西安交通大学数学研究生教学丛书

数值计算的算法与分析

张可村　赵英良　编著

科学出版社

北　京

内 容 简 介

本书主要介绍了数值计算中的经典、常用和一些最新方法，并且介绍了与构造算法和算法性能分析相关的理论及必要的数学基础.其算法主要包括如下四个方面：代数方程组数值解，如共轭斜量法、Newton 迭代法、区间迭代法等；常微和偏微分方程数值解法，如差分法、有限元法、区域分解算法等；数值逼近的各种方法，如插值法、逼近法、任意阶光滑逼近函数构造法、小波变换与逼近等；最优化方法，如非线性规划、线性规划、二次规划、几何规划、遗传算法、神经网络法等.

本书可作为研究生、博士生、计算数学和应用数学专业的本科生教材，也可供科技人员，特别是从事工程科学技术的工作人员参考.

图书在版编目(CIP)数据

数值计算的算法与分析/张可村，赵英良编著.—北京：科学出版社，2003
(西安交通大学数学研究生教学丛书)
ISBN 7-03-009364-X

Ⅰ.数… Ⅱ.①张…②赵… Ⅲ.数值计算-计算方法-研究生-教学参考资料 Ⅳ.O241

中国版本图书馆 CIP 数据核字(2001)第 025729 号

责任编辑：林 鹏 毕 颖/责任校对：包志虹
责任印制：安春生/封面设计：王 浩

科学出版社发行 各地新华书店经销

*

2003 年 1 月第 一 版 开本：B5(720×1000)
2004 年 6 月第二次印刷 印张：22 1/2
印数：3 001—5 000 字数：430 000

定价：33.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

前　言

高科技的突飞猛进,使计算机的硬件和软件应用于每个科学研究领域、生产部门和各级政府的管理决策部门,从而促进了科学技术的高速发展.然而如何才能有效地把计算机广泛应用于每一个研究领域和生产部门呢?这就要求每一个应用者熟练掌握和使用各种数值方法.正因为如此,现已有许多关于"计算方法"方面的著作和相应的计算机软件,但还不能满足现代科学技术发展的需要.因此必须要求不同领域的每一个科技工作者、管理人员,根据自己所研究问题的特点,构造相应的数值方法.不仅要善于从现有算法中选择相适应的方法,更重要的是如何根据实际问题的需要,改进和构造新的数值方法,弥补现有算法的不足,这就是我们写这本书的主要目的.我们尽量把常用的、经典的和最新的方法收集在书中,突出构思方法和构造算法常用技巧和手段,略去了一些古典的与算法构造关系不大的数学理论分析,在照顾到系统性、逻辑性和可读性的同时,特别注意收集了各领域中最新出现的数值方法和本书作者的一些最新研究成果,与近代科学技术相适应.

全书共分六章.第一章绪论,从宏观上简述了本书的目的、研究的内容和采用的手段;第二章介绍了构造数值方法常用的数学基础,使具有大学本科数学基础的读者能顺利读完本书;第三章介绍迭代技术在求解代数方程组中的应用;第四章,讨论离散化技术在求解常微和偏微分方程组中的各种应用;第五章介绍离散问题解析化技术在数值逼近论中的应用;第六章讨论优化技术在数学规划方法中的应用.第一、五、六章由张可村负责编写,第二、三、四章由赵英良负责编写.

感谢如下同志为本书撰写了他们最新、最好的研究成果:简金宝教授在约束规划中,撰写了强次可行方向法与快速算法;杨守志副教授在离散问题解析化技术中,撰写了小波变换与逼近;艾文宝副教授在线性规划中,撰写了内点算法;李换琴副教授在无约束规划方法中撰写了改进拟 Newton 法;王燕军博士撰写了不确定型优化算法、二次规划算法、信赖域法等.

还要感谢科学出版社有关领导同志,特别是本书的责任编辑对本书所付出的辛勤劳动.

由于我们水平所限,疏漏之处难以避免,恳请读者指正.

编者

2002.10

目　　录

第一章　绪　　论

§1　现代科学技术的一般过程

高科技的发展,迫使计算机科学的发展突飞猛进.由于计算机的高速、大容量、多功能,又为现代科学技术的发展提供了最优、最快的新途径,一般可按如下四个阶段进行.

1.1　工程问题数学化(数学建模)

采用恰当的数学语言,描述自然科学、社会科学、管理和决策科学各领域中关键而核心的问题,常称为数学建模.要建立一个好的数学模型,对于单方面的专家都是很困难的,必须由各相关领域的专家和数学工作者,特别是从事计算数学、应用数学研究工作的学者,紧密结合,相互取长补短才有可能.这是因为评价一个模型的优劣主要有两点:其一,用什么样的数学语言,才能真正反映工程实际;其二,所用数学语言,可否在计算机中实现,这二者缺一不可.因此,要求参与建模的工程专家必须精通专业,具有一定的数学和计算数学的基础知识,对于数学工作者,要掌握宽广的数学知识,还要了解该工程问题在国内外现状,面临的主要问题,采用哪种数学语言来描述此问题更为恰当.工程中的数学模型,一般可分为三类:其一,连续型(确定型),即能用数学解析式刻画工程问题;其二,离散型(统计型),找不到确定数学解析式来描述该工程问题;其三,不确定型(随机型).本书重点讨论连续型.

1.2　数学问题数值化(算法与分析)

从工程实际中抽象出的数学问题,绝大多数都不能直接用计算机语言来识别,因此,先进的计算工具——计算机,不能直接求解相应的数学问题,自然不能用于解决相应的工程问题.把数学问题数值化,就是如何根据不同的数学问题,寻求相应的方法,此方法(常称为数值方法)只能用四则运算和一些逻辑运算或者直接用计算机语言能描述相应的数学问题,便于用计算机求解的数学问题.此方法的优劣,直接关系到能否把计算机用于解决高科技问题.由此可知,数值计算在当代科技中的地位和作用,它直接关系到能否用现代的数学方法,最先进的计算工具去解决现代科学技术中的管理问题、规划和决策问题,各领域中高科技中的关键性问题.因此对数值计算的算法的构造,优劣的分析,是每一个科技工作者、决策者不可

缺少的基础知识,因此对于即将走上工作岗位的本科生,特别是硕士生、博士生是必须掌握的.

1.3 数值问题机器化(程序设计)

要求程序设计者,用最简练的机器语言,最快的速度,最少的存储量,设计软件,并获得准确的计算结果.要达到这些要求,程序设计者必须掌握数值方法的构思途径,算法的关键和难点;熟悉计算机软硬件的基础知识,能灵活应用某种机器语言,准确无误的描述每一个算法,并能以最快的速度发现并解决计算过程中出现的各种异常问题.这是检验程序设计者水平高低的客观标准,这也是衡量一个决策者,工程设计师,管理工作者水平高低的重要标志.程序设计,对于每一个科技工作者,管理工作者,都是必须具备的技能,掌握这种技能入门快,见效也快,也是年轻的科技工作者最喜欢干的工作.但要真正作一个高级程序设计者,也是十分困难的,必须具备丰富的想象力,总结归纳和设计的能力,具有总工程师,总设计师的能力和水平.

1.4 科学试验

前面三个阶段,只是为现代科学技术提供了一种途径,也可以说是捷径.但这种途径是否真正能解决科技中的问题,被生产实际部门直接采用,还必须将第三阶段获得的计算结果,在科学试验室进行检验,是否与工程实际相符? 是否能推广应用? 若不相符,分析其不符的根源何在? 确定返回到前三阶段的某一阶段重新开始,重复上述工作直到满意为止.这一步是必不可少的,且很成熟的,只要有相应的试验设备和原材料都可进行.前三阶段的实质就是把一个实际工程问题,置入计算机中,在计算机中可做大量的模拟试验,当基本上与实际问题相符时,再经过此步.这样可以减少实际试验次数,节省大量的原材料,缩短设计周期,且还能使性能达到最佳,是现代科学的必经途径.不采用现代科学技术的分析方法和手段只埋头做试验的方法,远远跟不上现代科学技术的发展.

