第一章　数值分析基础

数值分析

南京理工大学数统学院

1.1 数值分析课程简介

问题1：考虑线性方程组求解问题

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}$$

1.1 数值分析课程简介

问题1：考虑线性方程组求解问题

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}$$

理论上可依据 Grammer 法则求解，计算量为 $O(n \cdot n!)$。用美国Summit超级计算机，每秒20亿亿(2×10^{17})次运算，

$n = 20$ 时，需要243.29秒，约4分钟。

$n = 25$ 时，需要538583.68小时，约61.5年！

Grammer 法则理论上很完美，但实际工程问题中根本无法使用！

1.1 数值分析课程简介

问题1：考虑线性方程组求解问题

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}$$

理论上可依据 Grammer 法则求解，计算量为 $O(n{\cdot}n!)$。用美国Summit超级计算机，每秒20亿亿(2×10^{17})次运算，

$n = 20$ 时，需要243.29秒，约4分钟。

$n = 25$ 时，需要538583.68小时，约61.5年！

Grammer 法则理论上很完美，但实际工程问题中根本无法使用！

问题: 如何计算可节约计算量，设计实际工程可用的计算方法？

事实上，对于同一个问题，可设计具有不同计算量的算法。

例： 计算 $P(x) = 2x^4 + 3x^3 - 3x^2 + 5x + 1$

事实上，对于同一个问题，可设计具有不同计算量的算法。

例：计算 $P(x) = 2x^4 + 3x^3 - 3x^2 + 5x + 1$

1. 直接从左到右依次计算，需要 10 次乘法，4 次加法，共 14 次运算；

事实上，对于同一个问题，可设计具有不同计算量的算法。

例：计算 $P(x) = 2x^4 + 3x^3 - 3x^2 + 5x + 1$

1. 直接从左到右依次计算，需要 10 次乘法，4 次加法，共 14 次运算；
2. 先算 x^2，再算$x^3 = x^2 * x$，$x^4 = x^3 * x$，需要 7 次乘法，4 次加法，共 11 次运算；

事实上，对于同一个问题，可设计具有不同计算量的算法。

例：计算 $P(x)=2x^4+3x^3-3x^2+5x+1$

1. 直接从左到右依次计算，需要 10 次乘法，4 次加法，共 14 次运算；
2. 先算 x^2，再算$x^3=x^2*x$，$x^4=x^3*x$，需要 7 次乘法，4 次加法，共 11 次运算；
3. $P(x)=1+x*(5+x*(-3+x*(3+2*x)))$，只需要 4 次乘法，4 次加法，共只需 8 次运算。

启发： 合理地设计计算方法，可以简化计算，节省计算量！

问题2：考虑函数 $f(x) = \sin x$, $x \in [0, \pi]$ 的计算。

已知，$f(0) = 0, f(\frac{\pi}{6}) = \frac{1}{2}, f(\frac{\pi}{4}) = \frac{\sqrt{2}}{2}, \cdots$，问题：$f(\frac{\pi}{5}) =?$

问题2：考虑函数 $f(x) = \sin x$, $x \in [0, \pi]$ 的计算。

已知，$f(0) = 0, f(\frac{\pi}{6}) = \frac{1}{2}, f(\frac{\pi}{4}) = \frac{\sqrt{2}}{2}, \cdots$，问题：$f(\frac{\pi}{5}) =?$

可能的方法：

- 利用 Taylor 展开式 $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$，截取前 n 项近似计算。

问题2：考虑函数 $f(x)=\sin x$, $x\in[0,\pi]$ 的计算。

已知，$f(0)=0, f(\frac{\pi}{6})=\frac{1}{2}, f(\frac{\pi}{4})=\frac{\sqrt{2}}{2},\cdots$，问题：$f(\frac{\pi}{5})=?$

可能的方法：

- 利用 Taylor 展开式 $\sin x=x-\frac{x^3}{3!}+\frac{x^5}{5!}+\cdots$，截取前 n 项近似计算。
- 构造函数

$$P(x)=\frac{1}{2}\cdot\frac{(x-0)(x-\frac{\pi}{4})}{(\frac{\pi}{6}-0)(\frac{\pi}{6}-\frac{\pi}{4})}+\frac{\sqrt{2}}{2}\cdot\frac{(x-0)(x-\frac{\pi}{6})}{(\frac{\pi}{4}-0)(\frac{\pi}{4}-\frac{\pi}{6})}$$

则有 $f(0)=P(0), f(\frac{\pi}{6})=P(\frac{\pi}{6}), f(\frac{\pi}{4})=P(\frac{\pi}{4})$，由于在 $x=\frac{\pi}{5}$ 附近的已知节点上的函数值都正确，因此有理由相信 $f(\frac{\pi}{5})\approx P(\frac{\pi}{5})$。

问题： 哪种近似方法更好？更方便的计算，更小的误差...

更多的问题......

- 如何求方程 $x^3 - x + e^x = 0$ 的根？(非线性方程（组）求根)
- 如何计算 $\int_1^2 \frac{\sin x}{x} dx$？（数值积分）
- 如何求解描述大气运动的 Lorenz 微分方程？（常微分方程数值方法）

$$\begin{cases} \dot{x} = P(y - x) \\ \dot{y} = Rx - y - xz \\ \dot{z} = xy - bz \end{cases}$$

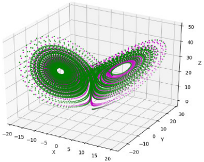


Figure: Lorenz方程吸引子

- 天气预测、石油开采、地下水污染等的数学模型求解（偏微分方程数值方法）

数值分析： 分析和研究数值计算方法的课程，数值计算也叫科学计算，已经和理论研究、实验研究一起，并列成为现代科学研究的三大研究手段之一。

Figure: 实际工程问题的解决途径

Top ten algorithms of the century: “We tried to assemble the 10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century” —Editors of IEEE Computational Science and Engineering, Jan. 2000 (后被 SIAM 转载)

1. 1946，Monte Carlo方法的 Metropolis 算法（机器学习重要基础之一）
2. 1947，解线性规划问题的单纯形算法（simplex method）
3. 1950，Krylov子空间迭代法（Lanczos过程、CG算法等，现代大规模科学计算的基础算法之一）
4. 1950’s，A. Householder形式化的矩阵分解方法（表示为矩阵分解）

Top ten algorithms of the century(Cont...)

5. 1957，Fortran 最优编译器（科学计算性能优于C/C++）
6. 1959-1961，能稳定计算矩阵特征值的 QR 算法（现代大规模科学计算的基础算法之一）
7. 1962，快速排序算法（Quicksort）
8. 1965，快速 Fourier 变换（FFT 及 IFFT 算法，现代信号分析核心基础算法之一）
9. 1977，整数关系侦察算法（Integer Relation Detection）
10. 1987，快速多极算法（fast multipole algorithm）

