



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

计算方法

大连理工大学应用数学系数学与应用数学专业基础课程



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

计

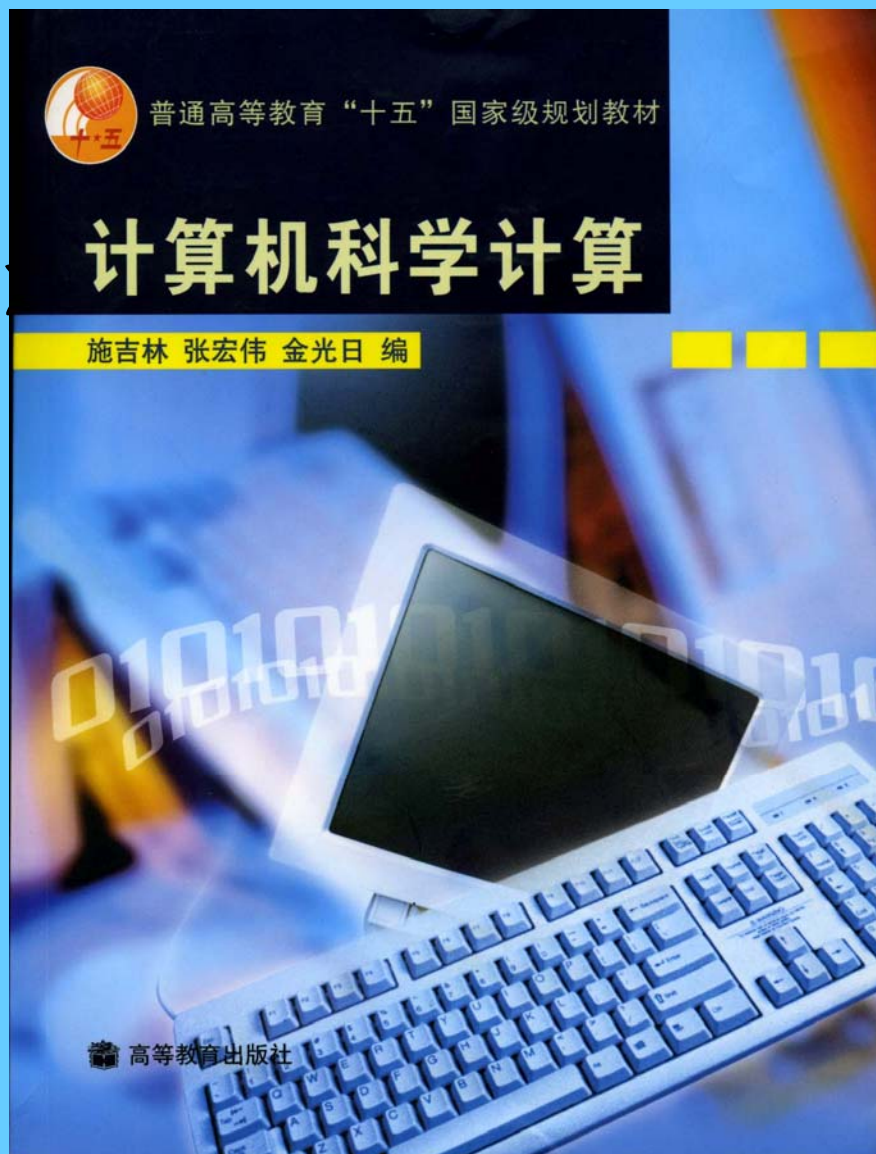


普通高等教育“十五”国家级规划教材

计算机科学计算

施吉林 张宏伟 金光日 编

算



高等教育出版社



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY



**知识，只有当它靠积极的
思考得来而不是凭记忆得来的
时候，才是真正的知识。**

——列夫·托尔斯泰



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

第1章 绪 论

1.1 计算机科学计算研究对象与特点



1.2 误差分析与数值方法的稳定性



1.3 向量与矩阵的范数





DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

1.1 计算机科学计算研究对象与特点

科学计算、理论计算和实验并列为三大科学方法。现代意义下的计算数学主要研究在计算机上计算的有效算法及其相关理论，从而使它成为一门新学科——科学计算。

本课程主要研究现代、行之有效数值方法



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

本课程主要研究用计算机求解各种数学问题的数值计算方法及其理论与软件实现

主要包括:

数值代数

$$Ax = b$$

$$f(x) = 0$$

数值逼近 (数值微分积分)

$$f(x) \quad f'(x)$$

$$\int_a^b \rho(x) f(x) dx$$

微分方程数值解法

$$u' = f(t, u), u(t_0) = u_0$$

矩阵分析简介



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

课程的特点:

一、构造计算机可行的有效算法

二、给出可靠的理论分析，即对任意逼近并达到精度要求，保证数值算法的收敛性和数值稳定性，并可进行误差分析。

三、有好的计算复杂性，既要时间复杂性好，是指节省时间，又要空间复杂性好，是指节省存储量，这也是建立算法要研究的问题，它关系到算法能否在计算机上实现。

四、数值实验，即任何一个算法除了从理论上要满足上述三点外，还要通过数值试验证明是行之有效的。



什么是有效算法？



考察，线性方程组的解法

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}$$

早在18世纪Cramer已给出了求解法则：

Cramer's Ruler



$$x_i = \frac{D_i}{D} \quad i = 1, 2, \dots, n \quad (D \neq 0)$$

$$D = \det(A) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

Cramer's
Ruler

第
 i
列

$$D_i = \det(A_i) = \begin{vmatrix} a_{11} & \cdots & b_1 & \cdots & a_{1n} \\ a_{21} & \cdots & b_2 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & b_n & \cdots & a_{nn} \end{vmatrix}$$



从理论上讲 **Cramer**法则是一个求线性方程组的数值方法，且对阶数不高的方程组行之有效。但是理论正确的数值方法在计算机上是否实际可行。

以求解**20**阶线性方程组为例，如果用**Cramer**法则求解，在算法中运用行列展开计算，则总的乘、除运算次数将达：

$$21! = 9.7 \times 10^{20} \text{ (次)}$$

若使用每秒一亿次的串行计算机计算，一年可进行的运算应为：

$$365(\text{天}) \times 24(\text{小时}) \times 3600(\text{秒}) \times 10^9 \approx 3.5 \times 10^{15} \text{ (次)}$$

共需要耗费时间为：

$$(9.7 \times 10^{20}) \div (3.5 \times 10^{15}) \approx 3.97 \times 10^5 \approx 30(\text{万年})$$



DUT

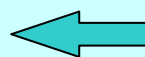
大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

Cramer 算法是“实际计算不了”的。为此，人们研究出著名的 **Gauss**消去数值方法，它的计算过程已作根本改进，使得上述例子的乘、除运算仅为3060次，这在任何一台电子计算机上都能完成。

随着科学技术的发展，出现的数学问题也越来越多样化，有些问题用消去法求解达不到精度，甚至算不出结果，从而促使人们对消去法进行改进，又出现了主元消去法，大大提高了消去法的计算精度。

寻求新的数值方法，这就是计算机科学计算生命力的来源。



返回本章



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

1.2 误差分析与数值方法的稳定性

1.2.1 误差来源与分类

1.2.2 误差的基本概念和有效数字

1.2.3 函数计算的误差估计

1.2.4 数值方法的稳定性和避免误差危害的基本原则



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

1.2.1 误差来源与分类

用计算机解决科学计算问题时经常采用的处理方式是将连续的问题离散化、用有限代替无限等，并且用数值分析所处理的一些数据，不论是原始数据，还是最终结果，绝大多数都是近似的，因此在此过程中，误差无处不在。

