信息与计算科学丛书—47

现代数值计算方法

刘继军　编著

科学出版社
www.sciencep.com

信息与计算科学丛书 47

现代数值计算方法

刘继军 编著

科学出版社

北京

内 容 简 介

本书是作者在东南大学讲授“现代数值计算方法”的讲稿的基础上形成的. 本书涵盖了经典的数值方法的大部分内容, 同时也包涵了近年来发展起来的一些新方法和对一些新的应用问题的处理, 如MATLAB的使用, 高维积分计算的统计方法等. 本书侧重算法的有效实现, 给出了很多算法的FORTRAN程序或者MATLAB程序, 并将它们用于处理一些具体的问题. 本书共分6章, 分别介绍数值计算的基本原理、矩阵分析基础、有限元方法的基本原理和应用、边界积分方程及其应用、积分计算的近代方法和快速Fourier变换和小波变换.

本书适合高等院校数学系研究生和工科相关专业研究生作为教材, 也可供大学教师和科研人员阅读参考.

图书在版编目(CIP)数据

现代数值计算方法/刘继军编著. —北京: 科学出版社, 2010
(信息与计算科学丛书; 47)
ISBN 978-7-03-027001-6
Ⅰ. ① 现… Ⅱ. ① 刘… Ⅲ. ① 数值计算-计算方法 Ⅳ. O241
中国版本图书馆CIP数据核字 (2010) 第043054号

责任编辑: 陈玉琢/责任校对: 钟 洋
责任印制: 钱玉芬/封面设计: 杜剑平

科学出版社出版
北京东黄城根北街16号
邮政编码: 100717
http://www.sciencep.com

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销
*
2010年3月第 一 版 开本: B5(720×1000)
2010年3月第一次印刷 印张: 16 3/4
印数: 1—3 000 字数: 320 000

定价: 58.00元

(如有印装质量问题, 我社负责调换)

《信息与计算科学丛书》序

20 世纪 70 年代末，由已故著名数学家冯康先生任主编，科学出版社出版了一套《计算方法丛书》，至今已逾 30 册. 这套丛书以介绍计算数学的前沿方向和科研成果为主旨，学术水平高、社会影响大，对计算数学的发展、学术交流及人才培养起到了重要的作用.

1998 年教育部进行学科调整，将计算数学及其应用软件、信息科学、运筹控制等专业合并，定名为“信息与计算科学专业”. 为适应新形势下学科发展的需要，科学出版社将《计算方法丛书》更名为《信息与计算科学丛书》，组建了新的编委会，并于2004年9月在北京召开了第一次会议，讨论并确定了丛书的宗旨、定位及方向等问题.

新的《信息与计算科学丛书》的宗旨是面向高等学校信息与计算科学专业的高年级学生、研究生以及从事这一行业的科技工作者，针对当前的学科前沿，介绍国内外优秀的科研成果. 强调科学性、系统性及学科交叉性，体现新的研究方向. 内容力求深入浅出，简明扼要.

原《计算方法丛书》的编委和编辑人员以及多位数学家曾为丛书的出版做了大量工作，在学术界赢得了很好的声誉，在此表示衷心的感谢. 我们诚挚地希望大家一如既往地关心和支持新丛书的出版，以期为信息与计算科学在新世纪的发展起到积极的推动作用.

石钟慈

2005 年 7 月

前　言

科学计算已经成为现代三大科学方法之一. 数值计算方法是数学研究的一个重要分支, 是科学计算的基础, 在现代科学技术和工程领域中的作用日趋重要. 从复杂过程的计算机模拟、天文探测、大范围的中长期数值天气预报, 到生命科学、核能的研究开发、人口发展趋势预测等广大的领域, 科学计算方法都起着基本的作用. 因此, 《现代数值计算方法》作为一门强调算法实现的基础课程, 在数学研究生的培养中具有重要的作用. 其任务是使得青年研究人员 (尤其是研究生) 具有数值计算的基本能力和素养, 尤其是自己动手编程计算的能力. 这种能力, 无论是对从事进一步数值计算的科学研究, 还是对从事应用领域的工作, 都是非常重要的.

