Iterative 8

王胤雅

25114020018

yinyawang25@m.fudan.edu.cn

2025 年 12 月 4 日

PROBLEM I Consider the global self–preconditioned MR iteration algorithm seen in Section 10.5.5. Define the acute angle between two matrices as

$$\cos\angle(X,Y) \equiv \frac{\langle X,Y\rangle}{\|X\|_F\,\|Y\|_F}.$$

1. Following what was done for the (standard) Minimal Residual algorithm seen in Chapter 5, establish that the matrices
$$B_k = AM_k, \qquad R_k = I - B_k$$
produced by global MR without dropping are such that
$$\|R_{k+1}\|_F \;\leq\; \|R_k\|_F \,\sin\angle(R_k, B_kR_k).$$

2. Let now $M_0 = \alpha A^T$ so that B_k is symmetric for all k (see Section 10.5.5). Assume that, at a given step k, the matrix B_k is positive definite. Show that
$$\cos\angle(R_k, B_kR_k) \;\geq\; \frac{\lambda_{\min}(B_k)}{\lambda_{\max}(B_k)},$$
in which $\lambda_{\min}(B_k)$ and $\lambda_{\max}(B_k)$ are, respectively, the smallest and largest eigenvalues of B_k.

SOLUTION. 1. 在 global MR 的第 k 步中，沿方向 B_kR_k 选取步长 α 更新残量：

$$R_{k+1} = R_k - \alpha B_kR_k. \tag{1}$$

最优步长 $\alpha^\star$ 定义为使 $\|R_{k+1}\|_F^2$ 最小的实数。写出范数平方并对 α 求导：

$$\begin{aligned}\|R_{k+1}\|_F^2 &= \|R_k - \alpha B_kR_k\|_F^2 \\ &= \|R_k\|_F^2 - 2\alpha\langle R_k,\, B_kR_k\rangle + \alpha^2\|B_kR_k\|_F^2.\end{aligned} \tag{2}$$

令 $\dfrac{d}{d\alpha}\|R_{k+1}\|_F^2 = 0$，得到

$$\alpha^\star = \frac{\langle R_k,\, B_kR_k\rangle}{\|B_kR_k\|_F^2}. \tag{3}$$

将 (3) 代入 (2)，得最小化以后的残量平方：

$$\begin{aligned}\|R_{k+1}\|_F^2 &= \|R_k\|_F^2 - \frac{\langle R_k,\, B_kR_k\rangle^2}{\|B_kR_k\|_F^2} \\ &= \|R_k\|_F^2\Big(1 - \Big(\frac{\langle R_k,\, B_kR_k\rangle}{\|R_k\|_F\,\|B_kR_k\|_F}\Big)^2\Big) \\ &= \|R_k\|_F^2\big(1-\cos^2\angle(R_k,\, B_kR_k)\big) \\ &= \|R_k\|_F^2\sin^2\angle(R_k,\, B_kR_k).\end{aligned} \tag{4}$$

注意括号内的分式正是锐角余弦，因此

$$\frac{\langle R_k,\, B_kR_k\rangle}{\|R_k\|_F\,\|B_kR_k\|_F} = \cos\angle(R_k,\, B_kR_k).$$

因此 (4) 可写成

$$\|R_{k+1}\|_F = \|R_k\|_F\,\sqrt{1-\cos^2\angle(R_k, B_kR_k)} = \|R_k\|_F\,\sin\angle(R_k, B_kR_k), \tag{5}$$

从而得到所需结论：

$$\|R_{k+1}\|_F \le \|R_k\|_F\,\sin\angle(R_k, B_kR_k).$$

2. 接下来证明对称正定情形下的角度估计。固定步 k，记 $B\equiv B_k$，假设 B 对称正定。对任意矩阵 X，有下列两条常用不等式：

$$\langle X, BX\rangle = \mathrm{tr}(X^TBX) \ge \lambda_{\min}(B)\,\|X\|_F^2 \tag{6}$$

$$\|BX\|_F \le \|B\|_2\,\|X\|_F = \lambda_{\max}(B)\,\|X\|_F. \tag{7}$$

式 (6) 来自于对称矩阵按正交矩阵对角化后对每一行向量分量分别乘以特征值求和的表示；(7) 则是矩阵谱范数的基本界。

令 $X=R_k$，把 (6) 与 (7) 代入余弦的定义：

$$\begin{aligned}\cos\angle(R_k,\, BR_k) &= \frac{\langle R_k,\, BR_k\rangle}{\|R_k\|_F\,\|BR_k\|_F} \\ &\ge \frac{\lambda_{\min}(B)\,\|R_k\|_F^2}{\|R_k\|_F\,(\lambda_{\max}(B)\,\|R_k\|_F)} = \frac{\lambda_{\min}(B)}{\lambda_{\max}(B)}.\end{aligned} \tag{8}$$

由此得到所需不等式：

$$\cos\angle(R_k,\, B_kR_k) \ge \frac{\lambda_{\min}(B_k)}{\lambda_{\max}(B_k)}. \tag{9}$$

□

PROBLEM II In the two-sided version of approximate inverse preconditioners, the option of minimizing

$$f(L,U) = \|I - LAU\|_F^2$$

was mentioned, where L is unit lower triangular and U is upper triangular.

a. What is the gradient of $f(L,U)$?

b. Formulate an algorithm based on minimizing this function globally.