误差的来源主要从以下几个方面：

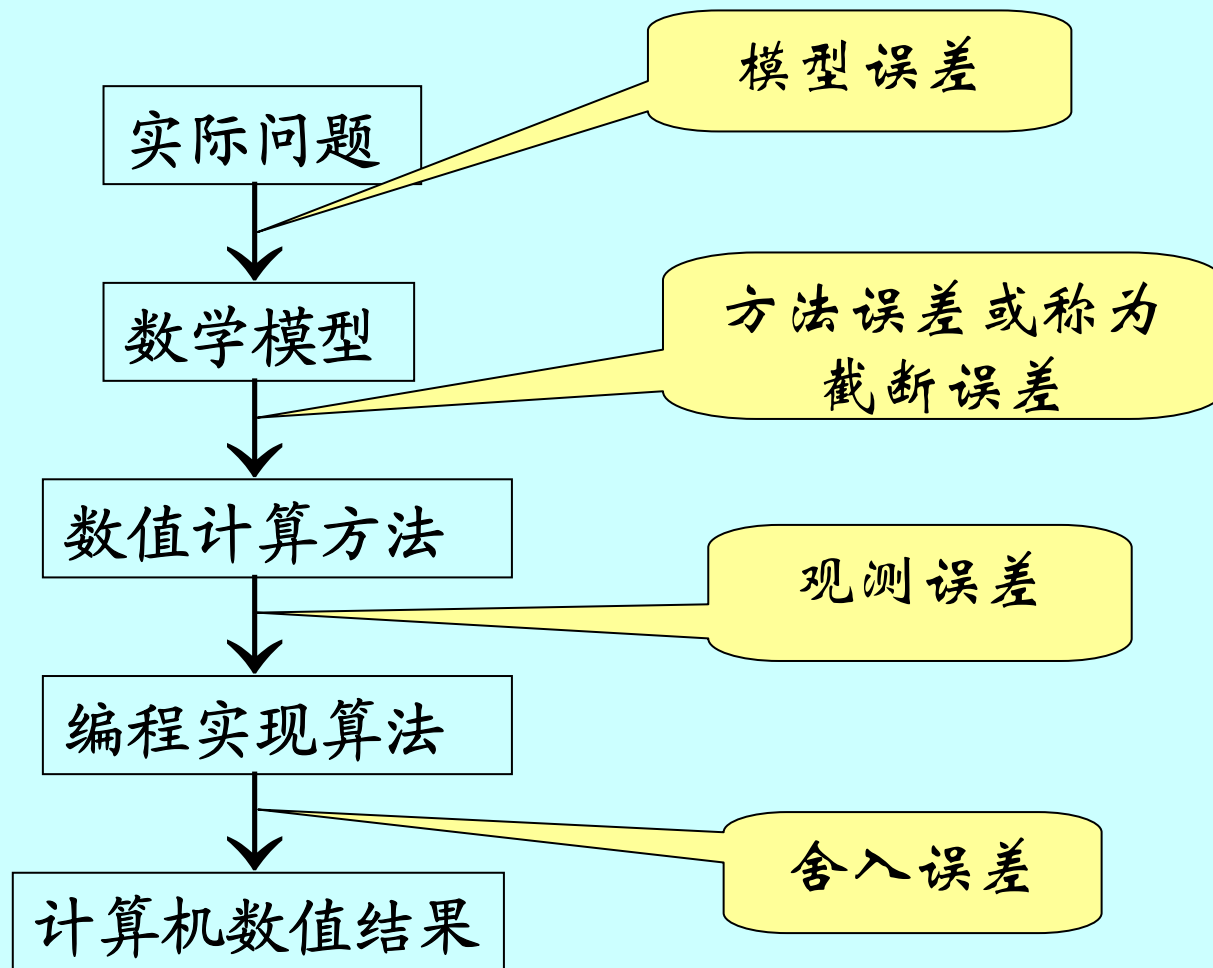


DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

计算机科学计算的流程图





DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

1. **模型误差** 由实际问题抽象出数学模型，要简化许多条件，这就不可避免地要产生误差。实际问题的解与数学模型的解之间的误差

2. **截断误差** 从数学问题转化为数值问题的算法时所产生的误差，如用有限代替无限的过程所产生的误差

截断误差通常是指用一个基本表达式替换一个相当复杂的算术表达式时所引起的误差。这一术语从用截断Taylor级数替换一个复杂的算术表达式的技术中衍生而来。



例如，给定 x 求 e^{x^2} 的值的运算，我们可用无穷级数：

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \cdots + \frac{x^{2n}}{n!} + \frac{x^{2(n+1)}}{(n+1)!} + \cdots$$

我们可用它的前 $n+1$ 项和

$$s(x) =$$

近似代替函数 e^{x^2} ，则数值方法的误差是

$$R_n(x) = e^{x^2} - s(x) = \frac{e^{(\theta x)^2}}{(n+1)!} x^{2(n+1)}, \quad 0 < \theta < 1$$

截断
误差



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

3. **观测误差** 初始数据大多数是由观测而得到的。由于观测手段的限制，得到的数据必然有误差

4. **舍入误差** 以计算机为工具进行数值运算时，由于计算机的字长有限，原始数据在计算机上的表示往往会有误差，在计算过程中也可能产生误差

例如，用 1.4121 近似代替 $\sqrt{2}$ ，产生的误差

$$E = \sqrt{2} - 1.41421 = 0.00000365 \dots$$

就是**舍入误差**。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

模型和观测两种误差不在本课程的讨论范围

这里主要讨论算法的截断误差与舍入误差，而截断误差将结合具体算法讨论

分析初始数据的误差通常也归结为舍入误差

研究计算结果的误差是否满足精度要求就是：

误差估计问题



返回本节



1.2.2 误差的基本概念和有效数字

定义1.4 设 x 为精确值， a 为 x 的一个近似值，称

$$x - a$$

绝对误差 (误差)

为近似值的**绝对误差**，简称**误差**。误差 $x-a$ 可正可负。

通常准确值 x 是未知的，因此误差 $x-a$ 也未知。

定义1.5 设 x 为精确值， a 为 x 的一个近似值，若有常数 e_a 使得

$$|x - a| \leq e_a$$

绝对误差界

(1-13)

则 e_a 叫做近似值的**误差界 (限)**。它总是正数。



例如，用毫米刻度的米尺测量一长度 x ，读出和该长度接近的刻度 a ， a 是 x 的近似值，它的误差界是 0.5mm ，于是有

$$|x - a| \leq 0.5 \text{ mm}$$

绝对误差界

如若读出的长度为 765mm ，则有，

$$|x - 765| \leq 0.5$$

虽然从这个不等式不能知道准确的 x 是多少，但可知

$$764.5 \leq x \leq 765.5,$$

结果说明 x 在区间 $[764.5, 765.5]$ 内。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

对于一般情形 $|x - a| \leq e_a$ ，即可以表示为

$$a - e_a \leq x \leq a + e_a,$$

也可以表示为

$$x = a \pm e_a。$$

但要注意的是，误差的大小并不能完全表示近似值的好坏。

**定义**若 $x \neq 0$ ，则将近似值的误差与准确值的比值

$$\frac{x - a}{x}$$

相对误差 (误差)

称为近似值 a 的**相对误差**。相对误差也可正可负。实际计算中，如果真值 x 未知时通常取

$$\frac{x - a}{x} \approx \frac{x - a}{a}$$

作为 a 的相对误差，条件是 $\frac{x - a}{x}$ 较小。



这是由于两者之差

$$\begin{aligned}\frac{x-a}{a} - \frac{x-a}{x} &= \frac{(x-a)^2}{x \cdot a} = \frac{(x-a)^2}{x \cdot (x - (x-a))} \\ &= \left(\frac{x-a}{x} \right)^2 \times \frac{1}{1 - \frac{x-a}{x}}\end{aligned}$$

是 $\frac{x-a}{x}$ 的平方项级，故可忽略不计。



例

有两个量 $x=3.000$, $x=3.100$, 则其绝对误差:

$$x - a = -0.1$$

绝对误差

其相对误差为:

$$\frac{x - a}{x} = \frac{-0.1}{3.00} = -0.333 \times 10^{-1},$$

相对误差

又有两个量 $x=300.0$, $a=310.0$, 则其绝对误差:

$$x - a = -0.1 \times 10^2,$$

绝对误差

其相对误差为:

$$\frac{x - a}{x} = \frac{-0.1 \times 10^2}{0.3 \times 10^4} = -0.333 \times 10^{-1}$$

相对误差



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

上例说明绝对误差有较大变化，相对误差相同。作为精确值的度量，绝对误差可能会引起误会，而相对误差由于考虑到准确值本身的大小而更有意义。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

相对误差的绝对值上界叫做**相对误差界（限）**，记为：

$$\left| \frac{x-a}{a} \right| \leq \frac{e_a}{|a|}$$

相对误差界（限）



例1

已知 $e = 2.71828182 \dots$ 其近似值 $a = 2.718$, 求 a 的绝对误差界和相对误差界。

解: $e - a = 0.00028182 \dots$, 因此其绝对误差界为:

$$|e - a| \leq 0.0003$$

相对误差界为:

$$\frac{|e - a|}{|a|} = \frac{0.0003}{2.718} \approx 0.0001110375 \leq 0.0002.$$

此例计算中不难发现, 绝对误差界和相对误差界并不是唯一的。我们要注意它们的作用





DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

误差界的取法

当准确值 x 位数比较多时，常常按四舍五入的原则得到 x 的前几位近似值 a ，例如

$$x = \pi = 3.14159265 \dots$$

取3位 $a_1 = 3.14$, $\pi - a_1 = 0.00159265 \dots$

取5位 $a_2 = 3.1416$, $\pi - a_2 = -0.00000735 \dots$

它们的误差界的取法应为：

$$|\pi - 3.14| \leq \frac{1}{2} \times 10^{-2}, \quad |\pi - 3.1416| \leq \frac{1}{2} \times 10^{-4}.$$



定义1.6 设 x 为精确值, a 为 x 的一个近似值, 表示为:

$$a = \pm 10^k \times 0.a_1a_2 \cdots a_n \cdots \quad (1-14)$$

可以是有限或无限小数形式, 其中 $a_i (i = 1, \cdots, n)$ 是0到9中的一个数字, $a_1 \neq 0$, k 为整数, n 为正整数, 如果其绝对误差界

$$|x - a| \leq \frac{1}{2} \times 10^{k-n}. \quad (1-15)$$

则称 a 为 x 的具有 n 位**有效数字**的近似值。



在例1中，由于

$$|e - a| < 0.0003 < \frac{1}{2} \times 10^{-3},$$

而 $a = 10^1 \times 0.2718$ ，那么， $k - n = -3 \Rightarrow n = 4$ ，即 a 是 $e = 2.71828182\dots$ 的具有4位有效字的近似值。如果取

$$a_1 = 2.7182 = 10^1 \times 0.27182$$

因

$$|e - a_1| < 0.00009 < \frac{1}{2} \times 10^{-3}$$

a_1 也只是 e 的具有4位有效数字的近似值。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

同样我们可以分析出 $a = 0.02718 = 10^{-1} \times 0.2718$ 作为

$$x = 0.0271828182\dots$$

的近似值，也具有4位有效数字。这是因为：

$$|x - a| < 0.000002 < \frac{1}{2} \times 10^{-5}$$

那么，有 $k - n = -5 \Rightarrow n = 4$ 。

这表明：有效数字位数与小数点的位置无关



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

如果一个近似值是由精确值经四舍五入得到的，那么，从这个近似值的末尾数向前数起直到再无非零数字止，所数到的数字均为有效数字

一般来说，绝对误差与小数位数有关，相对误差与有效数字位数有关



练习1

下列近似值的绝对误差限均为0.005，问它们各有几位有效数字？

$$a = 138.00, \quad b = -0.0312, \quad c = 0.86 \times 10^{-4}.$$

解：首先将它们表示成标准形式

$$a = 0.13800 \times 10^3, \quad b = -0.312 \times 10^{-1}, \quad c = 0.86 \times 10^{-4}.$$