当试验成功后,就可试制新产品,推广应用.实现,研制→生产→销售一体化,有助于提高产品质量,增强产品市场竞争能力,获得不可估量的社会效益和经济效益.

任何一个有竞争力的新产品,任何一项能产生社会和经济效益的科研课题,必须经过上述 4 个阶段.第一阶段是根本,模型是否反映工程实际问题的需要,取决于当代高级科技人员的理论水平;第二阶段是桥梁,所构造的算法是否能真正取代原数学模型,它的有效性和可实现性,取决于从事计算数学的研究人员的理论水平和创新能力;第三阶段是检验前两阶段工作的有效性和可行性的方法和手段;第四阶段检验该项研究成果是否有实际应用价值,同时也检验了前三阶段研究工作的

可靠性,也是能否见到经济效益和社会效益的关键.

§2 数值计算探讨的主要问题

数学来源于实践,从工程实际中抽象出来的连续型数学问题,从古到今大体上可分为如下四类:1.线性和非线性代数方程(组)求解;2.偏(常)微分方程、积分方程求解;3.逼近函数的构造方法(数值逼近);4.数学规划问题.除此之外,还有离散型、不确定(随机)型.在高科技中,还有许多的工程问题,很难用现有的数学语言来刻画.必须从工程实际需要出发,探索新的数学语言,构造相应的数值求解方法,不断地产生新的数学分支.这就需要众多的工程专家,数学工作者紧密结合,深入到科学研究前沿,去发现新问题,解决新问题,从而促进科学技术高速发展.在本书中,只对上述四类问题探索数值求解方法及其与算法有关的理论分析.为了使读者有一个宏观的印象,先对上述四类问题作如下简介.

2.1 线性和非线性代数方程组的数值解法

此类问题是后三类问题的基础,一般表示式如下:

$$f_j(x_1,x_2,\cdots,x_n)=b_j,\ j=1,2,\cdots,m$$

在一般情况下:$n\geqslant m$.

记 $x=(x_1,x_2,\cdots,x_n)^T$,$b=(b_1,b_2,\cdots,b_m)$,$F(x)=(f_1(x),f_2(x),\cdots,f_m(x))^T$.上方程组可简记为 $F(x)=b$.

当 $f_j(x)(j=1,2,\cdots,m)$ 均为线性函数:

$$f_j(x)=\sum_{i=1}^{n}a_{ji}x_i,\qquad j\leqslant 1,2,\cdots,m$$

时称 $F(x)=b$ 为线性方程组:$Ax=b$,其中 $A=(a_{ji})_{m\times n}$.此问题虽然是众所周知的,但当 $n>m$,且当 m,n 特别大时,常规的求解方法失效.

当 $f_j(x)(j=1,2,\cdots,m)$ 之一为非线性函数时,称 $F(x)=b$ 为非线方程组,此时,当 n 特别大时,求解十分困难.

又当 $m=n=1$,是大家熟知的非线性方程,其求解方法多而成熟,读者在"高等数学"或"数学分析"中都已学过.

此类问题应用极为广泛,在工程中,特别是管理科学中,许多问题都可归结为解线性或非线方程(组),且还有其他数学分支的数值求解过程,也要借助于这类问题的求解方法.惯用的解法是迭代法,详见本书第三章.

2.2 微分方程(组)的数值求解法

在最优控制,机械工程,能源动力工程,流体工程中有许多问题抽象出的数学

问题归结为求解常微分方程(组)的始值问题：

$$\begin{cases} f(x,y(x),y'(x),\cdots,y_{(x)}^{(n)})=0 \\ y_{(x_0)}^{(k)}=y_k,(k=0,1,\cdots,n-1) \end{cases}$$

其中，$y_{(x_0)}^{(o)}=y(x_0)$，$y_{(x_0)}^{(k)}$ 表示 $y(x)$ 的 k 阶导数在 x_0 处的值，y_k 为给定的初始条件，也可换成两端边值条件，如

$$y_{(a)}^{(k)}=y_k,(k=0,1,\cdots,l-1)$$
$$y_{(b)}^{(n-k)}=y_k,(k=n,n-1,\cdots,n-(l-1))$$

或偏微分方程的边值问题.常见的有抛物型方程，双曲型方程，椭圆型方程.

如，二阶线性抛物型方程：

$$\begin{cases} \dfrac{\partial u}{\partial t}-a(x,t)\dfrac{\partial^2 u}{\partial x^2}-2b(x,t)\dfrac{\partial u}{\partial x}+c(x,t)u(x,t)=d(x,t) \\ \text{初始条件：}u(x,0)=f(x),\quad 0<x<1 \\ \text{边值条件：}u(0,t)=f_1(t),u(1,t)=f_2(t),t>0 \end{cases}$$

其中，$a(x,t),b(x,t),c(x,t),d(x,t),f(x),f_1(x),f_2(x)$ 均为给定的函数，且 $a(x,t)>0$.

又如拟线性双曲型方程组：

$$\sum_{j=1}^{n}\left(a_{ij}\frac{\partial u_j}{\partial x}+b_{ij}\frac{\partial u_j}{\partial y}\right)-d_i=0,i=1,2,\cdots,n$$

其中 a_{ij},b_{ij},d_i 均为 $x,y,u_1,u_2,\cdots,u_n$ 的函数.如果 a_{ij},b_{ij},d_i 中均不含 $u_i,i=1,2,\cdots,n$，则称为线性双曲型方程组.上方程组再附加边界条件，方可求解.

再如一般二阶椭圆型方程第一边值问题：

$$\begin{cases} a(x,y)\dfrac{\partial^2 u}{\partial x^2}+b(x,y)\dfrac{\partial^2 u}{\partial y^2}+C(x,y)\dfrac{\partial u}{\partial x} \\ \qquad\qquad +d(x,y)\dfrac{\partial u}{\partial y}+e(x,y)u=f(x,y) \\ u(x,y)\big|_{\Gamma}=\varphi(x,y) \end{cases}$$

其中，$a(x,y),b(x,y),c(x,y),d(x,y),e(x,y),f(x,y),\varphi(x,y)$ 均为给定的函数，Γ 为区域的边界线.

上述方程都是经典的方程，要求精确的理论解，一般情况下是十分困难的.有的问题，解的存在性，惟一性，惟多性都很难断定，更何况近代科学技术中，抽象出的问题，比上述问题更为复杂，自然求理论解，解的存在性等问题更为困难，因此只能探索求近似解，目前惯用的方法是差分法，有限方法，边界方法等.换言之，采用不同的离散化技术就获得相应的数值求解方法，可见本书第四章介绍的方法.

2.3 逼近函数的构造法(数值逼近)

在科学试验,新产品性能分析和大规模的科学计算中,可获得大量的离散数据.在一维的情况下可用(x_i,y_i),$i=1,2,\cdots,n$ 表示;在多维的情况下可用(x^i,y_i),$i=1,2,\cdots,n$,表示,其中,$x^i=(x_1^i,x_2^i,\cdots,x_m^i)^T$,$x_i,y_i,x_j^i$ 均为实数.如何通过这些测试或计算获得的离散数据?获得原物体(产品)的各种性能指标?惯用的方法,是将离散数据解析化,即构造一个或多个函数(可以是多元函数),取代离散数据,此函数常称为逼近函数,通过对逼近函数的各种性能分析获得原物体相应的性能指标.如何构造逼近函数才能达到此目的呢?构造逼近函数的常规方法的难点在于如何根据离散数据的分布特性,选择相应的逼近函数所在的函数类.当逼近函数所在的函数类确定后,如何确定评价逼近程度,逼近优劣,从而获得相应的逼近方法.常用的逼近函数有:多项式,分段低次多项式,三角函数,小波函数等.常用的逼近方法有:插值法,最小二乘逼近法,最优一致逼近法等.常见逼近方法一般可归结为(一维问题)离散型:

$$\min_{a\in E^m} F(a)=\sum_{i=1}^{n}\left[(y_i-f(a,x_i))/\delta_i\right]^2$$

其中 $a=(a_1,a_2,\cdots,a_m)$,(x_i,y_i) 为离散点,δ_i 根据离散点的精确程度,给定常数,$f(a,x)=\sum_{j=1}^{m}a_j\varphi_j(x)$,$\varphi_1(x),\cdots,\varphi_m(x)$ 为线性无关的基函数,$x\in[\alpha,\beta]$,$x_i\in[\alpha,\beta]$.

