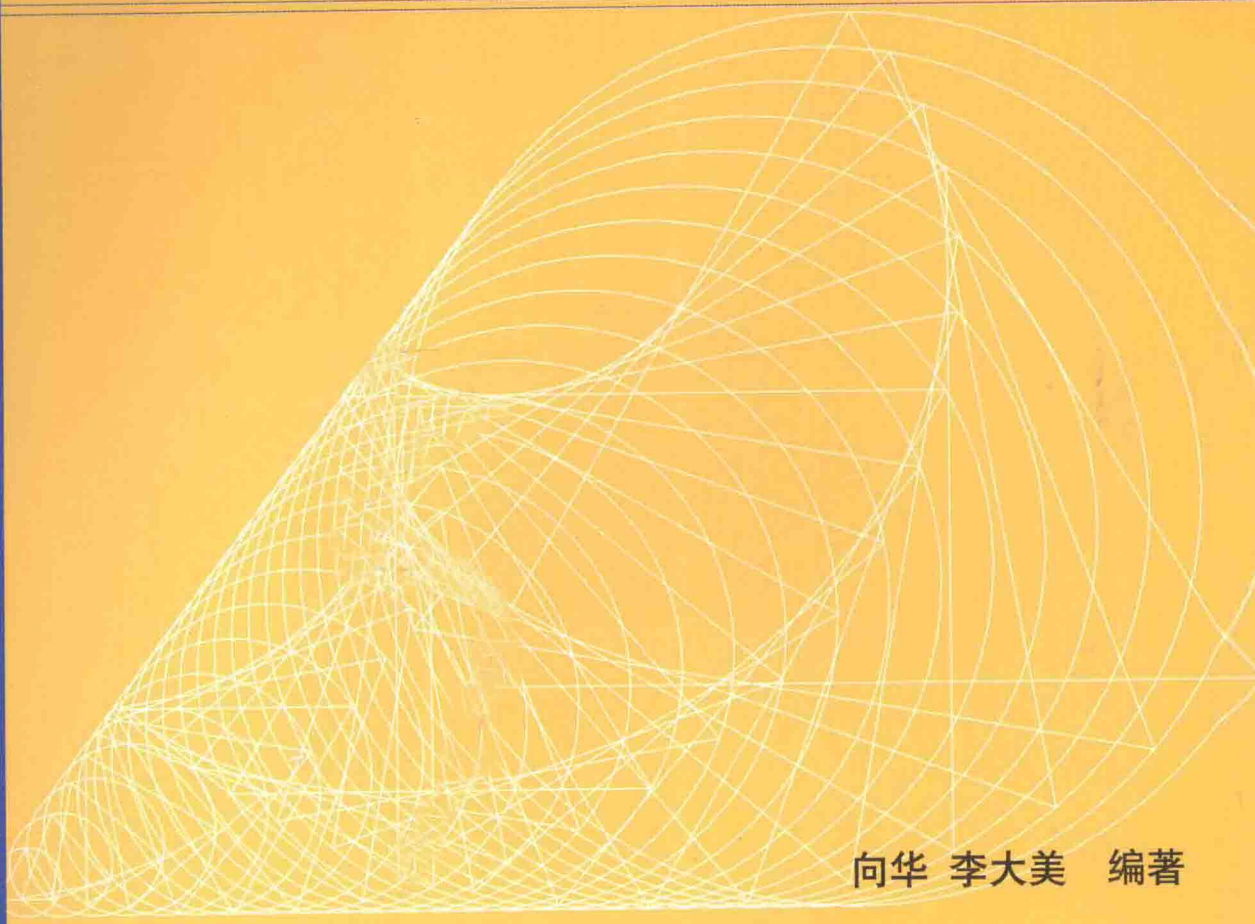
全国工程专业学位研究生教育国家级规划教材



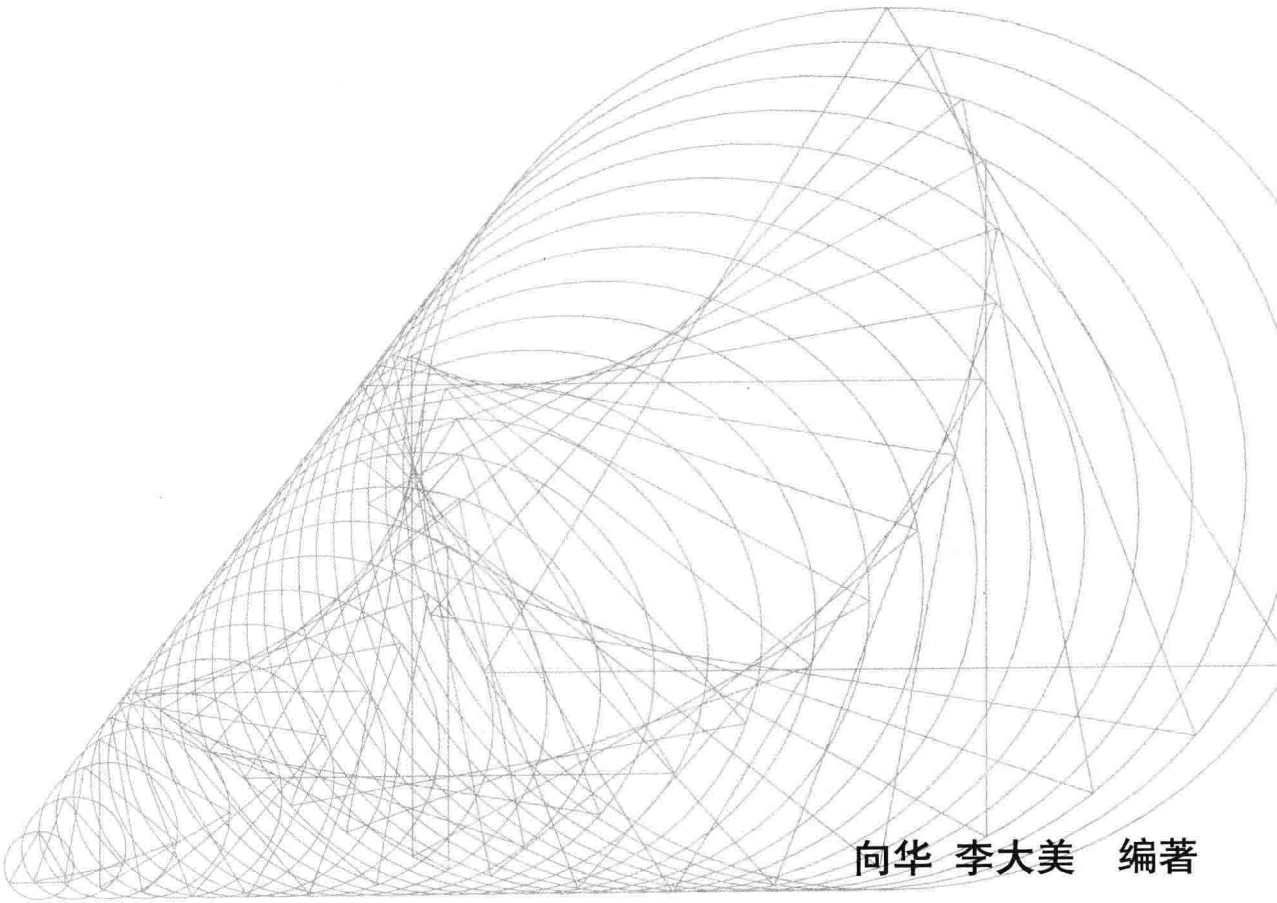
向华 李大美 编著

数值计算及其工程应用

http://www.tup.com.cn

清华大学出版社

全国工程[redacted]国家级规划教材



向华 李大美 编著

数值计算及其工程应用

清华大学出版社
北京

内 容 简 介

本教材主要是针对全国工程硕士专业学位研究生“数值分析”或“数值计算”课程的教学而编写的，特别针对各工程领域实际应用的特点，确定了教材的基本内容，其主导思想是：**“了解背景、掌握概念、注重原理、淡化推导、强调实现、突出应用.”**这也是该教材的主要特点，即介绍问题的工程背景，讲解基本概念和数学原理，介绍一般的数学理论和算法，淡化理论推导和纯粹的计算，重点讲授应用方法，借助于计算机和工具软实现算法，特别突出解决实际工程问题的实用性.

本书可作为相关各工程领域的工程硕士专业学位研究生“数值分析”或“数值计算”课程的教材，也可作为工科各专业的大学本科生和研究生的“数值分析”或“数值计算”课程教材或参考教材，也可供从事相关研究工作的工程技术人员参考之用.

图书在版编目(CIP)数据

数值计算及其工程应用/向华，李大美编著. --北京：清华大学出版社，2015
全国工程专业学位研究生教育国家级规划教材
ISBN 978-7-302-40177-3

Ⅰ. ①数… Ⅱ. ①向… ②李… Ⅲ. ①数值计算－研究生－教材 Ⅳ. ①O241

中国版本图书馆 CIP 数据核字(2015)第191784号

责任编辑：刘　颖
封面设计：何凤霞
责任校对：王淑云
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：15.5　　**字　　数**：377 千字
版　　次：2015 年 9 月第 1 版　　**印　　次**：2015 年 9 月第1 次印刷
印　　数：1～3000
定　　价：35.00 元

产品编号：057905-01

前言

