数值分析第一次上机练习实验报告

——线性代数方程组的数值解法

力 5 黄云帆

一、问题描述

设$H_n = [h_{ij}] \in \mathbb{R}^{n\times n}$是Hilbert 矩阵，即

$$h_{ij} = \frac{1}{i+j-1}.$$

对$\mathrm{n} = 2,3,4,\ldots$（根据计算机性能选取合适的n，建议算到$\mathrm{n} = 20$ 左右）

(a) 取$\mathbf{x} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \in \mathbb{R}^n$，及$\boldsymbol{b}_n = H_n\boldsymbol{x}$．再用Gauss消去法和Cholesky分解方法来解$H_n\boldsymbol{y} = \boldsymbol{b}_n$，并分析误差．

(b) 计算条件数：$\mathrm{cond}(H_n)_2$．

(c) 使用某种正则化方法改善(a)中的结果．

(d) 用SOR迭代方法和共轭梯度法求解$H_n\boldsymbol{x} = \boldsymbol{b}_n$，并与前面的直接方法做比较．

二、方法描述——线性方程组的典型直接解法与迭代解法

我们来分析上述问题．首先要利用精确解构造线性方程组，这样便于比较实际计算结果的误差；并通过计算Hilbert 矩阵的条件数，能够更清楚地理解不用预处理的直接解法导致误差爆炸的原因．在实际操作中，我们选取误差向量的 2-范数表征误差大小，选取矩阵的 2-范数来计算上述矩阵的条件数．由于Matlab已经提供了这两个函数，我们可以直接使用．

本次实验将要比较Gauss消去法和Cholesky分解两种直接方法（包括Tikhonov正则化预处理）以及SOR迭代方法和共轭梯度法两种迭代方法共五种求解方法对条件数较坏的Hilbert 矩阵的适用性．由于算例非常明确，且其中四种方法（不含预处理）在Matlab中均有已优化过的现成函数，因此下面我们以$\boldsymbol{A}\boldsymbol{x} = \mathbf{b}$为例给出具体操作方法，并侧重其中思想的描述．

1.直接方法及其预处理

[1]在处理只有一个右端向量的线性方程组时，Gauss消去法本质上与矩阵的LU分解方法相同，差别仅仅在于计算过程的选择．当然，在右端向量不止一个时，矩阵分解方法显然效率更高，这是由于先分解算子再代入具体问题能够使问题得到相当大程度上的简化．由于本次实验只有一个右端向量，我们就以LU分解的描述等价地代替Gauss消去法的描述．

LU分解方法理论上的可行性由非奇异矩阵的LU分解定理所保证．具体步骤为：

先写$\mathrm{A} \equiv \mathrm{LU}$　　(1)，其中L为单位下三角阵、U为上三角阵．接着依次解两个子问题$L\mathbf{y} = \mathbf{b}$，$U\boldsymbol{x} = \mathbf{y}$即可．其中$L, U$的具体表达可以通过比较(1)式两端结果得到．

上述方法实际中能够实现的必要条件是分解过程中$u_{rr} \neq 0$且不能够出现$|u_{rr}| \ll 1$的情况．实际操作中，我们需要采取选主元的方式修改上面的算法，即：

先写$\mathrm{PA} \equiv \mathrm{LU}$　　(2)，然后解$L\boldsymbol{y} = P\mathbf{b}$，$U\boldsymbol{x} = \mathbf{y}$．其中我们不直接利用(2)分解，而是把选取列主元的过程放在解第一个子问题过程中 0．

本次实验我们采用选主元LU分解这种方法进行求解．

[2] Cholesky分解适用于对称正定阵（Hilbert矩阵满足）的分解，由于矩阵结构变得简单，因此计算量大幅减少．具体步骤为：

先写$\mathrm{A} = \mathrm{L}L^T$　　(3)，再求解两个对称的方程$L\mathbf{y} = \mathbf{b}$，$L^T\boldsymbol{x} = \mathbf{y}$．