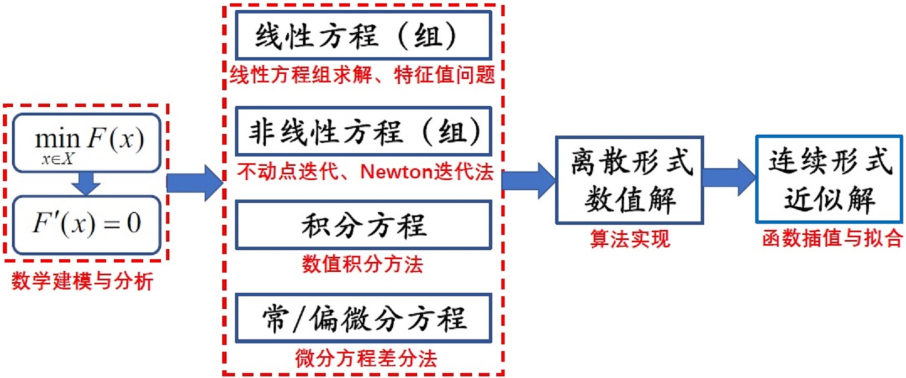
大多数算法都与数值计算直接相关！部分算法也属于本课程内容。

针对实际问题的数学模型求解，数值分析主要研究能够在计算机上进行实现的算法，其主要内容有：

1. 直接求解: 如何方便计算？减少计算量，甚至变不可能为可能
2. 近似求解，要解决的问题：
   (1) 用什么来近似？（近似的基本原理，研究算法设计原理）
   (2) 如何进行近似？（近似的具体方法，研究算法具体设计）
   (3) 如何度量近似的好坏？（近似的误差分析，研究算法优缺点）
   (4) 如何在计算机上实现？（近似的具体实现，上机实现具体算法）

最重要的概念：基的表示（逼近方法建模），算法设计，算法稳定性和收敛性（误差分析），算法复杂度

数学建模与数值分析主要内容：



1.2 误差分析基础

在实际问题中，误差几乎是不可避免的。误差根据其来源，可分为：

- **观测误差：** 由观测产生的误差（物理因素造成）
- **模型误差：** 由实际问题转化为数学模型时产生的误差（数学简化时的误差）；
- **截断误差：** 也被称为方法误差，是数值近似解与模型精确解之间的误差（构造数值算法时产生）；
- **舍入误差：** 实际计算中，计算机在计算过程中只能对有限位数的数字进行运算，对于超过位数的数字便自动四舍五入，产生舍入误差；

1.误差的度量

定义: 设 x^* 为精确值，x_a 为 x^* 的一个近似值，称 $|x_a - x^*|$ 为近似值的绝对误差（简称误差）。

一般称

$$\frac{|x_a - x^*|}{|x^*|} \quad (x^* \neq 0)$$

为相对误差，但实际中由于 x^* 是未知的，所以也可用

$$\frac{|x_a - x^*|}{|x_a|} \quad (x_a \neq 0)$$

作为相对误差。

定义: 设 x_a 为精确值 x^* 的一个近似值，如果存在 $\varepsilon_a > 0$，使得

$$|x_a - x^*| \leq \varepsilon_a$$

则称 ε_a 为近似值的一个绝对误差界，若 $x^* \neq 0$ ，称 $\dfrac{\varepsilon_a}{|x^*|}$ 为相对误差界。

定义: 设 $x_a = \pm 10^k \times 0.d_1 d_2 d_3 \cdots d_m \cdots$ 为精确值 x^* 的一个近似值，其中 $d_1 \neq 0$，x_a 有可能为有限或者无限小数形式，如果 n 是满足

$$|x_a - x^*| \leq 0.5 \times 10^{k-n}$$

的最大非负整数，则称 x_a 为 x^* 的具有 n 位十进制有效数字的近似值。

例： 对于圆周率 π 的近似（$\pi : 3.14159265358979323846\cdots$）

(1) $\pi_a = 3.14$，则 0.002 和 0.0006 分别为近似值的一个绝对误差界和一个相对误差界。

(2) $\pi_a = 3.14 = 10^1 \times 0.314$，由于 $|\pi - \pi_a| \le 0.5 \times 10^{-2} = 0.5 \times 10^{1-3}$，因此 3.14 近似 π 时有三位有效数字。

(3) $\pi_a = \frac{22}{7} = 10^1 \times 0.3142857\cdots$，由于 $|\pi - \pi_a| \le 0.5 \times 10^{-2} = 0.5 \times 10^{1-3}$，因此 $\frac{22}{7}$ 近似 π 时有三位有效数字。（祖冲之“约率”）

(4) $\pi_a = \frac{355}{113} = 10^1 \times 0.31415929\cdots$，由于 $|\pi - \pi_a| \le 0.5 \times 10^{-6} = 0.5 \times 10^{1-7}$，因此 $\frac{355}{113}$ 近似 π 时有七位有效数字。（祖冲之“密率”）

2. 求函数值的误差估计

(1) 假设一元函数f具有二阶连续导数, 自变量x的一个近似值为x_A, 由Taylor公式可以估计$f(x_A)$对$f(x)$近似的绝对误差:

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

其中ξ在x与x_A之间.

若$f'(x_A) \ne 0$, $|f''(\xi)|$与$f'(x_A)$相比不太大, 则上式右端第二项可忽略, 于是有$f(x_A)$绝对误差的如下近似估计:

$$|f(x_A) - f(x)| \le |f'(x_A)||x - x_A|.$$

(2) 若f是n元函数, 自变量$x_1, \cdots, x_n$的近似值分别为$x_{1A}, \cdots, x_{nA}$, 则有$f(\boldsymbol{x}_A)$绝对误差的如下近似估计:

$$|f(x_{1A}, \cdots, x_{nA}) - f(x_1, \cdots, x_n)| \leq \sum_{k=1}^{n} \left| \left(\frac{\partial f}{\partial x_k} \right)_A \right| |x_k - x_{kA}|.$$

3. 病态问题与条件数

若初始数据的微小扰动, 而使解严重失真, 则称该问题是病态的, 反之, 称为良态的.

例: $f(x)$的计算:

函数值相对误差:

$$\left|\frac{f(x+h)-f(x)}{f(x)}\right| \approx \left|\frac{hf'(x)}{f(x)}\right| = \left|\frac{xf'(x)}{f(x)}\right|\left|\frac{h}{x}\right|.$$

计算函数值问题的条件数: $C = \left|\dfrac{xf'(x)}{f(x)}\right|.$

4.数值稳定性

对于一个数值方法(公式), 若初始数据或计算过程中某一步的结果有微小改变, 而由此引起的最后结果的改变也是微小的, 则称该方法是数值稳定的, 否则称为数值不稳定的. (允许误差, 要求误差的传播可控)

如果一种数值方法有初始误差$\epsilon_0 > 0$, 由它引起此后运算n步的误差是ϵ_n(下面C是与n无关的常数)

(1) 如果$|\epsilon_n| \approx Cn\epsilon_0$, 称误差线性增长;

(2) 如果$|\epsilon_n| \approx C^n\epsilon_0$, 称误差指数增长.

例 建立

$$\mathrm{I}_n = \int_0^1 \frac{x^n}{x+25}\mathrm{d}x$$

的递推公式, 计算$\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3, \mathrm{I}_4$.

例 建立

$$\mathrm{I}_n = \int_0^1 \frac{x^n}{x+25}\mathrm{d}x$$

的递推公式，计算$\mathrm{I}_1, \mathrm{I}_2, \mathrm{I}_3, \mathrm{I}_4$.

解

$$\mathrm{I}_n + 25\mathrm{I}_{n-1} = \int_0^1 \frac{x^n + 25x^{n-1}}{x+25}\mathrm{d}x = \int_0^1 x^{n-1}\mathrm{d}x = \frac{1}{n},$$

$$\mathrm{I}_n = \frac{1}{n} - 25\mathrm{I}_{n-1}.$$

保留五位小数，计算得：

$$\mathrm{I}_0 = \int_0^1 \frac{1}{x+25}\mathrm{d}x = \ln(x+25)\Big|_0^1 \approx 0.03922,$$

$$\mathrm{I}_1 = 1 - 25\mathrm{I}_0 \approx 0.01950, \quad \mathrm{I}_2 = \frac{1}{2} - 25\mathrm{I}_1 \approx 0.01250,$$

$$\mathrm{I}_3 = \frac{1}{3} - 25\mathrm{I}_2 \approx 0.02083, \quad \mathrm{I}_4 = \frac{1}{4} - 25\mathrm{I}_3 \approx -0.27075.$$

$$\mathrm{I}_1 = 1 - 25\mathrm{I}_0 \approx 0.01950, \quad \mathrm{I}_2 = \frac{1}{2} - 25\mathrm{I}_1 \approx 0.01250,$$

$$\mathrm{I}_3 = \frac{1}{3} - 25\mathrm{I}_2 \approx 0.02083, \quad \mathrm{I}_4 = \frac{1}{4} - 25\mathrm{I}_3 \approx -0.27075.$$

$\Rightarrow$上述解法是有问题的!