全国工程专业学位研究生教育指导委员会根据各工程领域对专业人才的需求，提出了课程改革设想和指导性意见，旨在提高教学水平，提高学生解决工程实际问题的能力. 数值计算方法是一门重要的公共基础数学课程，其应用性很强，应用领域涉及各类工程、经济、管理、军事等，是许多工程技术人员和科研工作者的必备工具. 国内外已出版了不少优秀的教材，它们往往理论讲述得很深入. 本书注重理论和数值实验的结合，由浅入深地介绍算法的理论基础，同时也强调算法的实际应用，并配以具体的数值算例，引导初学者运用所学理论知识，尝试解决问题，逐步培养初学者的分析能力和动手能力. 俗话说，“光说不练假把式”，数值计算特别强调实践，一方面在计算机上编写程序实现算法，另一方面还要会运用所学的方法解决生产实践中的实际问题.

书中代码经过了精挑细选，几乎每行代码字斟句酌. “纸上得来终觉浅，绝知此事要躬行”，初学者最好对代码仔细阅读，用心体会，并且亲手在计算机上运行一遍，对结果有些定性分析. 在学习算例的基础上，逐步培养对算法和代码的感觉，培养分析问题、解决问题的能力，为后续工作打下一定基础. 本书代码主要用 MATLAB 编写，也有部分用 C 语言写成，我们也鼓励读者用 C 语言或 Fortran 语言完成算例.

全书分为 8 章. 第 1 章绪论是准备工作，涉及误差、范数和条件数等概念. 其他各章分别介绍各类问题的数值方法. 第 2 章讨论线性方程组的解法，包括列选主元 Gauss 消去法和经典迭代法，并介绍了以共轭梯度法为代表的近代方法；第 3 章主要是非线性方程(组)的 Newton 法；第 4 章介绍特征值问题的求解方法，特别是 QR 算法. 前面 4 章以数值代数为主，后面 4 章则主要是数值逼近与微分方程数值解方面的内容. 第 5 章讨论函数的插值与拟合，包括函数的最佳逼近；第 6 章介绍数值积分方法，如 Newton-Cotes 公式，Gauss 型求积公式等；第 7 章介绍常微分方程(组)的 Euler 方法，Runge-Kutta 法，线性多步法以及相容性、稳定性和收敛性等概念；第 8 章简要介绍三类偏微分方程的典型差分格式. 有少量内容的安排是为了满足部分基础扎实学生的学习要求. 由于课时的限制，可适当压缩一些内容，比如，第 2 章中 Gauss 消去法的误差分析、共轭梯度法、部分经典迭代法的收敛性证明，第 3 章中 Newton 法的变形，第 4 章中的隐式 Q 定理及相关内容，第 5 章中的最佳逼近，第 6 章中 Romberg 算法及二重积分，第 7 章中多步法稳定性的讨论，第 8 章中的一阶双曲方程组，等等.