则由已知条件，

$$|x - a| \leq \frac{1}{2} \times 10^{-2} \Rightarrow 3 - n = -2 \Rightarrow n = 5$$

即 a 有5位有效数字；



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

同理，由 $b = -0.312 \times 10^{-1}$,

$$\begin{aligned} |x - b| \leq \frac{1}{2} \times 10^{-2} &\Rightarrow -1 - n = -2 \\ &\Rightarrow n = 1 \end{aligned}$$

即 b 有 1 位有效数字;

由 $c = 0.86 \times 10^{-4}$,

$$\begin{aligned} |x - c| \leq \frac{1}{2} \times 10^{-2} &\Rightarrow -4 - n = -2 \\ &\Rightarrow n = -2 \end{aligned}$$

即 c 无有效数字。



定理 1.7 设实数 x 为某个精确值, a 为它的一个近似值,

其表达形式如 $a = \pm 10^k \times 0.a_1a_2 \cdots a_n \cdots$

(1) 如果 a 有 n 位有效数字, 则

$$\frac{|x-a|}{|a|} \leq \frac{1}{2a_1} \times 10^{1-n}, \quad (1-16)$$

(2) 如果

$$\frac{|x-a|}{|a|} \leq \frac{1}{2(a_1+1)} \times 10^{1-n}, \quad (1-17)$$

则 a 至少具有 n 位有效数字。

**证**

由 (1-14) 可得到

$$a_1 \times 10^{k-1} \leq |a| \leq (a_1 + 1) \times 10^{k-1} \quad (1-18)$$

所以如果 a 有 n 位有效数字, 那么

$$\frac{|x-a|}{|a|} = |x-a| \times \frac{1}{|a|} \leq \frac{1}{2} \times 10^{k-n} \times \frac{1}{a_1 \times 10^{k-1}} = \frac{1}{2a_1} \times 10^{1-n},$$

结论 (1) 成立。再由 (1-17) 和 (1-18)

$$\frac{|x-a|}{|a|} \leq \frac{|a|}{2(a_1+1)} \times 10^{1+n} \leq \frac{(a_1+1) \times 10^{k-1}}{2 \times (a_1+1)} \times 10^{1-n} = \frac{1}{2} \times 10^{k-n},$$

由定义 1.6 知, a 具有 n 位有效数字。
 返回本节



1.2.3 函数计算的误差估计

设一元函数 $f(x)$ 具有二阶连续导数, 自变量 x 的一个近似值为 a , $f(a)$ 作为 $f(x)$ 近似。我们用Taylor展开的方法来估计其误差。即有

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(\xi)(x-a)^2}{2},$$

进一步, 有

$$f(x) - f(a) = f'(a)(x-a) + \frac{f''(\xi)(x-a)^2}{2},$$

取绝对值, 由三角不等式, 得

$$|f(x) - f(a)| \leq |f'(a)||x-a| + \frac{|f''(\xi)||x-a|^2}{2},$$



注意不等式

$$|f(x) - f(a)| \leq |f'(a)| |x - a| + \frac{|f''(\xi)| |x - a|^2}{2},$$

其中 ξ 在 x 与 a 之间。如果 $f'(a) \neq 0$, $|f''(\xi)|$ 与 $|f'(a)|$ 相比不太大, 则可忽略 $|x - a|$ 的二次项, 得到 $f(a)$ 的一个近似误差

$$|f(x) - f(a)| \approx |f'(a)| |x - a|$$



如果 $f(x_1, x_2, \dots, x_n)$ 为 n 元函数，自变量 x_1, x_2, \dots, x_n 的近似值分别为 a_1, a_2, \dots, a_n ，则

$$f(x_1, x_2, \dots, x_n) - f(a_1, a_2, \dots, a_n) \approx \sum_{k=1}^n \left(\frac{\partial f}{\partial x_k} \right)_a (x_k - a_k) \quad (1-19)$$

其中 $\left(\frac{\partial f}{\partial x_k} \right)_a = \frac{\partial}{\partial x_k} f(a_1, a_2, \dots, a_n)$ ，所以可以估计到函数值的误差界，

$$|f(x_1, x_2, \dots, x_n) - f(a_1, a_2, \dots, a_n)| \leq \sum_{k=1}^n \left| \left(\frac{\partial f}{\partial x_k} \right)_a \right| \cdot |x_k - a_k| \quad (1-20)$$

现将估计式 (1-20) 应用到四则运算，即当 $n=2$ ，取

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



这时有

$$|f(x_1, x_2) - f(a_1, a_2)| \leq \left(\frac{\partial f}{\partial x_1} \right)_a \cdot |x_1 - a_1| + \left(\frac{\partial f}{\partial x_2} \right)_a \cdot |x_2 - a_2|$$

从而得到四则运算的估计式:

$$|(x_1 \pm x_2) - (a_1 \pm a_2)| \leq |x_1 - a_1| + |x_2 - a_2| \quad (1-21)$$

两个近似数相加减，其运算结果的精度不比原始数据的任何一个精度高。

$$|x_1 x_2 - a_1 a_2| \leq |a_2| |x_1 - a_1| + |a_1| |x_2 - a_2| \quad (1-22)$$

$$\left| \frac{x_1}{x_2} - \frac{a_1}{a_2} \right| \leq \frac{|a_2| |x_1 - a_1| + |a_1| |x_2 - a_2|}{|a_2|^2} \quad (1-23)$$

计算中应尽力避免小数作除数



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

练习2

已知 $a_1 = 1.21$, $a_2 = -3.65$, $a_3 = 9.81$ 均为有效数字,

求 $a_1 + a_2 \cdot a_3$ 的相对误差界。

解: 取 $f(x_1, x_2, x_3) = x_1 + x_2 \cdot x_3$, 根据 (1-20) 式, 得

$$\begin{aligned} \frac{|f(x_1, x_2, x_3) - f(a_1, a_2, a_3)|}{|f(a_1, a_2, a_3)|} &= \frac{|(x_1 + x_2 \cdot x_3) - (a_1 + a_2 \cdot a_3)|}{|a_1 + a_2 \cdot a_3|} \\ &\leq \frac{|x_1 - a_1| + |a_3| \cdot |x_2 - a_2| + |a_2| \cdot |x_3 - a_3|}{|a_1 + a_2 \cdot a_3|} \end{aligned}$$

由已知,

$$|x_1 - a_1| \leq \frac{1}{2} \times 10^{-2}, \quad |x_3 - a_3| \leq \frac{1}{2} \times 10^{-2}, \quad |x_2 - a_2| \leq \frac{1}{2} \times 10^{-2}, \quad \text{从而}$$

$$\frac{|(x_1 + x_2 \cdot x_3) - (a_1 + a_2 \cdot a_3)|}{|a_1 + a_2 \cdot a_3|} \leq \frac{1 + |a_3| + |a_2|}{|a_1 + a_2 \cdot a_3|} \times \frac{1}{2} \times 10^{-2} \approx 1.038 \times \frac{1}{2} \times 10^{-2} = 0.0052$$



观察

$$\frac{|(x_1 - x_2) - (a_1 - a_2)|}{|a_1 - a_2|} \leq \frac{|x_1 - a_1| + |x_2 - a_2|}{|a_1 - a_2|} \quad (1-24)$$

当 $x_1 \approx x_2$ 时，必有 $a_1 \approx a_2$ ，则 $a_1 - a_2 \approx 0$ ，进而 $\frac{1}{|a_1 - a_2|} \approx +\infty$ 。这时由 (1-24) 可知，计算的相对误差会很大，会导致计算值的有效数字的损失。

结论

在计算中应尽量避免出现两个相近的数相减



例4

一元二次方程 $ax^2 + 2bx + c = 0$ ($a \cdot c \neq 0$)

有两个根，其求根公式为

$$x_1 = \frac{-b + \sqrt{b^2 - ac}}{a}, \quad x_2 = \frac{-b - \sqrt{b^2 - ac}}{a}$$

如果 $b^2 \gg |ac|$ ，则 $\sqrt{b^2 - ac} \approx |b|$ ，用上述公式计算时如果 $b > 0$ ，则有 $x_1 \approx \frac{-b + |b|}{a} = 0$ 如果 $b < 0$ ，则有 $x_2 \approx \frac{-b - |b|}{a} = 0$

总之，两者其中之一必将会损失有效数字。



一般解二次方程 $ax^2 + 2bx + c = 0$ (设 a, b 均不为零), 应取

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

$$\operatorname{sgn}(b) = \begin{cases} 1, & \text{当 } b > 0 \text{ 时} \\ -1, & \text{当 } b < 0 \text{ 时} \end{cases} \quad \text{是 } b \text{ 的符号函数。}$$