有时也利用$\mathrm{A} = \mathrm{LD}L^T$　　(4)分解，这样做的好处是可以避免平方根的计算．

前者称为平方根法，后者称为改进的平方根法．两种方法中L的具体表达可通过(3)，(4)两式得到．

本次实验我们采用平方根法这种方法进行求解．

[3]Hilbert 矩阵在 2-范数意义下的条件数很坏，因此上述两种直接解法当n比较大时，残差的 2-范数会非常大．这里我们采取Tikhonov正则化预处理方法来改善矩阵的条件数．

此方法涉及矩阵的奇异值．我们知道，实对称矩阵$\mathrm{M} = \bar{A}^T A$的特征值称为矩阵A的奇异值．我们设$\{\boldsymbol{u}_i\}_{i=1}^n$是M的特征向量，并把$\boldsymbol{v_j} = \frac{A\boldsymbol{u}_j}{\mu_j}(1 \leq \mathrm{j} \leq \mathrm{r})(\mu_j \neq 0)$扩充成$\mathbb{C}^n$中的一组规范正交基$\{\boldsymbol{v}_i\}_{i=1}^n$．利用上面引入的记号，我们可以写出$A\boldsymbol{x} = \mathbf{b}$的解为：

$$\mathbf{x} = \sum_{j=1}^{r} \frac{1}{\mu_j} (\boldsymbol{b}, \boldsymbol{v}_j)\, \boldsymbol{u}_j . \qquad (5)$$

Tikhonov正则化预处理是将$\frac{1}{\mu_j}$乘上一个因子$\frac{\mu_j^2}{\alpha+\mu_j^2}$，其中α称为正则化参数．相当于解方程$(\alpha I + \bar{A}^T A)\boldsymbol{x}_\alpha = \bar{A}^T\mathbf{b}$，其解的表达为：

$$\boldsymbol{x}_\alpha = \sum_{j=1}^{r} \frac{\mu_j}{\alpha+\mu_j^2} (\boldsymbol{b}, \boldsymbol{v}_j)\, \boldsymbol{u}_j . \quad (6)$$

之所以进行正则化，是因为当矩阵A的奇异值远小于 1 时，由(5)式知便会带来较大的舍入误差．可以看到，当$\alpha \to 0$时，(6)将收敛于(5)，从而成为一个好的近似．可以预见，问题的核心在于α的选取.

i.为了使$\mathrm{cond(A)}_2 > \mathrm{cond}(\alpha I + \bar{A}^T A)_2$，需要

$$\alpha > \mu_1 \mu_n \quad (7).$$

一般$\mu_1 = \mathcal{O}(1)$，因此一般要求$\alpha > 10\mu_n$.

ii.通过先验误差估计，为使右端向量等价的扰动误差与预处理产生的近似误差相近，粗略地可选为

$$\alpha \sim \mathcal{O}(\mu_1^2 \delta) \quad (8)\text{-}1.$$

其中δ为扰动误差．经更精确的估计，若

$$\alpha \sim \mu_1 \delta^{\frac{2}{3}} \quad (8)\text{-}2$$

则误差比较理想．当δ取为计算机精度（10^{-14}），$\mu_1 = \mathcal{O}(1)$时，α可取在10^{-10}量级．

本次实验将首先通过奇异值分解与先验估计预先估计正则化参数的范围，然后经过预处理后，再用Gauss消去法和Cholesky分解方法求解，从而比较预处理前后误差的变化情况．

2.迭代方法及其预处理

相比于直接法，迭代法更适用于处理科学计算中经常遇到的大型带有一定结构的稀疏矩阵，因此应用更为广泛．

[1]简单迭代是指具有单步的、线性的、常倍乘矩阵的迭代格式的迭代法．SOR迭代格式是G − S迭代格式的推广版本．前者不仅如同后者充分利用每一次迭代过程中的最新结果，而且能够根据需要来调整使用前次迭代与最新迭代结果的比例，以求达到更加快速的收敛．具体迭代格式为：

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{B}_{SOR}^{\boldsymbol{\omega}} \boldsymbol{x}^{(k)} + \boldsymbol{f}_{SOR} \quad (9)$$

其中，$\boldsymbol{B}_{SOR}^{\boldsymbol{\omega}} = (D - \omega L)^{-\mathbf{1}}[(1-\omega)D + \omega U]$，$\boldsymbol{f}_{SOR} = (D - \omega L)^{-\mathbf{1}}\omega\mathbf{b}$.