本书的第 5,6 章主要由李大美撰写；第 1～4 章，第 7,8 章及 5.7 节由向华撰写，全书数值算例由向华选编并调试. 感谢研究生刘兵、申海伦、许雪敏、尹纯辉和张仕洋帮助输入了部分文字，感谢刘颖编辑大量细致的工作. 由于编者水平有限，书中难免有错误之处，欢迎广大读者批评指正.

向华　李大美

2015 年 1 月于珞珈山

目 录

第1章

绪　论

理论分析、科学试验与科学计算为认识自然的三大主要方式手段. 理论上，我们已经得到了描述宏观世界、微观世界和宇观世界运动的方程，比如：

$$F = ma,$$

$$\mathrm{i}\hbar \frac{\partial \psi}{\partial t} = H\psi$$

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \kappa T_{\mu\nu}.$$

它们分别对应于 Newton 定律、Schrodinger 方程和 Einstein 场方程，其中 $H = -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + V(x, y, z)$ 为 Hamilton 算符，$R_{\mu\nu}$ 和 $T_{\mu\nu}$ 分别为 Ricci 张量和能动张量，其他详见文献[21,19,44]. 这些方程极其优美，除少数特殊情形外，均不能解析求解，一般需要求助于数值方法在计算机上近似求解. 工程和科学领域的诸多问题越来越依赖于数值计算方法.

科学计算的基础为数值计算方法. 数值计算方法亦称为计算方法或数值分析，是研究用计算机解决数学问题的理论与方法. 针对生产实践和科学实验建立数学模型后，构造求解问题的数值方法，对方法的收敛性、稳定性和误差等进行理论分析，编写代码实现算法并在计算机上算出结果，最后对计算结果进行分析，总结规律或解释现象.

数值计算的一般策略为“化整为零，截弯取直，以简驭繁，化难于易”. 设法将复杂问题转化为相近的简单问题，方案包括：用线性问题代替非线性问题，如 Newton 法；用有限维空间代替无限维空间，如有限元法；用有限过程代替无限过程，如导数的差分近似；用简单函数代替复杂函数，如多项式逼近；用代数方程代替微分方程；用低阶方程组代替高阶方程组，等等.

按研究内容，数值分析传统上分为数值代数、数值逼近和微分方程数值解三大块. 随着计算机技术和数值方法的发展，科学计算领域也取得了长足的进展，比如在大规模和多尺度计算方面. 经过半个多世纪的探索，涌现了一大批优秀算法，尤其是 2000 年评出的 20 世纪十大算法，它们分别为：

(1) 1946 年，Monte Carlo 方法；

(2) 1947 年，单纯形法；

(3) 1950 年，Krylov 子空间方法；

(4) 1951 年,矩阵分解方法;

(5) 1957 年,优化的 Fortran 编译器;

(6) 1959—1961 年,矩阵特征值的 QR 算法;

(7) 1962 年,快速排序算法;

(8) 1965 年,快速 Fourier 变换(FFT);

(9) 1977 年,整数关系探测算法;

(10) 1987 年,快速多极算法(FMM).

这些优秀的算法对科学研究乃至人们日常生活都产生了深远的影响,其中一些算法将在本书中介绍.

1.1 误差的基本概念

数值方法涉及对原问题的近似,自然就引出误差的概念.广义的误差涉及模型误差、观测误差、截断误差和舍入误差,其中截断误差和舍入误差又称为计算误差,是我们这里主要关心的.

1. 计算误差

截断误差是数学模型的精确解与数值方法的近似解之间的误差,如用近似公式 $e^x \approx 1+x$ 计算指数函数时的截断误差为 $\frac{1}{2!}x^2 e^{\theta x}$ $(0<\theta<1)$. 舍入误差是由机器表示数时产生的,由于计算机字长有限,需对超过存储位数的数字进行舍入而产生的误差. 在数值代数问题中,舍入误差占主要地位;在微分方程数值解中主要考虑截断误差.

关于截断误差和舍入误差,下面以例子说明. 用有限差分近似导数,即

$$f'(x) \approx \frac{f(x+h)-f(x)}{h}, \tag{1.1}$$

直观上说步长 h 越小,近似越准确. 步长 h 能否趋于 0 呢?

由 Taylor 展开,有

$$f(x+h) = f(x) + f'(x)h + \frac{h^2}{2}f''(\theta), \quad \theta \in (x, x+h)$$

忽略其中的 $O(h^2)$ 项,可得近似式(1.1). 设 $M=\max|f''(x)|$,则(1.1)式的截断误差界为 $\frac{Mh}{2}$. 设计算单个函数值时的舍入误差为 ε,则(1.1)式的舍入误差界为 $\frac{2\varepsilon}{h}$. 结合此二者,总的计算误差界为

$$\frac{Mh}{2} + \frac{2\varepsilon}{h}.$$

当 h 较大时,截断误差较大;当 h 太小时,舍入误差占主要成分. 当 $h=2\sqrt{\frac{\varepsilon}{M}}$ 时,总的计算误差最小.

另外还经常用到绝对误差、相对误差和有效数字的概念. 设 x^* 为准确值,x 为 x^* 的近似. 定义绝对误差 $\Delta x = x^* - x$,相对误差 $\Delta_r x = \frac{x^*-x}{x^*}$;有时也将 $\frac{x^*-x}{x}$ 作为相对误差. 相

对误差是无量纲数，通常用百分比表示. 一般精确值 x^* 未知，Δx 不能给出；但往往可以估计其范围，确定一正数 ε，使得 $|\Delta x| \leqslant \varepsilon$，此时称 ε 为绝对误差限/界. 同样，若能找到正数 ε_r，使得 $|\Delta_r x| \leqslant \varepsilon_r$，则称 ε_r 为相对误差限.