例

方程 $x^2 - 16x + 1 = 0$ 的根为: $x_1 = 8 + \sqrt{63}$, $x_2 = 8 - \sqrt{63}$

若取三位有效数字计算, 有 $\sqrt{63} \approx 7.94$, 则由习惯的公式

$x_1 = 8 + \sqrt{63} \approx 8.0 + 7.94 = 15.9$, 有三位有效数字。而

$x_2 = 8 - \sqrt{63} \approx 8.00 - 7.94 = 0.06$, 只有一位有效数字。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

其原因为在计算 x_2 时发生了两个相近数相减, 造成有效数字损失。 x_2 的精确值是 $0.062746 \dots$, 如果改用公式:

$x_2 = \frac{1}{x_1}$, 计算得 $x_2 \approx 0.0629$, 具有三位有效数字。

← 返回本节



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

1.2.4 数值方法的稳定性和避免误差危害的基本原则

1. 数值方法的稳定性

用某一种数值方法求一个问题的数值解，如果在方法的计算过程中舍入误差在一定条件下能够得到控制（或者说舍入误差的增长不影响产生可靠的结果），则称该方法是数值稳定的；否则，出现与数值稳定相反的情况，则称之为数值不稳定的。



练习1 计算积分 $I_n = \int_0^1 \frac{x^n}{x+5} dx, n = 0, 1, 2, \dots, 7$

解： 由于

$$\begin{aligned} I_n + 5I_{n-1} &= \int_0^1 \frac{x^n}{x+5} dx + 5 \int_0^1 \frac{x^{n-1}}{x+5} dx = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5} dx \\ &= \int_0^1 x^{n-1} dx = \frac{1}{n} \end{aligned}$$

则递归算法如下：

1. $I_n = \frac{1}{n} - 5I_{n-1}$, 由 $I_0 = \ln \frac{6}{5}$ 计算出 I_1, \dots, I_7
2. $I_{n-1} = \frac{1}{5} \left(\frac{1}{n} - I_n \right)$, 由 $I_7 = 0.0210$ 计算出 I_7, \dots, I_0



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

n	I_n	方法1	方法2
0	0.1820	0.1820	0.1820
1	0.0880	0.0900	0.0880
2	0.0580	0.0500	0.0580
3	0.0431	0.0830	0.0431
4	0.0343	-0.0165	0.0343
5	0.0284	1.0250	0.0284
6	0.0240	-4.9580	0.0240
7	0.0210	24.933	0.0210



设 I_0 的近似值为 \bar{I}_0 ，然后按方法1计算 I_1, \dots, I_7 的近似值 $\bar{I}_1, \dots, \bar{I}_7$ ，如果最初计算时误差为： $E_0 = I_0 - \bar{I}_0$ 递推过程的舍入误差不记，并记 $E_n = I_n - \bar{I}_n$ ，则有

$$E_7 = I_7 - \bar{I}_7 = (-5)E_6 = (-5) \cdot (-5)E_5 = \dots = (-5)^7 E_0$$

由此可见，用该方法计算 I_1, \dots, I_7 时，当计算 I_0 时产生的舍入误差为 E_0 ，那么计算 I_7 时产生的舍入误差放大了 $5^7 = 78,125$ 倍，因此，该方法是数值不稳定的。

按方法2计算时，记初始误差为 $E_7 = I_7 - \bar{I}_7$ ，则有

$$E_0 = I_0 - \bar{I}_0 = \left(-\frac{1}{5}\right)E_1 = \left(-\frac{1}{5}\right) \cdot \left(-\frac{1}{5}\right)E_2 = \dots = \left(-\frac{1}{5}\right)^7 E_7$$

由此可知，使用公式2计算时不会放大舍入误差。因此，该方法是数值稳定的。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

2、避免误差危害的基本原则

为了用数值方法求得数值问题满意的近似解，在数值运算中应注意下面两个基本原则。

(I) 避免有效数字的损失

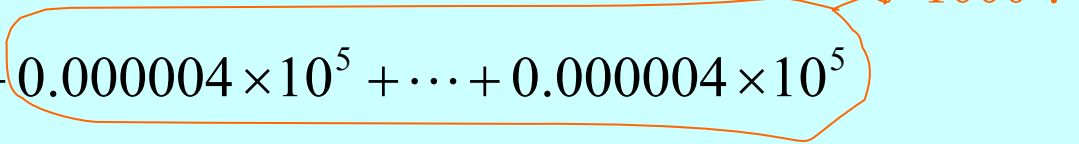
在四则运算中为避免有效数值的损失，应注意以下事项：

- (1) 在做加法运算时，应防止“大数吃小数”；
- (2) 避免两个相近数相减；
- (3) 避免小数做除数或大数做乘数。



例2 在五位十进制的计算机上计算 $x = 63015 + \sum_{i=1}^{1000} \delta_i$, $\delta_i = 0.4$

解 计算机作加减法时, 先将所相加数阶码对齐, 根据字长舍入, 再加减。如果用63015依次加各个 δ_i , 那么上式用规范化和阶码对齐后的数表示为:

$$x = 0.63015 \times 10^5 + 0.000004 \times 10^5 + \cdots + 0.000004 \times 10^5$$


因其中 0.000004×10^5 的舍入结果为0, 所以上式的计算结果是 0.63015×10^5 。这种现象被称为“大数吃小数”。如果改变运算

次序, 先把1000个 δ_i 相加, 再和63015相加, 即

$$\begin{aligned} x &= \underbrace{0.4 + 0.4 + \cdots + 0.4}_{1000} + 0.63015 \times 10^5 = 0.4 \times 10^3 + 0.63015 \times 10^5 \\ &= 0.004 \times 10^5 + 0.63015 \times 10^5 = 0.63415 \times 10^5 \end{aligned}$$

后一种方法的结果是正确的, 前一种方法的舍入误差影响太大。



又例如, 在八位十进制计算机上, 计算 $3.712 + \frac{2}{10^{-8}}$

$$\begin{aligned} 3.712 + \frac{2}{10^{-8}} &= 3.712 + 2 \times 10^8 \\ &= (0.000000003712 + 0.2) \times 10^9 \\ &= 2 \times 10^8 \end{aligned}$$

3.712 与 0.2×10^9 在计算机上做和时, 3.712由于阶码升为9位尾数左移变成机器零, 这便说明用 **小数做除数** 或用 **大数做乘数** 时, 容易产生大的舍入误差, 应尽量避免。



(II) 减少运算次数

例如，多项式求值运算，设 $p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$

如果直接逐项求和计算，需要大约 $2n$ 次乘法运算 即

$$x \cdot x \rightarrow x^2 \cdot x \rightarrow x^3 \cdot x \rightarrow \cdots \rightarrow x^{n-1} \cdot x \rightarrow x^n \quad \bullet \quad \bullet \quad \bullet \quad n-1 \text{ 次}$$

$$a_n \cdot x^n, a_{n-1} \cdot x^{n-1}, \cdots, a_1 \cdot x \quad \rightarrow n \text{ 次}$$

若取 $t_k = x^k$, $u_k = a_0 + a_1 x + \cdots + a_k x^k$,

则有递推公式:

$$\begin{cases} t_k = x \cdot t_{k-1} \\ u_k = u_{k-1} + a_k \cdot t_k \end{cases} \quad k = 1, 2, \cdots, n, \quad \begin{cases} t_0 = 1 \\ u_0 = a_0 \end{cases}$$

$p_n(x) = u_n$ 就是所求的值。总的计算量需进行 $2n$ 次乘法。



若将公式变成如下递推公式，即令

$$\begin{aligned} p_n(x) &= (a_n x + a_{n-1}) x^{n-1} + \cdots + a_1 x + a_0 \\ &= ((a_n x + a_{n-1}) x + a_{n-2}) x^{n-2} + a_{n-3} x^{n-3} \cdots + a_1 x + a_0 \\ &= \cdots = \\ &= (\cdots (a_n x + a_{n-1}) x + a_{n-2}) x + \cdots + a_2) x + a_1) x + a_0 \end{aligned}$$

若令 $s_k = (\cdots (a_n x + a_{n-1}) x + a_{n-2}) x + \cdots + a_{k+1}) x + a_k$

则有递推公式:

$$\begin{cases} s_n = a_n \\ s_k = x \cdot s_{k+1} + a_k \end{cases} \quad k = n-1, n-2, \cdots, 2, 1, 0$$

$p_n(x) = s_0$ 就是所求的值。总的计算量为 n 次乘法。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

$$P_5(x) = 5x^5 + 0x^4 + 31x^3 + 3x^2 + 1x - 1$$



$$P_5(x) = (((((5x + 0)x + 1)x - 3)x + 1)x - 1$$

$$5 \quad 0 \quad 1 \quad -3 \quad 1 \quad -1$$

令 $x=2$

$$10 \quad 20 \quad 42 \quad 39 \quad 138 \quad 157 = P_5(2)$$

以上计算过程称之为 秦九韶算法



例6

利用 $\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$ 计算 $\ln 2$ ，若要精确到 10^{-5} ，要计算十万项的和，计算量很大，另一方面舍入误差的积累也十分严重。

如果改用级数

$$\ln \frac{1+x}{1-x} = 2 \left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots + \frac{x^{2n+1}}{2n+1} + \cdots \right)$$

$$\text{取 } x = \frac{1}{3}, \quad \ln 2 = \ln \frac{1+\frac{1}{3}}{1-\frac{1}{3}} = 2 \left(\frac{1}{3} + \frac{\left(\frac{1}{3}\right)^3}{3} + \frac{\left(\frac{1}{3}\right)^5}{5} + \cdots + \frac{\left(\frac{1}{3}\right)^{2n+1}}{2n+1} + \cdots \right)$$