基于这种考虑, 在 2005 年东南大学的研究生课程建设讨论会上, 决定在数学系应用数学博士点开设这样一门博士学位课程, 同时也面向数学系全体专业方向硕士研究生及工科研究生. 经过近三年的精心准备, 该门课程现在已经立项为东南大学研究生精品课程并加以建设, 本书就是在该门课程的讲稿基础上形成的. 现有的数值计算方法的教材虽然很多, 但大都侧重于专门的计算领域, 理论性较强, 对算法的具体实现及实现的有效性也涉及不多. 本书涵盖了经典的数值方法的大部分内容, 同时也包含了近年来发展起来的一些新的方法和对一些新的应用问题的处理. 例如, MATLAB 的使用, 高维积分计算的统计方法等. 我们特别强调算法的有效实现, 为此给出了很多算法的 FORTRAN 程序或者 MATLAB 程序, 并将它们用于处理一些具体的问题. 我们希望, 通过这样一种以计算方法的数值实现过程和计算机编程能力为重点的课程内容设计, 再通过自己的程序实现, 使读者真正掌握现代计算技术的一些基本方法, 而不仅仅是满足于模型问题、小型问题.

由于这门课程的覆盖面比较广泛, 本书尽可能地包含了计算方法的一些重要内容. 但是, 由于计算科学内容的广博和精深, 要在一门课程中包含所有的方法及其理论基础, 是不可能的. 因此本书的内容选择在很大程度上是根据个人的研究兴趣来组织的, 有些重要的内容如优化方法、逼近论等都没有包含. 读者可以通过其他专著对这些重要方法加以研究. 在本书的写作过程中, 我们也参考了一些现有的教材和专著, 把它们统一列在本书的参考文献中.

本书的完成是东南大学研究生精品课程建设项目的一项成果. 同时本书的出版得到了东南大学科技出版基金和国家自然科学基金 (No.10771033) 的支持, 谨此致谢. 我的研究生王海兵等对书稿的校对付出了辛勤的劳动, 在此一并致谢. 同时还

要感谢科学出版社的编辑为此书付出的劳动. 本书中的不妥和疏漏之处在所难免, 恳请有关学者和同行不吝指正.

刘继军

2009 年 12 月于东南大学

目　录

第 1 章　数值计算的基本原理

对给定的数学模型, 现代数值计算的基本任务就是通过数值离散构造出合适的计算格式求出原来物理模型的某种近似解. 由于问题的离散和对离散问题的数值求解过程中, 会引起各种各样的误差, 因此为了保证最后得到的数值解确实是原来问题解的某种近似, 在数值计算中必须遵循一些基本的原理. 这些基本的要求是构造有效的数值方法的基础. 我们在一个统一的框架下来介绍这些基本概念和结果.

1.1　问题的适定性和条件数

考虑一般的关于 x 的问题

$$F(x,d)=0 \tag{1.1.1}$$

的求解, 其中 d 是给定的数据集合, F 是描述待定量 x 和已知数据 d 的关系的一个泛函. 对不同的问题, x,d 可以是数、向量或者函数, F 可以是有限维或者无限维的泛函. 如果 F,d 是给定的, 上述问题的求解就是典型的正问题. 有时也有可能 x,F 是已知的, 要求输入数据, 这就是所谓的反问题, 或者 x,d 是已知的要求泛函关系 F, 这称为辨识问题. 需要指出的是, 正问题或者反问题是相对的. 但是通常也有一个公认的判别标准.

如果问题 (1.1.1) 存在唯一的连续依赖于输入数据 d 的解 x, 就称该问题的解是适定的, 否则问题就称为是不适定的. 这里的连续依赖性和度量数据大小的模的选择有关系. 但对给定的物理模型, 度量数据的模一般不能任意选择, 由问题的物理背景而定. 对于不适定的问题, 在数值求解以前, 必须先进行正则化以转化为一个近似的适定问题. 问题的不适定性来源于问题的本质, 任何数值方法本身都无法避开这个困难. 但可以通过正则化考虑该问题的一个近似的适定问题. 这是计算方法的另外一个专门的研究领域.