分析原因: 计算I_0时只保留了小数点后五位,

$$I_1 = 1 - 25I_0 \approx 0.01950, \quad I_2 = \frac{1}{2} - 25I_1 \approx 0.01250,$$
$$I_3 = \frac{1}{3} - 25I_2 \approx 0.02083, \quad I_4 = \frac{1}{4} - 25I_3 \approx -0.27075.$$

$\Rightarrow$上述解法是有问题的!

分析原因: 计算I_0时只保留了小数点后五位,

上述解法每计算一步, 将前一步误差放大25倍!

解 将递推公式写为:

$$\mathrm{I}_{n-1} = \frac{1}{25}\left(\frac{1}{n} - \mathrm{I}_n\right),$$

$$\left|\mathrm{I}_{n-1} - \widetilde{\mathrm{I}}_{n-1}\right| = \frac{1}{25}\left|\mathrm{I}_n - \widetilde{\mathrm{I}}_n\right|.$$

另一方面, 注意到:

$$0 < \int_0^1 \frac{x^n}{x+25}\mathrm{d}x < \int_0^1 \frac{x^n}{25}\mathrm{d}x = \frac{1}{25}\frac{1}{n+1} \to 0(n \to \infty),$$

因此, $\mathrm{I}_n \to 0(n \to \infty)$, 可取$\widetilde{\mathrm{I}}_7 = 0$, 由递推公式计算得:

$\widetilde{\mathrm{I}}_4 = 0.00774$, $\widetilde{\mathrm{I}}_3 = \cdots\cdots$

5.误差的危害及处理

危害一：损失有效数字（很接近的数做减法时出现）

例： 求解 $x^2 - 16x + 1 = 0$，很容易得到解

$$x_1 = 8 + \sqrt{63} \approx 15.9372539332\cdots, \quad x_2 = 8 - \sqrt{63} \approx 0.0627460668\cdots$$

计算机只能对有限数字进行运算，对于无理数只能采用近似值计算。采用三位十进制数计算，取 $\sqrt{63} \approx 7.94$，有三位有效数字。

5.误差的危害及处理

危害一：损失有效数字（很接近的数做减法时出现）

例： 求解 $x^2-16x+1=0$，很容易得到解

$$x_1=8+\sqrt{63}\approx 15.9372539332\cdots, \quad x_2=8-\sqrt{63}\approx 0.0627460668\cdots$$

计算机只能对有限数字进行运算，对于无理数只能采用近似值计算。采用三位十进制数计算，取 $\sqrt{63}\approx 7.94$，有三位有效数字。

(1) 若直接计算，则 $x_1=8+\sqrt{63}\approx 15.9$, $\quad x_2=8-\sqrt{63}\approx 0.06$，$x_1$ 有三位有效数字，而x_2 只有一位有效数字。

5.误差的危害及处理

危害一：损失有效数字（很接近的数做减法时出现）

例： 求解 $x^2 - 16x + 1 = 0$，很容易得到解

$$x_1 = 8 + \sqrt{63} \approx 15.9372539332\cdots, \quad x_2 = 8 - \sqrt{63} \approx 0.0627460668\cdots$$

计算机只能对有限数字进行运算，对于无理数只能采用近似值计算。采用三位十进制数计算，取 $\sqrt{63} \approx 7.94$，有三位有效数字。

(1) 若直接计算，则 $x_1 = 8+\sqrt{63} \approx 15.9$， $x_2 = 8-\sqrt{63} \approx 0.06$，$x_1$ 有三位有效数字，而x_2 只有一位有效数字。

(2) 若用$x_2 = \frac{1}{x_1}$，则 $x_2 \approx 0.0629$，具有两位有效数字。

计算原则：避免很接近的数字作减法运算

危害二：大数“吃”小数（相差很大的数作加法时出现）

例： 求用三位十进制有效数字计算

$$x = 101 + \delta_1 + \delta_2 + \cdots + \delta_{100}$$

其中 $\delta_i \in [0.1, 0.4]$, $i = 1, 2, \cdots, 100$。

危害二：大数“吃”小数（相差很大的数作加法时出现）

例： 求用三位十进制有效数字计算

$$x = 101 + \delta_1 + \delta_2 + \cdots + \delta_{100}$$

其中 $\delta_i \in [0.1, 0.4]$， $i = 1, 2, \cdots, 100$。

(1) 若直接从左到右计算，则每次计算由于有效位数的限制，小数部分全部丢失（大数“吃”小数），最后得到 $x \approx 101$。

危害二：大数“吃”小数（相差很大的数作加法时出现）

例： 求用三位十进制有效数字计算

$$x = 101 + \delta_1 + \delta_2 + \cdots + \delta_{100}$$

其中 $\delta_i \in [0.1, 0.4]$， $i = 1, 2, \cdots, 100$。

(1) 若直接从左到右计算，则每次计算由于有效位数的限制，小数部分全部丢失（大数“吃”小数），最后得到 $x \approx 101$。

(2) 若先把小数部分全部加起来（此时小数部分之和在[10,40]之间），最后再加101，则可得到 $x \in [111, 141]$，相对要准确得多。

计算原则：避免相差很大的数字作加法运算

危害三：误差的放大（问题病态或采用不稳定方法时出现）

例： 对于方程组

$$\begin{pmatrix} 1 & 0 \\ 0 & 10^{-16} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

此时得到精确解：$x_1 = x_2 = 0$。

若方程的右端有了很小的扰动

$$\begin{pmatrix} 1 & 0 \\ 0 & 10^{-16} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 10^{-10} \\ 10^{-10} \end{pmatrix}$$

此时得到解：$x_1 = 10^{-10}/1 \approx 0,\ x_2 = 10^{-10}/10^{-16} = 10^6$!

计算原则：避免很小的数做分母（避免系统病态/ill-posed）

1.3 数值分析基础知识

1、线性空间与基的表达

定义：（线性空间）设 X 为一非空集合，K 为一数域（实数域 $\mathbb{R}$ 或复数域 $\mathbb{C}$），若对于任意元素 $x, y \in X$ 和数 $k, l \in K$，都有 $kx + ly \in X$，则称 X 为数域 K 上的线性空间。（对于加法和数乘封闭）

(1) $C^n[a,b] = \{f \,|\, f$为定义在 $[a,b]$ 上的具有 n 阶连续导数的实值函数$\}$；

(2) $P_n[a,b] = \{f \,|\, f$为定义在 $[a,b]$ 上的所有次数不超过 n 的代数多项式$\}$；

(3) $L^p[a,b] = \left\{ f(x) \,\middle|\, \left[\int_a^b |f(x)|^p dx\right]^{1/p} < +\infty, x \in [a,b] \right\}$;

(4) $\mathbb{R}^n / \mathbb{C}^n = \left\{ x \,\middle|\, x = (x_1, x_2, \cdots, x_n)^T, x_i \in \mathbb{R}/\mathbb{C}, i = 1, 2, \cdots, n \right\}$;

(5) $\mathbb{R}^{n\times m} / \mathbb{C}^{n\times m} = \{ A \,|\, A = (a_{ij})_{n\times m},\ a_{ij} \in \mathbb{R}/\mathbb{C},\ i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m \}$