近似值参加运算后所得的结果也是近似的，含有误差，这就是误差的传播. 考虑可微函数的求值，精确值为 $y^* = f(x_1^*, x_2^*, \cdots, x_n^*)$，自变量用近似值 $x_1, x_2, \cdots, x_n$ 进行计算，得函数值的近似值 $y = f(x_1, x_2, \cdots, x_n)$. 由多元函数 Taylor 公式，得绝对误差和相对误差分别为

$$\Delta y = y^* - y \approx \sum_{i=1}^{n} f_i'(x_1, x_2, \cdots, x_n)\Delta x_i, \quad \Delta x_i = x_i^* - x_i,$$

$$\Delta_r y = \Delta y / y \approx \sum_{i=1}^{n} \frac{x_i f_i'}{y} \frac{\Delta x_i}{x_i} = \sum_{i=1}^{n} \frac{x_i f_i'}{y} \Delta_r x_i, \quad \Delta_r x_i = \Delta x_i / x_i.$$

故有如下的误差限估计：

$$\varepsilon(y) \approx \sum_{i=1}^{n} |f_i'(x_1, x_2, \cdots, x_n)| \varepsilon(x_i),$$

$$\varepsilon_r(y) \approx \sum_{i=1}^{n} \left| \frac{x_i}{y} f_i'(x_1, x_2, \cdots, x_n) \right| \varepsilon_r(x_i).$$

例如，$\varepsilon\left(\frac{x_1}{x_2}\right) \approx \frac{1}{|x_2|}\varepsilon(x_1) + \frac{|x_1|}{x_2^2}\varepsilon(x_2)(x_2 \neq 0)$；$\varepsilon_r\left(\frac{x_1}{x_2}\right) \approx \varepsilon_r(x_1) + \varepsilon_r(x_2)(x_1 x_2 \neq 0)$.

有效数字为从第一个非零数字开始至末尾的所有数字，其位数与小数点位置无关. 下面说明有效数字的个数反映精确程度(简称精度). 设实数 x^*，经四舍五入后的近似值 x 有 n 位有效数字，表示为如下标准形式：

$$x = \pm 0.a_1 a_2 \cdots a_n \times 10^m = \pm (a_1 \times 10^{-1} + a_2 \times 10^{-2} + \cdots + a_n \times 10^{-n}) \times 10^m,$$

其中 $a_1 \neq 0, a_i \in \{0, 1, 2, \cdots, 9\}, i = 1, 2, \cdots, n$.

x 近似 x^* 的绝对误差、相对误差与有效数字有如下关系：注意到最后一位 a_n 是四舍五入得到的，故绝对误差限

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

从而，相对误差限

$$|\Delta_r x| = \frac{|x^* - x|}{|x^*|} \leqslant \frac{\frac{1}{2} \times 10^{m-n}}{0.a_1 \times 10^m} = \frac{1}{2a_1} \times 10^{-(n-1)}.$$

故有效数字越多，相对误差限就越小，近似数的精度越高.

若 x 的相对误差限为

$$|\Delta_r x| \leqslant \frac{1}{2(a_1 + 1)} \times 10^{-(n-1)},$$

则绝对误差限

$$|\Delta x| = |x^*| \cdot |\Delta_r x| \leqslant |(0.a_1 + 0.1) \times 10^m| \frac{1}{2(a_1 + 1)} \times 10^{-(n-1)} = \frac{1}{2} \times 10^{m-n}.$$

所以 x 至少具有 n 位有效数字.

2. 舍入误差

由于计算机只能有限精度地表示实数，从而产生舍入误差. 具体地说，实数在计算机中

用浮点数以如下形式表示：

$$\pm d_0.d_1d_2\cdots d_{p-1}\times\beta^E,$$

这里 β 为基底(一般 $\beta=2$，即二进制数)，p 反映精度(下文将解释)，指数 E 的范围为 $L\leqslant E\leqslant U$. 约定 $d_0=1$，无须存储；如此表示的浮点数称为正规化数，可由四个整数表征，记为 $F(\beta,p,L,U)$.

浮点数分为单精度和双精度. IEEE 754 标准中单精度浮点数由 32 位(4 个字节)存储，其中1 位存符号，8 位存指数部分，其余 23 位存尾数部分 $d_i(1\leqslant i\leqslant p-1)$；双精度浮点数由 64 位(8 个字节)存储，其中 1 位存符号，11 位存指数部分，52 位存尾数(注意指数部分存储时要加上偏移量：单精度浮点数为 127，双精度浮点数为 1023). IEEE 754 标准中单精度浮点数与双精度浮点数对应的4 个参数如表 1.1 所示.

表 1.1 IEEE 754 标准中单精度浮点数与双精度浮点数的对应参数

	β	p	L	U
单精度	2	24	−126	127
双精度	2	53	−1022	1023

显然，能准确表示的实数是有限的，我们可导出所能精确表示的数的个数，以及能表示的最大或最小实数；并且容易知道在区间$[2^e,2^{e+1}]$上的浮点数是等距的，间距 $\Delta x=2^{-p+1}\times 2^e$. 特别地，1 与右边第 1 个数的距离 $\text{ulp}=2^{-p+1}$. 双精度数 $\text{ulp}=2^{-52}\approx 2.22\times 10^{-16}$，相当于 MATLAB 中的常量 eps.

表示一个实数 $x\in(2^e,2^{e+1})$，我们可以有两种方式. 第一种是截断，亦称向零舍入. 这时表示 x 的相对误差为

$$\frac{\Delta x}{x}<2^{-p+1}=\text{ulp}.$$

第二种方式是最近舍入，取与 x 最接近的浮点数(四舍五入). 这时表示 x 的相对误差为

$$\frac{\frac{\Delta x}{2}}{x}<2^{-p}=\frac{1}{2}\text{ulp}.$$

设 $fl(x)$为表示 x 的浮点数，定义机器精度为用浮点数表示一个非零实数 x 的最大可能相对误差，即

$$\frac{|fl(x)-x|}{|x|}\leqslant\varepsilon_{\text{mach}}.$$

当用最近舍入时 $\varepsilon_{\text{mach}}=2^{-p}$；用截断时 $\varepsilon_{\text{mach}}=2^{-p+1}$. 我们有时也用下面的式子：

$$fl(x)=x(1+\delta),\quad |\delta|\leqslant\varepsilon_{\text{mach}}.$$

上面提到的仅是正规化数，另外次正规化数，± 0，$\pm\infty$，以及各类中断需特殊表示. 图 1.1 标出了 $F(2,3,-1,1)$表示的正规化数(图中 0 和 4 除外).

