对初稿提了宝贵意见，宋永忠、李治林同志设计了部分习题，罗亮生、李治林等同志在使用原讲义进行教学时，提了不少有益的建议，朱云霖同志誊清了原稿，赵金熙副教授、罗亮生、李治林、陈金如等同志帮助校订了清样，在此一并表示诚挚的谢意。

限于水平，书中错误在所难免，恳请读者不吝指正。

编 著 者

1988年12月于南京

目　录

第 1 章

基本浮点运算的误差分析

如所周知，微积分学中严格的极限理论是建立在实数理论基础上的．实数系是一个无限的、稠密的、连续的集合，通称它为连续统．在实数系中，实数的算术运算满足交换律、结合律和分配律．

然而，我们在某一数字电子计算机上算题时，只能使用该计算机所能表示的数的全体，它是一个有限的、离散的且其元素分布不均匀的集合．尤其要指出的是，在此集合中，实数的算术运算律往往不复成立．

为了弄清一些数值方法在计算机上的实现情况，以及分析计算结果的误差，本章中我们首先介绍机器数系，然后分析基本浮点运算的舍入误差．本章内容是分析各种算法在计算机上实现时产生的舍入误差的基础．

本书只讨论施行浮点运算的有关算法的误差分析．

§1 机器数系

任一计算机的字长 t 均为一确定的正整数．在 β 进制计算机中，浮点数 x 可以表示成下列形式

$$x = \pm\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t}\right)\beta^b, \qquad (1.1)$$

其中，b称为阶码，$0.d_1d_2\cdots d_t$称为尾数，诸$d_i(i=1,2,\cdots,t)$在集合$\{0,1,2,\cdots,\beta-1\}$中取值。t是尾数的位数，即该计算机的字长。对于(1.1)中的非零数x，当$d_1\neq 0$时称为规格化浮点数。随计算机的不同，阶码b有不同的上限U与下限L：

$$L\leqslant b\leqslant U.$$

定义1.1 计算机中形如(1.1)的浮点数x的全体组成的集合，称为该计算机的机器数系，以F记之。

显然F是由该计算机的四个参数β，t，L，U所决定的，即$F=F(\beta,t,L,U)$。

本书规定，数零用零尾数表示，即$d_1=d_2=\cdots=d_t=0$。

为使数的浮点表示唯一和保证所用数的精度，以后我们常将非零数表示成规格化浮点数。

有趣的是，在集合F中存在非零数，将其规格化后不再属于F（本章习题1）。

在计算机上算题时，要求所用初始数据及计算过程的中间结果和最后结果的绝对值，均不大于F中的最大正数。否则，计算机会因无法按规定表示它们，而发生所谓的溢出。

§2 数据的机器数近似

对给定的计算问题，在计算机上进行计算时，要求所给的初始数据和中间结果的绝对值，均不得超过F中的最大正数。即使如此，由于F是一个离散的有限集合，上述数据仍有可能不属于F，所以存在着必须用F中的浮点数近似表示它们的问题。设实数$x\in F$，在F中取x的机器数近似，

记之为 $\mathrm{fl}(x)$。显然，我们希望 $\mathrm{fl}(x)$ 具有最佳逼近的性质：

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

对十进制系统，当计算机字长为 t 时，常常可以用四舍五入法来确定非零实数 x 的机器数近似 $\mathrm{fl}(x)$。设

$$x=10^b\cdot a,\qquad \frac{1}{10}\leqslant|a|<1,$$

其中

$$|a|=0.d_1d_2\cdots d_td_{t+1}\cdots,\quad d_1\in\{1,2,\cdots,9\},$$
$$d_i\in\{0,1,\cdots,9\},\qquad i=2,3,\cdots.$$

取

$$\mathrm{fl}(x)=\mathrm{sign}(x)a'\cdot 10^b,$$

$$a'=\begin{cases}0.d_1d_2\cdots d_t, & \text{若 } 0\leqslant d_{t+1}\leqslant 4,\\ 0.d_1d_2\cdots d_t+10^{-t}, & \text{若 } d_{t+1}\geqslant 5.\end{cases}$$