例 1.1.1　求多项式的实根的问题是一个典型的不适定问题. 考虑

$$p(x)=x^4-x^2(2a-1)+a(a-1).$$

当实数 a 连续变化时, 该多项式的实根的个数是振荡的：$a\geqslant 1$ 时有 4 个实根, $a\in[0,1)$ 时有两个实根, $a<0$ 时无实根.

解 x 对输入数据 d 的连续依赖性意味着输入数据的小的扰动引起的解的改变也是很小的. 用 $\delta d, \delta x$ 分别表示数据 d 的允许的扰动及对应的解 x 的扰动. 这意味着

$$F(x+\delta x, d+\delta d)=0. \tag{1.1.2}$$

从而问题的稳定性意味着对任意的 d, 存在 $\eta_0=\eta_0(d)>0, K_0=K_0(\eta_0)>0$, 使得对满足 $\|\delta d\| \leqslant \eta_0$ 的任意扰动 δd, 有

$$\|\delta x\| \leqslant K_0\|\delta d\|. \tag{1.1.3}$$

这里对数据和解的模可能是不同的. 为了给出上述过程的一个定量的描述, 给出下面的定义.

定义 1.1.1　对问题 (1.1.1), 定义相对条件数

$$K(d)=\sup_{\delta d\in D}\frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|}, \tag{1.1.4}$$

其中 D 是以 d 为心的一个邻域, 它表示数据 d 的允许扰动的集合以使得扰动的问题 (1.1.2) 有意义. 如果 $d=0$ 或者 $x=0$, 定义绝对条件数

$$K_{\mathrm{abs}}(d)=\sup_{\delta d\in D}\frac{\|\delta x\|}{\|\delta d\|}. \tag{1.1.5}$$

如果 $K(d)$ 对任意的输入数据 "很大", 就称问题 (1.1.1) 是病态的. "大"、"小" 的程度依赖于考虑的问题.

问题的非病态性和用于求解它的数值方法是无关的. 对非病态的问题, 既可以构造出稳定的数值方法, 也有不稳定的数值方法. 算法稳定性的概念和前面定义的问题的稳定性的概念是类似的.

注解 1.1.2　即使在问题的条件数不存在的情况下 (通常情况的描述是无穷大), 问题也不一定就是病态的. 可以构造出一个适定的问题 (如代数方程的多重根), 其条件数是无穷大, 但它可以等价于 (即有相同的解) 一个具有有限条件数的适定问题.

如果 (1.1.1) 有唯一解, 则在数据集和解集之间有一个映射 G 满足

$$x=G(d) \quad \text{使得} \quad F(G(d),d)=0. \tag{1.1.6}$$

在此定义下, (1.1.2) 就产生了 $x+\delta x=G(d+\delta d)$. 假定 G 在 d 是可导的, 记关于 d 的导数为 $G'(d)$, 则由 G 的一阶 Taylor 展开得到在 $\delta d\to 0$ 时有

$$G(d+\delta d)-G(d)=G'(d)\delta d+o(\|\delta d\|).$$

不计高阶无穷小, 近似得到

$$K(d) \approx \|G'(d)\| \frac{\|d\|}{\|G(d)\|}, \quad K_{\mathrm{abs}}(d) \approx \|G'(d)\|. \tag{1.1.7}$$

该估计在分析形如 (1.1.6) 的问题时非常有用, 见下面的例子.

例 1.1.2 二次代数方程的根. 在 $p \geqslant 1$ 时方程 $x^2 - 2px + 1 = 0$ 的根是 $x_\pm = p \pm \sqrt{p^2-1}$. 此时 $F(x,p) = x^2 - 2px + 1$, 数据 d 就是系数 p, 解 x 是向量 $\{x_+, x_-\}$. 至于条件数, 只要取 $G : \mathbb{R} \to \mathbb{R}^2, G(p) = \{x_+, x_-\}$, (1.1.6) 就是成立的. 记 $G_\pm(p) = x_\pm$, 则 $G'_\pm(p) = 1 \pm p/\sqrt{p^2-1}$. 取 (1.1.7) 中的模为 $\|\cdot\|_2$ 就得到