上述要探索的问题,都属于离散问题解析化技术,详见第五章.

2.4 数学规划方法

数学规划方法是运筹学的重要分支,常称为优化方法,它渗透到科学技术的每一个领域以及各级政府的管理部门,决策规划部门,甚至到每一个企业,每一个人.这是因为,优化方法的实质,就是用最小的投入取得最大的收获.

优化方法也可分为连续型和离散型:如线性规划,二次规划,几何规划,非线性规划等都属于连续型;又如整数规划,0,1 规划,组合优化都属离散型:除此之外,还有不确定型(随机型),如随机规划等.

连续型规划问题,一般可写成如下形式:

$$\min_{x\in X} f(x)$$

其中 $X\subseteq E^n$,$x=(x_1,x_2,\cdots,x_n)^T$,当 $X=E^n$ 称无约束优化,当 $X\subset E^n$ 称为约束优化.在一般情况下 X 可表示成如下形式:$X=\{x\mid h_j(x)=0,j\in J\}$,或 $X=\{x\mid g_i(x)\leqslant 0,i\in I\}$ 或 $X=\{x\mid h_j(x)=0,j\in J,g_i(x)\leqslant 0,i\in I\}$.其中 I,J 为有限指标集.

$f(x)$称为目标函数，$h_j(x)$，$g_i(x)$称为约束函数，X称约束区域.

当$f(x)$，$h_j(x)$，$g_i(x)$给不同的函数关系式就对应不同的规划，如$f(x)$，$h_j(x)$，$g_i(x)$中有一个函数为非线性函数，常称为非线性规划.当$h_j(x)$，$g_i(x)$为线性函数，目标函数是线性函数时称为线性规划.若目标函数时是二次函数时称为二次规划.优化方法，就是探索如何求出上述问题的数值近似解，若n较大，$f(x)$的非线性程度较高时，是十分困难的.本书第六章将介绍各种求解方法.

§3 误差的来源及其对算法的影响

采用数值方法求解§2中所提出的数学问题，获得的解是近似解.若近似程度满足不了实际问题的需要，该方法将失效.因此在建立数学模型，构造数值求解方法时，必须注意误差的影响.

3.1 误差的来源

1.模型误差

用不同的数学语言描述、刻画工程问题时，对应不同的数学模型，此模型与原工程问题，总是有差异的，这种差异常称为模型误差，建模的关键问题是如何选择恰当的数学语言去刻画该工程问题，使其模型误差最小.若模型误差超出了允许范围，采用任何方法求解此数学模型都将失去任何实用价值.要真正定性特别是定量分析模型误差范围是很困难的，有待于在实际中，针对不同工程问题去摸索相应的方法.

2.模型参数误差

在一般情况下，建模时总有很多假设和难以确定的因素，这些假设和因素在模型中往往以参数形式出现.若参数有误差，必将影响计算结果.如含参变量的积分：$\int_0^1 f(x,y,t)\mathrm{d}t$，当参变量$x$，$y$有误差时，也会给积分值带来误差.

3.方法误差(截断误差)

同一数学问题，采用不同数值方法，有不同的误差称方法误差.例如读者熟知的同一积分，采用矩形公式、梯形公式、抛物线等数值积分公式，误差明显不一样.又如

$$\sin x = \sum_{k=0}^{\infty}(-1)^k \frac{x^{2k+1}}{(2k+1)!} \approx \sum_{k=0}^{n}(-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

取不同的n，得到的误差不一样，其方法误差为$\left|\sum_{k=n+1}^{\infty}(-1)^k \frac{x^{2k+1}}{(2k+1)!}\right| \leqslant \frac{|x|^{2k+1}}{(2k+1)!}$，常称为截断误差.

4.计算误差(舍入误差)

当算法给定时,要实现算法,就要对数进行四则运算.在计算机中,存放一个数只能是有限位数,在计算过程,必须进行舍入,例如 $10 \div 3 \doteq 3.33\cdots3$,大量的计算必带来舍入误差的积累和传递.

如何建立数学模型,构造相应的数值求解方法,使前述四种误差均能达到最小,是算法构造者和建模者的核心任务.

3.2　误差的种类及求法

1.误差一般可分为如下两种:

绝对误差

$$\Delta y = \text{精确值减去近似值} \triangleq y - \tilde{y}$$

相对误差

$$\delta_y = \frac{\Delta y}{y} \overset{\text{或}}{=} \frac{\Delta y}{\tilde{y}}$$

当参加运算的数,数量级相差较大时,一般采用相对误差.

2.函数误差及四则运算误差的求法

设函数 $y = f(x_1, x_2, \cdots, x_n)$,当自变量 x_i 有绝对误差 Δx_i 和相对误差 $\delta(x_i)$ 时,如何求函数的绝对误差 Δy 和相对误差 δy?

由全微分是函数增量的线性主部,易知绝对误差:

$$\begin{aligned}\Delta y &= f(x_1 + \Delta x_1, \cdots, x_i + \Delta x_i, \cdots, x_n + \Delta x_n) - f(x_1, x_2, \cdots, x_n)\\ &\approx dy = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i\end{aligned}$$

由此可知函数在某一点的绝对误差,不仅与该点自变量的绝对误差有关,还依赖于函数在该点对每个变量的偏导数,且当函数对某个变量的偏导数发生巨变时,直接使绝对误差发生巨变.

相对误差:

$$\delta_y = \frac{\Delta y}{y} = \sum_{i=1}^{n} \frac{x_i}{y} \frac{\partial f}{\partial x_i} \frac{\Delta x_i}{x_i} = \sum_{i=1}^{h} \frac{x_i}{y} \frac{\partial f}{\partial x_i} \delta(x_i)$$

由此可知,函数的相对误差和自变量相对误差的关系,不仅依赖于偏导数,还依赖于每个自变量的值与函数值之比,当自变量不接近零时而函数值接近零,会使函数的相对误差发生巨变.

四则运算的绝对和相对误差,只是上述函数的绝对与相对误差的特殊情况.即在上述公式中取 $y = f(x_1, x_2)$,f 可分别取 x_1, x_2 的和、差、积、商,可得和、差、积、商的误差.

和、差的绝对和相对误差，取 $y = f(x_1, x_2) = x_1 \pm x_2$

$$\Delta y = \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 = \Delta x_1 \pm \Delta x_2$$

$$\delta y = \frac{\Delta y}{y} = \frac{\Delta x_1 \pm \Delta x_2}{x_1 \pm x_2}$$

由此可得到代数和的绝对误差等于绝对误差的代数和.从上式可知应尽量避免两个绝对值很接近的同号的两个数相减，异号的两个数相加，否则影响和、差的相对误差.

商的绝对和相对误差：

$$f(x_1, x_2) = \frac{x_1}{x_2}$$

取

$$\Delta y = \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 = \frac{\Delta x_1}{x_2} - \frac{x_1}{x_2^2}\Delta x_2 = \frac{x_2\Delta x_1 - x_1\Delta x_2}{x_2^2}$$

$$\delta y = \frac{\Delta y}{x_1 / x_2} = \Delta y \cdot \frac{x_2}{x_1} = \frac{x_2\Delta x_1 - x_1\Delta x_2}{x_2 x_1} = \delta(x_1) - \delta(x_2)$$

由此可知，商的相对误差等于分子的相对误差与分母的相对误差之差.应避免绝对值很小的数作除数，否则其商的绝对误差发生巨变.