显然，用 $\mathrm{fl}(x)$ 近似表示 x，其绝对误差 $\mathrm{fl}(x)-x$ 和相对误差 $[\mathrm{fl}(x)-x]/x$ 分别满足不等式

$$|\mathrm{fl}(x)-x|\leqslant\frac{1}{2}\cdot 10^{b-t},\tag{1.2}$$

$$|\varepsilon|=\left|\frac{\mathrm{fl}(x)-x}{x}\right|\leqslant\frac{\frac{1}{2}\cdot 10^{b-t}}{10^b\cdot|a|}\leqslant 5\cdot 10^{-t}.\tag{1.3}$$

(1.3)可改写成

$$\mathrm{fl}(x)\equiv x(1+\varepsilon),\qquad |\varepsilon|\leqslant 5\cdot 10^{-t},\tag{1.4}$$

其中 ε 是 $\mathrm{fl}(x)$ 的相对误差。我们通常用相对误差来刻划计算结果的精度。

对二进制系统，设

$$x = 2^b \cdot a, \quad \frac{1}{2} \leqslant |a| < 1,$$

$$|a| = 0.d_1 d_2 \cdots d_t d_{t+1} \cdots, \quad d_1 = 1, \quad d_i = 0 \text{或} 1, \quad i = 2, 3, \cdots$$

取

$$\mathrm{fl}(x) = \mathrm{sign}(x) a' \cdot 2^b,$$

其中

$$a' = \begin{cases} 0.d_1 d_2 \cdots d_t, & \text{若} d_{t+1} = 0, \\ 0.d_1 d_2 \cdots d_t + 2^{-t}, & \text{若} d_{t+1} = 1. \end{cases}$$

此时有

$$\mathrm{fl}(x) = x(1+\varepsilon), \quad |\varepsilon| \leqslant 2^{-t}, \tag{1.5}$$

其中ε是用$\mathrm{fl}(x)$近似表示x时的相对误差.

定义1.2 称$\mathrm{eps} = \frac{1}{2}\beta^{1-t}$为计算机的相对精度.

§3 基本浮点算术运算的舍入误差

基本浮点算术运算在计算机上的实现，随计算机而异. 本节，我们假定计算机字长为t，且具有双精度（双字长）累加器. 下面分别讨论在计算机上对两个规格化浮点数

$$x_i = 2^{b_i} \cdot a_i (\text{或} 10^{b_i} \cdot a_i), \quad i = 1, 2. \tag{1.6}$$

或零进行基本浮点算术运算的情况. 以记号$\mathrm{fl}(x_1 * x_2)$表示x_1与x_2在计算机上进行$*$运算（$* \in \{+, -, \times, /\}$）算得的结果，并分析其误差.

3.1 加法（包括减法）

在计算机中，x_1和x_2相加时，若$x_1 x_2 \neq 0$，先将它们的小数点对齐，我们称此为对阶. 不失一般性，可设x_1是模较大

的数．此时，需计算整数b_1-b_2．

（1）若$b_1-b_2>t$，这时x_2对精确和x_1+x_2的前$t+1$位数字没有任何影响，此时，计算机规定

$$\mathrm{fl}(x_1+x_2)\equiv x_1. \tag{1.7}$$

$\mathrm{fl}(x_1+x_2)$的绝对误差为

$$\Delta\equiv\mathrm{fl}(x_1+x_2)-(x_1+x_2)=-x_2,$$

其相对误差为

$$\varepsilon\equiv\frac{\Delta}{x_1+x_2}=\frac{-x_2}{x_1+x_2}.$$

显然，两者的界分别为

$$|\Delta|\leqslant 2^{b_1}\cdot\frac{1}{2}\cdot 2^{-t}\quad(\text{或}10^{b_1}\cdot\frac{1}{2}\cdot 10^{-t}),$$

$$|\varepsilon|\leqslant 2^{-t}(\text{或}\frac{1}{2}\cdot 10^{1-t}).$$

例1　设$t=4$，计算$10^2(0.4512)+10^{-8}(0.6973)$．

此时$b_1-b_2>t$，按(1.7)，有

$$\mathrm{fl}[10^2(0.4512)+10^{-8}(0.6973)]\equiv 10^2(0.4512).$$

（2）若$b_1-b_2\leqslant t$，先将x_2的尾数a_2向右移b_1-b_2位，即对阶．然后精确地计算$a_1+2^{b_2-b_1}a_2$（或$a_1+10^{b_2-b_1}a_2$）．注意，运算是在双精度累加器中进行的．此精确和的尾数a满足关系

$$0\leqslant|a|\leqslant|a_1|+2^{b_2-b_1}|a_2|$$

$$(\text{或}|a_1|+10^{b_2-b_1}|a_2|)<1+1.$$

若$|a|\geqslant 1$，将此尾数乘以2^{-1}（或10^{-1}），同时精确和的阶码加1，以使其成为$2t$位的规格化浮点数；若$0<|a|<\frac{1}{2}$（或