图 1.1

3. 减少误差的原则

由于涉及计算误差，数学(理论)上等价的问题，数值上并不等价. 在实际计算中要注意控制误差，下面是减少误差的几个原则.

(1) 避免两个相近的数相减.

两个相近数相减可造成有效数字大量丢失，这就是所谓的灾难性相消(castraphy cancellation).

比如，计算

$$f(x)=\frac{1-\cos x}{x^2},$$

取 $x=1.2\times10^{-5}$，保留 10 位有效数字，$c=\cos x=0.99\cdots99$，$1-c=0.00\cdots01$，

$$\frac{1-c}{x^2}=\frac{10^{-10}}{1.44\times10^{-10}}=0.6944.$$

利用 $\cos x=1-2\sin^2\frac{x}{2}$，将计算式改写为

$$f(x)=\frac{1}{2}\left(\frac{\sin\frac{x}{2}}{\frac{x}{2}}\right)^2,$$

可得 $f(1.2\times10^{-5})\approx0.5$，按此式计算可避免相近的数相减.

又如，一元二次方程 $ax^2+bx+c=0$ 的求根问题. 当 $b^2\gg4|ac|$ 时，$\sqrt{b^2-4ac}\approx|b|$，可按以下公式计算：

$$x_1=\frac{-b-\operatorname{sgn}(b)\sqrt{b^2-4ac}}{2a},\quad x_2=\frac{c}{ax_1}.$$

否则，有一个根的计算不可靠.

又如，求标准误差的公式

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. 此公式需两次遍历数据：一次求平均，一次计算标准差. 如果改用数学上等价的公式

$$S_n^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right],$$

则仅需一次遍历，但是其中相减的两个量一般都较大且接近，从而产生严重相消(甚至为负，使开方运算失效).

(2) 避免大数吃小数. 比如，计算 $s=A+\delta_1+\delta_2+\cdots+\delta_n$，这里 A 的绝对值很大而 $\delta_i(i=1,2,\cdots,n)$ 的绝对值很小. 比如，$A=1000$，$\delta_i=0.001(i=1,2,\cdots,1000)$，假设用十进制机器，以 4 位尾数和 1 位指数存储数据. 如果从左至右依次相加，则每个 δ_i 均被吃掉；更好的方法是改变运算次序，先加 δ_i，最后加 A. 同样的原则我们可以计算出

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi}{6}=1.644934066848\cdots$$

(3) 避免绝对值小的数作除数. 这时容易放大被除数中的微小误差，甚至导致向上溢

出. 后文将看到,用 Gauss 消去法求解线性方程组时,选主元策略可以避免出现小除数.

(4) 简化计算步骤. 比如用秦九韶(Horner)算法计算多项式 $p_n(x)=a_nx^n+\cdots+a_1x+a_0$ 的函数值. 定义

$$S_n=a_n,\quad S_k=xS_{k+1}+a_k(k=n-1,\cdots,0),$$

则 $p_n(x)=S_0$.

(5) 选用数值稳定的计算公式. 如计算定积分 $I_n=\mathrm{e}^{-1}\int_0^1 x^n\mathrm{e}^x\mathrm{d}x(n=0,1,\cdots,100)$. 利用分部积分法得递推公式

$$I_n=1-nI_{n-1}(n=1,2,\cdots,100),\quad I_0=1-\mathrm{e}^{-1}.$$

取 $\widetilde{I}_0$ 为初始值,比如 $\widetilde{I}_0=0.6321$,与 I_0 的误差不超过 $\frac{1}{2}\times10^{-4}$. 按上述递推式计算,结果记为 $\widetilde{I}_n$,则满足关系

$$\widetilde{I}_n-I_n=-n(\widetilde{I}_n-I_{n-1})=\cdots=(-1)^n n!(\widetilde{I}_0-I_0).$$

初始误差会随着计算步数的增加而迅速扩大,最终将湮没真实解. 该计算式不能控制误差的传播,是数值不稳定的. 注意到

$$\frac{\mathrm{e}^{-1}}{n+1}=\mathrm{e}^{-1}(\min_{0\leqslant x\leqslant1}\mathrm{e}^x)\int_0^1 x^n\mathrm{d}x<I_n<\mathrm{e}^{-1}(\max_{0\leqslant x\leqslant1}\mathrm{e}^x)\int_0^1 x^n\mathrm{d}x=\frac{1}{n+1}.$$

取估计式 $\widetilde{I}_n=\frac{1+\mathrm{e}^{-1}}{2(n+1)}$,按 $I_{n-1}=\frac{1}{n}(1-I_n)$ 递推则是数值稳定的.

1.2 向量范数与矩阵范数

后文针对线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$,讨论系数矩阵和右端向量的误差对解向量的影响,需要对向量和矩阵进行度量,故需引入向量范数和矩阵范数的概念.

线性空间 $\mathbb{R}^n$ 上向量 $\boldsymbol{x}$ 的范数是满足下面 3 个条件的非负实数 $\|\cdot\|$:

(1) $\|\boldsymbol{x}\|\geqslant0$,当且仅当 $\boldsymbol{x}=\boldsymbol{0}$ 时,$\|\boldsymbol{x}\|=0$;

(2) $\|\alpha\boldsymbol{x}\|=|\alpha|\,\|\boldsymbol{x}\|$,$\forall\alpha\in\mathbb{R}$;

(3) $\|\boldsymbol{x}+\boldsymbol{y}\|\leqslant\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$,$\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$.