$$K(d) \approx \frac{|p|}{\sqrt{p^2-1}}, \quad p > 1. \tag{1.1.8}$$

由 (1.1.8) 知道, 在方程有相异根时 ($p \geqslant \sqrt{2}$), 问题 $F(x,p) = 0$ 是良态的. 然而在方程有重根 ($p = 1$) 时, 方程的行为有了戏剧性的变化. 首先, $G_\pm(p) = p \pm \sqrt{p^2-1}$ 在 $p = 1$ 不可导, 从而 (1.1.8) 无意义. 另一方面, (1.1.8) 似乎表明, $p \to 1+$ 时处理的问题是病态的. 然而所求的问题其实不是病态的. 事实上, 我们可以把此问题重新叙述为一个等价的形式

$$F(x,t) = x^2 - \frac{1+t^2}{t}x + 1 = 0, \quad t = p + \sqrt{p^2-1}.$$

该问题的根是 $x_- = t, x_+ = 1/t$, 它们在 $t = 1$ 时是相同的. 此变换从而消除了原来问题以 p 为参数时的奇性, 现在的两个根 $x_\pm = x_\pm(t)$ 在 $t = 1$ 附近是连续的, 用 (1.1.7) 来求条件数时我们得到在 $t = 1$ 附近 $K(t) \approx 1$, 从而变换后的问题是良态的.

例 1.1.3 线性代数方程组 $Ax = b, x, b \in \mathbb{R}^n$ 的求解, A 是非奇异的 $n \times n$ 矩阵. 此时容易解出 $x = G(b) = A^{-1}b$, 从而 $G'(b) = A^{-1}$. (1.1.7) 变为 (数据 $d = b$)

$$K(d) \approx \frac{\|A^{-1}\|\|b\|}{\|A^{-1}b\|} = \frac{\|Ax\|}{\|x\|}\|A^{-1}\| \leqslant \|A\|\|A^{-1}\| = K(A), \tag{1.1.9}$$

其中 $K(A)$ 是矩阵 A 的条件数. 因此如果 A 是非病态的, 求解 $Ax = b$ 关于右端项 b 的扰动是稳定的.

例 1.1.4 $f : \mathbb{R} \to \mathbb{R}$ 是 C^1 类的. 考虑非线性方程

$$F(x,d) = f(x) = \varphi(x) - d = 0,$$

其中 $\varphi : \mathbb{R} \to \mathbb{R}$ 是一个已知的函数, d 是数据. 该方程只有当 φ 在 d 的邻域内可逆时才有解 $x = \varphi^{-1}(d)$. 由于 $(\varphi^{-1})'(d) = (\varphi'(x))^{-1}$, (1.1.7) 的第一个关系在 $d \neq 0$ 时变为

$$K(d) \approx \frac{|d|}{|x|}|(\varphi'(x))^{-1}|, \tag{1.1.10}$$

而在 $d=0$ 时变为

$$K_{\mathrm{abs}}(d)\approx|(\varphi'(x))^{-1}|. \tag{1.1.11}$$

因此在 x 是 $\varphi(x)-d=0$ 的多重根时是不适定的. 在 $\varphi'(x)$ 比较小时是病态的.

根据 (1.1.7), $\|G'(d)\|$ 是 $K_{\mathrm{abs}}(d)$ 的一个近似, 我们有时称之为一阶绝对条件数. 这个近似并不总是能给出 $K_{\mathrm{abs}}(d)$ 的一个有效的估计. 例如, 当 $G'(d)=0$ 而在 d 的邻域内 $G(d)$ 不恒为零时. 一个具体的例子是 $x=G(d)=\cos d-1, d\in(-\pi/2,\pi/2)$. $G'(0)=0$ 但是 $K_{\mathrm{abs}}(d)=2/\pi$.