关于收敛性，由于Hilbert 矩阵是对称正定矩阵，其SOR迭代格式当且仅当$\omega \in (0,2)$ 时收敛．因此可以预见，问题的核心在于ω的选取.

对于相容次序矩阵，已经有结论：

定理(Young)

若B_J只有实特征值，且$\Lambda = \rho(B_J) < 1$，则SOR 迭代对于$\omega \in (0,2)$收敛，且在

$$\omega = \frac{2}{1+\sqrt{1-\Lambda^2}} \qquad (10)$$

时达到最小值.

尽管Hilbert 矩阵不满足次序相容的条件，但由于其Jacobi矩阵为实对称故特征值全为实数，并且其Jocabi迭代收敛，因此作为尝试，可选取(10)式求出的ω作为初步试验值，并逐步作修正.

本次实验采用SOR 迭代进行求解，ω的取值方式如上段所述.

[2]在求解大规模系数矩阵时常常使用极小化残差的方法，即把线性方程组的求解问题转化为在特定函数空间中极小化残差的优化问题．共轭梯度法适用于对称正定阵，它是从整体来寻找最佳的搜索方向，其第一步与最速下降法相同，后面各步不但要求与负梯度向量共面，还要求与上一步的搜索向量共面，这样便大大提高了搜索的效率及有效性．课上已经证明，如果不考虑舍入误差，利用共轭梯度法至多n步便可得到精确解.

然而实际上舍入误差不可避免．如果系数矩阵的条件数很大，上述方法收敛速度还会很慢，因此考虑通过预处理在不改变矩阵的对称正定性的同时改善其条件数．取$S \in \mathbb{R}^{n\times n}$，将原方程组改写为

$$S^{-1}A(S^{-1})^T\boldsymbol{u} = S^{-1}\boldsymbol{b}，\text{其中 } \boldsymbol{u} = S^T\boldsymbol{x}.$$

这相当于对向量空间做了一个变换．可以预见，问题的核心在于S的选取.

置$M = SS^T$，课上已经证明，若选取$M^{-1}A \approx I$就可以大大降低条件数．又希望M对称正定且稀疏，故不妨写$A = M - N$，其中N的份额尽可能小．一般说来，预处理有简单迭代法与不完全Cholesky分解法两种方式，这里不作展开.

本次实验采用共轭梯度法与Jacobi迭代预处理的共轭梯度法进行比较.

三、方案设计

本次实验主要是考察各种计算方法的精确性与有效性，并认识预处理的重要作用．为了各个方法的整合与比较，我们利用Matlab 编写了包括Gauss消去法及其预处理(Gauss_H.m)、Cholesky分解方法及其预处理(Cholesky_H.m)、SOR 迭代方法(SOR_H.m)、共轭梯度法及其预处理(Conjgrad_H.m)在内的 4 个主要程序，以及用来比较预处理对共轭

梯度法结果误差之影响的独立程序(Conjgrad.m)．其中，条件数的计算、奇异值的计算以及ω的试验过程穿插在上述程序之中．另外，考虑到问题对实际的指导意义，我们只选取$n \geq 9$时的方程组作为算例．

四、计算结果及其分析

经过上述程序的运行，我们得到了如下面 3 个表格中列出的各算法的计算指标．

表格 1：直接法的结果

n	Gauss 消去法误差		Cholesky 分解法误差		较优 α 值
	预处理前	预处理后	预处理前	预处理后	
9	2.8981E-05	8.5694E-04	2.6037E-05	9.8425E-04	10^(-12)
10	5.4998E-04	1.2356E-03	6.6640E-04	1.3297E-03	10^(-12)
11	1.3100E-02	1.6300E-03	1.5449E-01	1.6241E-03	10^(-12)
12	5.1220E-01	2.3083E-03	6.4198E-01	2.1163E-03	10^(-12)
13	1.4053E+01	2.2117E-03	2.2726E+01	3.4855E-03	10^(-12)
14	1.2173E+01	2.1023E-03	NaN	2.1155E-03	10^(-12)
15	9.3946E+00	1.7701E-03	NaN	1.7189E-03	10^(-12)
16	1.1012E+02	1.7015E-03	NaN	1.8052E-03	10^(-12)
17	6.6641E+01	2.1090E-03	NaN	2.0895E-03	10^(-12)
18	2.3551E+01	2.3313E-03	NaN	2.5890E-03	10^(-12)
19	9.2681E+01	2.4199E-03	NaN	2.4798E-03	10^(-12)
20	1.0552E+02	2.3118E-03	NaN	2.3414E-03	10^(-12)