积的误差：

$$f(x_1, x_2) = x_1 x_2$$

取

$$\Delta y = \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 = x_2\Delta x_1 + x_1\Delta x_2$$

$$\delta y = \frac{\Delta y}{x_1 x_2} = \delta(x_1) + \delta(x_2)$$

由此可知，积的相对误差等于各因子相对误差之和.

3.3　误差对算法的影响

模型误差、方法误差是客观存在的，在一般情况下，要作定量分析是很困难的，因此不在此深究.当算法确定以后，在计算机上实现此算法时，计算误差的传递的快慢和积累的多少，直接影响算法的优劣.误差传递慢或不传递且积累少的算法称为稳定算法，否则称为不稳定的算法.不稳定的算法得到的数值解，往往被误差所淹没而失去任何价值.一般地讨论算法的稳定性也是很困难的，也不在本书深究，对于具体算法，若在应用中发现是不稳定，可作具体分析.在此只是告诉读者，误差对算法是有影响的，甚至影响很大，在作四则运算时，应尽量避免和减少误差的产生，以保证计算结果的准确性和可靠性.

§4 构造算法的途径

各类数学问题的数值求解方法是无法统计的,难以找到恰当的语言描述,特别是当计算机软件和硬件广泛应用于每一个领域的今天更是如此.但从宏观上分析,各种算法的构造,可按如下四种途径进行.

1.迭代技术;

2.离散化技术;

3.离散问题解析化技术;

4.优化技术.

每一种途径,一般是针对某一类数学问题提出的,但它们之间又有内在联系,可以相互渗透.例如,迭代技术可用于其他三种技术,优化技术也可用于另外三种技术.下面简介各种技术的基本思想和适用范围.

4.1 迭代技术

迭代技术也称为迭代法,常用于求解线性或非线方程(组)的解,在优化技术中可用于求最优解.迭代法的实质,就是逐步达到最终的目标,且后一步比前一步更加接近目标,只要找到本步和前一步或多步的关系式且本步更加接近目标,该迭代方法就构成了.例如,求方程 $f(x)=0$,在$[a,b]$上的实根,可按如下方法构造迭代式.

$\forall\, x_0 \in [a,b]$,$f(x)$在 x_0 的附近用一阶 Taylor 公式取代,即上方程变为

$$f(x_0)+f'(x_0)(x-x_0)=0$$

从此式解出 x,记为 x_1

$$x_1 = x_0-(f'(x_0))^{-1}f(x_0)$$

再将 $f(x)$在 x_1 处展开一阶 Taylor 公式可得

$$x_2 = x_1-(f'(x_1))^{-1}f(x_1)$$

如此继续下去,得到迭代通式

$$x_{n+1} = x_n-(f'(x_n))^{-1}f(x_n),\quad n=0,1,\cdots$$

可以证明:当 $f(x)$在$[a,b]$有根,且 $f''(x)$不变号,只要取初始点 x_0 满足 $f(x_0)f''(x_0)>0$,由上迭代式产生的数列$\{x_n\}$收敛于原方程的根 x^*,即$\lim x_n = x_*$.当 n足够大时,可取 $x_* \approx x_n$为近似解,此方法可以推广到解线性和非线性方程组,详见本书第三章.

4.2 离散化技术

离散化技术常用于求解常微分方程和偏微分方程的数值解.此技术的关键在

于如何剖分(离散化)曲线(曲面)所在的区域为若干小区域,再将方程中未知函数的各阶导数或偏导数,用相应离散点处未知函数的差商取代后,将微分方程转化为代数方程组,通过解代数方程组获得未知函数在离散处的离散值,作为原微分方程的数值解,此解能否取代原方程的解,取决于剖分的方法和稠密的程度.分割的技巧在于如何处理下述矛盾:分割太细,计算工作量大,计算误差的积累和传递影响数值解;分割粗了,获得的数值解又不能完全反映原问题的解.由此可知均分的方法并不是好方法,它难以解决这对矛盾.本书第四章是为解决这对矛盾而设置的.

下面仅以一阶常微分方程初始值问题为例,说明离散技术的基本思想.

求解一阶方程 $y'=f(x,y)$ 在 $[a,b]$ 上满足初始条件 $y(a)=y_0$(已知)的解

若将区间等分成 n 个小区间:

$$(x_i,x_{i+1}),x_i=a+ih,i=0,1,\cdots,n,h=\frac{b-a}{n}$$

$$y'(x_i)=\frac{y(x_{i+1})-y(x_i)}{h}=f(x_i,y(x_i))$$

即

$$y(x_{i+1})=y(x_i)+hf(x_i,y_i),i=0,1,\cdots,n-1$$

由 $y(x_0)=y(a)=y_0$ 及上式可求得原方程的数值解

$$y(a),y(x_i),\cdots,y(b)$$

这就是读者熟知的欧拉(Euler)法.

若将区间任意分成 n 个小区间:

$$(x_i,x_{i+1}),i=0,1,2,\cdots,n-1$$

并记

$$\Delta h_i=x_{i+1}-x_i\quad y(x_i)=y_i\quad y(x_0)=y(a)=y_0,y(x_n)=y(b).$$

原微分方程可转化成如下代数方程组:

$$y_{i+1}-y_i-\Delta h_i f(x_i,y_i)=0,\quad i=0,1,\cdots,n-1$$

此方程组是以 $y_1,y_2,\cdots,y_n$ 为未知数的方程组,当 f 为线性函数时,此方程为线性方程组;当 f 为非线性函数时,此方程为非线性代数方程组.解此方程组可得到原微分方程组的数值解 $y_1,y_2,\cdots,y_n$.

仿此可讨论求解偏微分方程的离散化技术,那时分割区域可能是平面或空间上的区域,分割的方法更多,技巧更强,详见本书第四章.

4.3 离散问题解析化技术

离散问题解析化技术系指如何根据平面或空间上若干离散点的特性,构造一解析函数来取代,使其绝大多数离散点在解析式所表示曲线或曲面上,或在其附近.此技巧一般可以归结为求解线性或非线性最小二乘问题,从几何上讲,可称为

曲线或曲面拟合问题,常采用分段折线,二次曲线,三次曲线去拟合平面上的离散点,用分片的平面,二次曲面,三次曲面去拟合三维空间上的离散点.

最简单的方法是用多项式或多元多项式去取代离散点.设平面上离散点为 (x_i, y_i), $i=1,2,\cdots,n$,拟合的多项式为 $p_m(x)=\sum_{j=0}^{m} a_j x^j$,或广义多项式 $q_m(x)=\sum_{j=0}^{m} \alpha_j x^{\beta_j}$,其中 a_j, β_j 均为待定实数,a_j 可归结为求如下规划问题:

$$\min_{a\in E^{m+1}} F(a)=\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m} a_j x_i^j\right)^2$$

α_j, β_j 可归结为求如下规划问题:

$$\min_{\substack{\alpha\in E^m\\ \beta\in E^m}} F(\alpha,\beta)=\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m} \alpha_j x_i^{\beta_j}\right)^2$$

其中

$$a=(a_0,a_1,\cdots,a_m)^T,\ \alpha=(\alpha_1,\alpha_2,\cdots,\alpha_m)^T,\ \beta=(\beta_1,\cdots,\beta_m)^T$$

由此可知,优化技术在离散问题解析化技术中的地位和作用.详细讨论可见第五章.

4.4　优化技术

严格地说优化技术包括两个部分,其一建立优化模型,其二提供求解优化模型的有效方法,应用范围极广.如何从不同的问题中,抽象出相应的模型,再根据模型的特点和工程实际问题的需要构造可行的求解方法.这里所说的不同问题,包括工程问题,几何问题,数学问题,日常生活中的问题.例如,线性或非线性方程组的求解,微分方程的近似解析解,最佳逼近函数的求法均可归结为一类特殊的优化问题——线性或非线性最小二乘问题.