设 $\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$ 为线性空间 X 的一组基，则对于任意 $\alpha\in X$，有

基的表示： $\alpha = k_1\varphi_1+k_2\varphi_2+\cdots+k_d\varphi_d$

其中 $k_1,k_2,\cdots,k_d$ 为表示系数，此时称 X 是由$\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$ 张成的，记为

$$X=\mathrm{span}\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$$

设 $\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$ 为线性空间 X 的一组基，则对于任意 $\alpha \in X$，有

$$\text{基的表示：}\quad \alpha = k_1\varphi_1+k_2\varphi_2+\cdots+k_d\varphi_d$$

其中 $k_1, k_2, \cdots, k_d$ 为表示系数，此时称 X 是由$\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$ 张成的，记为

$$X = \text{span}\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$$

例如，$P_n[a,b] = \text{span}\{1, x, x^2, \cdots, x^n\}$，对任意 $p_n(x) \in P_n[a,b]$，有

$$p_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \quad \forall x \in [a,b]$$

设 $\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$ 为线性空间 X 的一组基，则对于任意 $\alpha\in X$，有

$$\text{基的表示：}\quad \alpha=k_1\varphi_1+k_2\varphi_2+\cdots+k_d\varphi_d$$

其中 $k_1,k_2,\cdots,k_d$ 为表示系数，此时称 X 是由$\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$ 张成的，记为

$$X=\mathrm{span}\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$$

例如，$P_n[a,b]=\mathrm{span}\{1,x,x^2,\cdots,x^n\}$，对任意 $p_n(x)\in P_n[a,b]$，有

$$p_n(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0,\quad \forall x\in[a,b]$$

注意：一个空间的基是不唯一的。

对于 $C_{m\times n} = A_{m\times l}B_{l\times n}$ （矩阵的分解，对矩阵 C 进行分解），

$$[\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n] = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_l] \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{l1} & b_{l2} & \cdots & b_{ln} \end{bmatrix}$$

因此，$c_j = b_{1,j}\mathbf{a}_1 + b_{2,j}\mathbf{a}_2 + \cdots + b_{l,j}\mathbf{a}_l, \quad j = 1, 2, \cdots, n$。

将 $\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_l\}$ 看作子空间的“基”，则 $c_j \in \mathrm{span}\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_l\}$。

对于 $C_{m\times n}=A_{m\times l}B_{l\times n}$（矩阵的分解，对矩阵 C 进行分解），

$$[\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_n]=[\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_l]\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{l1} & b_{l2} & \cdots & b_{ln} \end{bmatrix}$$

因此，$c_j=b_{1,j}\mathbf{a}_1+b_{2,j}\mathbf{a}_2+\cdots+b_{l,j}\mathbf{a}_l,\quad j=1,2,\cdots,n$。

将 $\{\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_l\}$ 看作子空间的“基”，则 $c_j\in\mathrm{span}\{\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_l\}$。

需要指出的是，在严格数学定义上，构成“基”的向量组必须是线性无关的。但在实际工程问题中，即使向量组中有部分向量线性相关，由于不影响对元素的线性表示，往往还是被称为基（冗余基或冗余字典）。

如果矩阵 A 可以正交对角化，即存在正交矩阵 P，使得

$$\begin{aligned}
A &= P\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)P^T \\
&= [p_1, p_2, \cdots, p_n]\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)[p_1, p_2, \cdots, p_n]^T \\
&= \lambda_1 \cdot p_1 p_1^T + \lambda_2 \cdot p_2 p_2^T + \cdots + \lambda_n \cdot p_n p_n^T
\end{aligned}$$

这里 $p_i p_i^T, i = 1, 2, \cdots, n$ 都是由 A 的特征向量生成的秩 1 矩阵，则 A 可看作是矩阵子空间 $\mathrm{span}\{p_1p_1^T, p_2p_2^T, \cdots, p_np_n^T\}$ 中的一个元素，而表示系数则是对应的特征值。

如果矩阵 A 可以正交对角化，即存在正交矩阵 P，使得

$$\begin{aligned} A &= P\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)P^T \\ &= [p_1, p_2, \cdots, p_n]\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)[p_1, p_2, \cdots, p_n]^T \\ &= \lambda_1 \cdot p_1 p_1^T + \lambda_2 \cdot p_2 p_2^T + \cdots + \lambda_n \cdot p_n p_n^T \end{aligned}$$

这里 $p_i p_i^T, i = 1, 2, \cdots, n$ 都是由 A 的特征向量生成的秩 1 矩阵，则 A 可看作是矩阵子空间 $\mathrm{span}\{p_1 p_1^T, p_2 p_2^T, \cdots, p_n p_n^T\}$ 中的一个元素，而表示系数则是对应的特征值。

若 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$，且对于$k > K, K < n$时 λ_k “充分小”，则

$$A \approx \tilde{A} = \lambda_1 \cdot p_1 p_1^T + \lambda_2 \cdot p_2 p_2^T + \cdots + \lambda_K \cdot p_K p_K^T$$

这种近似在当前的大数据降维中经常使用（$K \ll n$）。

2、内积及其应用

定义：（内积） 设 X 是数域 K 上的线性空间，若存在映射 $(\cdot,\cdot):X\times X\to\mathbb{R}$ 满足如下条件:

(1) 对任意 $x\in X$，有 $(x,x)\geq 0$ ，当且仅当 $x=0$ 时等号成立;

(2) 对任意 x, $y\in X$，有 $(x,y)=\overline{(y,x)}$;

(3) 对任意 x, $y,z\in X$ 和 $k,l\in K$，有 $(kx+ly,z)=k(x,z)+l(y,z)$;

则称映射 $(\cdot,\cdot):X\times X\to\mathbb{R}$ 为线性空间 X 上的内积，定义了内积运算后的线性空间 X 称为内积空间。

内积的性质

(1) $(x, y) = \overline{(y, x)}$;

(2) $(\lambda x, y) = \lambda(x, y)$;

(3) $(x + y, z) = (x, z) + (y, z)$;

(4) $(x, x) \geq 0$，仅当$x = 0$ 时才有$(x, x) = 0$;

(5) $(x, y)(y, x) \leq (x, x)(y, y)$ （Cauchy-Schwarz 不等式）.

内积的性质

(1) $(x,y)=\overline{(y,x)}$;

(2) $(\lambda x,y)=\lambda(x,y)$;

(3) $(x+y,z)=(x,z)+(y,z)$;

(4) $(x,x)\geq 0$，仅当$x=0$ 时才有$(x,x)=0$;

(5) $(x,y)(y,x)\leq(x,x)(y,y)$ （Cauchy-Schwarz 不等式）.