表格 2：奇异值与条件数指标变化

n	条件数		奇异值 μ_1	奇异值 μ_n
	预处理前	预处理后		
9	4.9315E+11	2.9787E+12	1.7259E+00	3.4997E-12
10	1.6025E+13	3.0693E+12	1.7519E+00	1.0932E-13
11	5.2215E+14	3.1502E+12	1.7748E+00	3.3992E-15
12	1.6515E+16	3.2234E+12	1.7954E+00	1.1005E-16
13	8.1661E+17	3.2900E+11	1.8138E+00	2.2212E-18
14	2.2272E+17	3.3512E+12	1.8306E+00	8.2194E-18
15	2.6388E+17	3.4075E+12	1.8459E+00	6.9953E-18
16	4.5445E+18	3.4598E+12	1.8600E+00	4.0929E-19
17	5.3394E+17	3.5085E+12	1.8731E+00	3.5510E-18
18	8.2095E+17	3.5542E+12	1.8852E+00	2.2964E-18
19	7.3816E+17	3.5970E+12	1.8965E+00	2.5693E-18
20	2.5077E+18	3.6375E+12	1.9071E+00	7.6067E-19

结合表格1、2我们可以看到，对于较高阶数的线性方程组，经预处理的直接法相比原来在减小舍入误差方面有了很大改善．其中有两点值得注意：

1）从表格 2 中注意到，随着方程组阶数 n 的增大，最小的奇异值大致徘徊在10^{-19}量级，因此此时矩阵 A 的性质是很坏的，以至于 Cholesky 分解法已经无法应用（错误提示为“矩阵不正定”）．经过预处理，不但 Cholesky 分解法得以应用，并且其最终误差也与 Gauss 消去法的相当．其中的原因可以从表 2 条件数的变化清晰地看到．

2）考察Tikhonov正则化使用的较优 α 参数值以及改善后的条件数，发现在求解阶数较高的线性方程组时这两个指标在数量级上基本不变．考虑到Hilbert 矩阵最大的奇异值稳定在$\mathcal{O}(1)$数量级，上述结果在一定程度上便证明了(8)-1、(8)-2 估计式的有效性．这同时表明，条件数的改善不依赖于具体方程组，而是由计算机的运算性能所决定．

需要说明的是，**通过更大阶数(直到 n=10000)的试验，我们发现即使阶数继续不断增大，考察预处理后的上述两种直接方法得到的结果，其误差 2-范数仍不超过 1.5，分量的误差几乎全部在 0.01 之内**．因此，预处理在更大阶数的范围内减小舍入误差的效果仍是非常显著的．