例　在已知的三角形内求一点 p 使其到三边的距离之乘积最大.

解　此问题可归结为如下优化问题:

$$\begin{cases} \text{极大化函数}: f(x,y,z)=xyz \\ \text{约束条件}: ax+by+cz=2s,\ x>0,\ y>0,\ z>0 \end{cases}$$

其中 a,b,c 是三角形的三个边的边长,s 为其面积,x,y,z 为三角形内任意一点 p 到三个边 a,b,c 的距离.在此为优化变量,求解此问题的方法很多,最简单的方法是利用几何平均值不超过算术平均值.

事实上,

$$f(x,y,z)=xyz=(ax)(by)(cz)/abc\leqslant\frac{1}{abc}\left(\frac{ax+by+cz}{3}\right)^3$$

$$\leqslant \left(\frac{2s}{3}\right)^3 / abc$$

等号成立仅当 $ax = by = cz$，由约束条件解得，最大点 $x^* = \frac{2s}{3a}, y^* = \frac{2s}{3b}, z^* = \frac{2s}{3c}$，最大值 $= \frac{8s^3}{27abc}$.

从最大点的取值，由初等几何可得 p 点位置.

在工程中抽象出的优化模型，采用上例的方法求出精确解，一般是十分困难的，甚至是不可能的，只有求近似数值解，即采用各种手段，构造序列 $\{x^n\}$ 使 $f(x^n)$ 逐步逼近 $f(x)$ 在全空间或在某一给定区域上最小点或局部最小点 x^*.采用不同的技巧，构造序列 $\{x^n\}$ 就对应不同的优化方法，技巧的高低，看其 x^n 趋于 x^* 和 $f(x^n)$ 趋于 $f(x^*)$ 的速度的快慢和精度的高低.

数值计算的上述四条途径，多数用于确定型连续的数学问题，但在工程实际中还有大量的问题是不确定型且离散问题，是随机问题，解决这类问题需要概率统计的理论与方法，本书不深入讨论，只在书末把解决这类问题当前常用到的遗传算法，神经网络方法，作了一些简介.

第二章　理论基础

§1　矩　　阵

矩阵是数值计算的基本工具,其基本概念和基本理论在后面的章节中有广泛应用.

1.1　特殊矩阵

将 $m\times n$ 个数(复数或实数)写成矩阵的形式

$$A=(a_{ij})=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

则称 A 为 $m\times n$ 阶矩阵,a_{ij},$i=1,\cdots,m$,$j=1,\cdots,n$ 为矩阵的元素.如果 $m=n$,则简称为 n 阶矩阵.

n 阶矩阵 A 的主对角元素之和称为矩阵的迹:

$$\operatorname{trace}(A)=\sum_{i=1}^{n}a_{ii}$$

$$\operatorname{trace}(A+B)=\operatorname{trace}A+\operatorname{trace}B$$

设 A 为 n 阶矩阵,如果存在 $\lambda\in C$ 和非零向量 $x\in R^n\neq 0$,使得

$$Ax=\lambda x$$

则称 λ 为矩阵 A 的特征值,x 为 A 相应于特征值 λ 的特征向量.

A 的所有特征值的和等于 A 的迹

$$\operatorname{trace}A=\sum_{i=1}^{n}\lambda_i=\sum_{i=1}^{n}a_{ii}$$

其中 λ_i,$i=1,\cdots,n$ 为 A 的特征值.A 的所有特征值的积等于 A 的行列式

$$\det A=\prod_{i=1}^{n}\lambda_i$$

若 A 为 n 阶矩阵,当 $i\neq j$ 时,$a_{ij}=0$,则称 A 为对角矩阵,记为 $\operatorname{diag}(a_{ii})$.$a_{ii}=1$,$i=1,\cdots,n$ 的对角矩阵称为单位矩阵,用 I 或 E 表示;如果对 $i<j$,$(i\leqslant j)$ 有 $a_{ij}=0$,称 A 为下三角(严格下三角)矩阵,常用 L 表示;如果对 $i>j$,$(i\geqslant j)$ 有 $a_{ij}=0$,则称 A 为上三角(严格上三角)矩阵,一般用 R 或 U 表示.主对角线元素全

为 1 的上(下)三角矩阵称为单位上(下)三角矩阵.

对于 n 阶矩阵 A 若存在 n 阶矩阵 B 使

$$AB = BA = I$$

称矩阵 A 可逆或非奇异,B 称为矩阵 A 的逆矩阵,记为 $B = A^{-1}$,否则称 A 为奇异的.n 阶矩阵 A 非奇异的充要条件是 A 的行列式 $\det A \neq 0$.

两个 $m \times n$ 阶矩阵 A 和 B,若存在非奇异的 m 阶矩阵 P 和非奇异的 n 阶矩阵 Q,使得

$$B = PAQ$$

则称 A 和 B 等价,等价矩阵可以通过一系列初等变换相互转换.

对于两个 n 阶矩阵 A 和 B,若存在非奇异矩阵 P 使得

$$B = P^{-1}AP$$

则称矩阵 A 和 B 相似.相似矩阵具有相同的特征多项式和特征值,相似矩阵迹相同.任何矩阵 都相似于一个若当(Jordan)标准型.

定理 1.1 设 A 为一 $n \times n$ 的复矩阵,则存在非奇异矩阵 P,使得

$$A = P^{-1}JP$$

其中

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & 0 \\ & & \ddots & \\ & 0 & & J_s \end{pmatrix}$$

称为若当标准型 $J_i(i = 1, \cdots, s)$,有如下形式

$$J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & 0 & & \lambda_i \end{pmatrix}$$

交换矩阵 A 的行和列所得的矩阵叫做 A 的转置矩阵,用 A^T 表示.如果 $A = A^T$,则称 A 为对称矩阵.若 A 为实对称矩阵,则 A 的所有特征值都是实数,并且存在正交矩阵 Q,使得 $A = Q^T DQ$,其中 $D = \mathrm{diag}(\lambda_i)$,$\lambda_i(i = 1, \cdots, n)$ 为 A 的特征值.

如果对称矩阵 A 的各阶顺序主子式均取正值,则称 A 为对称正定矩阵,对称正定矩阵有如下等价条件:设 $A = A^T$.

1)若 $x \in R^n$ 不为零,则有 $x^T Ax > 0$.

2) A 的所有特征值均大于零.

把矩阵 A 的元素 a_{ij} 用它的共轭复数代替所得的矩阵,称为 A 的共轭矩阵,记为 $\bar{A}$.矩阵 A 的共轭矩阵 $\bar{A}$ 的转置叫做 A 的共轭转置,记为 A^*.

$$A^* = (\bar{A})^T = (\bar{A^T})$$

$$A = (A^*)^* = (A^T)^T = \bar{\bar{A}}$$

当 A 为实矩阵时，$A^* = A^T$.

对于 n 阶矩阵 A，若有 $A^* = A$，则称 A 为 Hermite 矩阵 .

在实数域中，若 n 阶矩阵 A 有 $AA^T = I$，则称 A 为正交矩阵，一般用 Q 表示，而在复数域中，对 n 阶矩阵 A，若有 $AA^* = I$，则称 A 为酉矩阵，对酉矩阵、正交矩阵有 $|\det A| = 1$，$A^* = A^{-1}$.

1.2 矩阵分解

一个矩阵通常可以分解成几个矩阵的乘积的形式，在矩阵理论的研究与应用中，把矩阵分解为一些特殊因子的乘积有着重要的意义，这些特殊的分解形式可以反映原矩阵的某些特性，如矩阵的秩、行列式、特征值或奇异值等，同时矩阵分解在矩阵的计算和理论分析中也有重要应用 .