只须计算前9项的和，截断误差便小于 10^{-10}

 返回本章



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

1.3 向量与矩阵范数

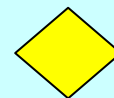
1.3.1 向量范数



1.3.2 范数的等价性



1.3.3 矩阵范数



1.3.4 矩阵范数的性质





DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

我们在讨论实数、复数的大小、误差时，把任何一个实数变量或复数复变量与一个非负实数联系起来，即在 $f(x)=|x|$ 意义下，这个实数提供了实数变量或复数复变量大小的度量。

注意到 $f(x)=|x|$ 满足以下三个条件：

(1) 非负性 $|x| \geq 0$ 当且仅当 $x=0$ 时 $|x|=0$

(2) 齐次性 $|\alpha x| = |\alpha| \cdot |x|$

(3) 三角不等式 $|x + y| \leq |x| + |y|$



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

把任何一个向量或矩阵与一个非负实数联系起来，在某种意义下，这个实数提供了向量和矩阵的大小的度量。

取不同的向量或矩阵，就对应有一类向量或矩阵函数，其中每一个函数都可以看作向量或矩阵大小的一种度量。

范数的主要的应用：

一、研究这些矩阵和向量的误差估计

二、研究矩阵和向量的序列以及级数的收敛准则



1.3.1 向量范数

定义1.1 定义在 \mathbb{C}^n (n 维复向量空间) 上的一个非负实值函数, 记为 $f(x) = \|x\|$, 若该函数满足以下三个条件:
即对任意向量 x 和 y 以及任意复常数 $\alpha \in \mathbb{C}$

(1) 非负性 $\|x\| \geq 0$ 当且仅当 $x = 0$ 时 $\|x\| = 0$

(2) 齐次性 $\|\alpha x\| = |\alpha| \cdot \|x\|$

(3) 三角不等式 $\|x + y\| \leq \|x\| + \|y\|$

则称函数 $\|\cdot\|$ 为 \mathbb{C}^n 上的一个向量范数。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

设任意 n 维向量 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ (\mathbf{x}^T 为向量 \mathbf{x} 的转置)。

常用的向量范数有：

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \quad (1-1)$$

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} = \sqrt{\mathbf{x}^H \cdot \mathbf{x}} \quad (\mathbf{x}^H \text{为向量 } \mathbf{x} \text{ 的共轭转置}) \quad (1-2)$$

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| \quad (1-3)$$

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad 1 \leq p < +\infty \quad (1-4)$$

$|x_i|$ 表示 x_i 的模。上述四种范数分别称为 1，2， ∞ 范数和 p -范数



前面三种范数都为 p -范数，当 $p = 1, 2, \infty$ 时的范数。

注意，当时 $p \rightarrow \infty$ ， $\|\mathbf{x}\|_p \rightarrow \|\mathbf{x}\|_\infty$ 。事实上，

$$\|\mathbf{x}\|_\infty^p = \max_{1 \leq i \leq n} |\mathbf{x}_i|^p \leq \sum_{i=1}^n |\mathbf{x}_i|^p \leq n \cdot \max_{1 \leq i \leq n} |\mathbf{x}_i|^p = n \cdot \|\mathbf{x}\|_\infty^p$$

两边开 p 次方得

$$\|\mathbf{x}\|_\infty \leq \left(\sum_{i=1}^n |\mathbf{x}_i|^p \right)^{\frac{1}{p}} \leq n^{\frac{1}{p}} \cdot \|\mathbf{x}\|_\infty, \text{ 由于 } \lim_{p \rightarrow \infty} \sqrt[p]{n} = 1, \text{ 故 } \|\mathbf{x}\|_p \rightarrow \|\mathbf{x}\|_\infty$$

容易验证以上三种范数均满足范数定义中的三个条件。

下面我们分析向量的 $1, 2$ 和 ∞ -范数的几何意义，

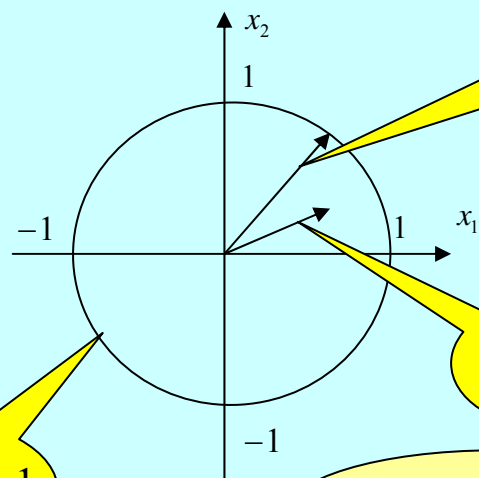
为此，不妨设 $\mathbf{x} = (x_1, x_2)^T \in \mathbf{R}^2$ ， $\|\mathbf{x}\|_2 \leq 1$ 、 $\|\mathbf{x}\|_\infty \leq 1$ 和 $\|\mathbf{x}\|_1 \leq 1$ 。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY



$$\|x\|_2 = 1$$

$$\|x\|_2 < 1$$

$$x_1^2 + x_2^2 = 1$$

$$\max\{x_1 < 1, x_2 = 1\}$$

$$\max\{x_1 < 1, x_2 < 1\}$$

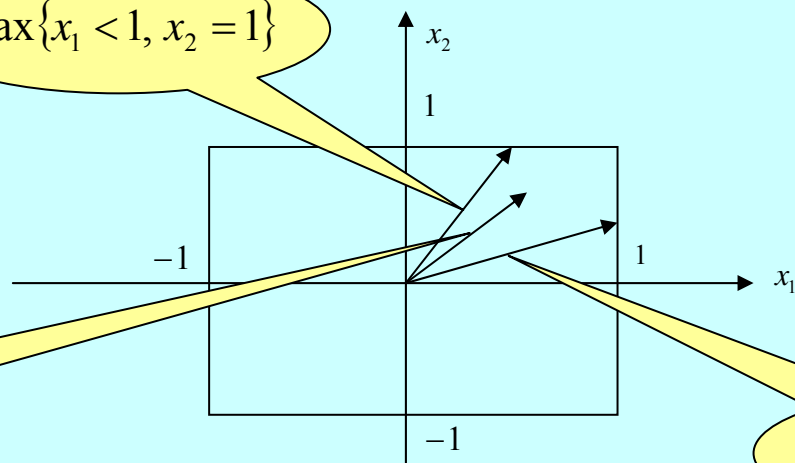
$$-x_1 + x_2 = 1$$

$$-x_1 - x_2 = 1$$

$$x_1 + x_2 = 1$$

$$|x_1| + |x_2| < 1$$

$$x_1 - x_2 = 1$$



$$\max\{x_1 = 1, x_2 < 1\}$$



一般情况下，对给定的任意一种向量范数 $\|\cdot\|$ ，可定义出加权的范数： $\|\mathbf{x}\|_w = \|\mathbf{W}\mathbf{x}\|$ 其中 \mathbf{W} 为对角矩阵，其对角元素是它的每一个分量的权系数。例如，对任 $\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbf{C}^3$

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -2 \end{pmatrix} \in \mathbf{C}^{3 \times 3},$$

加权的1-范数为：

$$\|\mathbf{x}\|_w = \|\mathbf{W}\mathbf{x}\|_1 = \left\| \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \right\|_1 = |x_1| + 3|x_2| + 2|x_3|$$

加权的2-范数为：

$$\|\mathbf{x}\|_w = \|\mathbf{W}\mathbf{x}\|_2 = \left(|x_1|^2 + 9|x_2|^2 + 4|x_3|^2 \right)^{1/2}$$



例 对任给 $x = (x_1, x_2, x_3)^T \in C^3$, 试问如下实值函数是否构成向量范数?

- ~~1.~~ $|x_1| + |2x_2 + x_3|$, ~~2.~~ $|x_1| + |2x_2| - 5|x_3|$, ~~3.~~ $|x_1|^4 + |x_2|^4 + |x_3|^4$,
✓ 4. $|x_1| + 3|x_2| + 2|x_3|$

答: 1. 和2. 不满足非负性条件, 1. 中取 $x_1 = 0, x_3 = -2x_2$

2. 中取 $x_1 = 0, x_3 = \frac{2}{5}x_2$

3. 不满足齐次性条件都不是向量范数;

4. 满足加权向量范数的定义, 故构成向量范数。



对任给 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in C^n$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T \in C^n$, 恒有

$$|(x, y)| \leq \sqrt{(x, x)} \cdot \sqrt{(y, y)} = \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2$$

称之为**Cauchy-Schwarz**不等式。

证:

$$\begin{aligned} 0 &\leq (x + \lambda y, x + \lambda y) = (x, x) + \lambda(x, y) + \lambda(y, x) + \lambda^2(y, y) \\ &\leq (x, x) + 2\lambda|(x, y)| + \lambda^2(y, y) = f(\lambda) \end{aligned}$$

由二次函数的判别式: $4|(x, y)|^2 - 4(x, x) \cdot (y, y) \leq 0$ 得证。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