向量范数 $\|\boldsymbol{x}\|$ 是 $\mathbb{R}^n$ 上向量 $\boldsymbol{x}$ 的连续函数,一个重要性质是 $\mathbb{R}^n$ 上所有向量范数等价. 向量范数 $\|\cdot\|$ 和 $\|\cdot\|_*$ 等价的含义是,对 $\forall\boldsymbol{x}\in\mathbb{R}^n$,存在常数 $c_1,c_2>0$,使

$$c_1\|\boldsymbol{x}\|\leqslant\|\boldsymbol{x}\|_*\leqslant c_2\|\boldsymbol{x}\|.$$

如下是几种常用的向量范数:

2-范数,$\|\boldsymbol{x}\|_2=\left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}}$;

1-范数,$\|\boldsymbol{x}\|_1=\sum_{i=1}^n|x_i|$;

∞-范数,$\|\boldsymbol{x}\|_\infty=\max_{0\leqslant i\leqslant n}|x_i|$;

p-范数(Hölder 范数),$\|\boldsymbol{x}\|_p=\left(\sum_{i=1}^n|x_i|^p\right)^{\frac{1}{p}}$,$1\leqslant p<\infty$;

能量范数 $\|\boldsymbol{x}\|_A=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Ax})^{\frac{1}{2}}$,其中 $\boldsymbol{A}$ 为对称正定矩阵.

$\mathbb{R}^{n\times n}$中矩阵范数为满足以下条件的非负实数$\|\cdot\|$：

(1) $\|\boldsymbol{A}\|\geqslant 0,\forall \boldsymbol{A}\in\mathbb{R}^{n\times n}$；当且仅当$\boldsymbol{A}=\boldsymbol{0}$时，$\|\boldsymbol{A}\|=0$.

(2) $\|\alpha\boldsymbol{A}\|=|\alpha|\,\|\boldsymbol{A}\|,\forall \boldsymbol{A}\in\mathbb{R}^{n\times n},\alpha\in\mathbb{R}$.

(3) $\|\boldsymbol{A}+\boldsymbol{B}\|\leqslant\|\boldsymbol{A}\|+\|\boldsymbol{B}\|,\forall \boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{n\times n}$.

由这个定义不难推广至$\mathbb{R}^{m\times n}$. 如果除此以外还满足下面第(4)条，则称为相容的矩阵范数.

(4) $\|\boldsymbol{AB}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{B}\|,\forall \boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{n\times n}$.

下面通过已知向量范数来定义矩阵范数

$$\|\boldsymbol{A}\|=\max_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|}=\max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{Ax}\|,$$

称为诱导的矩阵范数(又称为从属于向量范数的矩阵范数)，它自然满足上述矩阵范数定义中的4条.

我们有下面3种常用的矩阵范数：

$$\|\boldsymbol{A}\|_\infty=\max_i\sum_{j=1}^n|a_{ij}|,$$

$$\|\boldsymbol{A}\|_1=\max_j\sum_{i=1}^n|a_{ij}|,$$

$$\|\boldsymbol{A}\|_2=\sqrt{\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})},$$

这里$\lambda_{\max}(\cdot)$表示矩阵的最大特征值.

下面导出∞-范数的表达式. 令$\mu=\max\limits_{1\leqslant i\leqslant n}\sum\limits_{j=1}^n|a_{ij}|$，由向量范数的定义，有

$$\|\boldsymbol{Ax}\|_\infty=\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^n a_{ij}x_j\right|\leqslant\max_{1\leqslant i\leqslant n}\sum_{j=1}^n|a_{ij}||x_j|$$

$$\leqslant\max_{1\leqslant i\leqslant n}\left(\max_{1\leqslant j\leqslant n}|x_j|\sum_{j=1}^n|a_{ij}|\right)=\max_{1\leqslant j\leqslant n}|x_j|\max_{1\leqslant i\leqslant n}\left(\sum_{j=1}^n|a_{ij}|\right)=\|\boldsymbol{x}\|_\infty\mu.$$

由∞-范数的定义，$\|\boldsymbol{A}\|_\infty=\max\limits_{\|\boldsymbol{x}\|_\infty=1}\|\boldsymbol{Ax}\|_\infty$，所以$\|\boldsymbol{A}\|_\infty\leqslant\mu$. 下面说明等号可以取到. 设$\boldsymbol{A}$的第$k$行元素绝对值之和等于$\mu$，即

$$\sum_{j=1}^n|a_{kj}|=\max_{1\leqslant i\leqslant n}\sum_{j=1}^n|a_{ij}|=\mu.$$

按如下方式取向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$：当$a_{kj}\geqslant 0$时，取$x_j=1$；当$a_{kj}<0$时，取$x_j=-1$. 这样，$\|\boldsymbol{x}\|_\infty=1$，且$\|\boldsymbol{Ax}\|_\infty=\sum\limits_{j=1}^n|a_{kj}|=\mu$.

由此可知，$\|\boldsymbol{A}\|_\infty=\mu=\max\limits_{1\leqslant i\leqslant n}\sum\limits_{j=1}^n|a_{ij}|$.

下面导出2-范数的表达式. 注意到$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$对称，可设其特征对为$(\lambda_i,\boldsymbol{u}_i)$，$i=1,2,\cdots,n$；且$\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}$，$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$. 对任意$\boldsymbol{x}\in\mathbb{R}^n$，按此特征向量系展开，有$\boldsymbol{x}=\sum\limits_{i=1}^n\beta_i\boldsymbol{u}_i$，则容易计算下面的向量2-范数：

$$\|\boldsymbol{x}\|_2=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}=\sum_{i=1}^n\beta_i^2,$$

$$\| \boldsymbol{Ax} \|^2 = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Ax} = \sum_{i=1}^{n}\lambda_i\beta_i^2.$$

所以

$$\frac{\| \boldsymbol{Ax} \|^2}{\| \boldsymbol{x} \|^2} = \frac{\sum_{i=1}^{n}\lambda_i\beta_i^2}{\sum_{i=1}^{n}\beta_i^2} \leqslant \frac{\sum_{i=1}^{n}\lambda_1\beta_i^2}{\sum_{i=1}^{n}\beta_i^2} = \lambda_1.$$

故 $\| \boldsymbol{A} \|_2^2 = \max\limits_{\boldsymbol{x}\neq\boldsymbol{0}} \dfrac{\| \boldsymbol{Ax} \|^2}{\| \boldsymbol{x} \|^2} \leqslant \lambda_1$；当 $\boldsymbol{x}=\boldsymbol{u}_1$ 时取等号. 故 $\| \boldsymbol{A} \|_2 = \sqrt{\lambda_1}$.