出现较多的分解为矩阵的三角分解 .如果 n 阶矩阵 A 的各阶顺序主子式均不为零，那么 A 可分解成下三角矩阵 L 和上三角矩阵 R 的乘积：

$$A = LR \tag{1.1}$$

这种分解一般不惟一，但如果要求其中的一个三角矩阵为单位三角矩阵，则上述分解是惟一的 . R 为单位上三角矩阵时的分解叫做 Crout 分解，L 为单位下三角矩阵时的分解为 Doolittle 分解，在要求 L 与 R 同为单位三角矩阵时，把 L 或 R 中不为 1 的对角元提出来，分解式(1.1)成为

$$A = LDR \tag{1.2}$$

其中，D 为一对角矩阵，且非奇异 .当 A 为实对称正定矩阵时，A 可分解为

$$A = LL^T \tag{1.3}$$

其中，L 为主对角线元素全为正的非单位下三角矩阵，并称为 Cholesky 分解 .如要求 L 为单位下三角矩阵，类似于(1.2)式有

$$A = LDL^T \tag{1.4}$$

这时 D 为主对角线元素全为正的对角矩阵 .

在矩阵分解中一种稳定的分解为正交分解 .如果 A 为一个非奇异的 n 阶实矩阵，则存在正交矩阵 Q 以及一个主对角线元素全为正的上三角矩阵 R 使

$$A = QR \tag{1.5}$$

称为矩阵 A 的正交分解 .而当 A 为任意非奇的 n 阶矩阵时，存在酉矩阵 U 和主对角线元素全为正的上三角矩阵 R 使

$$A = UR \tag{1.6}$$

当 A 为对称正定矩阵时，则存在正交矩阵 Q 使得

$$A = QDQ^T, D = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \tag{1.7}$$

称为正交分解,其中 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 为矩阵 A 的特征值.如果 A 为复的 Hermite 矩阵,则 A 有分解式

$$A = UDU^*, \quad D = \mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n) \tag{1.8}$$

其中 U 为酉矩阵.若 A 为一个其特征值全为实数的实 n 阶矩阵时,A 可以分解为

$$A = QRQ^T \tag{1.9}$$

其中,Q 为正交矩阵,R 为上三角矩阵.事实上对任意 n 阶非奇异矩阵 A,总存在正交矩阵与 Q_1,Q_2,使

$$A = Q_1 DQ_2^T \tag{1.10}$$

其中,D 为主对角线元素全为正的对角矩阵.

对于 $m\times n$ 阶矩阵 A,如果 $\mathrm{rank}A = n$,分解式(1.6)成为

$$A = UR = U\begin{pmatrix} R_1 \\ 0 \end{pmatrix} \tag{1.11}$$

其中,R_1 为 n 阶上三角矩阵,0 为 $(m-n)\times n$ 阶零矩阵.如果 $\mathrm{rank}A = r(<n)$,矩阵 A 的分解式为

$$A = U\begin{pmatrix} R_{11} & 0 \\ 0 & 0 \end{pmatrix} V^* \tag{1.12}$$

其中,R_{11} 为秩为 r 的 r 阶上三角矩阵,其主对角线元素全为正,V 为 n 阶酉矩阵.事实上 $\mathrm{rank}A = r$ 的 $m\times n$ 阶矩阵 A 还可得到更简单的分解式

$$A = U\begin{pmatrix} \Omega & 0 \\ 0 & 0 \end{pmatrix} V^* \tag{1.13}$$

其中,$\Omega = \mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$,$\sigma_i^2 = \lambda_i$,$i=1,2,\cdots,r$,$\lambda_i$ 为矩阵 A^*A 的全部非零特征值,而在 A 为实矩阵时,(1.13)式为

$$A = P\begin{pmatrix} \Omega & 0 \\ 0 & 0 \end{pmatrix} Q^T \tag{1.14}$$

其中,P,Q 为正交矩阵,$\sigma_i>0$,$i=1,2,\cdots,r$,为矩阵 A 的奇异值,而(1.14)式称为矩阵 A 的奇异值分解.

§2 向量和矩阵的范数

在数值计算中,我们需要衡量向量的大小、两个向量点之间的距离,判别 n 维向量点、列矩阵序列的收敛性,这就需要引入范数的概念.

2.1 向量范数

定义 2.1 设 C^n 是复数域 P 上的线性空间,如果函数 $\|\cdot\|: C^n \to R^+$ 满足

(1) 正定性：$\| x \| \geqslant 0, x \in E^n$，当且仅当 $x = 0$ 时，$\| x \| = 0$；

(2) 齐次性：$\| \alpha x \| = | \alpha | \| x \|, x \in C^n, \alpha \in C$；

(3) 三角不等式 $\| x + y \| \leqslant \| x \| + \| y \|, x, y \in C^n$，则称 $\| \cdot \|$ 为 C^n 上的范数.

常用的范数为 p-范数(也叫 l_p 范数)

$$\| x \|_p = \left(\sum_{i=1}^{n} | x_i |^p \right)^{1/p}, \qquad 1 \leqslant p < \infty$$

当 $p = 1, 2, \infty$ 时，分别称为1-范数、2-范数和 ∞-范数或 l_1 范数、l_2 范数和 l_∞ 范数，记为

$$\| x \|_1 = \sum_{i=1}^{n} | x_i |$$

$$\| x \|_2 = \left(\sum_{i=1}^{n} | x_i |^2 \right)^{1/2}$$

$$\| x \|_\infty = \max_{1 \leqslant i \leqslant n} | x_i |$$

另一种常见的向量范数为加权范数或椭圆范数，它依据对称正定矩阵 A 来定义：

$$\| x \|_G = [x^T G^T G x]^{1/2} = (x^T A x^T)^{1/2} = \| Gx \|_2$$

其中 G 为非奇矩阵且满足 $G^T G = A$.

可以证明，向量范数是其各分量的连续函数，范数是向量大小的一种度量.同一个向量，用不同的范数得到的数值是不同的，但它们之间有着密切的联系.

定理 2.1(范数等价性定理) 设 $\| \cdot \|_A$ 和 $\| \cdot \|_B$ 是定义于 C^n 上的两种范数，则存在常数 c_1 和 c_2，使得对任意 $x \in C^n$，都有

$$c_1 \| x \|_A \leqslant \| x \|_B \leqslant c_2 \| x \|_A \tag{2.1}$$

例如，对 l_1, l_2, l_∞ 范数，有

$$\| x \|_\infty \leqslant \| x \|_1 \leqslant n \| x \|_\infty$$

$$\| x \|_\infty \leqslant \| x \|_2 \leqslant \sqrt{n} \| x \|_\infty$$

$$\frac{1}{\sqrt{n}} \| x \|_1 \leqslant \| x \|_2 \leqslant \| x \|_1$$

2.2 矩阵范数

与向量范数对应，矩阵范数是定义在全体 $m \times n$ 矩阵或 n 阶矩阵集合上的实值函数.

2.2.1 范数的定义

定义 2.2(矩阵范数) 对任意 n 阶复矩阵 A 和 B，如果函数 $\| \cdot \| : C^{m \times n} \to$

R^+ 满足

(1) 正定性：$\| A \| \geqslant 0$，当且仅当 $A = 0$ 时，$\| A \| = 0$；

(2) 齐次性：$\| \alpha A \| = |\alpha| \| A \|$，$\alpha \in C, A \in C^{m\times n}$；

(3) 三角不等式：$\| A + B \| \leqslant \| A \| + \| B \|$；

(4) 相容性：$\| AB \| \leqslant \| A \| \cdot \| B \|$.

则称 $\| \cdot \|$ 为复线性空间 $C^{m\times n}$ 上的矩阵范数，也记为 $\| A \|$.

常用的矩阵范数是 Frobenius 范数或叫 Euclid 范数：

$$\begin{aligned} \| A \|_F &= \left(\sum_{j=1}^{n} \sum_{i=1}^{n} |a_{i,j}|^2 \right)^{1/2} \\ &= [\text{trace}(A^* A)]^{1/2} \\ &= \left[\sum_{i=1}^{n} \| A v_i \|^2 \right]^{1/2} \end{aligned}$$

其中 a_{ij} 为 A 的第 i 行、j 列元素，$v_1, v_2, \cdots, v_n$ 是 n 维空间中的一组标准正交基.