$\frac{1}{10}$），则将此尾数乘以2(或10)的适当次幂，同时阶码作相应的改变，以使此精确和成为2t位的规格化浮点数，最后，将此精确和舍入成t位规格化浮点数，此数即为x_1与x_2在计算机上算得的和$\mathrm{fl}(x_1+x_2)$；若$a=0$，则$\mathrm{fl}(x_1+x_2)\equiv 0$。

现再举几例，仍设计算机字长为4。

例2　$10^4(0.9628)+10^3(0.4976)$。

如上所述，加法按以下方式进行

$$\begin{array}{r} 10^4(0.9628\ 0000) \\ +)\,10^4(0.0497\ 6000) \\ \hline 10^4(1.0125\ 6000) \end{array}$$

将此精确和$10^4(1.0125\ 6000)$规格化，得$10^5(0.1012\ 5600)$，再将其尾数舍入到t位，则算得的和

$$\mathrm{fl}(x_1+x_2)\equiv 10^5(0.1013).$$

例3　$10^4(0.9628)+10^3(0.3716)$。

此时有

$$\begin{array}{r} 10^4(0.9628\ 0000) \\ +)\,10^4(0.0371\ 6000) \\ \hline 10^4(0.9999\ 6000) \end{array}$$

精确的和为$10^4(0.9999\ 6)$，而算得的和$\mathrm{fl}(x_1+x_2)$为$10^5(0.1000)$。注意，$\mathrm{fl}(x_1+x_2)$的阶码比x_1+x_2的大1。

例4　$10^{-3}(0.1003)+10^{-4}(-0.9974)$。

因为

$$\begin{array}{r} 10^{-3}(0.1003\ 0000) \\ -)\,10^{-3}(0.0997\ 4000) \\ \hline 10^{-3}(0.0005\ 6000) \end{array}$$

将精确和规格化，得$10^{-6}(0.5600\ 0000)$，再将其尾数舍入成t位，于是算得的和$\mathrm{fl}(x_1+x_2)$为$10^{-6}(0.5600)$。

由此例可看出，两个符号相反而绝对值相近的数相加

时，和可能损失一定数位的有效数字，我们称这种现象为相消。但是在例4中，$fl(x_1+x_2)\equiv x_1+x_2$，也就是说，算得的和$fl(x_1+x_2)$仍是精确的。其原因是使用了双精度累加器。我们强调这一点，是为了指出，若只用单精度累加器，相消往往导致精度的降低（参见本章§5）。

若x_1,x_2中有一为零，不妨设$x_1=0$，此时$fl(x_1+x_2)\equiv x_2$。若x_1,x_2均为零，则$fl(x_1+x_2)\equiv 0$。

下面分析$fl(x_1+x_2)$的误差。

定理1.1 设计算机字长为t，并有双精度累加器，又设x_1,x_2为F中的规格化浮点数或零，则有

$$fl(x_1+x_2)\equiv(x_1+x_2)(1+\varepsilon), \tag{1.8}$$

其中ε是满足条件

$$|\varepsilon|\leqslant 2^{-t}\ (\text{或}\ \frac{1}{2}\cdot 10^{1-t}) \tag{1.9}$$

的某个数。

证明 若x_1,x_2有一或全为零，则有

$$fl(x_1+x_2)\equiv x_1+x_2,$$

(1.8)，(1.9)成立，且$\varepsilon=0$。

对非零数x_1,x_2，不妨仍设$|x_1|\geqslant|x_2|$。

（1）当$b_1-b_2>t$时，定理显然成立。

（2）当$b_1-b_2\leqslant t$时，如上所述，只在将规格化后的精确和

$$x_1+x_2=2^{b_3}\cdot a_3\ (\text{或}10^{b_3}\cdot a_3)$$

舍入成t位规格化浮点数$fl(x_1+x_2)$时有误差，其绝对误差界为

$$2^{b_3}\cdot\frac{1}{2}\cdot 2^{-t}\ (\text{或}10^{b_3}\cdot\frac{1}{2}\cdot 10^{-t}).$$