注：内积可以认为是线性空间中两个元素之间的“相关性”（方向的一致性）的一种度量（以 $x,y\in\mathbb{R}^n$ 为例，$(x,y)=|x|\cdot|y|\cdot\cos(x,y)$）。

例如，利用内积最大可以对图像处理中的梯度场匹配问题建模。

➢ 数字图像可以表示为以像素为元素的矩阵

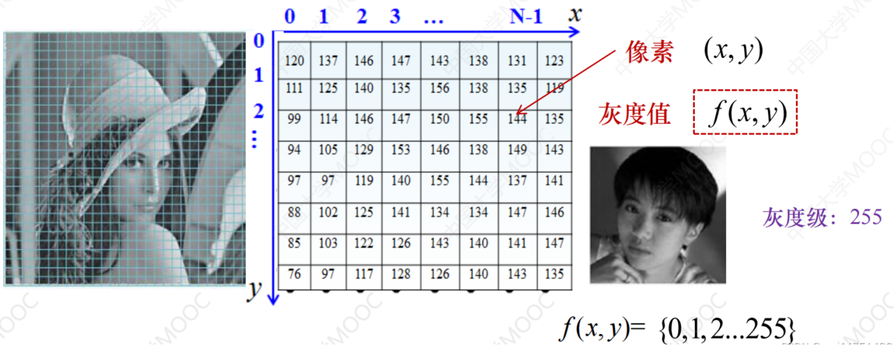


Figure: 图像在计算机中的表示，数字图像

常用的内积：

(1) 当 $X=\mathbb{C}^n$，则对任意 $x=(x_1,x_2,\ldots,x_n)^T, y=(y_1,y_2,\cdots,y_n)^T\in\mathbb{C}^n$

$$(x,y)=y^H x=\sum_{i=1}^{n} x_i\bar{y}_i,\quad (x,y)_w=\sum_{i=1}^{n} w_i\cdot x_i\bar{y}_i$$

其中 $w_i>0,\ \ i=1,2,\cdots,n$ 为加权系数。

(2) 当 $X=L^2[a,b]$，则对于任意 $f,g\in L^2[a,b]$，有

$$(f,g)=\int_a^b f\cdot\bar{g}\,dx,\quad (f,g)_\rho=\int_a^b \rho(x)f(x)\bar{g}(x)\,dx,$$

其中 $\rho(x)\geq 0, x\in[a,b]$，且在任意子区间上 $\rho(x)$ 不恒为零，称 $\rho(x)$ 为权函数。

定义: 设 $x \in \mathbb{C}^n$，非负实数 $\sqrt{(x,x)}$ 称为向量 x 的长度，通常称为**模**或者**范数**，记为 $\|x\|$ 或 $|x|$。若 $\|x\|=1$，则称 x 为单位向量。

定义: 设$x, y \in \mathbb{C}^n$，则

(1) 若 $(x,y)=0$，则称 x 与 y **正交**，记作 $x \perp y$.

(2) 若非零向量组 $x_1, \cdots, x_m$ 两两正交，即当 $i \neq j$ 时，$(x_i, x_j)=0$，当 $i=j$ 时，$(x_i, x_i)=c_i>0$，则称该向量组为**正交向量组**。

(3) 若正交向量组 $x_1, \cdots, x_m$ 还满足 $\|x_i\|=1, i=1,2,\cdots,m$，则称其为**标准正交向量组**。

定义： 设 $A\in\mathbb{C}^{n\times n}$，若有 $A^HA=AA^H=I$ 或 $A^{-1}=A^H$，称 A 为**酉矩阵**（在实数意义下，就是正交矩阵）。

酉矩阵的性质：

(1) 若 A 是酉矩阵，则 $A^{-1},\bar{A},A^T,A^H,A^k(k=1,2,\cdots)$ 也是酉矩阵。(定义验证)

(2) 若 A,B 是酉矩阵，则 AB 也是酉矩阵。(定义验证)

(3) 若 A 是酉矩阵，则 $|\det A|=1$。

(4) A 是酉矩阵的充要条件是它的 n 个列（行）向量是标准正交向量组。

(5) A 是酉矩阵的充要条件是对任意 $x,y\in\mathbb{C}^n$，有 $(Ax,Ay)=(x,y)$。

(6) 若 A 是酉矩阵，λ 为 A 的特征值，则 $|\lambda|=1$。

3.范数及其应用

定义： 设 X 是数域 K 上的线性空间，若映射 $||\cdot||: X \to \mathbb{R}$ 同时满足下列条件：

(1) （非负性）对任意 $x \in X$，有 $||x|| \geq 0$，当且仅当 $x = 0$ 时等号成立;

(2) （齐次性）对任意 $x \in X$ 和任意 $\lambda \in K$，有 $||\lambda x|| = |\lambda| \cdot ||x||$;

(3) （三角不等式）对任意 $x, y \in X$，有 $||x + y|| \leq ||x|| + ||y||$;

则称 $||\cdot||: X \to \mathbb{R}$ 为线性空间 X 上的一个范数，此时称 $(X, ||\cdot||)$ 为赋范线性空间（简称 X 为赋范线性空间）。

3.范数及其应用

定义： 设 X 是数域 K 上的线性空间，若映射 $||\cdot||: X \to \mathbb{R}$ 同时满足下列条件：

(1) （非负性）对任意 $x \in X$，有 $||x|| \geq 0$，当且仅当 $x = 0$ 时等号成立;

(2) （齐次性）对任意 $x \in X$ 和任意 $\lambda \in K$，有 $||\lambda x|| = |\lambda| \cdot ||x||$;

(3) （三角不等式）对任意 $x, y \in X$，有 $||x + y|| \leq ||x|| + ||y||$;

则称 $||\cdot||: X \to \mathbb{R}$ 为线性空间 X 上的一个范数，此时称 $(X, ||\cdot||)$ 为赋范线性空间（简称 X 为赋范线性空间）。

注： 范数可以度量元素的大小，是一种距离的定义，事实上 $||x|| = ||x - 0||$，即 x 到 0 的距离。因此前面误差的概念都可以在范数意义下重新定义，可以对 $x, y \in X$ 之间的误差进行度量。

设 $(X, ||\cdot||)$ 是一个赋范线性空间，如果对于 X 中的任一 Cauchy 序列 $\{x_n\}$（即满足 $||x_m - x_n|| \to 0$ $(m,\ n \to +\infty)$ 的序列），必存在 $x \in X$，使得 $||x_n - x|| \to 0$ ，则称$(X, ||\cdot||)$ 为完备的赋范线性空间，通常称为 Banach 空间。

设 $(X, ||\cdot||)$ 是一个赋范线性空间，如果对于 X 中的任一 Cauchy 序列 $\{x_n\}$（即满足 $||x_m - x_n|| \to 0$ $(m,\ n \to +\infty)$ 的序列），必存在 $x \in X$，使得 $||x_n - x|| \to 0$ ，则称$(X, ||\cdot||)$ 为完备的赋范线性空间，通常称为 Banach 空间。

注1: 设 X 是一个内积空间，则 $||x|| = \sqrt{(x,\ x)}$ 定义了 X 上的一种范数，因此内积空间一定是赋范线性空间，但反之未必;

设 $(X, ||\cdot||)$ 是一个赋范线性空间，如果对于 X 中的任一 Cauchy 序列 $\{x_n\}$（即满足 $||x_m - x_n|| \to 0$ $(m,\ n \to +\infty)$ 的序列），必存在 $x \in X$，使得 $||x_n - x|| \to 0$，则称$(X, ||\cdot||)$ 为完备的赋范线性空间，通常称为 Banach 空间。

注1: 设 X 是一个内积空间，则 $||x|| = \sqrt{(x,\ x)}$ 定义了 X 上的一种范数，因此内积空间一定是赋范线性空间，但反之未必;

注2: 若内积空间 X 在范数 $||x|| = \sqrt{(x,\ x)}$ 意义下是 Banach 空间，则称 X 是一个完备的内积空间，通常称为 Hilbert 空间;

设 $(X, ||\cdot||)$ 是一个赋范线性空间，如果对于 X 中的任一 Cauchy 序列 $\{x_n\}$（即满足 $||x_m - x_n|| \to 0$ $(m,\ n \to +\infty)$ 的序列），必存在 $x \in X$，使得 $||x_n - x|| \to 0$ ，则称$(X, ||\cdot||)$ 为完备的赋范线性空间，通常称为 Banach 空间。

注1: 设 X 是一个内积空间，则 $||x|| = \sqrt{(x,\ x)}$ 定义了 X 上的一种范数，因此内积空间一定是赋范线性空间，但反之未必;

注2: 若内积空间 X 在范数 $||x|| = \sqrt{(x,\ x)}$ 意义下是 Banach 空间，则称 X 是一个完备的内积空间，通常称为 Hilbert 空间;