1.2　数值方法的稳定性

下面假定 (1.1.1) 是适定的. 求 (1.1.1) 的近似解的数值方法一般而言是考虑一系列的近似问题

$$F_n(x_n,d_n)=0,\quad n\geqslant 1, \tag{1.2.1}$$

其中 n 是一个参数, 依赖于具体问题. 很显然我们希望 $n\to\infty$ 时有 $x_n\to x$, 即数值解收敛于问题的精确解. 为此只要 $d_n\to d, F_n\to F$ 就可以了. 如果 (1.1.1) 的数据 d 也可以作为 F_n 的数据, 并且在 $n\to\infty$ 时成立

$$F_n(x,d)=F_n(x,d)-F(x,d)\to 0, \tag{1.2.2}$$

我们称 (1.2.1) 是相容的, 其中 x 是 (1.1.1) 对应于数据 d 的精确解. 如果 $F_n(x,d)=0$ 对所有的 n 成立, 该方法称为是强相容的.

有时候 (在使用迭代方法时), 近似问题 (1.2.1) 可能取下面的形式:

$$F_n(x_n,x_{n-1},\cdots,x_{n-q},d)=0,\quad n\geqslant q, \tag{1.2.3}$$

其中 $x_0,x_1,\cdots,x_{q-1}$ 是已知的. 这时强相容是指 $F_n(x,x,\cdots,x,d)=0$ 对一切的 $n\geqslant q$ 都成立.

例 1.2.1　记 $f:\mathbb{R}\to\mathbb{R}$. 考虑求方程 $f(x)=0$ 的单根的 Newton 迭代法

$$\text{给定 } x_0,\quad x_n=x_{n-1}-\frac{f(x_{n-1})}{f'(x_{n-1})},\quad n\geqslant 1. \tag{1.2.4}$$

只要取 $F_n(x_n,x_{n-1},f)=x_n-x_{n-1}+\dfrac{f(x_{n-1})}{f'(x_{n-1})}$, (1.2.4) 就是 (1.2.3) 的形式, 因此它是强相容的, 因为 $F_n(\alpha,\alpha,f)=0$ 对所有的 n 成立.

再如, 对连续函数 $f(t)$ 在 $[a,b]$ 上的定积分 $x=\int_a^b f(t)\mathrm{d}t$, 采用复合中点公式计算的格式

$$x_n = h\sum_{k=1}^{n} f\left(\frac{t_k + t_{k-1}}{2}\right),$$

其中 $h=\dfrac{b-a}{n}, t_k = a+(k-1)h$, 该算法是相容的. 对 $[a,b]$ 上连续的分片线性函数 f, 它还是强相容的.

一般而言, 通过对数学问题的极限过程 (积分、导数、级数) 取截断得到的数值方法不可能是强相容的.

和对问题 (1.1.1) 的适定性要求一样, 为了使数值方法是稳定的, 我们也要求对任何固定的 n, 对应于数据 d_n 存在唯一的解 x_n, 并且 x_n 要连续依赖于 d_n. 即

$$\begin{aligned}&\forall d_n, \exists \eta_0 = \eta_0(d_n), K_0 = K_0(\eta_0) \text{ 使得} \|\delta d_n\| \leqslant \eta_0\\ &\quad \Rightarrow \|\delta x_n\| \leqslant K_0\|\delta d_n\| \text{ 对一切的} \|\delta d_n\| \leqslant \eta_0 \text{成立.}\end{aligned} \tag{1.2.5}$$

和 (1.1.4) 的处理一样, 对序列 (1.2.1) 中的每一个问题, 引进

$$K_n(d_n) = \sup_{\delta d_n \in D_n} \frac{\|\delta x_n\|/\|x_n\|}{\|\delta d_n\|/\|d_n\|}, \quad K_{\mathrm{abs},n}(d_n) = \sup_{\delta d_n \in D_n} \frac{\|\delta x_n\|}{\|\delta d_n\|}. \tag{1.2.6}$$

如果 $K_n(d_n)$ 对任何允许的数据 d_n 比较小, 数值方法就称为是良态的, 否则称为是病态的.