SOLUTION. 固定矩阵 $A \in \mathbb{R}^{n\times n}$。定义误差矩阵

$$E := I - LAU.$$

记 Frobenius 范数为 $\|X\|_F^2 = \operatorname{tr}(X^TX)$。当写到矩阵微分时，使用迹运算的循环性与内积表示 $\langle X, Y\rangle = \operatorname{tr}(X^TY)$。

1. 目标函数

$$f(L,U) = \|E\|_F^2 = \operatorname{tr}(E^TE).$$

对 L 的变分:
令 $L \mapsto L + \mathrm{d}L$，有

$$\mathrm{d}E = -\,\mathrm{d}L\,AU.$$

因此

$$\mathrm{d}f = 2\operatorname{tr}(E^T\mathrm{d}E) = -2\operatorname{tr}\left(E^T(\mathrm{d}L\,AU)\right).$$

利用迹的循环性

$$\mathrm{d}f = -2\operatorname{tr}\left((AUE^T)\,\mathrm{d}L\right).$$

由内积识别 $\mathrm{d}f = \langle \nabla_L f, \mathrm{d}L\rangle$，因此无结构约束下的梯度为

$$\nabla_L f = -2\,(AUE^T)^T = -2\,EU^TA^T.$$

等价写法代回 $E = I - LAU$:

$$\nabla_L f = 2\,(LAU - I)\,U^TA^T$$

对 U 的变分:
令 $U \mapsto U + \mathrm{d}U$ 保持 L 固定，有

$$\mathrm{d}E = -\,LA\,\mathrm{d}U,$$

从而

$$\mathrm{d}f = 2\operatorname{tr}(E^T\mathrm{d}E) = -2\operatorname{tr}\left(E^T(LA\,\mathrm{d}U)\right).$$

再次利用迹的循环性:

$$\mathrm{d}f = -2\operatorname{tr}\left((A^TL^TE)\,\mathrm{d}U\right).$$

由此得到无结构约束下的梯度

$$\nabla_U f = -2\,A^TL^TE.$$

等价写法:

$$\nabla_U f = 2\,A^TL^T(LAU - I)$$

2. 在下面算法描述中，记投影算子为 $\mathcal{P}_{\mathrm{L}}(\cdot)$ 与 $\mathcal{P}_{\mathrm{U}}(\cdot)$。下面给出两类常用且实用的算法思路：交替最小化与 投影梯度法。二者均会保持 L、U 的结构约束，并可扩展到带稀疏模式限制的情形。

   方法一：交替最小化:
   思路：固定 U 求关于 L 的最佳解，固定 L 求关于 U 的最佳解，交替进行直到收敛。

   无结构时，将梯度设为零得到正规方程：

$$(LAU - I)\,U^T A^T = 0 \quad \Longrightarrow \quad L\,(AUU^T A^T) = U^T A^T.$$

   同理对 U：

$$A^T L^T L A\,U = A^T L^T.$$

   实际实现时，为保持三角结构，对上述“原始解”需要做结构投影并避免直接求逆以保证数值稳定——建议采用按列或按行逐列最小二乘并在每列/每行上只更新允许的自由分量。

   下面是伪代码：

Algorithm 1 交替最小化构造 L, U

1: 初始化 $L^{(0)} = I$, $U^{(0)} = I$ 或其它符合结构的初值
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: 固定 $U^{(k)}$，解关于 L 的正规方程

$$L\,M_{U^{(k)}} = B_{U^{(k)}}$$

其中 $M_U = AUU^T A^T$, $B_U = U^T A^T$
4: 将解做结构投影：$L^{(k+1)} = \mathcal{P}_{\mathrm{unit\_lower}}(L_{\mathrm{raw}})$
5: 固定 $L^{(k+1)}$，解关于 U 的正规方程

$$(A^T L^{(k+1)T} L^{(k+1)} A)\,U = A^T L^{(k+1)T}$$

6: 将解做结构投影：$U^{(k+1)} = \mathcal{P}_{\mathrm{upper}}(U_{\mathrm{raw}})$
7: 若 $\|I - L^{(k+1)} A U^{(k+1)}\|_F$ 小于容忍度则停止
8: **end for**

   实现要点:

   - 不要直接求大矩阵的逆，优先用分解或逐列最小二乘以保证稳定性。
   - 若希望保持稀疏模板，则在求解子问题时仅解对应自由未知量，其他位置固定为零。
   - 交替最小化每一步通常能显著降低目标，但可能收敛到局部极小点。

