计算物理作业 Ex01

物理学院 *Bryan*[*] 学号：1500066666

1 数值误差的避免

1.1 求平均的误差

N 数平均的误差来源于求和、除以 N 两个过程；在 N 较大时，除以大数所引入的误差相对较小，此时求和的误差占主要成分。

两数相加时，引入的相对误差为机器精度 $\frac{\epsilon_M}{2}$; 记 $x_0 = \max|x_i|$, 考虑最坏的情况，即可能的误差最大值，这一情形在每个 $x_i \to x_0$ 时取到。不妨设 x_i 均为正数，此时求和的上限为：

$$f \circ f \circ \cdots \circ f(x_0) \equiv f^{N-1} \circ (x_0), \quad f(x) = (x + x_0)\left(1 + \frac{\epsilon_M}{2}\right) \tag{1.1}$$

这里 f 是每次数值求和操作的函数表示。

作用于 x_0	$\mathcal{O}(1)$ 项	$\mathcal{O}(\frac{\epsilon_M}{2})$ 系数
$f^0 = \mathbb{1}$	x_0	0
f^1	$2x_0$	$2x_0$
f^2	$3x_0$	$5x_0$
$\vdots$	$\vdots$	$\vdots$
f^k	$x_0 + kx_0$	c_k

f 的作用规律：先加 x_0，再乘以 $(1 + \frac{\epsilon_M}{2})$

考察 $\frac{\epsilon_M}{2}$ 的系数，设 f^k 作用后的 $\frac{\epsilon_M}{2}$ 系数为 c_k, 则不难发现：

$$c_k = c_{k-1} + x_0 + kx_0 \tag{1.2}$$

其中 kx_0 源于前一步 $\mathcal{O}(1)$ 项的系数。已知 $c_0 = 0$, 展开此递推关系，即得：

$$\begin{aligned} c_{N-1} &= \frac{(N+2)(N-1)}{2}\, x_0, \\ \text{均值的误差限：}\quad \frac{1}{N} \cdot c_{N-1} \frac{\epsilon_M}{2} &= \frac{(N+2)(N-1)}{2N} \frac{\epsilon_M}{2} \max|x_i| \sim \frac{N}{2}\frac{\epsilon_M}{2} \max|x_i| \end{aligned} \tag{1.3}$$

[*] 邮箱：pls_contact_via_github@fake_email.com

1.2 方差计算的稳定性

两种方差计算公式如下：

$$S^2 = \frac{1}{N-1}\left\{\sum_i x_i^2 - N\bar{x}^2\right\} \tag{1.4a}$$

$$= \frac{1}{N-1}\sum_i (x_i-\bar{x})^2 \tag{1.4b}$$

沿用前文给出的估计办法，可以给出两式的误差限；有：

$$\begin{aligned} e_{(a)} &\sim S^2\frac{\epsilon_M}{2} + \left\{\frac{N}{2}\frac{\epsilon_M}{2}\max|x_i|^2 + \frac{N}{2}\frac{\epsilon_M}{2}\max|x_i|\times 2\right\} \\ &= \left(S^2 + \frac{N}{2}\max|x_i|^2 + N\max|x_i|\right)\frac{\epsilon_M}{2}, \\ e_{(b)} &\sim \frac{N}{2}\frac{\epsilon_M}{2}\max|x_i-\bar{x}|^2 \end{aligned} \tag{1.5}$$

可见，多数情况下第一式 (1.4a) 将带来较大误差；特别是在 x_i 很大但方差却很小的情况下，此时将产生大数相消，从而大量损失有效数字。相比之下，第二式 (1.4b) 较为稳定和准确。

1.3 递归计算的稳定性

考察 $I_n = \int_0^1 \mathrm{d}x\,\frac{x^n}{x+5}$, 首先有 $I_0 = \ln(x+5)\big|_0^1 = \ln\frac{6}{5}$, 而：

$$I_k + 5I_{k-1} = \int_0^1 \mathrm{d}x\,\frac{x^k+5x^{k-1}}{x+5} = \int_0^1 \mathrm{d}x\,x^{k-1} = \frac{x^k}{k}\bigg|_0^1 = \frac{1}{k}, \quad k=1,2,\ldots \tag{1.6}$$

从而可以递归地给出 I_k 的值。

关注这一过程的误差传递，设计算值 $\hat{I}_{k-1} = I_{k-1} + \epsilon_{k-1}$, 则相应地：

$$\begin{aligned} \hat{I}_k &\sim \left(\frac{1}{k} - 5\hat{I}_{k-1}\right)\left(1+\frac{\epsilon_M}{2}\right) \\ &\sim I_k - 5\epsilon_{k-1} + I_k\frac{\epsilon_M}{2} \\ \text{即有：}\quad \epsilon_k &\sim -\left({}^{5}\!/_{I_k}\right)\epsilon_{k-1} + \frac{\epsilon_M}{2} \end{aligned} \tag{1.7}$$