2.2.2 相容范数和导出范数

矩阵范数可以理解为矩阵大小的一种度量，当矩阵 A 只有一列或一行元素时，A 成为一个向量，如果矩阵范数 $\| A \|$ 和向量范数 $\| x \|$ 满足

$$\| Ax \| \leqslant \| A \| \cdot \| x \|$$

则称 $\| A \|$ 是与向量范数 $\| x \|$ 相容的矩阵范数.

对于矩阵范数，可以通过已知的向量范数来定义与之相容的矩阵范数.设 $\| x \|$ 为一种向量范数，对任意 n 阶矩阵 A，把向量 Ax 的范数在单位球面 $\| x \| = 1$ 上的最大值定义为矩阵 A 的范数，即

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

可以证明，这样定义的 $\| A \|$ 满足矩阵范数定义的四个条件，并称这种矩阵范数是由向量范数导出的矩阵范数，并且它与 $\| x \|$ 是相容的.

与向量的 l_1, l_2, l_m 范数对应，由它们导出的矩阵范数分别为

$$\| A \|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|$$

$$\| A \|_2 = \sqrt{\lambda_n}$$

$$\| A \|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|$$

其中 λ_n 为 $A^* A$ 的最大特征值.另一种常用且与 l_2 向量范数相容的矩阵范数是 F-范数：

$$\begin{aligned}\| A \|_F &= \Big(\sum_{j=1}^{n} \sum_{i=1}^{n} | a_{i,j} |^2 \Big)^{1/2} \\ &= [\operatorname{trace}(A^* A)]^{1/2} \\ &= \Big[\sum_{i=1}^{n} \| A v_i \|^2 \Big]^{1/2}\end{aligned}$$

其中 $v_1, v_2, \cdots, v_n$ 是 n 维空间中的一组标准正交基.它不但比 l_2 范数便于计算,而且还有一个重要的性质 —— 酉不变性,即对任意 n 阶酉矩阵 P 和 Q,有

$$\| A \|_F = \| PA \|_F = \| AQ \|_F = \| PAQ \|_F$$

且满足

$$\| AB \|_F \leqslant \min\{ \| A \|_F \| B \|_2, \| A \|_2 \cdot \| B \|_F \}$$

与加权向量范数相容的矩阵范数为

$$\| A \|_G = \| GAG^{-1} \|_2$$

其中 G 为非奇异矩阵,这种加权的矩阵范数在数值计算的误差分析中经常使用.

2.2.3 矩阵范数的等价性

与向量范数类似,各种不同的矩阵范数之间也存在等阶性.

定理 2.2 对于 n 阶矩阵集合上任意两个由向量范数导出的矩阵范数 $\| \cdot \|_A$ 和 $\| \cdot \|_B$ 有

$$c_1^2 \| A \|_A \leqslant c_1 c_2 \| A \|_B \leqslant c_2^2 \| A \|_A$$

其中 c_1, c_2 由(2.1)确定.

2.2.4 矩阵的条件数

在数值计算中,矩阵的奇异性常常给问题的求解带来很大的困难,而矩阵的条件数则可以衡量矩阵的奇异程度.

定义 2.3 对于一个非奇异 n 阶矩阵 A,称

$$k(A) = \| A \| \cdot \| A^{-1} \|$$

为矩阵 A 的条件数.

由矩阵范数的相容性,有

$$k(A) = \| A \| \cdot \| A^{-1} \| \geqslant \| AA^{-1} \| = \| I \| = 1$$

即矩阵的条件数总不小于 1.条件数越小,则 A 的非奇异程度越高,称 A 为良态的;条件数越大,则 A 的非奇异程度越差,称 A 为病态的.如果 λ_1, λ_n 分别是 A 的模最小和模最大的特征值,则

$$k(A) = | \lambda_n | / | \lambda_1 |$$

2.2.5 谱半径

矩阵的谱半径是另一个衡量矩阵“大小”的量.

定义 2.4 设 A 是一个 $n\times n$ 的复矩阵,$\lambda_i(i=1,2,\cdots,n)$ 是 A 的特征值,则

$$\rho(A)=\max_{1\leqslant i\leqslant n}\{|\lambda_i|\}$$

称为 A 的谱半径.

(1) 对任何一种相容的矩阵范数 $\|A\|$ 均有

$$\rho(A)\leqslant\|A\|$$

(2) 对于任意一个 $\varepsilon>0$,总存在相容的矩阵范数 $\|A\|$,使得

$$\|A\|\leqslant\rho(A)+\varepsilon$$

(3) 对任意 n 阶矩阵 A,

$$\|A\|_2=[\rho(A^*A)]^{\frac{1}{2}}$$

(4) 如果 A 为一个正规矩阵或 Hermite 矩阵,有

$$\|A\|_2=\rho(A)$$

(5) 对任意 n 阶矩阵 A 以及任意 $k=0,1,2,\cdots$,有

$$\rho(A^k)=\rho^k(A)$$

(6) 如果 $\rho(A)<1$,当且仅当 $\lim\limits_{k\to 0}A^k=0$ 且 $(I-A)$ 可逆,其逆为

$$(I-A)^{-1}=I+A+A^2+\cdots$$

如果 $\|A\|<1$,则还可得

$$\|(I-A)^{-1}\|\leqslant\sum_{i=0}^{\infty}\|A\|^i=\frac{1}{1-\|A\|}$$

注意,虽然谱半径可以理解为矩阵大小的一种度量,但它不是范数.

§3 集合的基本概念

3.1 开集与闭集

定义 3.1 $x\in R^n,\varepsilon>0$,集合

$$S(x_0,\varepsilon)=\{x\in R^n\mid\|x-x_0\|<\varepsilon\}$$

称为以 x_0 为中心,以 ε 为半径的开球,或点 x_0 的 ε 邻域.

定义 3.2 设 Ω 是 R^n 的子集.若 $x\in\Omega$,且存在开球 $S(x,\varepsilon)\subset\Omega$,则称 x 为 A 的一个内点;若 $x\in R^n\setminus\Omega$,且存在开球 $S(x,\varepsilon)\in R^n\setminus\Omega$,则称 x 为 Ω 的外点;若 $x\in R^n$ 既非 Ω 的内点也非外点,则称 x 为 Ω 的边界点.Ω 的内点的全体称为 Ω 的内部,记作 $\operatorname{int}\Omega$;Ω 的外点的全体称为 Ω 的外部,A 的边界点的全体称为 Ω 的边界,记作 $\partial\Omega$.

定义 3.3 设 $\Omega \subset R^n$.若 $\Omega = \operatorname{int}\Omega$,即对任意 $x \in \Omega$,存在 $S(x,\varepsilon) \subset \Omega$,则称 Ω 是 R 中的开集;若 $R^n \setminus \Omega$ 是 R^n 中的开集,则 Ω 为 R^n 中的闭集.

在 R^1 中,开区间是开集,闭区间是闭集,半开半闭区间即非开集也非闭集.R^n,Φ 单点子集既是开集又是闭集.

定义 3.4 Ω 为 R^n 中的一个点集,若存在常数 C 使得对所有 $x \in \Omega$,均有 $\| x \| \leqslant C$,则称 Ω 为有界集,其中 $\| \cdot \|$ 为任何一种范数.

定理 3.1 有限多个开集的交集是开集,有限多个闭集的并集是闭集.任意多个开集的并集是开集,任意多个闭集的交集仍是闭集.