假定每一个离散的问题 (1.2.1) 定义了一个数值数据和解之间的映射

$$x_n = G_n(d_n), \text{ i.e., } F_n(G_n(d_n), d_n) = 0. \tag{1.2.7}$$

如果 G_n 还是可导的, 则由 (1.2.7) 得到

$$K_n(d_n) \approx \|G_n'(d_n)\| \frac{\|d_n\|}{\|G_n(d_n)\|}, \quad K_{\mathrm{abs},n}(d_n) \approx \|G_n'(d_n)\|. \tag{1.2.8}$$

由此可以看出, 如果 (1.1.1) 和 (1.2.1) 的允许数据集是相同的, 我们就可以把 (1.2.6) 和 (1.2.7) 中的 d_n 用 d 来代替. 这样一来, 我们就可以对给定的数据 d 定义 "相对渐近条件数" 和 "绝对渐近条件数"

$$K^{\mathrm{num}}(d) = \lim_{k\to\infty} \sup_{n\geqslant k} K_n(d), \quad K_{\mathrm{abs}}^{\mathrm{num}}(d) = \lim_{k\to\infty} \sup_{n\geqslant k} K_{\mathrm{abs},n}(d).$$

例 1.2.2 加法运算和减法运算. $f:\mathbb{R}^2\to\mathbb{R}, f(a,b)=a+b$ 是一个线性映射, $f'(a,b)=(1,1)^{\mathrm{T}}$. 利用向量模得到 $K(a,b)\approx(|a|+|b|)/(|a+b|)$. 因此具有相同符

号的两个数的加法运算是适定的, 因为 $K(a,b)\approx 1$. 但是, 如果 a,b 具有相反符号, 即考虑减法运算, 几乎都是不适定的, 因为 $|a+b|\ll |a|+|b|$. 这就是在数值计算中的有效数字的抵消现象, 要尽量避免.

例 1.2.3　再次考虑前面的例 1.2. 当 $p>1$ 时, 问题是适定的. 但是如果用 $x_-=p-\sqrt{p^2-1}$ 来计算 x_-, 这是一个不稳定的算法. 它是由于有效数字的抵消现象引起的. 解决此问题的一个可行办法是先计算 $x_+=p+\sqrt{p^2-1}$, 再用 $x_-=\dfrac{1}{x_+}$ 来计算 x_-. 另一方面, 也可以用 Newton 迭代法来求方程 $F(x,p)=x^2-2px+1=0$ 的根

$$\text{给定 } x_0,\quad x_n=x_{n-1}-\frac{x_{n-1}^2-2px_{n-1}+1}{2x_{n-1}-2p}:=f_n(p),\quad n\geqslant 1.$$

由 (1.2.8) 得到 $p>1$ 时 $K_n(p)\approx |p|/|x_n-p|$. 为计算 $K^{\mathrm{num}}(p)$, 我们注意到在算法收敛时, x_n 收敛于 x_+,x_- 其中之一, 从而 $|x_n-p|\to\sqrt{p^2-1}$, 即 $K_n(p)\to K^{\mathrm{num}}(p)\approx |p|/\sqrt{p^2-1}$. 它和精确问题的条件数 (1.1.8) 的值非常接近. 因此我们知道, 如果 $|p|\approx 1$, 用 Newton 方法求二次代数方程的单根是病态的, 其他情况则是良态的.

数值方法的最终目的是通过对 (1.2.1) 的一系列问题的求解, 在 n 很大时用 x_n 来近似 (1.1.1) 的精确解 x. 此概念的准确描述是下面的

定义 1.2.1　(1.2.1) *的数值方法称为是收敛的, 是指*

$$\begin{aligned}&\forall\varepsilon>0,\exists n_0=n_0(\varepsilon),\delta=\delta(n_0,\varepsilon)\ \text{使得}\\&\forall n>n_0(\varepsilon),\forall\delta d_n:\|\delta d_n\|\leqslant\delta\Rightarrow\|x(d)-x_n(d+\delta d_n)\|\leqslant\varepsilon,\end{aligned}\qquad(1.2.9)$$

其中 d 是问题 (1.1.1) 的允许数据, $x(d)$ 是对应的解, $x_n(d+\delta d_n)$ 是问题 (1.2.1) 对应于数据 $d+\delta d_n$ 的解.