表格 3：迭代法的结果

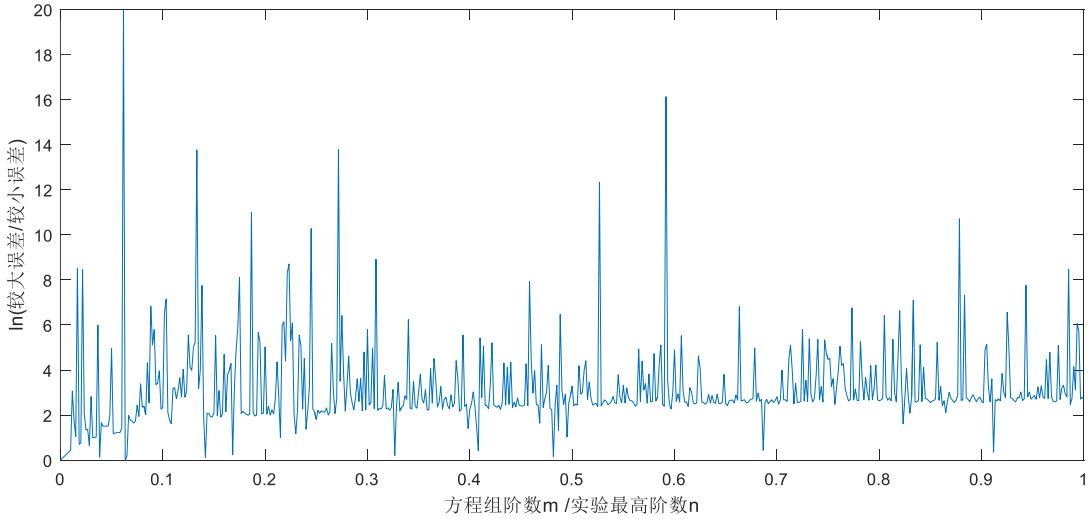
n	SOR		较优 ω 值（精确到 0.01）	共轭梯度法误差	
	迭代次数	结果误差		无预处理	有预处理
9	1.0208E+06	2.6221E-03	1.75	7.5330E-04	2.1337E-03
10	7.2792E+05	3.6415E-03	1.35	**9.8515E+01**	**1.9896E-02**
11	5.7430E+05	3.8668E-03	0.95	2.5921E-04	5.2563E-04
12	5.0373E+05	3.7813E-03	0.65	3.8300E-04	8.2531E-04
13	4.6200E+05	3.6233E-03	0.45	**1.1983E+03**	**2.5360E-01**
14	4.3769E+05	3.4762E-03	0.30	1.2289E-04	1.5630E-05
15	4.2299E+05	3.1809E-03	0.26	1.6758E-04	6.4328E-04
16	4.1502E+05	2.9460E-03	0.20	2.1998E-04	8.5874E-04
17	4.0956E+05	2.7130E-03	0.16	2.8013E-04	5.3648E-04
18	4.0665E+05	2.4890E-03	0.13	**1.6585E-03**	**9.8847E-05**
19	4.0035E+05	2.2854E-03	0.11	8.6539E-05	2.3351E-04
20	4.0449E+05	2.0021E-03	0.09	1.0744E-04	2.9444E-04

注：eps = 1.0e-8，M = 3.0e+6．

表格 3 左半部分表明的是在收敛精度 eps 与限制迭代次数 M 的要求下，所需的最低迭代次数(对应较优 ω 值)及其收敛到的极限所对应误差的 2-范数．从结果误差可以清楚地看出舍入误差对方程求解的影响，也可以看出较优 ω 值随 n 的变化趋势．需要说明的有两

点：一是上述较优 ω 值是针对结果误差的，在结果误差相差不大的时候，应着重降低迭代次数；二是在实验过程中还尝试了 SSOR 迭代方法，结果发现其迭代次数急速增加，效率便急剧降低，分析这是由于运算步骤翻倍造成的.

表格 3 右半部分反映了预处理前后共轭梯度法部分迭代结果的误差．从表中可以看出，预处理前后误差的变化并没有统一的规律，除标有黑体的数据对外，其他的均没有出现较大的相对偏差．**进一步考察更高阶数的线性方程组，可以得到偏差的变化规律．我们用二者比值(较大者/较小者)的对数表征其变化规律：**(图中 n=600)



从图中可以看到，随着方程组阶数 n 的增大，二者误差比的对数的“支点”似乎在一个数值附近．初步分析，这种规律似乎暗示了预处理的共轭梯度方法的某种数学特征，但由于数学基础限制这里不做深入讨论，权当一种启发.

五、结论

本次实验通过求解以条件数较差的Hilbert 矩阵为系数矩阵的线性方程组，比较了各种直接解法与迭代解法，尤其着重考察了奇异值、条件数两个指标和预处理方案对于结果误差的影响．从实验中可以看到，奇异值、条件数大小与误差大小正相关；预处理因为其很好地改善了矩阵的条件数，因此无论是对于直接法还是迭代法，尤其针对含有条件数较坏的系数矩阵的方程组，这种方案都是至关重要的．同时，迭代法尽管理论上会收敛于真解，但由于舍入误差影响，实际上会收敛但不可能收敛到真解——这从 SOR 迭代、共轭梯度法两个例子中可以察觉.

因此，在解线性方程组时，进行必要的预处理、选择与问题相容的迭代方法是两个至关重要的问题．这可能是今后在求解过程中需要着重注意的.