系数 $\kappa = |5/I_k|$ 是关键；若 $\kappa < 1$, 则误差将得到控制，不会进一步放大。

然而，不幸的是，本问题中的 $I_n < I_0 < 1$, 即始终有 $\kappa > 1$, 初始误差 ϵ 将随递归过程不断（指数）放大，可见这一算法是不稳定的。

2 矩阵的模与条件数

2.1 矩阵 A 的基本性质

考虑矩阵 A, 有:

$$(A-\mathbb{1})_{ij}=\begin{cases}-1, & \text{for } \ i<j,\\ 0, & \text{for } \ i\geq j,\end{cases} \tag{2.1}$$

计算 n 阶 A_n 的行列式，注意有递归关系 $A_{n+1}=\begin{pmatrix} 1 & [-1]\times n \\ & A_n\end{pmatrix}$ —— 这里借用了 python 的记号：$[-1]\times n$ 表示长为 n 的常数列表。按第一行展开（Laplace expansion），注意到 A_{1j} 元素的代数余子式（minor）均有一列零元，故其对行列式的贡献为零，从而：

$$\det A_n = 1\times\det A_{n-1}=\cdots=\det A_1=1 \tag{2.2}$$

事实上，上述过程可推广到任何三角矩阵，由此得到三角矩阵的本征值：

$$\det A_n = a_{11}\det A_{n-1}=\cdots=\prod_i a_{ii} \tag{2.3}$$

即其对角元素的乘积。

2.2 A^{-1} 的形式

承接上文，记 A_n 的某代数余子式为 $\left|\tilde{A}_{ij}^{(n)}\right|$, $|\cdot|$ 为行列式的简记符号；从定义出发，有：

$$A^{-1}=\frac{\operatorname{adj}A}{\det A} \tag{2.4}$$

其中 $\operatorname{adj}A$ 的元素为 $(-1)^{i+j}\,|\tilde{A}_{ji}|\,/|A|$, 注意指标有交换，即应当取一个额外的转置。

类似前面的讨论，对一般的上三角矩阵，均有：

$$\left|\tilde{A}_{ij}^{(n)}\right|=\begin{cases}0, & \text{for } \ i<j,\\ |A_n|/a_{ii}, & \text{for } \ i=j,\end{cases} \tag{2.5}$$

再复合上一个转置，可得上（下）三角矩阵的逆依然是上（下）三角矩阵。

进一步，$\left|\tilde{A}_{ij}^{(n)}\right|$, $i>j$ 的情形较为复杂，这里同样采用递归的办法。不难发现，有：

$$\begin{aligned}\left|\tilde{A}_{ij}^{(n+1)}\right| &= \begin{vmatrix} 1 & [-1]\times(n-1) \\ & \tilde{A}_{i-1,j-1}^{(n)}\end{vmatrix} = \left|\tilde{A}_{i-1,j-1}^{(n)}\right|, \quad i>j>1,\\ &= \begin{vmatrix} \tilde{A}_{i,j}^{(n)} & \vdots \\ & 1\end{vmatrix} = \left|\tilde{A}_{ij}^{(n)}\right|, \quad i>j,\end{aligned} \tag{2.6}$$

上述化简利用了 A 的上三角特性。

由此可见，A^{-1} 具有平行于主对角线的带状结构，且：

$$A_n^{-1} = \begin{pmatrix} A_{n-1}^{-1} & \vdots \\ & 1 \end{pmatrix} = \begin{pmatrix} 1 & \cdots \\ & A_{n-1}^{-1} \end{pmatrix} \tag{2.7}$$

综上，A^{-1} 的形态已经基本确定，唯一未定的元素只剩下最右上角的 $A_{1,n}^{-1}$ 了。可以方便地以待定系数的方法给出 $A_{1,n}^{-1}$; 利用 $A^{-1}A = \mathbb{1}$, 不难得到：

$$A_{1,n}^{-1} = \sum_{j<n} A_{1,j}^{-1} = \sum_{i>1} A_{i,n}^{-1} \tag{2.8}$$

即它是第 1 行（或第 n 列）除去其自身以外其他元素的总和。如此，便可以递归地得到：

$$A_1^{-1} = (\,1\,), \quad A_2^{-1} = \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}, \quad A_3^{-1} = \begin{pmatrix} 1 & 1 & 2 \\ & 1 & 1 \\ & & 1 \end{pmatrix}, \quad A_4^{-1} = \begin{pmatrix} 1 & 1 & 2 & 4 \\ & 1 & 1 & 2 \\ & & 1 & 1 \\ & & & 1 \end{pmatrix}, \quad \cdots$$
$$A_{ij}^{-1} = \begin{cases} 0, & \text{for} \quad i > j, \\ 1, & \text{for} \quad i = j, \\ 2^{j-i-1}, & \text{for} \quad i < j, \end{cases} \tag{2.9}$$