为了验证 (1.2.9) 的意思, 只要检查在相同的条件下成立

$$\|x(d+\delta d_n)-x_n(d+\delta d_n)\|\leqslant\frac{\varepsilon}{2}.\qquad(1.2.10)$$

事实上, 由 (1.1.3) 有

$$\begin{aligned}\|x(d)-x_n(d+\delta d_n)\|&\leqslant\|x(d)-x(d+\delta d_n)\|+\|x(d+\delta d_n)-x_n(d+\delta d_n)\|\\&\leqslant K_0\|\delta d_n\|+\frac{\varepsilon}{2}.\end{aligned}$$

取 $\delta=\min\{\eta_0,\varepsilon/(2K_0)\}$ 就得到 (1.2.9).

x_n 向 x 的收敛性可以用绝对误差或者相对误差来描述, 分别定义为

$$E(x_n)=|x-x_n|,\quad E_{\mathrm{rel}}(x_n)=\frac{|x-x_n|}{|x|},\quad \text{对}x\neq 0.\qquad(1.2.11)$$

如果 x, x_n 是向量或者是矩阵, 除了用 (1.2.11) 来定义外 (其中取适当的模), 还可以用分量的相对误差

$$E_{\mathrm{rel}}^{c}(x_n) = \max_{i,j} \frac{|(x - x_n)_{i,j}|}{|x_{i,j}|} \tag{1.2.12}$$

来描述.

下面来讨论稳定性和收敛性的关系.

这两个概念是有密切联系的. 首先, 如果 (1.1.1) 是适定的, 则数值问题 (1.2.1) 是收敛的一个必要条件就是它应该是稳定的.

假定 (1.2.1) 是收敛的, 我们通过确定 $\|\delta x_n\|$ 的界来证明该方法是稳定的. 利用 (1.1.3) 和 (1.2.10) 得到

$$\begin{aligned}\|\delta x_n\| &= \|x_n(d + \delta d_n) - x_n(d)\| \\ &\leqslant \|x_n(d) - x(d)\| + \|x(d) - x(d + \delta d_n)\| + \|x(d + \delta d_n) - x_n(d + \delta d_n)\| \\ &\leqslant K(\delta(n_0, \varepsilon), d)\|\delta d_n\| + \varepsilon. \end{aligned} \tag{1.2.13}$$

由此估计知道, 对充分大的 n, $\|\delta x_n\|/\|\delta d_n\|$ 可以被一个与 $K(\delta(n_0, \varepsilon), d)$ 同阶的常数界住, 因此方法是稳定的. 这就是为什么我们总是讨论稳定的数值方法的原因, 因为它是收敛的必要条件.

如果 (1.2.1) 和原来的问题 (1.1.1) 是相容的, 则数值方法的稳定性也是 (1.2.1) 的收敛性的一个充分条件. 事实上, 在此假定下有

$$\begin{aligned}\|x(d + \delta d_n) - x_n(d + \delta d_n)\| \leqslant{}& \|x(d + \delta d_n) - x(d)\| \\ &+ \|x(d) - x_n(d)\| + \|x_n(d) - x_n(d + \delta d_n)\|.\end{aligned}$$

由 (1.1.3), 右端第一项可以用 $\|\delta d_n\|$ 界住. 由稳定性条件 (1.2.5), 第三项也可以用 $\|\delta d_n\|$ 界住. 对第二项, 如果 F_n 是关于 x 可导的, 由 Taylor 展开得

$$F_n(x(d), d) - F_n(x_n(d), d) = \left.\frac{\partial F_n}{\partial x}\right|_{(\bar{x}, d)} (x(d) - x_n(d)),$$

其中 $\bar{x}$ 介于 $x(d), x_n(d)$ 之间. 再假定 $\partial F_n/\partial x$ 是可逆的, 就有

$$x(d) - x_n(d) = \left.\left(\frac{\partial F_n}{\partial x}\right)^{-1}\right|_{(\bar{x}, d)} [F_n(x(d), d) - F_n(x_n(d), d)]. \tag{1.2.14}$$

注意到 $F_n(x_n(d), d) = F(x(d), d) = 0$, 式 (1.2.14) 化为

$$\|x(d) - x_n(d)\| = \left\|\left.\left(\frac{\partial F_n}{\partial x}\right)^{-1}\right|_{(\bar{x}, d)}\right\| \|F_n(x(d), d) - F(x(d), d)\|.$$

由 (1.2.2) 就有 $n \to \infty$ 时 $\|x(d) - x_n(d)\| \to 0$.

