计算物理学(A)第一次作业

物理学院　陈伟杰 1500011335

April 4, 2018

1　数值误差的避免

1.1　N个数求平均

假设每一步加法都会引入ϵ_i的误差，记机器加法为$\oplus$。则有

$$
\begin{aligned}
\bar{x}^* &= \frac{1}{N}(\ldots((x_1 \oplus x_2) \oplus x_3) \oplus \ldots \oplus x_N) \\
&= \frac{1}{N}(\ldots((x_1 + x_2)(1+\epsilon_1) + x_3)(1+\epsilon_2) + \ldots x_N)(1+\epsilon_{N-1}) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} x_i + \sum_{k=1}^{N-1}\epsilon_k \sum_{i=1}^{k+1} x_i + O(\epsilon^2)\right)
\end{aligned}
$$

略去$O(\epsilon^2)$项，计算机器算出的平均值与理论平均值的误差：

$$
\begin{aligned}
|\bar{x}^* - \bar{x}| = \frac{1}{N}\left|\sum_{k=1}^{N-1}\epsilon_k \sum_{i=1}^{k+1} x_i\right| &\le \frac{1}{N}\sum_{k=1}^{N-1}|\epsilon_k|\sum_{i=1}^{k+1}|x_i| \\
&\le \frac{1}{N}\sum_{k=1}^{N-1}\left((k+1)\frac{\epsilon_M}{2}\cdot|x_M|\right) = \frac{\epsilon_M \cdot |x_M|}{4}\left(\frac{N^2+N-2}{N}\right) \\
&\sim \frac{N}{4}\epsilon_M|x_M|
\end{aligned}
$$

上式中$|x_M|$表示$\{x_i\}$的绝对值最大的元素。

假定N很大，则可以认为$|x_M|$与$\bar{x}$量级相同，因此计算$\bar{x}$的相对舍入误差大约是$\boldsymbol{\frac{N}{4}\epsilon_M}$

1.2　计算方差的稳定性和准确性

两个公式的运算次数都是$O(N)$量级，因此可以认为这个过程的机器舍入造成的误差大致相同。但是考虑到求和的数据比较多，那么可能有：

$$
\frac{1}{N-1}\sum_{i=1}^{N} x_i^2 \sim \frac{N}{N-1}\bar{x}^2 \gg S^2
$$

使得在两项相减时的机器舍入误差非常显著，对结果造成很大的不稳定性。

在这个层面上，虽然第二个公式的运算次数比第一个公式大致多N次，但是这样的N个$(x_i-\bar{x})^2$大致在相同的量级，并且都小于S^2，运算中积累的舍入误差就比较小，结果也更稳定。

综上，第二个公式在计算中更稳定和准确。

1.3　积分递推的稳定性

$n=0$时

$$
I_0 = \int_0^1 \frac{1}{x+5}dx = ln(6/5)
$$

$n=k\geq 1$时

$$I_k+5I_{k-1}=\int_0^1\left(\frac{x^k}{x+5}+\frac{5x^{k-1}}{x+5}\right)dx=\int_0^1 x^{k-1}dx=1/k$$

则对于$n\gg 1$的I_n由递推关系得：

$$I_n=\sum_{k=1}^{n}\frac{(-5)^{n-k}}{k}+(-5)^n I_0$$

因此如果计算I_0有一微小误差ϵ，则由递推关系会有 $\boldsymbol{(-5)^n\epsilon}$ 的误差，因此由递推关系求I_n是不稳定的。

2 矩阵的模与条件数

已知A是一$n\times n$的上三角实矩阵，矩阵A的元素定义如下：

$$A_{ij}=\begin{cases}1 & i=j\\ -1 & i<j\\ 0 & i>j\end{cases}\qquad A=\begin{bmatrix}1 & -1 & -1 & \dots & -1\\ & 1 & -1 & \dots & -1\\ & & 1 & \dots & -1\\ & & & \ddots & -1\\ & & & & 1\end{bmatrix}$$

2.1 A的行列式

对于上三角矩阵，容易得到$det(A)=1$。由于$det(A)\neq 0$，因此A是非奇异矩阵。

2.2 A的逆矩阵

逆矩阵A^{-1}的元素定义和形式如下：

$$A^{-1}_{ij}=\begin{cases}1 & i=j\\ 2^{j-i-1} & i<j\\ 0 & i>j\end{cases}\qquad A^{-1}=\begin{bmatrix}1 & 1 & 2 & 4 & \dots & 2^{n-3} & 2^{n-2}\\ & 1 & 1 & 2 & \dots & 2^{n-4} & 2^{n-3}\\ & & 1 & 1 & \dots & 2^{n-5} & 2^{n-4}\\ & & & \ddots & \ddots & \ddots & 2^{n-5}\\ & & & & \ddots & \ddots & \vdots\\ & & & & & 1 & 1\\ & & & & & & 1\end{bmatrix}$$