还有一个常用的范数是 F-范数(Frobenius 范数)，定义为

$$\| \boldsymbol{A} \|_{\mathrm{F}} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}} = \sqrt{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}.$$

可根据矩阵范数定义证明 2-范数和 F-范数有下面的酉不变性. 设 $\boldsymbol{U},\boldsymbol{V}$ 为酉矩阵(即 $\boldsymbol{U}^{\mathrm{H}}\boldsymbol{U}=\boldsymbol{I},\boldsymbol{V}^{\mathrm{H}}\boldsymbol{V}=\boldsymbol{I}$，这里上标 $\boldsymbol{H}$ 表示共轭转置)，则

$$\| \boldsymbol{UA} \|_2 = \| \boldsymbol{AV} \|_2 = \| \boldsymbol{UAV} \|_2 = \| \boldsymbol{A} \|_2,$$

$$\| \boldsymbol{UA} \|_{\mathrm{F}} = \| \boldsymbol{AV} \|_{\mathrm{F}} = \| \boldsymbol{UAV} \|_{\mathrm{F}} = \| \boldsymbol{A} \|_{\mathrm{F}}.$$

由矩阵范数可讨论矩阵序列的收敛性.

设 $\boldsymbol{A}_k=(a_{ij}^{(k)})(k=1,2,\cdots)$，若 $\lim\limits_{k\to\infty} a_{ij}^{(k)} = a_{ij}\ (i,j=1,2,\cdots,n)$，则称 $\boldsymbol{A}=(a_{ij})$ 为矩阵序列 $\{\boldsymbol{A}_k\}$ 的极限，记为 $\lim\limits_{k\to\infty}\boldsymbol{A}_k=\boldsymbol{A}$. 容易验证矩阵序列 $\{\boldsymbol{A}_k\}$ 收敛于 $\boldsymbol{A}$ 等价于 $\lim\limits_{k\to\infty} \| \boldsymbol{A}-\boldsymbol{A}_k \| = 0$.

在线性方程组的迭代法中，我们特别关注矩阵序列 $\{\boldsymbol{A}^k\}$ 是否趋于零矩阵. 这里需要引入另一个概念. 定义谱半径 $\rho(\boldsymbol{A})$ 为矩阵 $\boldsymbol{A}$ 的特征值的最大模，即

$$\rho(\boldsymbol{A}) = \max\{|\lambda| \mid \boldsymbol{Ax}=\lambda\boldsymbol{x}, \boldsymbol{x}\neq\boldsymbol{0}\}.$$

若 $\| \cdot \|$ 是诱导的矩阵范数，则 $\rho(\boldsymbol{A}) \leqslant \| \boldsymbol{A} \|$；同时可证明，$\forall \varepsilon>0$，存在诱导的矩阵范数 $\| \cdot \|$，使得 $\| \boldsymbol{A} \| \leqslant \rho(\boldsymbol{A})+\varepsilon$.

下面的定理在后面迭代法的收敛性证明中要用到.

定理 1.1 $\lim\limits_{k\to\infty}\boldsymbol{A}^k=\boldsymbol{0}$ 等价于谱半径 $\rho(\boldsymbol{A})<1$.

证 (必要性)已知 $\lim\limits_{k\to\infty}\boldsymbol{A}^k=\boldsymbol{0}$，假设 $\rho(\boldsymbol{A})\geqslant 1$，则 $\boldsymbol{A}$ 的按模最大的特征值 λ 满足：$|\lambda|\geqslant 1$. 设对应的特征向量为 $\boldsymbol{x}$，即 $\boldsymbol{Ax}=\lambda\boldsymbol{x}$，则 $\boldsymbol{A}^k\boldsymbol{x}=\lambda^k\boldsymbol{x}$. 两边取范数有 $\| \boldsymbol{A}^k\boldsymbol{x} \| = |\lambda|^k \| \boldsymbol{x} \| \geqslant \| \boldsymbol{x} \|$. 故 $\| \boldsymbol{x} \| \leqslant \| \boldsymbol{A}^k\boldsymbol{x} \| \leqslant \| \boldsymbol{A}^k \| \ \| \boldsymbol{x} \|$，从而 $\| \boldsymbol{A}^k \| \geqslant 1$，与 $\lim\limits_{k\to\infty} \| \boldsymbol{A}^k \| = 0$ 矛盾.

(充分性)设 $\rho(\boldsymbol{A})<1$，则存在 $\varepsilon>0$，使 $\rho(\boldsymbol{A})+\varepsilon<1$. 对 $\varepsilon>0$，存在矩阵范数 $\| \cdot \|$，使得 $\| \boldsymbol{A} \| < \rho(\boldsymbol{A})+\varepsilon<1$. 故 $\lim\limits_{k\to\infty} \| \boldsymbol{A}^k \| = 0$，即 $\lim\limits_{k\to\infty}\boldsymbol{A}^k=\boldsymbol{0}$. □

此定理的证明亦可以用其他方法，如 Jordan 分解，请读者自行练习.

1.3 向后误差和条件数

下面以线性代数方程组为例介绍向后误差和条件数等概念，并考察线性代数方程组的性态，求解线性方程组的具体算法在下一章讨论.

设求解线性方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，得到计算解 $\boldsymbol{y}$. 这里 $\boldsymbol{y}$ 一般不可能是准确解 $\boldsymbol{x}$，设 $\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{\delta}_x$；

计算解 $\boldsymbol{y}$ 一般不满足原方程，但是满足一个与原方程相近的方程：

$$(\boldsymbol{A}+\boldsymbol{\delta}_A)\boldsymbol{y}=\boldsymbol{b}+\boldsymbol{\delta}_b.$$

显然关系式中$\boldsymbol{\delta}_A$ 和$\boldsymbol{\delta}_b$（以范数衡量）越小，表明计算解越准确. $\boldsymbol{\delta}_A$ 和$\boldsymbol{\delta}_b$ 视为向后误差（具体讲还需区分范数型向后误差和分量型向后误差，请参考文献[48,49]及其中参考文献）.