例：求向量 $x = (-1, 2, 4)^T$ 的 1, 2和 ∞ -范数。

解：

$$\|x\|_1 = |-1| + 2 + 4 = 7;$$

$$\|x\|_2 = \sqrt{|-1|^2 + 2^2 + 4^2} = \sqrt{21}$$

$$\|x\|_\infty = \max\{|-1|, 2, 4\} = 4。$$

← 返回本节



1.3.2 范数的等价性

在 C^n 上可以定义各种向量范数，其数值大小一般不同。但是在各种向量范数之间存在下述重要的关系。根据向量范数的定义可以验证：

$$1) \quad \|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_{\infty}$$

$$2) \quad \frac{1}{\sqrt{n}}\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$$

$$3) \quad \frac{1}{\sqrt{n}}\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_2$$

以3) 为例证明之，事实上， $\|\mathbf{x}\|_{\infty}^2 = \max_{1 \leq i \leq n} |x_i|^2 \leq \sum_{i=1}^n |x_i|^2 = \|\mathbf{x}\|_2^2$

又 $\|\mathbf{x}\|_2^2 \leq \sum_{i=1}^n \max_{1 \leq i \leq n} |x_i|^2 = n \cdot \|\mathbf{x}\|_{\infty}^2 \Rightarrow \frac{1}{\sqrt{n}}\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\infty}$ ，故3) 得证。



更一般地有，

定理1.1

（向量范数的等价性定理）设 $\|\cdot\|_\beta$ 和 $\|\cdot\|_\alpha$ 为 \mathbf{C}^n 上的任意两种向量范数，则存在两个与向量无关的正常数 $c_1 > 0$ 和 $c_2 > 0$ ，使得下面的不等式成立

$$c_1 \|\mathbf{x}\|_\beta \leq \|\mathbf{x}\|_\alpha \leq c_2 \|\mathbf{x}\|_\beta \quad (1-6)$$

并称 $\|\cdot\|_\alpha$ 和 $\|\cdot\|_\beta$ 为 \mathbf{C}^n 上的等价范数。



1.3.3 矩阵范数

定义1.2

定义在 $\mathbf{C}^{n \times n}$ 上的一个非负实值函数，记为 $f(A) = \|A\|$ ，若该函数满足以下条件：

即对任意矩阵 A 、 B 以及任意复常数 $\alpha \in \mathbf{C}$

(1) 非负性 $\|A\| \geq 0$ 当且仅当 $A = 0_{n \times n}$ 时 $\|A\| = 0$

(2) 齐次性 $\|\alpha A\| = |\alpha| \cdot \|A\|$

(3) 三角不等式 $\|A + B\| \leq \|A\| + \|B\|$

(4) 相容性 $\|AB\| \leq \|A\| \cdot \|B\|$

则称函数 $\|\cdot\|$ 为 $\mathbf{C}^{n \times n}$ 上的一个矩阵范数。



$$\|A\|_{m_1} = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}| \quad (1-7)$$

$$\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{\frac{1}{2}} \quad (1-8)$$

显然上述两个函数均满足矩阵范数定义中的（1）—（3）

我们分别称由（1-7）和（1-8）所定义的范数为矩阵的

m_1 -范数和Frobenius范数（简称F-范数）。



下面证明(1-7)满足相容条件， 证： 由定义

$$\begin{aligned}\| \mathbf{AB} \|_{m_1} &= \sum_{i=1}^m \sum_{j=1}^n \left| \sum_{k=1}^l a_{ik} \cdot b_{kj} \right| \\&= \sum_{i=1}^m \sum_{j=1}^n \left| a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{il}b_{lj} \right| \\&\leq \sum_{i=1}^m \sum_{j=1}^n \left(|a_{i1}b_{1j}| + |a_{i2}b_{2j}| + \cdots + |a_{il}b_{lj}| \right) \\&\leq \sum_{i=1}^m \sum_{j=1}^n \left(|a_{i1}| + |a_{i2}| + \cdots + |a_{il}| \right) \cdot \left(|b_{1j}| + |b_{2j}| + \cdots + |b_{lj}| \right) \\&= \left(\sum_{i=1}^m \sum_{k=1}^l |a_{ik}| \right) \left(\sum_{k=1}^l \sum_{j=1}^n |b_{kj}| \right) = \| \mathbf{A} \|_{m_1} \| \mathbf{B} \|_{m_1}\end{aligned}$$



下面证明(1-8)满足相容条件， 证： 记

$$AB = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{pmatrix} (B_1 \quad B_2 \quad \cdots \quad B_n) = \begin{pmatrix} A_1 B_1 & A_1 B_2 & \cdots & A_1 B_n \\ A_2 B_1 & A_2 B_2 & & A_2 B_n \\ \vdots & \vdots & \ddots & \vdots \\ A_m B_1 & A_m B_2 & \cdots & A_m B_n \end{pmatrix}$$

其中 $A_i = (a_{i1}, a_{i2}, \dots, a_{in}) \quad i = 1, 2, \dots, m$ $(A_i^T B_j) = A_i B_j = \sum_{k=1}^l a_{ik} b_{kj}$
 $B_j = (b_{1j}, b_{2j}, \dots, b_{nj})^T \quad j = 1, 2, \dots, n$

$$\begin{aligned} \|AB\|_F &= \sqrt{\sum_{i=1}^m \sum_{j=1}^n \left| \sum_{k=1}^l a_{ik} b_{kj} \right|^2} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_i B_j|^2} \leq \sqrt{\sum_{i=1}^m \sum_{j=1}^n \left(\|A_i\|_2^2 \|B_j\|_2^2 \right)} \\ &= \sqrt{\sum_{i=1}^m \sum_{j=1}^n \left[\left(\sum_{k=1}^l |a_{ik}|^2 \right) \left(\sum_{k=1}^l |b_{kj}|^2 \right) \right]} = \sqrt{\sum_{i=1}^m \sum_{k=1}^l |a_{ik}|^2} \sqrt{\sum_{j=1}^n \sum_{k=1}^l |b_{kj}|^2} = \|A\|_F \|B\|_F \end{aligned}$$



例 设 $A = (a_{ij})_{m \times n} \in \mathbf{C}^{m \times n}$, $f(A) = \max_{ij} |a_{ij}|$, 问是否构成A的一种范数?

解: 取 $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, 那么, $AB = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$,

则可得出

$$f(A) = f(B) = 1, \quad f(AB) = 2, \quad \text{从而有}$$

$$f(AB) > f(A) \cdot f(B), \quad \text{故不构成} A \text{的一种范数。}$$

若定义实值函数: $\|A\| = \sqrt{m \cdot n} \cdot \max_{ij} |a_{ij}|$, 则可验证其构成A的一种范数。



由于矩阵与向量在实际运算中常同时出现，所以矩阵范数与向量范数也会同时出现，因此应该建立矩阵范数与向量范数的联系，即相容性的问题。

定义1 对于一种矩阵范数 $\|\cdot\|_M$ 和一种向量范数 $\|\cdot\|_V$ 如果对任意 $m \times n$ 矩阵 A 和任意 n 维向量 x ，满足

$$\|Ax\|_V \leq \|A\|_M \|x\|_V$$

则称矩阵范数 $\|\cdot\|_M$ 与向量范数 $\|\cdot\|_V$ 是相容的。

矩阵 m_1 范数与向量的 p -范数是相容的，即 $\|Ax\|_p \leq \|A\|_{m_1} \|x\|_p$

而矩阵的 F -范数与向量的 2 -范数是相容的，即 $\|Ax\|_2 \leq \|A\|_F \|x\|_2$

可以证明任意一种矩阵范数必然存在与之相容的向量范数



定义

称如下集合为矩阵 $A \in C^{n \times n}$ 的谱

$$\sigma(A) = \{ \lambda \mid \det(\lambda I - A) = 0 \}$$

称如下实数为矩阵 $A \in C^{n \times n}$ 的谱半径

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

注：若矩阵为： $A = (a_{ij})_{m \times n}$ ，则其复共轭矩阵为： $A^H = (\bar{a}_{ji})_{m \times n}$

酉矩阵为： $A^H A = A A^H = I$

例如： $A = \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}$ ， $A^H = \begin{pmatrix} -i & 0 \\ 0 & -i \end{pmatrix}$ ，则此矩阵为酉阵。



2. 算子范数

定理1.2 若定义

$$\|A\|_M = \max_{x \neq 0} \frac{\|Ax\|_V}{\|x\|_V} = \max_{\|x\|_V=1} \|Ax\|_V \quad (1-9)$$

则 $\|A\|_M$ 是一种矩阵范数。

我们称由关系式（1-9）定义的矩阵范数为从属向量范数的矩阵范数简称从属范数或**算子范数**。



证 首先注意到,

$$\max_{x \neq 0} \frac{\|Ax\|_V}{\|x\|_V} = \max_{x \neq 0} \left[\frac{1}{\|x\|_V} \cdot \|Ax\|_V \right] = \max_{x \neq 0} \left\| A \left(\frac{x}{\|x\|_V} \right) \right\|_V = \max_{\|y\|_V=1} \|Ay\|_V ,$$

其中 $y = \frac{x}{\|x\|_V}$, $\|y\|_V = \left\| \frac{x}{\|x\|_V} \right\|_V = \frac{\|x\|_V}{\|x\|_V} = 1$

则由(1-9)立刻可得到 $\frac{\|Ax\|_V}{\|x\|_V} \leq \|A\|_M$ 即 $\|Ax\|_V \leq \|A\|_M \|x\|_V$

即相容性成立。由于 $\|Ax\|_V$ 是 \mathbf{C}^n 中的有界闭集 $\mathbf{D} = \{x \mid \|x\|_V = 1, x \in \mathbf{C}^n\}$ 上的连续函数, 故对每一个矩阵 A 而言, 都能够找到向量 x_0 , 使得

$\|x_0\|_V = 1$, 而且 $\|A\|_M = \|Ax_0\|_V = \max_{\|x\|_V=1} \|Ax\|_V$



下面证明 $\|A\|_M$ 是一种矩阵范数.