若$x_1+x_2\neq 0$，因为$|x_1+x_2|\geqslant 2^{b_3}\cdot\frac{1}{2}$(或$10^{b_3}\cdot\frac{1}{10}$)，故$\mathrm{fl}(x_1+x_2)$的相对误差界为

$$2^{-t}\text{(或}\frac{1}{2}\cdot 10^{1-t}\text{)},$$

若$x_1+x_2=0$，显然$\varepsilon=0$．定理证毕．

我们将上面讨论的结果重述如下．x_1与x_2算得的和$\mathrm{fl}(x_1+x_2)$是$x_1(1+\varepsilon)$与$x_2(1+\varepsilon)$两数的精确和，其中ε是满足(1.9)的某个数．换言之，$\mathrm{fl}(x_1+x_2)$是原计算问题x_1+x_2的某个扰动问题$x_1(1+\varepsilon)+x_2(1+\varepsilon)$的精确和，而$\varepsilon$为对$x_1,x_2$的相对扰动．这种分析误差的方法，实质上是将计算过程中舍入误差对计算结果的影响，等价返回到对初始数据的相对扰动．注意，此处ε既为算得的和$\mathrm{fl}(x_1+x_2)$的相对误差，又为对x_1,x_2的相对扰动．

3.2 乘法

在计算机上，x_1与x_2的乘法运算如下：先将阶码b_1与b_2相加得b_3，然后计算尾数a_1与a_2的精确的$2t$位积a_1a_2，它满足

$$\frac{1}{2^2}\leqslant|a_1a_2|<1\quad\left(\text{或}\frac{1}{10^2}\leqslant|a_1a_2|<1\right).\qquad(1.10)$$

如有必要，在左移尾数的同时改变阶码，以使所得的精确积规格化，最后，将它舍入成t位规格化浮点数，此数即为由x_1，x_2算得的积$\mathrm{fl}(x_1\times x_2)$．

举两例如下，仍设计算机字长为4．

例5　$10^{-2}(0.7631)\times 10^{4}(0.4512)$．

两阶码相加得2，两尾数相乘的精确积为0.3443　1072，再将尾数舍入成四位数，即得算得的积

$$10^{2}(0.3443).$$

例6　$10^4(0.1314)\times 10^{-1}(0.1026)$.

此时

$$b_1+b_2=3,\quad a_1\times a_2=0.0134\ 8164,$$

于是

$$\mathrm{fl}[(10^4\times 0.1314)\times(10^{-1}\times 0.1026)]\equiv 10^2(0.1348).$$

若x_1,x_2中有一或全为零，那么

$$\mathrm{fl}(x_1\times x_2)\equiv 0.$$

类似地，我们有

定理1.2　设计算机字长为t，且有双精度累加器，又设x_1,x_2为F中的规格化浮点数或零，则有

$$\mathrm{fl}(x_1\times x_2)\equiv x_1\times x_2(1+\varepsilon),\qquad (1.11)$$

其中ε为满足条件

$$|\varepsilon|\leqslant 2^{-t}\quad (\text{或}\frac{1}{2}\cdot 10^{1-t})\qquad (1.12)$$

的某个数.

现将该定理重述如下：算得的积$\mathrm{fl}(x_1\times x_2)$是$x_1(1+\varepsilon)$与$x_2$的，或$x_1$与$x_2(1+\varepsilon)$的，或$x_1(1+\varepsilon)^{\frac{1}{2}}$与$x_2(1+\varepsilon)^{\frac{1}{2}}$的精确积，其中$\varepsilon$是满足(1.12)的某个值. 我们可以用被认为是方便的任何方式，将因子$1+\varepsilon$分配给x_1或x_2或同时分配给它们二者. 这就是说，$\mathrm{fl}(x_1\times x_2)$是原计算问题$x_1\times x_2$的某一扰动问题的精确积.

3.3　除法

和通常一样，在计算机上做除法也不允许除数为零. x_1除以x_2是按下列方式实现的.

计算b_1减b_2得b_3，再将a_1置于双精度累加器中的前t位，后t位上置零. 比较$|a_1|$与$|a_2|$，若$|a_1|\geqslant|a_2|$，先将累加