2.3 A的∞模

首先矢量的∞模定义如下(x_i为矢量x的分量)

$$\|x\|_\infty=\max_i|x_i|$$

则对矩阵A的∞模，按定义有：

$$\begin{aligned}\|A\|_\infty&=\sup_{x\neq 0}\frac{\|Ax\|_\infty}{\|x\|_\infty}=\sup_{\|x\|_\infty=1}\|Ax\|_\infty=\sup_{\|x\|_\infty=1}\left(\max_i\sum_{j=1}^{n}|A_{ij}||x_j|\right)\\&=\max_i\left(\sup_{\|x\|_\infty=1}\sum_{j=1}^{n}|A_{ij}||x_j|\right)=\max_i\sum_{j=1}^{n}|A_{ij}|\end{aligned}$$

2.4 A的欧氏模

对于酉矩阵$U \in \mathbb{C}^{n\times n}$，由欧氏模的定义得：

$$\|U\|_2 = \sup_{\|x\|_2=1} \|Ux\|_2 = \sup_{\|x\|_2=1} \sqrt{x^\dagger U^\dagger U x} = \sup_{\|x\|_2=1} \|x\|_2 = 1$$

同理对$U^\dagger$有：

$$\left\|U^\dagger\right\|_2 = \sup_{\|x\|_2=1} \left\|U^\dagger x\right\|_2 = \sup_{\|x\|_2=1} \sqrt{x^\dagger U U^\dagger x} = \sup_{\|x\|_2=1} \|x\|_2 = 1$$

因此$\left\|U^\dagger\right\|_2 = \|U\|_2 = 1$。

对任意$A \in \mathbb{C}^{n\times n}$，计算$\|UA\|_2$：

$$\|UA\|_2 = \sup_{\|x\|_2=1} \|UAx\|_2 = \sup_{\|x\|_2=1} \sqrt{x^\dagger A^\dagger U^\dagger U A x} = \sup_{\|x\|_2=1} \|Ax\|_2 = \|A\|_2$$

即$\|UA\|_2 = \|A\|_2$。

首先要证明$\|(UA)^{-1}\|_2 = \|A^{-1}\|_2$：

$$\begin{aligned}
\left\|(UA)^{-1}\right\|_2 &= \sup_{x\neq 0} \frac{\|(UA)^{-1}x\|_2}{\|x\|_2} = \sup_{\|x\|_2=1} \frac{\sqrt{x^\dagger U (A^{-1})^\dagger A^{-1} U^\dagger x}}{\|x\|_2} \\
&= \sup_{\|x\|_2=1} \frac{\sqrt{x^\dagger U (A^{-1})^\dagger A^{-1} U^\dagger x}}{\|U^\dagger x\|_2} = \sup_{\|y\|_2=\|U^\dagger x\|_2=1} \frac{\sqrt{y^\dagger (A^{-1})^\dagger A^{-1} y}}{\|y\|_2} \\
&= \sup_{\|y\|_2=\|U^\dagger x\|_2=1} \left\|A^{-1}y\right\|_2 = \left\|A^{-1}\right\|_2
\end{aligned}$$

由条件数的定义，计算$K_2(UA)$：

$$K_2(UA) = \|(UA)\|_2 \cdot \left\|(UA)^{-1}\right\|_2 = \|A\|_2 \cdot \left\|A^{-1}\right\|_2 = K_2(A)$$

2.5 A在∞模下的条件数

由前面给出的A与A^{-1}的形式得：

$$\|A\|_\infty = n \qquad \left\|A^{-1}\right\|_\infty = 2^{n-1} \quad \Rightarrow \quad K_\infty(A) = \|A\|_\infty \cdot \left\|A^{-1}\right\|_\infty = n2^{n-1}$$

3 Hilbert矩阵

3.1 H_n的矩阵元和矢量b

对于D的表达式可以改写成：

$$D = \int_0^1 dx \left(\sum_{i=1}^n c_i x^{i-1} - f(x)\right)\left(\sum_{j=1}^n c_j x^{j-1} - f(x)\right)$$

将上式对c_i做变分：

$$\frac{\delta D}{\delta c_i} = \int_0^1 x^{i-1}\left(\sum_{j=1}^n c_j x^{j-1} - f(x)\right) dx$$

由极小值条件，即有

$$\begin{aligned}
&\int_0^1 x^{i-1}\left(\sum_{j=1}^n c_j x^{j-1} - f(x)\right)dx = 0\\
\Rightarrow\quad &\int_0^1 x^{i-1}\left(\sum_{j=1}^n c_j x^{j-1}\right)dx = \int_0^1 x^{i-1} f(x)dx\\
\Rightarrow\quad &\sum_{j=1}^n c_j \int_0^1 x^{i+j-2}dx = \int_0^1 x^{i-1} f(x)dx\\
\Rightarrow\quad &\sum_{j=1}^n c_j H_{ij} = b_i
\end{aligned}$$