(1) 非负性, 当 $A=O$ 时, $\|A\|_M=0$; 而当 $A \neq O$ 时, 存在 $x_0 \in \mathbf{C}^n$, 使 $Ax_0 \neq 0$, 从而 $\|A\|_M \geq \frac{\|Ax_0\|_V}{\|x_0\|_V} > 0$

(2) 齐次性, 对任意的 $\alpha \in \mathbf{C}$, 有

$$\|A\|_M = \max_{x \neq 0} \frac{\|(\alpha A)x\|_V}{\|x\|_V} = |\alpha| \cdot \max_{x \neq 0} \frac{\|Ax\|_V}{\|x\|_V} = |\alpha| \cdot \|A\|_M$$

(3) 三角不等式, 假设 A 和 B 分别为 $m \times n$ 阶矩阵, 则

$$\begin{aligned} \|A+B\|_M &= \max_{\|x\|_V=1} \|(A+B)x\|_V = \max_{\|x\|_V=1} \|Ax+Bx\|_V \leq \max_{\|x\|_V=1} \|Ax\|_V + \max_{\|x\|_V=1} \|Bx\|_V \\ &\leq \|A\|_M + \|B\|_M \end{aligned}$$



(4) 相容性, $AB=O$ 时显然。设 $AB \neq O$, 且又假设 A 和 B 分别为 $m \times l$ 和 $l \times n$ 阶矩阵, $\mathbf{y} = B\mathbf{x} \neq 0$, 因此

$$\begin{aligned}\|AB\|_M &= \max_{\|\mathbf{x}\|_V=1} \|(AB)\mathbf{x}\|_V = \max_{\|\mathbf{x}\|_V=1} \frac{\|(AB)\mathbf{x}\|_V}{\|B\mathbf{x}\|_V} \|B\mathbf{x}\|_V \\ &\leq \max_{\mathbf{y} \neq 0} \frac{\|A\mathbf{y}\|_V}{\|\mathbf{y}\|_V} \max_{\|\mathbf{x}\|_V=1} \|B\mathbf{x}\|_M \leq \|A\|_M \|B\|_M\end{aligned}$$

因此, $\|\cdot\|_M$ 是一种矩阵范数, 并且是一种与向量范数 $\|\cdot\|_V$ 相容的矩阵范数。



在向量范数中，最常用的范数为向量的1-范数、2-范数和 ∞ -范数，下面分别给出从属这三种向量范数的矩阵范数。

定理1.3

几种常用的算子范数

$$(1) \quad \|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \quad (\text{列和范数})$$

$$(2) \quad \|A\|_2 = \sqrt{\lambda_{\max}(A^H A)} \quad (\text{谱范数})$$

其中 $\lambda_{\max}(A^H A)$ 表示矩阵 $A^H A$ 的最大特征值；（或 $\sqrt{\rho(A^T A)}$ ）

$$(3) \quad \|A\|_{\infty} = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}| \quad (\text{行和范数})$$



证 在此只给出 (1) 的证明。设 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, 则

$$\begin{aligned}\|\mathbf{Ax}\|_1 &= \left\| \begin{pmatrix} \sum_{j=1}^n a_{1j}x_j \\ \sum_{j=1}^n a_{2j}x_j \\ \vdots \\ \sum_{j=1}^n a_{nj}x_j \end{pmatrix} \right\|_1 = \sum_{i=1}^m \left| \sum_{j=1}^n a_{ij}x_j \right| \leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij}| |x_j| = \sum_{j=1}^n \left(\sum_{i=1}^m |a_{ij}| \right) |x_j| \\ &\leq \left(\max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \right) \sum_{j=1}^n |x_j| = \left(\max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \right) \|\mathbf{x}\|_1\end{aligned}$$

因此, 有

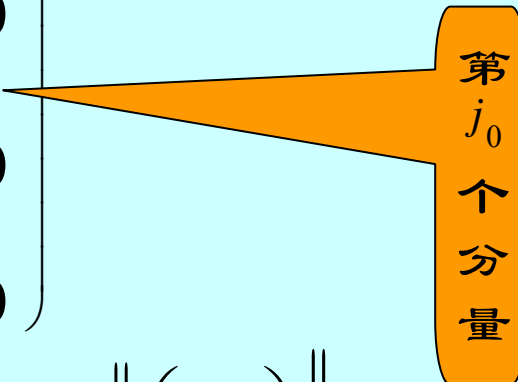
$$\|\mathbf{A}\|_1 = \max_{\mathbf{x} \neq 0} \frac{\|\mathbf{Ax}\|_1}{\|\mathbf{x}\|_1} \leq \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$$

为了证明 (1), 只需找到 \mathbf{x}_0 , 使得 $\frac{\|\mathbf{Ax}_0\|_1}{\|\mathbf{x}_0\|_1} = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$



下面讨论 \mathbf{x}_0 的选取。如果存在某个 j_0 使得 $\max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| = \sum_{i=1}^m |a_{ij_0}|$

则可取

$$\mathbf{x}_0 = \mathbf{e}_{j_0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$


第 j_0 个分量

那么, 即有

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \|A\mathbf{x}_0\|_1 = \|A\mathbf{e}_{j_0}\|_1 = \left\| \begin{pmatrix} a_{1j_0} \\ a_{2j_0} \\ \vdots \\ a_{mj_0} \end{pmatrix} \right\|_1 = \sum_{i=1}^m |a_{ij_0}|$$



推论 对任何算子范数，单位矩阵 $I \in \mathbf{R}^{n \times n}$ 的范数值为1，即

$$\|I\| = 1。$$

事实上，

$$\|I\| = \max_{x \neq 0} \frac{\|Ix\|}{\|x\|} = \max_{x \neq 0} \frac{\|x\|}{\|x\|} = 1$$

特别地， $\|A\|_{m_1}$ 、 $\|A\|_F$ 不是算子范数。

事实上，

$$\|I\|_{m_1} = \sum_{i=1}^n \sum_{j=1}^n |a_{ij}| = \sum_{i=1}^n 1 = n \neq 1 \quad \|I\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\sum_{i=1}^n 1^2} = \sqrt{n} \neq 1$$



例2 设 $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 4 \\ 0 & -2 & 4 \end{pmatrix}$, 求 $\|A\|_1$ 、 $\|A\|_\infty$ 、 $\|A\|_2$ 、 $\|A\|_{m_1}$ 、 $\|A\|_F$ 。

解: $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^3 |a_{ij}| = \max\{1, 4, 8\} = 8$

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^3 |a_{ij}| = \max\{1, 6, 6\} = 6$$

$$\|A\|_{m_1} = \sum_{i=1}^3 \sum_{j=1}^3 |a_{ij}| = 1 + 2 + 4 + |-2| + 4 = 13$$

$$\|A\|_F = \sqrt{\sum_{i=1}^3 \sum_{j=1}^3 |a_{ij}|^2} = \sqrt{1 + 2^2 + 4^2 + |-2|^2 + 4^2} = \sqrt{41}$$



$$\mathbf{A}^T \mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & 4 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 4 \\ 0 & -2 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 32 \end{pmatrix}$$

令

$$\det(\lambda I - \mathbf{A}^T \mathbf{A}) = \begin{vmatrix} \lambda - 1 & 0 & 0 \\ 0 & \lambda - 8 & 0 \\ 0 & 0 & \lambda - 32 \end{vmatrix} = (\lambda - 1)(\lambda - 8)(\lambda - 32) = 0$$

得，

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T \mathbf{A})} = \sqrt{32} = 4\sqrt{2}$$



一般地有，

定理

（矩阵范数的等价性定理）设 $\|\cdot\|_\beta$ 和 $\|\cdot\|_\alpha$ 为 $\mathbb{C}^{m \times n}$

上的任意两种矩阵范数，则存在两个与矩阵无关的正常数 $C_1 > 0$ 和 $C_2 > 0$ ，使得下面的不等式成立

$$c_1 \|A\|_\beta \leq \|A\|_\alpha \leq c_2 \|A\|_\beta$$

并称 $\|\cdot\|_\alpha$ 和 $\|\cdot\|_\beta$ 为 $\mathbb{C}^{m \times n}$ 上的等价范数。



可以证明：

1. 任意给定的矩阵范数必然存在与之相容的向量范数；任意给定的向量范数必然存在与之相容的矩阵范数（如从属范数）。
2. 一个矩阵范数可以与多种向量范数相容（矩阵的 m_1 -范数与向量的 p -范数相容）；多种矩阵范数可以与一个向量范数相容（矩阵的 F -范数、2-范数与向量的2-范数相容）。
3. 从属范数一定与所定义的向量范数相容，但是矩阵范数与向量范数相容却未必有从属关系。（矩阵的 F -范数与向量的2-相容，但无从属关系）。
4. 并非任意的矩阵范数与任意的向量范数相容。



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

矩阵范数与向量范数不相容的例子：

取 $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$, $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, 则有 $\|A\|_1 = 1$, $\|x\|_\infty = 1$,