2.3 矩阵的 ∞ 模

已知矢量 p 模：

$$\|x\|_p = \left\{ \sum_i |x_i|^p \right\}^{\frac{1}{p}} \xrightarrow{p\to\infty} \max|x_i| \lim_{p\to\infty} \left\{ 1 + \sum_{|x_i|<|x_{\max}|} \left| \frac{x_i}{x_{\max}} \right|^p \right\}^{\frac{1}{p}} = \max|x_i| \tag{2.10}$$

考虑相应的矩阵模，首先有：

$$\frac{\|Ax\|_p}{\|x\|_p} \xrightarrow{p\to\infty} \frac{\max\left|A_{ij}x^j\right|}{\max|x_i|} \tag{2.11}$$

$\forall\, x \neq 0$, 取上界，即得到 $\|A\|$. 注意数乘不改变 $\frac{\|Ax\|}{\|x\|}$, 故不妨限制 $\|x\| = 1$, 从而：

$$\|Ax\|_p \xrightarrow{p\to\infty} \max\left|A_{ij}x^j\right| \leq \max_i \sum_j |A_{ij}| \tag{2.12}$$

当 $x^j = \operatorname{sign} A_{ij}$ 时取到等号。也就是说，$\|A\|_\infty$ 即为矩阵的行和最大值。

2.4 矩阵的欧式模

欧式模由于和线性空间上的标准内积一致，因此有额外的优良性质。例如，对 $\mathbb{C}$ 上的幺正矩阵 U 而言，$\forall\, x$，均有：

$$\begin{aligned} p = 2, \quad \|Ux\|^2 = x^\dagger U^\dagger U x = x^\dagger U U^\dagger x &= \left\|U^\dagger x\right\|^2, \\ &= x^\dagger x = \|x\|^2, \end{aligned} \tag{2.13}$$

因此 $\|U\|_2 = \left\|U^\dagger\right\|_2 = 1$. 类似有 $\|(UA)\,x\|_2 = \|U(Ax)\|_2 = \|Ax\|_2$, 故 $\|UA\|_2 = \|U\|_2$. 而矩阵的条件数可一般性地表示为 $K_p(A) = \|A\|_p \left\|A^{-1}\right\|_p$, 故：

$$K_2(A) = K_2(UA) \tag{2.14}$$

2.5 A 的 ∞ 模条件数 $K_\infty(A)$

对于前述例子 A, 其行和最大值在第一行取到，即：$\|A\|_\infty = n$. 类似地，$\left\|A^{-1}\right\|_\infty = 1 + 2^{n-1} - 1 = 2^{n-1}$, 从而有：

$$K_\infty(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty = n\,2^{n-1} \tag{2.15}$$

3 Hilbert 矩阵

3.1 H_n 的具体形式

多项式近似的残差 $D = \int_0^1 \mathrm{d}x \left\{\sum_{i=1}^n c_i x^{i-1} - f(x)\right\}^2$, 为使其尽可能小，对参量 c_j 微分，即有：

$$\begin{aligned} 0 = \frac{\partial D}{\partial c_j} &= \int_0^1 \mathrm{d}x\ \frac{\partial}{\partial c_j} \left\{\sum_{i=1}^n c_i x^{i-1} - f(x)\right\}^2 \\ &= \int_0^1 \mathrm{d}x \cdot 2 \left\{\sum_{i=1}^n c_i x^{i-1} - f(x)\right\} \sum_{i=1}^n \delta_i^j x^{i-1} \\ &= \int_0^1 \mathrm{d}x \cdot 2x^{j-1} \left\{\sum_{i=1}^n c_i x^{i-1} - f(x)\right\} \end{aligned} \tag{3.1}$$

积分，得 $0 = \sum_{i=1}^n c_i \, \frac{1}{i+j-1} - \int_0^1 \mathrm{d}x\, x^{j-1} f(x)$, 即有：

$$\begin{aligned} H_n \cdot c &= b, \\ (H_n)_{ij} = \frac{1}{i+j-1}, &\quad b_j = \int_0^1 \mathrm{d}x\, x^{j-1} f(x) \end{aligned} \tag{3.2}$$

3.2 H_n 的特征

由 $(H_n)_{ij} = \frac{1}{i+j-1} = (H_n)_{ji}$, 可见 H_n 为对称矩阵；此外，参见上文，有：

$$c^T H_n c = \sum_{i,j=1}^{n} c_i c_j \int_0^1 x^{i+j-2}\,\mathrm{d}x = \int_0^1 \mathrm{d}x \left\{ \sum_{i=1}^{n} c_i x^{i-1} \right\}^2 \geq 0 \tag{3.3}$$