注3: 对于 L^p 空间，仅当 $p = 2$ 时， L^p 是 Hilbert 空间，其余情况均为 Banach 空间。

常用的范数：

当 $X = C[a,b]$，对于任意 $f \in C[a,b]$，可定义连续意义下的范数：

(1) $||f||_1 = \int_a^b |f(x)|\, dx$（1范数）

(2) $||f||_2 = \left(\int_a^b |f(x)|^2 dx \right)^{1/2}$（2范数）

(3) $||f||_\infty = \max\limits_{x \in [a,b]} |f(x)|$（$\infty$范数）

(4) $||f||_p = \left(\int_a^b |f(x)|^p dx \right)^{1/p} \ (1 \le p < \infty)$（$p$范数）

（可以证明：$\lim\limits_{p \to +\infty} ||f||_p = ||f||_\infty$）

向量范数

当 $X=\mathbb{C}^n$，设 $x=(\xi_1,\xi_2,\cdots,\xi_n)^T\in\mathbb{C}^n$，规定

(1) $\|x\|_1=\sum\limits_{k=1}^{n}|\xi_k|$ （1范数）

(2) $\|x\|_2=\sqrt{\sum\limits_{k=1}^{n}|\xi_k|^2}=\sqrt{x^Hx}$ （2范数）

(3) $\|x\|_\infty=\max\limits_{k}|\xi_k|$ （∞范数）

(4) $\|x\|_p=\left(\sum\limits_{k=1}^{n}|\xi_k|^p\right)^{1/p}$ $(1\le p<\infty)$
（p范数）

（可以证明 $\lim\limits_{p\to+\infty}||x||_p=||x||_\infty$）

向量范数

当 $X=\mathbb{C}^n$，设 $x=(\xi_1,\xi_2,\cdots,\xi_n)^T\in\mathbb{C}^n$，规定

(1) $\|x\|_1=\sum_{k=1}^{n}|\xi_k|$ （1范数）

(2) $\|x\|_2=\sqrt{\sum_{k=1}^{n}|\xi_k|^2}=\sqrt{x^Hx}$ （2范数）

(3) $\|x\|_\infty=\max_k|\xi_k|$ （∞范数）

(4) $\|x\|_p=\left(\sum_{k=1}^{n}|\xi_k|^p\right)^{1/p}$ $(1\le p<\infty)$ （p范数）

（可以证明 $\lim_{p\to+\infty}||x||_p=||x||_\infty$）

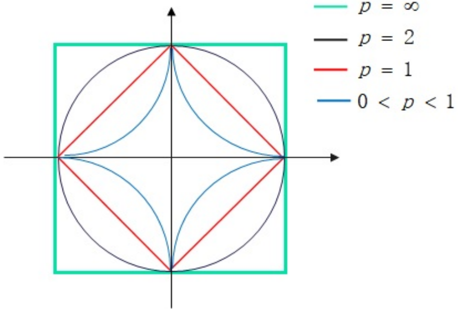


Figure: $||x||_p$ 意义下的单位圆

向量范数的基本性质：

(1) 任何向量范数都是 $\mathbb{C}^n$ 上的连续函数 $\left(\lim\limits_{||\Delta x||\to 0}||x+\Delta x|| = ||x||\right)$。

(2) 设 $\|\cdot\|_\alpha$ 与 $\|\cdot\|_\beta$ 是 $\mathbb{C}^n$ 上的两个向量范数，如存在 $a, b > 0$ 使得

$$a\|x\|_\beta \le \|x\|_\alpha \le b\|x\|_\beta, \qquad (\forall x \in \mathbb{C}^n)$$

则称向量范数 $\|\cdot\|_\alpha$ 与 $\|\cdot\|_\beta$ 等价。$\mathbb{C}^n$ 上的所有向量范数等价。

(3) 设 $A \in \mathbb{C}_n^{n\times n}$，$\|\cdot\|_\alpha$ 是 $\mathbb{C}^n$ 上一个向量范数，则 $\|x\|_\beta = \|Ax\|_\alpha$ 也是 $\mathbb{C}^n$ 上向量范数。

(4) 设 A 是 n 阶 Hermite 正定矩阵，对任意 $x \in \mathbb{C}^n$，规定 $\|x\|_A = \sqrt{x^H Ax}$，则 $\|x\|_A$ 是一个向量范数（椭圆范数）。

给定 $\mathbb{C}^n$ 中的向量序列 $\{x^{(k)}\}$，其中 $x^{(k)}=(\xi_1^{(k)},\cdots,\xi_n^{(k)})^T\ (k=1,2,\cdots)$，如果

$$\lim_{k\to\infty}\xi_j^{(k)}=\xi_j \qquad (j=1,2,\cdots,n)$$

则称向量序列 $\{x^{(k)}\}$ 收敛于 $x=(\xi_1,\cdots,\xi_n)^T$，简称 $\{x^{(k)}\}$ 收敛，记为

$$\lim_{k\to\infty}x^{(k)}=x \quad \text{或} \quad x^{(k)}\to x\ (k\to+\infty)$$

不收敛的向量序列称为是发散的。

如果当 $k\to\infty$ 时，有 $\|x^{(k)}-x\|\to 0$，则称 $\{x^{(k)}\}$ 依范数 $\|\cdot\|$ 收敛于向量 $x\in\mathbb{C}^n$。简记为

$$x^{(k)}\xrightarrow{\|\cdot\|}x \quad (k\to\infty)\ \text{或}\ \lim_{k\to\infty}x^{(k)}=x\ (\text{依范数}\|\cdot\|)。$$

矩阵范数

定义： 如果 $\mathbb{C}^{n\times n}$ 上的一个实函数$\|\cdot\|:\mathbb{C}^{n\times n}\to\mathbb{R}$ 满足以下条件：

(1) （非负性）$\|A\|\geq 0$，当且仅当 $A=\mathbf{0}$ 时，$\|A\|=0$

(2) （齐次性）对 $\lambda\in\mathbb{C}$，$\|\lambda A\|=|\lambda|\,\|A\|$

(3) （三角不等式）$\|A+B\|\leq\|A\|+\|B\|$

(4) （相容性）$\|AB\|\leq\|A\|\,\|B\|$

则称 $\|\cdot\|$ 为是 $\mathbb{C}^{n\times n}$ 上的一个矩阵范数，而 $\|A\|$ 为 $\mathbb{C}^{n\times n}$ 上 A 的范数（矩阵范数）。

常用的矩阵范数：设 $A=(a_{ij})_{n\times n}\in\mathbb{C}^{n\times n}$，规定

$$\|A\|_{m_1}=\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}| \qquad (m_1\text{范数})$$

$$\|A\|_{m_\infty}=\color{red}{n}\max_{i,j}|a_{ij}| \qquad (m_\infty\text{范数})$$

$$\|A\|_F=(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2)^{1/2}=\sqrt{tr(A^HA)} \quad (\text{Frobenius 范数，}F\text{范数})$$

F 范数的酉不变性：对酉矩阵 U, V，有

$$\|UA\|_F=\|AV\|_F=\|UAV\|_F=\|A\|_F$$

单位矩阵在矩阵乘法中的作用类似于 1 在数的乘法中的作用，但对于 m_1-范数，F-范数和 m_∞-范数，有 $\|I_n\|_{m_1} = n$, $\|I_n\|_F = \sqrt{n}$, $\|I_n\|_{m_\infty} = n$，能否构造出使 $\|I_n\| = 1$ 的范数呢？