其中H_{ij}和b_i形式如下：

$$H_{ij} = \int_0^1 x^{i+j-2}dx = \frac{1}{i+j-1} \qquad\qquad b_i = \int_0^1 x^{i-1} f(x)dx$$

3.2 H_n的性质

由上述H_{ij}表达式容易得到$H_{ij} = H_{ji}$，即H_n是对称矩阵。

另一方面对于$\forall c \in \mathbb{R}^n$，考虑二次型$c^T H_n c$

$$c^T H_n c = \sum_{i,j=1}^n c_i c_j H_{ij} = \int_0^1 dx\left(\sum_{i,j=1}^n c_i c_j x^{i+j-2}\right) = \int_0^1 dx\left(\sum_{i=1}^n c_i x^{i-1}\right)^2 = \int_0^1 dx P_n(x)^2 \geq 0$$

因此，H_n至少是半正定的，而由于$\int_0^1 dx P_n(x)^2 = 0$仅当$P_n(x) = 0$(即对于$\forall 1 \leq i \leq n$，$c_i = 0$)成立，所以H_n是正定矩阵。

由于正定实矩阵的性质，H_n的任意阶主子式的行列式都大于0，因此$det(H_n) > 0$，即H_n非奇异。

3.3 H_n的行列式

根据$det(H_n)$的严格表达式，可以写出递推关系：

$$\begin{aligned}
det(H_{n+1}) &= det(H_n)\frac{(n!)^4}{(2n)!(2n+1)!} = det(H_n)\frac{(\Gamma(n+1))^4}{\Gamma(2n+1)\Gamma(2n+2)} = \frac{det(H_n)}{2^{4n+1}}\frac{\pi(\Gamma(n+1))^2}{\Gamma(n+\frac{1}{2})\Gamma(n+\frac{3}{2})}\\
&= \frac{det(H_n)}{2^{4n}}\prod_{k=1}^n\left(\frac{k}{k-\frac{1}{2}}\right)\left(\frac{k}{k+\frac{1}{2}}\right) = \frac{det(H_n)}{2^{4n}}\prod_{k=1}^n\left(1+\frac{1}{(2k+1)(2k-1)}\right)
\end{aligned}$$

取对数得：

$$ln(det(H_{n+1})) = ln(det(H_n)) - 4nln(2) + \sum_{k=1}^n ln\left(1+\frac{1}{(2k+1)(2k-1)}\right)$$

递推计算得$1 \leq n \leq 10$的$ln(det(H_n))$的值：

n	1	2	3	4	5
$ln(det(H_n))$	0	-2.4849	-7.6779	-15.6152	-26.3095
$det(H_n)$	1	8.3333e-02	4.6296e-4	1.6534e-07	3.7493e-12
n	6	7	8	9	10
$ln(det(H_n))$	-39.7662	-55.9886	-74.9784	-96.7369	-121.2650
$det(H_n)$	5.3673e-18	4.8358e-25	2.7371e-33	9.7202e-43	2.1642e-53

3.4 比较GEM与Cholesky分解

当n比较小时，GEM与Cholesky分解都能准确地给出$H_n x = b$的解。当n=10时，GEM和Cholesky分解分别给出解如下(参考值由Mathematica给出)：

	1	2	3	4	5
GEM	-9.99807293e+00	9.89833719e+02	-2.37564620e+04	2.40207863e+05	-1.26110682e+06
Cholesky	-9.99872776e+00	9.89892770e+02	-2.37577605e+04	2.40219965e+05	-1.26116570e+06
参考值	-10	990	-23760	240240	-1261260
	6	7	8	9	10
GEM	3.78335918e+06	-6.72602994e+06	7.00061346e+06	-3.93787022e+06	9.23703109e+05
Cholesky	3.78352368e+06	-6.72630350e+06	7.00088086e+06	-3.93801199e+06	9.23734551e+05
参考值	3783780	-6726720	700280	-3938220	923780

两种方法的解比较接近，大致在小数点后4-5位左右才有差别。与参考值比较，发现**Cholesky分解**更为准确。

根据条件数$K(A)$的定义：

$$K(A) = \|A\| \cdot \left\|A^{-1}\right\| = \frac{\|\delta x\| \,/\, \|x\|}{\|\delta Ax\| \,/\, \|Ax\|}$$

由上式得，假若由运算产生的$\|\delta Ax\| \,/\, \|Ax\|$认为是机器舍入精度$\epsilon_M$，则解的相对误差约为$\epsilon_M K(A)$。对于GEM，由于$H_n$的条件数极端大，因此数值解只有几位的准确数字。

而对于Cholesky分解，由于$H_n = A^\dagger A$，因此可以认为$K(A) \sim \sqrt{K(H_n)}$，所以尽管Cholesky分解的运算次数较多，但误差总体是线性积累的，而条件数比GEM小，所以最终结果比GEM更准确和稳定。