3.2 极限与收敛

定义 3.5 设 $\{x_n\}$ 是 R^n 中的点列,$x^* \in R^n$.若对任意的 $\varepsilon > 0$,存在自然的 N,当 $n > N$ 时,$\| x_n - x^* \| < \varepsilon$,则称 x^* 为 $\{x_n\}$ 的一个极限点,或说 $\{x_n\}$ 收敛到点 x^*,记作 $\lim\limits_{n\to\infty} x_n = x^*$.

定理 3.2 R^n 中任一个有界无穷点列(集)必有收敛子列.

定义 3.6 设 $\{x_n\}$ 是 R^n 中的一个子列,对于任意给定的 $\varepsilon > 0$,存在自然数 $N > 0$,对任意 $n > N$ 和任意正整数 m 有

$$\| x_{n+m} - x_n \| < \varepsilon$$

成立,则称 $\{x_n\}$ 为一 Cauchy 序列.

Cauchy 序列是判别一个序列收敛的有利工具.

定理 3.3 在 R^n 空间中 $\{x_n\}$ 收敛的充分必要条件是 $\{x_n\}_{n=1}^{\infty}$ 是 Cauchy 序列.

§4 凸集与凸函数

4.1 凸集

定义 4.1 设 S 为 R^n 中的一个集合,若对 S 中任意两点 x,y,都有

$$\lambda x + (1-\lambda) y \in S, \quad 0 \leqslant \lambda \leqslant 1$$

则称 S 为凸集,$\lambda x + (1-\lambda) y$ 称为 x,y 的凸组合.

例 4.1 超平面 $H \overset{\Delta}{=} \{x \mid p^T x = \alpha, p \in R^n\}$ 为凸集.

例 4.2 半空间 $H^- \overset{\Delta}{=} \{x \mid p^T x \leqslant \alpha, p \in R^n, \alpha \in R\}$ 为凸集.

定义 4.2 设 M 是 R^n 中的任一集合,包含 M 的 R^n 中的最小凸集称为 M 的凸包,记为 $\operatorname{Co} M$.

设 x_1,x_2 是 R^n 中两个不同的点,则集合 $\{x_1, x_2\}$ 的凸包是以 x_1,x_2 为端点的

线段.若 x_1, x_2, x_3 是 R^n 中不共线的3点,则集合$\{x_1, x_2, x_3\}$的凸包是以 x_1, x_2, x_3 为顶点的三角形.一般地,设$\{x_0, x_1, \cdots, x_k\}$是 R^n 中 $k+1$ 个仿射无关的点(即 $x_1 - x_0, \cdots, x_k - x_{k-1}$ 线性无关),则点集$\{x_0, \cdots, x_k\}$的凸包 $\sigma_k = \mathrm{Co}\{x_0, \cdots, x_k\}$ 称为一个 k 维单纯形.

4.2　凸函数

定义 4.3　设 D 是 R^n 中的非空凸集,函数 $f: D \subset R^n \to R^1 = \{-\infty, +\infty]$称为是凸的,如果对所有 $x, y \in D$,及 $0 < \lambda < 1$,

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y)$$

如果当 $x \neq y$ 时,上述严格不等式成立,则称 $f(x)$ 是严格凸的.如果存在正常数 $c > 0$,使对所有 $x, y \in D$,$0 < \lambda < 1$,有

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y) - c\lambda(1-\lambda) \| x - y \|^2$$

则称 $f(x)$ 是一致凸的.显然凸,严格凸和一致凸有下列关系:

$$\text{一致凸} \Rightarrow \text{严格凸} \Rightarrow \text{凸}$$

有限个凸函数的非负线性组合是凸函数.

定理 4.1　设 S 是 R^n 中的一个非空凸集,f 是定义在 S 上的凸函数,则水平集 $S_\alpha = \{x \mid x \in R^n, f(x) \leqslant \alpha\}$ 是凸集.

定理 4.2　设 f 是凸集 D 上的凸函数,且 $f \in C^1$,那么,对任何 $x \in D$,$y \in D$,有

$$f(y) \geqslant f(x) + (y-x)^T \nabla f(x)$$

从几何上看,对于连续可微的凸函数 f,其函数图形位于其上任一点 x 处的切平面的上方.

若 $\nabla f(x)$ 不存在,我们也可以通过引入次梯度的概念来刻画凸函数的几何特性.

定义 4.4　对于凸函数 $f: R^n \to R^1 \cup \{+\infty\}$,在给定点 $x \in R^n$,若方向 $\xi \in R^n$ 满足

$$f(y) \geqslant f(x) + \xi^T (y-x), \forall\, y \in R^n$$

则称 ξ 为凸函数 f 在点 x 处的次梯度,称这些次梯度构成的集合为 f 在 x 处的次微分,记为

$$\partial f(x) \triangleq \{\xi \mid \xi \in R^n, \forall\, y \in R^n, f(y) \geqslant f(x) + \xi^T (y-x)\}$$

在几何上,曲面 $z = f(y)$ 位于超平面

$$z = f(x) + \xi^T (y-x)$$

上方,且与超平面交于 $y = x$ 点.我们称这样的超平面为凸函数 f 的支撑超平面.

定理 4.3　设 D 是 R^n 中的非空开凸集,$f(x)$ 是定义在 D 上的可微函数,则

$f(x)$为凸函数的充要条件是对任意两点 $x, y \in D$,成立

$$f(y) \geqslant f(x) + (y-x)^T \nabla f(x)$$

而 $f(x)$为严格凸函数的充要条件是上式不等号严格成立.

定理 4.4 设 D 是 R^n 中的非空开凸集,$f(x)$定义于 D 且 $f \in C^2$,则 $f(x)$为凸函数的充要条件是对任何 $x \in D$,二阶导数矩阵 $\nabla^2 f(x)$ 是半正定的.

定理 4.5 设 D 是 R^n 中的非空开凸集,$f(x)$是定义在 S 上的二次可微函数,如果在每一点 $x \in D$,$\nabla^2 f(x)$ 是正定的,则 $f(x)$为严格凸函数.

注意:定理 4.5 的逆定理不成立.

凸函数在优化中有着极好的性质.

定理 4.6 设 D 是 R^n 中的非空凸集,$f(x)$是定义在 D 上的凸函数,则 $f(x)$在 S 上的局部极小点是整体极小点,且极小点的集合是凸集.

§5 多元函数

设 $f: D \subset R^n \to R$,则称 f 是定义在 D 上的 n 元函数.

5.1 多元函数的连续性

定义 5.1 连续函数.设 $f: D \subset R^n \to R$, $x_0 \in D$.如果对于任意给定的 $\varepsilon > 0$,存在 $\delta(\varepsilon)$,当 $\| x - x_0 \| \leqslant \delta(\varepsilon)$ 时,有

$$| f(x) - f(x_0) | < \varepsilon$$

成立,则称 $f(x)$在 x_0 连续.

若 $f(x)$在 D 上每个点连续,则称 $f(x)$为 D 上的连续函数.

定义 5.2 一致连续.设 $f: D \subset R^n \to R$ 在 D 上连续,如果对任意 $\varepsilon > 0$,存在 $\delta(\varepsilon) > 0$,对任何 $x, y \in D$,当 $\| x - y \| < \delta(\varepsilon)$ 时,都有

$$| f(x) - f(y) | < \varepsilon$$

则称 $f(x)$在 D 上是一致连续的.

一致连续必定连续,但连续不一定一致连续,但如果连续函数定义在有界闭区域上,则有

定理 5.1 任何定义在有界闭区域上的连续函数是一致连续的.

比一致连续更强的连续条件是 Hölder 连续和 Lipschitz 连续.

定义5.3 设 $f: \bar{D} \subset R^n \to R$ 为定义在闭区域上的一个多元函数,若存在常数 $L > 0$,使得对任意 $x, y \in \bar{D}$,有

$$| f(x) - f(y) | \leqslant L \| x - y \|^p$$

其中 $0 < p \leqslant 1$,则称 $f(x)$在 $\bar{D}$ 上是 Hölder 连续的,如果 $p = 1$,则称 $f(x)$在 $\bar{D}$