单位矩阵在矩阵乘法中的作用类似于 1 在数的乘法中的作用，但对于 m_1-范数，F-范数和 m_∞-范数，有 $\|I_n\|_{m_1}=n$，$\|I_n\|_F=\sqrt{n}$，$\|I_n\|_{m_\infty}=n$，能否构造出使 $\|I_n\|=1$ 的范数呢？

定义： 已知 $\mathbb{C}^n$ 上的一个向量范数 $\|\cdot\|_v$，对 $A\in\mathbb{C}^{n\times n}$，规定

$$\|A\|=\max_{x\neq 0}\frac{\|Ax\|_v}{\|x\|_v}\quad\left(=\max_{\|x\|_v=1}\|Ax\|_v\right)$$

则 $\|\cdot\|$ 是 $\mathbb{C}^{n\times n}$ 上与向量范数 $\|\cdot\|_v$ 相容的矩阵范数，且满足 $\|I_n\|=1$，称之为由向量范数 $\|\cdot\|_v$ 诱导出的矩阵范数或从属于向量范数 $\|\cdot\|_v$ 的矩阵范数，简称诱导范数、从属范数或算子范数。

由向量 1，2，∞ 范数导出的矩阵范数

(1) $\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|$，　（极大列和范数）

(2) $\|A\|_2 = \sqrt{\lambda_1}$，$\lambda_1$ 为$A^H A$或AA^H 的最大特征值（谱范数，2-范数）

(3) $\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|$，　（极大行和范数）

矩阵2-范数的性质

1. 酉不变性：设 U, V 为酉矩阵，则 $\|UA\|_2 = \|AV\|_2 = \|UAV\|_2 = \|A\|_2$
2. 若 $A^H A = AA^H$，则称 A 为正规矩阵。

   若 A 是正规矩阵，且 $\lambda_1, \lambda_2, \cdots, \lambda_n$ 是 A 的 n 个特征值，则

$$\|A\|_2 = \max_k |\lambda_k|$$

例： 设 U 是酉矩阵，$A=\begin{pmatrix} -1 & i & 0 \\ -i & 0 & -i \\ 0 & i & -1 \end{pmatrix}$，试计算 $\max_{||x||_1=4}||Ax||_1$ $\max_{||Ux||_2=3}||Ax||_2$.

例： 设 U 是酉矩阵，$A = \begin{pmatrix} -1 & i & 0 \\ -i & 0 & -i \\ 0 & i & -1 \end{pmatrix}$，试计算 $\max_{||x||_1=4}||Ax||_1$
$\max_{||Ux||_2=3}||Ax||_2$.

解： (1)首先设 $x = 4y$，则

$$\max_{||x||_1=4}||Ax||_1 = 4 \cdot \max_{||y||_1=1}||Ay||_1 = 4||A||_1 = 8$$

（2）设 $x = 3U^H y$，则

$$\max_{\|Ux\|_2=3} \|Ax\|_2 = \max_{\|y\|_2=1} 3\|AU^H y\|_2 = 3\|AU^H\|_2 = 3\|A\|_2.$$

而 A 是 Hermite 矩阵，其特征值为 $\lambda_1 = -2$，$\lambda_1 = -1$，$\lambda_1 = 1$。所以

$$\max_{\|Ux\|_2=3} \|Ax\|_2 = 3\|A\|_2 = 3|\lambda_1| = 6.$$

矩阵范数与向量范数的关系

定义: （**相容**）对于 $\mathbb{C}^n$ 上给定的一种向量范数 $||x||_v$ 和 $\mathbb{C}^{n\times n}$ 上给定的一种矩阵范数 $||A||_m$，如果

$$||Ax||_v \le ||A||_m \ ||x||_v,$$

则称上述矩阵范数和向量范数是相容的。

定理： 任何向量范数都存在与之相容的矩阵范数（算子范数），反之设$\|\cdot\|_m$ 是一个矩阵范数，则必存在与之相容的向量范数。

证明： 对 $x \in \mathbb{C}^n$，记 $X = (x, x, \cdots, x) \in \mathbb{C}^{n\times n}$。定义 $\mathbb{C}^n$ 上函数 $\|x\|_v = \|X\|_m$。易证 $\|x\|_v$ 是 $\mathbb{C}^n$ 上的范数，并且对任意 $A \in \mathbb{C}^{n\times n}$，有

$$\|Ax\|_v = \|(Ax, Ax, \cdots, Ax)\|_m = \|AX\|_m \le \|A\|_m\|X\|_m = \|A\|_m\|x\|_v$$

例： 向量 1 范数与矩阵 m_1 范数相容。

例： 向量 1 范数与矩阵 m_1 范数相容。

证明： 设 $A=(a_{ij})_{n\times n}$，$x=(\xi_1,\cdots,\xi_n)^T$，则

$$
\begin{aligned}
\|Ax\|_1 &= \sum_{i=1}^{n}\left|\sum_{k=1}^{n} a_{ik}\xi_k\right| \le \sum_{i=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}|\,|\xi_k|\right) \\
&\le \sum_{i=1}^{n}\left[\left(\sum_{k=1}^{n}|a_{ik}|\right)\left(\sum_{k=1}^{n}|\xi_k|\right)\right] \\
&= \left(\sum_{i=1}^{n}\sum_{k=1}^{n}|a_{ik}|\right)\left(\sum_{k=1}^{n}|\xi_k|\right) \\
&= \|A\|_{m_1}\|x\|_1
\end{aligned}
$$

例： 向量2 范数与矩阵F 范数相容。

例： 向量2 范数与矩阵F 范数相容。

$$
\begin{aligned}
\|Ax\|_2 &= \sqrt{\sum_{i=1}^{n} \left| \sum_{k=1}^{n} a_{ik}\xi_k \right|^2} \\
&\le \sqrt{\sum_{i=1}^{n} \left(\sum_{k=1}^{n} |a_{ik}|\,|\xi_k| \right)^2} \\
&\le \sqrt{\sum_{i=1}^{n} \left[\left(\sum_{k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{k=1}^{n} |\xi_k|^2 \right) \right]} \\
&= \|A\|_F \|x\|_2
\end{aligned}
$$

可证明：向量 1，2，∞ 范数均与矩阵 m_∞ 范数相容。

矩阵的谱半径

定义： 设 $A \in \mathbb{C}^{n\times n}$，$\lambda_1, \lambda_2, \cdots, \lambda_n$ 为 A 的 n 个特征值，称

$$\rho(A) = \max_j |\lambda_j|$$

为 A 的谱半径。

矩阵的谱半径

定义： 设 $A \in \mathbb{C}^{n\times n}$，$\lambda_1, \lambda_2, \cdots, \lambda_n$ 为 A 的 n 个特征值，称

$$\rho(A) = \max_j |\lambda_j|$$

为 A 的谱半径。

注：$\rho(A)$ 是 $\mathbb{C}^{n\times n}$ 上的一个函数，但它不是矩阵范数。

例：设

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},\ E = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

则$A \neq 0$，但$\rho(A) = 0$；（不满足非负性条件）

$\rho(A+B) = 1 > 0 = \rho(A) + \rho(B)$；（不满足三角不等式）

$\rho(EF) = (1+\sqrt{5})/2 > 1 = \rho(E)\rho(F)$。（不满足相容性）

定理： 设 $A \in \mathbb{C}^{n\times n}$，则 $\rho(A) \leq \|A\|$.