我们还可以从另一个角度审视：

$$(\boldsymbol{A}+\boldsymbol{\delta}_A)(\boldsymbol{x}+\boldsymbol{\delta}_x)=\boldsymbol{b}+\boldsymbol{\delta}_b. \tag{1.2}$$

上式表明，当数据$(\boldsymbol{A},\boldsymbol{b})$有扰动$(\boldsymbol{\delta}_A,\boldsymbol{\delta}_b)$时，解向量有扰动$\boldsymbol{\delta}_x$. 下面考察扰动量$\boldsymbol{\delta}_x$ 的大小. 先介绍下面的引理.

引理 1.1 设 $\|\cdot\|$ 为诱导矩阵范数，且 $\|\boldsymbol{A}\|<1$，则 $\|(\boldsymbol{I}-\boldsymbol{A})^{-1}\|\leqslant\dfrac{1}{1-\|\boldsymbol{A}\|}$.

证 先证 $\boldsymbol{I}-\boldsymbol{A}$ 非奇异. 假设 $\boldsymbol{I}-\boldsymbol{A}$ 奇异，则有非零向量 $\boldsymbol{x}\neq\boldsymbol{0}$，使得$(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{x}=\boldsymbol{0}$，即 $\boldsymbol{x}=\boldsymbol{A}\boldsymbol{x}$，于是 $\|\boldsymbol{x}\|=\|\boldsymbol{A}\boldsymbol{x}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|$，故 $\|\boldsymbol{A}\|\geqslant1$，矛盾.

再由$(\boldsymbol{I}-\boldsymbol{A})^{-1}(\boldsymbol{I}-\boldsymbol{A})=\boldsymbol{I}$得，$(\boldsymbol{I}-\boldsymbol{A})^{-1}=\boldsymbol{I}+(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{A}$. 两边取范数，得 $\|(\boldsymbol{I}-\boldsymbol{A})^{-1}\|\leqslant1+\|(\boldsymbol{I}-\boldsymbol{A})^{-1}\|\,\|\boldsymbol{A}\|$，故 $\|(\boldsymbol{I}-\boldsymbol{A})^{-1}\|\leqslant\dfrac{1}{1-\|\boldsymbol{A}\|}$. □

定理 1.2 设 $\boldsymbol{A}$ 为非奇异方阵，$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}\neq\boldsymbol{0}$，且(1.2)式成立，则当 $\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{\delta}_A\|<1$ 时有

$$\frac{\|\boldsymbol{\delta}_x\|}{\|\boldsymbol{x}\|}\leqslant\frac{\kappa(\boldsymbol{A})}{1-\kappa(\boldsymbol{A})\dfrac{\|\boldsymbol{\delta}_A\|}{\|\boldsymbol{A}\|}}\left(\frac{\|\boldsymbol{\delta}_b\|}{\|\boldsymbol{b}\|}+\frac{\|\boldsymbol{\delta}_A\|}{\|\boldsymbol{A}\|}\right), \tag{1.3}$$

这里 $\kappa(\boldsymbol{A})=\|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|$.

证 由(1.2)式展开，并利用 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，得

$$\boldsymbol{A}\boldsymbol{\delta}_x+\boldsymbol{\delta}_A\boldsymbol{x}+\boldsymbol{\delta}_A\boldsymbol{\delta}_x=\boldsymbol{\delta}_b,$$

$$(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\delta}_A)\boldsymbol{\delta}_x=\boldsymbol{A}^{-1}(\boldsymbol{\delta}_b-\boldsymbol{\delta}_A\boldsymbol{x}).$$

因设扰动量很小，满足 $\|\boldsymbol{A}^{-1}\boldsymbol{\delta}_A\|\leqslant\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{\delta}_A\|<1$，故 $\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\delta}_A$ 可逆，可解出

$$\boldsymbol{\delta}_x=(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\delta}_A)^{-1}\boldsymbol{A}^{-1}(\boldsymbol{\delta}_b-\boldsymbol{\delta}_A\boldsymbol{x}).$$

要衡量$\boldsymbol{\delta}_x$ 的大小，需两边取范数，得

$$\begin{aligned}\|\boldsymbol{\delta}_x\|&\leqslant\frac{1}{1-\|\boldsymbol{A}^{-1}\boldsymbol{\delta}_A\|}\|\boldsymbol{A}^{-1}\|(\|\boldsymbol{\delta}_b\|+\|\boldsymbol{\delta}_A\|\,\|\boldsymbol{x}\|)\\&=\frac{\|\boldsymbol{x}\|}{1-\|\boldsymbol{A}^{-1}\boldsymbol{\delta}_A\|}\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|\left(\frac{\|\boldsymbol{\delta}_b\|}{\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|}+\frac{\|\boldsymbol{\delta}_A\|}{\|\boldsymbol{A}\|}\right).\end{aligned}$$

再利用 $\|\boldsymbol{b}\|=\|\boldsymbol{A}\boldsymbol{x}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|$，得

$$\frac{\|\boldsymbol{\delta}_x\|}{\|\boldsymbol{x}\|}\leqslant\frac{\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|}{1-\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{\delta}_A\|}\left(\frac{\|\boldsymbol{\delta}_b\|}{\|\boldsymbol{b}\|}+\frac{\|\boldsymbol{\delta}_A\|}{\|\boldsymbol{A}\|}\right).$$

□

(1.3)式左端$\dfrac{\|\boldsymbol{\delta}_x\|}{\|\boldsymbol{x}\|}$是解的相对扰动，右端$\dfrac{\|\boldsymbol{\delta}_b\|}{\|\boldsymbol{b}\|}$，$\dfrac{\|\boldsymbol{\delta}_A\|}{\|\boldsymbol{A}\|}$是数据的相对扰动，该相对扰动被放大 $\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ 倍. 定义 $\mathrm{cond}(\boldsymbol{A})=\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ 为条件数. 当条件数大时，输入数据相对小的扰动导致解较大的改变，解对扰动敏感，问题是病态的. 条件数由问题本身决定，而向后误差则取决于算法.

尽管数值分析专家很早就注意到“坏条件”(ill-conditioned)的情形，但第一个正式使用