而 $\|Ax\|_\infty = 2 > \|A\|_1 \cdot \|x\|_\infty$

故矩阵的 $\|\cdot\|_1$ 与向量的 $\|\cdot\|_\infty$ 不相容。



对任给 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in C^n$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T \in C^n$,

向量的内积的定义为:

$$(\mathbf{x}, \mathbf{y}) = \mathbf{y}^H \mathbf{x} = (\bar{y}_1 \ \bar{y}_2 \ \cdots \ \bar{y}_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^n x_i \bar{y}_i$$

特别地,

$$(\mathbf{x}, \mathbf{x}) = \mathbf{x}^H \mathbf{x} = \sum_{i=1}^n x_i \bar{x}_i = \sum_{i=1}^n |x_i|^2 = \|\mathbf{x}\|_2^2$$

那么,

$$\begin{aligned} (U\mathbf{x}, U\mathbf{x}) &= (U\mathbf{x})^H (U\mathbf{x}) = \mathbf{x}^H U^H U\mathbf{x} = \mathbf{x}^H (U^H U\mathbf{x}) \\ &= (U^H U\mathbf{x}, \mathbf{x}) \end{aligned}$$



对于酉矩阵 $U^H U = U U^H = I$ ，我们可有如下的结论：

$$\|U\|_2 = 1, \|AU\|_2 = \|UA\|_2 = \|A\|_2 \quad (\text{酉矩阵的范数不变性})$$

事实上，

$$\|U\|_2^2 = \max_{x \neq 0} \frac{\|Ux\|_2^2}{\|x\|_2^2} = \max_{x \neq 0} \frac{(Ux, Ux)}{(x, x)} = \max_{x \neq 0} \frac{(U^H U x, x)}{(x, x)} = \max_{x \neq 0} \frac{(x, x)}{(x, x)} = 1$$

$$\|UA\|_2^2 = \max_{x \neq 0} \frac{\|(UA)x\|_2^2}{\|x\|_2^2} = \max_{x \neq 0} \frac{((UA)x, (UA)x)}{(x, x)} = \max_{x \neq 0} \frac{((UA)^H (UA)x, x)}{(x, x)}$$

$$= \max_{x \neq 0} \frac{((A^H U^H UA)x, x)}{(x, x)} = \max_{x \neq 0} \frac{((A^H A)x, x)}{(x, x)} = \max_{x \neq 0} \frac{(Ax, Ax)}{(x, x)}$$

$$= \max_{x \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \|A\|_2^2$$



$$\begin{aligned}\|AU\|_2^2 &= \max_{x \neq 0} \frac{\|(AU)x\|_2^2}{\|x\|_2^2} = \max_{x \neq 0} \frac{((AU)x, (AU)x)}{(x, x)} = \max_{x \neq 0} \frac{(A(Ux), A(Ux))}{(x, x)} \\ &= \max_{y \neq 0} \frac{(Ay, Ay)}{((U^H y), (U^H y))} = \max_{y \neq 0} \frac{\|Ay\|_2^2}{\|U^H y\|_2^2} = \max_{y \neq 0} \frac{\|Ay\|_2^2}{\|y\|_2^2} = \|A\|_2^2\end{aligned}$$

注意，这里取 $y = Ux \Leftrightarrow x = U^H y$ 是一一对应关系。



1.3.3 矩阵范数的性质

定理1.4

设 $\|\cdot\|_M$ 为 $C^{n \times n}$ 矩阵空间的任一矩阵范数，
则对任意的 n 阶方阵 A 均有

$$\rho(A) \leq \|A\|_M \quad (1-10)$$

其中 $\rho(A)$ 为方阵 A 的谱半径。

证 设 $|\lambda| = \rho(A)$ ，则存在 $x \neq 0$ ，满足 $Ax = \lambda x$ ，从而

$$|\lambda| \|x\|_M = \|\lambda x\|_M = \|Ax\|_M \leq \|A\|_M \|x\|_M$$

故得到

$$\rho(A) = |\lambda| \leq \|A\|_M$$



注意：当 $A = A^T$ 时， $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sqrt{\lambda_{\max}(A^2)} = \lambda_{\max}(A) = \rho(A)$

问实值函数 $\rho(A)$ 可不可以作为的一种范数？

取 $A = B^T = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ ，则有 $\rho(A) = 0$ ， $\rho(B) = 0$ ，而

$\rho(A + B) = 1$ ，即有 $\rho(A) + \rho(B) = 0$ ；从而

$$\rho(A + B) > \rho(A) + \rho(B) = 0$$

故不可以作为的一种范数。



定理1.5

对于任给的 $\varepsilon > 0$, 则存在 $\mathbb{C}^{n \times n}$ 上的一种算子范数

$\|\cdot\|_M$ (依赖矩阵 A 和常数 ε) , 使得

$$\|A\|_M \leq \rho(A) + \varepsilon \quad (1-11)$$

注: 定理1.5 中的矩阵范数 $\|\cdot\|_M$ 与给定的矩阵 A 有关。针对矩阵 A 构造的矩阵范数 $\|\cdot\|_M$ 对于另一个矩阵 B , 不等式

$$\|B\|_M \leq \rho(B) + \varepsilon$$

不一定成立。

**定理1.6**

$\mathbf{C}^{n \times n}$ 上的一种算子矩阵范数 $\|\cdot\|$, 如果 $A \in \mathbf{C}^{n \times n}$ 且

$\|A\| < 1$, 则 $I \pm A$ 可逆, 且

$$\|(I \pm A)^{-1}\| \leq \frac{1}{1 - \|A\|} \quad (1-12)$$

证 由定理1.4可得, $\rho(A) \leq \|A\| < 1$ 。设 λ_i 为矩阵 A 的任意非零特征值
 则矩阵 $I \pm A$ 的特征值为: $\mu_i = 1 \pm \lambda_i \neq 0$ 从而可知 $\det(I \pm A) = \prod_{i=1}^n \mu_i \neq 0$
 即 $I \pm A$ 可逆。进一步

$$(I \pm A)^{-1} (I \pm A) = I \Rightarrow (I \pm A)^{-1} (I \pm A) A = I A$$

$$\Rightarrow \|(I \pm A)^{-1}\| = \|I \mp (I \pm A)^{-1} A\| \Rightarrow \|(I \pm A)^{-1}\| \leq \|I\| + \|(I \pm A)^{-1}\| \|A\| = 1 + \|(I \pm A)^{-1}\| \|A\|$$

整理后便可得:

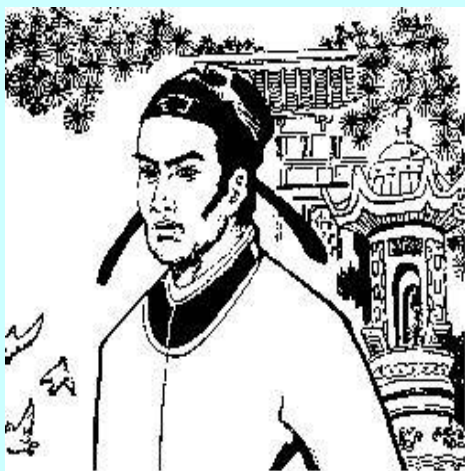




DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY



秦九韶（公元1202 - 1261），字道古，安岳人。南宋大数学家秦九韶与李冶、杨辉、朱世杰并称宋元数学四大家。

秦九韶聪敏勤学。宋绍定四年（1231），秦九韶考中进士，先后担任县尉、通判、参议官、州守、同农、寺丞等职。后在湖北、安徽、江苏、浙江等地做官，1261年左右被贬至梅州（今广东梅县），不久死于任所。

秦九韶是一位既重视理论又重视实践，既善于继承又勇于创新的数学家。他所提出的大衍求一术和正负开方术及其名著《数书九章》，是中国数学史上光彩夺目的一页，对后世数学发展产生了广泛的影响。美国著名科学史家G. 萨顿 (Sarton, 1884 - 1956) 说过，秦九韶是“他那个民族，他那个时代，并且确实也是所有时代最伟大的数学家之一”。

秦九韶的数学成就及对世界数学的贡献主要表现在以下方面：



- 1、秦九韶的《数书九章》是一部划时代的巨著
- 2、秦九韶的“大衍求一术”，领先高斯554年，被康托尔称为“最幸运的天才”
- 3、秦九韶的任意次方程的数值解领先英国人霍纳（W·G·Horner，1786—1837年）

572年



DUT

大连理工大学

DALIAN UNIVERSITY OF TECHNOLOGY

克萊姆 (Gabriel Cramer)

生于： 公元1704年 瑞士 日内瓦

卒于： 公元1752年 法国 巴尼奥勒

十八世纪瑞士数学家，精于数学和几何学。早年在日内瓦读书，1724年起在日内瓦加尔文学院任教，1734年成为几何学教授，1750年任哲学教授。

他一生未婚，专心治学，平易近人且德高望重，先后当选为伦敦皇家学会、柏林研究院和法国、意大利等学会的成员。

主要著作是1750年出版《代数曲线的分析引论》，定义了正则、非正则、超越曲线和无理曲线等概念，第一次正式引入坐标系的纵轴（y轴）。

为了确定经过5个点的一般二次曲线

$$A + Bx + Cy + Dy^2 + Exy + x^2 = 0$$

的系数，应用了著名的“克莱姆法则”（Cramer's Rule），即由线性方程组的系数确定方程组解的表达式。

该法则于1729年由英国数学家马克劳林（Maclaurin）得到，1748年发表，但克莱姆的优越符号使之流传。