证明： 设 λ 是 A 的特征值，$\|\cdot\|_v$ 为 $\mathbb{C}^n$ 上与 $\|\cdot\|$ 相容的向量范数。由 $Ax = \lambda x$ 得

$$|\lambda|\,\|x\|_v = \|\lambda x\|_v = \|Ax\|_v \leq \|A\|\,\|x\|_v$$

由于 $\|x\|_v \neq 0$，从而 $|\lambda| \leq \|A\|$，故 $\rho(A) \leq \|A\|$

注：设 $A \in \mathbb{C}^{n\times n}$，则

(1) $\|A\|_2 \leq \max\{\|A\|_\infty, \|A\|_1\}$

(2) 当 A 是正规矩阵时，$\rho(A) = \|A\|_2$。

定理： 设 $A \in \mathbb{C}^{n\times n}$，对任意 $\varepsilon > 0$，存在矩阵范数 $\|\cdot\|_m$ 使

$$\|A\|_m \leq \rho(A) + \varepsilon$$

定理： 设 $A \in \mathbb{C}^{n\times n}$，对任意 $\varepsilon > 0$，存在矩阵范数 $\|\cdot\|_m$ 使

$$\|A\|_m \le \rho(A) + \varepsilon$$

证明： 设 $P \in \mathbb{C}_n^{n\times n}$ 使得

$$P^{-1}AP = J = \begin{pmatrix} \lambda_1 & \delta_1 & & \\ & \lambda_2 & \ddots & \\ & & \ddots & \delta_{n-1} \\ & & & \lambda_n \end{pmatrix}, \quad \delta_i = 0 \text{ 或 } 1$$

令 $D = diag(1, \varepsilon, \varepsilon^2, \cdots, \varepsilon^{n-1})$，则有

$$D^{-1}P^{-1}APD = D^{-1}JD = \begin{pmatrix} \lambda_1 & \varepsilon\delta_1 & & & \\ & \lambda_2 & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon\delta_{n-1} \\ & & & & \lambda_n \end{pmatrix}$$

于是有$\|D^{-1}P^{-1}APD\|_\infty \le \max_j(|\lambda_j| + \varepsilon) = \rho(A) + \varepsilon.$

规定

$$\|B\|_m = \|D^{-1}P^{-1}BPD\|_\infty, \quad B \in C^{n\times n}$$

则$\|\cdot\|_m$ 是$C^{n\times n}$ 上的一种矩阵范数且有

$$\|A\|_m = \|D^{-1}P^{-1}APD\|_\infty \le \rho(A) + \varepsilon$$

扰动方程与矩阵的条件数

实际工程问题很多都归结于求解线性方程

$$Ax = b$$

而在实际中，误差往往是不可避免的，因此需要研究下面的扰动方程组

$$(A + \delta A)(x + \delta x) = b + \delta b$$

其中 δx 是由系数矩阵的扰动 δA 和右端扰动 δb 引起的解的误差。

我们需要在有误差 δA 和 δb 的情况下，研究解的误差。

引理： 设 $P \in \mathbb{C}^{n\times n}$，若对 $\mathbb{C}^{n\times n}$ 上的某一矩阵范数 $\|\cdot\|$ 有 $\|P\| < 1$，则 $\mathrm{I} - P$ 可逆。

证明： 若 $(\mathrm{I} - P)$ 不可逆，则 $(\mathrm{I} - P)x = 0$ 有非零解 $x^{(0)}$，即有

$$(\mathrm{I} - P)x^{(0)} = 0，\text{或} x^{(0)} = Px^{(0)}.$$

设$\|\cdot\|_v$ 是 C^n 上与矩阵范数 $\|\cdot\|$ 相容的向量范数，则

$$\|x^{(0)}\|_v = \|Px^{(0)}\|_v \le \|P\|\,\|x^{(0)}\|_v$$

即有 $\|P\| \ge 1$，与条件矛盾。

定理： 设 $A \in \mathbb{C}_n^{n\times n}$，$\delta A \in \mathbb{C}^{n\times n}$，$b$, $\delta b \in \mathbb{C}^n$。若对 $\mathbb{C}^{n\times n}$ 上某矩阵范数 $\|\cdot\|$ 有$\|A^{-1}\|\,\|\delta A\| < 1$，则

$$Ax = b \quad \text{与} \quad (A+\delta A)(x+\delta x) = b + \delta b$$

的解满足（相对误差）

$$\frac{\|\delta x\|_v}{\|x\|_v} \le \frac{\|A\|\,\|A^{-1}\|}{1 - \|A\|\,\|A^{-1}\|\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|_v}{\|b\|_v}\right)$$

其中 $\|\cdot\|_v$ 是 $\mathbb{C}^n$ 上与矩阵范数 $\|\cdot\|$ 相容的向量范数。

证明： 首先，$A+\delta A = A(\mathrm{I}+A^{-1}\delta A)$，根据引理 $A+\delta A$ 可逆。将 $(A+\delta A)(x+\delta x)=b+\delta b$ 展开，得

$$A\delta x + (\delta A)x + (\delta A)\delta x = \delta b$$

$$\delta x = -A^{-1}(\delta A)x - A^{-1}(\delta A)\delta x + A^{-1}\delta b$$

$$\|\delta x\|_v \le \|A^{-1}\|\,\|\delta A\|\,\|x\|_v + \|A^{-1}\|\,\|\delta A\|\,\|\delta x\|_v$$

$$+\|A^{-1}\|\,\|\delta b\|_v$$

又$\|b\|_v = \|Ax\|_v \le \|A\|\,\|x\|_v$，得

$$\frac{\|\delta x\|_v}{\|x\|_v} \le \frac{\|A\|\,\|A^{-1}\|}{1-\|A\|\,\|A^{-1}\|\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|_v}{\|b\|_v}\right)$$

定义：设 $A \in \mathbb{C}_n^{n\times n}$，$\|\cdot\|$ 是 $\mathbb{C}^{n\times n}$ 上的矩阵范数，称

$$cond(A) = \|A\|\,\|A^{-1}\|$$

为**矩阵 A 的条件数**。

如果方程组系数矩阵条件数大，称该方程组是病态的。

常用的矩阵条件数：

(a) $cond_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$;

(b) $cond_2(A) = \|A\|_2 \|A^{-1}\|_2 = \sqrt{\dfrac{\mu_1}{\mu_n}}$，其中 μ_1, μ_n 分别为 $A^H A$ 的最大与最小特征值；

(c) 当 A 为正规矩阵时，$cond_2(A) = \left|\dfrac{\lambda_1}{\lambda_n}\right|$，其中 λ_1, λ_n 分别是 A 的按模最大和最小特征值。

例： 设

$$A = \begin{pmatrix} -1 & i & 0 \\ -i & 0 & -i \\ 0 & i & -1 \end{pmatrix}, \delta(A) \in \mathbb{C}^{3\times 3}, 0 \neq b \in \mathbb{C}^3.$$

为使 $Ax = b$ 的解 x 与 $(A + \delta(A))x = b$ 的解 $\hat{x}$ 的相对误差 $\dfrac{\|\hat{x} - x\|_2}{\|x\|_2} \leq 10^{-4}$，问 $\dfrac{\|\delta(A)\|_2}{\|A\|_2}$ 应满足什么条件（不超过何值）？

解： 因为 $|\lambda \mathrm{I} - A| = (\lambda+1)(\lambda-1)(\lambda+2)$，所以

$$\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2 = 2$$

从而

$$\frac{\|\hat{x}-x\|_2}{\|x\|_2} \le \frac{\mathrm{cond}_2(A)}{1-\mathrm{cond}_2(A)\frac{\|\delta A\|_2}{\|A\|_2}} \frac{\|\delta A\|_2}{\|A\|_2} \le 10^{-4},$$

$$\frac{2}{1-2\frac{\|\delta A\|_2}{\|A\|_2}} \frac{\|\delta A\|_2}{\|A\|_2} \le 10^{-4},$$

从而得到

$$\frac{\|\delta A\|_2}{\|A\|_2} \le \frac{1}{2.0002} \times 10^{-4} \le \frac{1}{2} \times 10^{-4} = 5 \times 10^{-5}.$$