上述证明的结果, 是数值分析的基础, 称之为等价性定理或者 Lax-Richtmyer 定理: 对相容的数值方法, 稳定性等价于收敛性.

1.3　误差的先验和后验估计

数值方法的稳定性可以通过下面两种不同的策略来分析:

(1) 前向分析 (forward analysis). 它利用输入扰动数据的误差和数值方法本身的误差给出解的误差 $\|\delta x_n\|$ 的界.

(2) 后向分析 (backward analysis). 为了得到在算法是完全准确的假定下的计算结果, 该方法对给定问题, 寻找对应输入数据的误差的界. 对给定某个计算的解 $\hat{x}_n$, 后向分析寻找满足 $F_n(\hat{x}_n, d_n + \delta d_n) = 0$ 的数据扰动 δd_n 的界. 注意在进行此估计时, 产生 $\hat{x}_n$ 的方式是不予考虑的.

前向分析和后向分析是先验估计的两种例子. 后者不仅可以用于讨论数值方法的稳定性, 还可以用于讨论收敛性, 称之为误差先验估计.

误差先验估计不同于误差后验估计, 后者的目的是利用由具体的数值方法得到的数值结果来估计近似解的误差. 如果用 $\hat{x}_n$ 来表示计算得到的 x 的近似的数值解, 后验误差估计利用残差 $r_n := F(\hat{x}_n, d)$ 来估计误差 $\hat{x}_n - x$.

例 1.3.1　考虑 n 次多项式 $p_n(x) = \sum_{k=0}^{n} a_k x^k$ 的根 $\alpha_1, \cdots, \alpha_n$ 的求解问题. 假定 $\tilde{p}_n(x) = \sum_{k=0}^{n} \tilde{a}_k x^k$ 是一个扰动的多项式, 其根为 $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_n$. 前向分析的任务是用系数的误差 $\tilde{a}_k - a_k$ 来估计根的误差 $\tilde{\alpha}_i - \alpha_i (i = 1, \cdots, n)$. 另一方面, 假定 $\{\hat{\alpha}_i\}$ 是准确多项式 $p_n(x)$ 用某种数值方法计算得到的给定的近似根, 后向分析提供了满足 $\sum_{k=0}^{n} (a_k + \delta a_k)\hat{\alpha}_k = 0$ 的系数 a_k 应该有的扰动 δa_k 的一个估计. 而后验误差估计则是要用残数 $p_n(\hat{\alpha}_i)$ 的某个函数来估计 $\alpha_i - \hat{\alpha}_i$.

例 1.3.2　对非奇异矩阵 $A \in \mathbb{R}^{n \times n}$, 考虑 $Ax = b$ 的求解问题. 考虑扰动问题 $\tilde{A}\tilde{x} = \tilde{b}$. 前向分析的任务是用误差 $\tilde{A} - A, \tilde{b} - b$ 来估计解的误差 $\tilde{x} - x$. 后向分析则提供了满足 $(A + \delta A)\hat{x}_n = b + \delta b$ 的系统扰动 $\delta A, \delta b$ 的一个估计, 其中 $\hat{x}_n$ 是由某种数值方法得到的解. 后验误差估计则是要用残数 $r_n = b - A\hat{x}_n$ 的某个函数来估计 $x - \hat{x}_n$.

需要指出, 后验误差估计在自适应的误差控制策略的设计中有重要的作用. 该策略利用后验估计适当改变离散参数 (如数值积分或微分方程中空间步长) 来保证误差不超过给定的水平.

利用自适应误差控制的数值方法称为自适应方法. 在实际计算过程中, 这类方