当且仅当 $\sum_{i=1}^{n} c_i x^{i-1} \equiv 0$ 即 $c = 0$ 时取到等号。可见 Hilbert 矩阵是对称正定矩阵。

此外，由对称正定性还可知 Hilbert 矩阵必定是非奇异的。事实上，对任一对称正定矩阵 A 而言，若它同时是奇异矩阵，则 $\det A = 0$, 故存在 $c \neq 0$ 使得 $Ac = 0$, 进而导致 $c^T Ac = 0$, 这与对称正定性矛盾。因此对称正定矩阵均非奇异。

3.3 $\det H_n$ 的行为

已知：

$$\det H_n = \frac{c_n^4}{c_{2n}}, \quad c_n = 1! \cdot 2! \dots (n-1)! \tag{3.4}$$

为估计 $\det H_n$ 的大小，取对数，注意到 $\ln c_n = \sum_{m=1}^{n-1} \ln m!$, 可得：

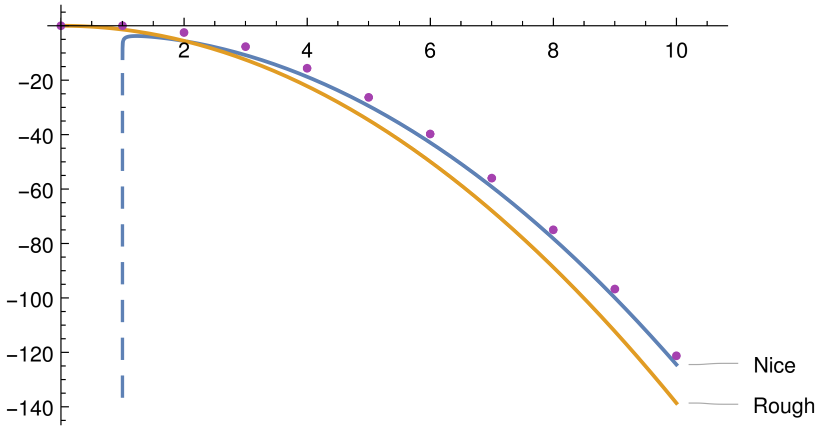
$$\ln \det H_n = 4 \ln c_n - \ln c_{2n} = \left\{ 4 \sum_{m=1}^{n-1} - \sum_{m=1}^{2n-1} \right\} \ln m! \tag{3.5}$$

下面尝试用 Stirling 近似给出上式子的一个近似。已知：

$$m! \sim \sqrt{2\pi m} \left(\frac{m}{e} \right)^m \tag{3.6}$$

带入上式，初步化简后得到：

$$\begin{aligned}
\ln \det H_n \sim & \left\{ 4 \sum_{m=1}^{n-1} - \sum_{m=1}^{2n-1} \right\} m \ln m \\
& + n + \left(n - \frac{3}{2} \right) \ln (2\pi) + \frac{1}{2} \big(4 \ln (n-1)! - \ln (2n-1)! \big) \\
\sim & \left\{ 4 \sum_{m=1}^{n-1} - \sum_{m=1}^{2n-1} \right\} m \ln m \\
& + (2n-1) \ln (n-1) - \left(n - \frac{1}{4} \right) \ln (2n-1) \\
& + \left(n - \frac{9}{4} \right) \ln (2\pi) + \frac{3}{2}
\end{aligned} \tag{3.7}$$



$\ln\det H_n$ **的渐进行为**

散点为精确值，Nice 为完整的渐进表达式，Rough 为 $-n^2\ln 4$.

注：Nice 表达式在 $n=1$ 处发散，故只适用于 $n\geq 2$.

进一步，考虑积分的几何意义，对 $\sum m\ln m$ 可采用如下近似：

$$\sum_{m=1}^{k} m\ln m \sim \int_{\frac{3}{2}}^{n+\frac{1}{2}} \mathrm{d}x\, x\ln x \tag{3.8}$$

即可得到 $\ln\det H_n$ 的完整近似形式。截取最高阶项，我们得到：

$$\boxed{\ln\det H_n \sim -2n^2\ln 2, \quad \det H_n \sim 4^{-n^2}, \quad n\to\infty} \tag{3.9}$$

由此可见，$\det H_n$ 随 n 增大而指数地减小，即 H_n 迅速地接近于一个奇异矩阵。注意，4^{-n^2} 给出的粗略近似（后面标记为 Rough）适用于大宗量 n 的情形；对于较小的 n 值而言，由于略去了过多的低阶项，结果可能不甚理想。