   方法二：投影梯度法:
   思路：利用前面得到的梯度，按投影梯度方向更新 L、U，并在每步后投影回结构域。

   伪代码如下：

Algorithm 2 投影梯度构造 L, U

初始化 $L^{(0)}$，$U^{(0)}$

for $k = 0, 1, 2, \ldots$ **do**

计算无结构梯度

$$G_L = \nabla_L f(L^{(k)}, U^{(k)}) = -2\, E^{(k)} U^{(k)T} A^T, \qquad G_U = \nabla_U f(L^{(k)}, U^{(k)}) = -2\, A^T L^{(k)T} E^{(k)},$$

其中 $E^{(k)} = I - L^{(k)} A U^{(k)}$

投影梯度到自由度：

$$\tilde{G}_L = \mathcal{P}_{\text{strict\_lower}}(G_L), \qquad \tilde{G}_U = \mathcal{P}_{\text{upper}}(G_U)$$

步长选择得到 α_k，β_k

更新并投影结构：

$$\tilde{L} = L^{(k)} - \alpha_k \tilde{G}_L, \qquad L^{(k+1)} = \mathcal{P}_{\text{unit\_lower}}(\tilde{L})$$

$$\tilde{U} = U^{(k)} - \beta_k \tilde{G}_U, \qquad U^{(k+1)} = \mathcal{P}_{\text{upper}}(\tilde{U})$$

若 $\|I - L^{(k+1)} A U^{(k+1)}\|_F$ 小于阈值则停止

end for

实现要点

- 步长可以采用简单的常数步长，也可以做单变量线搜索或最速下降以加速收敛。
- 投影操作非常简单：对 L 将上三角清零并把主对角置为 1；对 U 将下三角清零。
- 若需要稀疏模式限制，则在投影时同时把不在模式上的元素置零。

数值稳定性与实际注意事项:

- 在正规方程中出现对称正定矩阵时，优先使用 Cholesky 分解而非直接求逆。
- 对于大型稀疏问题，应采用逐列/逐行局部最小二乘来保持稀疏性和数值稳定性。
- 交替最小化通常收敛较快，但可能陷入局部最优；投影梯度法更简单、更灵活，便于加入正则化或稀疏约束。
- 若对目标函数有额外权重或只在给定模式上最小化，在梯度与正规方程中只保留对应自由位置项。

□

PROBLEM III With the standard splitting $A = D - E - F$, in which D is the diagonal of A and $-E, -F$ its lower- and upper-triangular parts, respectively, we associate the factored approximate inverse factorization

$$(I + ED^{-1})\, A\, (I + D^{-1}F) = D + R. \tag{10.85}$$

1. Determine R and show that it consists of second–order terms, i.e., terms involving products of at least two matrices from the pair E, F.

2. Now use the previous approximation for $D + R \equiv D_1 - E_1 - F_1$,

$$(I + E_1 D_1^{-1})(D + R)(I + D_1^{-1} F_1) = D_1 + R_1.$$

Show how the approximate inverse factorization (10.85) can be improved using this new approximation. What is the order of the resulting approximation?

SOLUTION. 设方阵 A 的标准分裂为

$$A = D - E - F,$$

其中 D 为 A 的对角矩阵，$-E$ 为 A 的严格下三角部分，$-F$ 为 A 的严格上三角部分。考察近似逆因式分解

$$(I + ED^{-1})\, A\, (I + D^{-1}F) = D + R.$$

1. 把左边逐步展开。首先用 $A = D - E - F$ 得

$$(I + ED^{-1})(D - E - F).$$

按项展开并利用 $ED^{-1}D = E$：

$$\begin{aligned}(I + ED^{-1})(D - E - F) &= (D - E - F) + ED^{-1}(D - E - F) \\ &= D - E - F + E - ED^{-1}E - ED^{-1}F \\ &= D - F - ED^{-1}E - ED^{-1}F.\end{aligned}$$

于是

$$(I + ED^{-1})A = D - F - ED^{-1}E - ED^{-1}F.$$

再右乘 $(I + D^{-1}F)$：

$$\begin{aligned}&(D - F - ED^{-1}E - ED^{-1}F)(I + D^{-1}F) \\ =\ &(D - F - ED^{-1}E - ED^{-1}F) \\ &+ (D - F - ED^{-1}E - ED^{-1}F)D^{-1}F.\end{aligned}$$

计算第二个括号的四项，逐项乘以 $D^{-1}F$：

$$DD^{-1}F = F, \qquad -FD^{-1}F, \qquad -(ED^{-1}E)D^{-1}F, \qquad -(ED^{-1}F)D^{-1}F.$$

把这些项代回并合并同类项，因此可写成

$$(I + ED^{-1})A(I + D^{-1}F) = D + R,$$

其中

$$\begin{aligned}R = &- ED^{-1}E \ - \ ED^{-1}F \ - \ FD^{-1}F \\ &- \ (ED^{-1}E)D^{-1}F \ - \ (ED^{-1}F)D^{-1}F.\end{aligned} \tag{10}$$

在表达式 (10) 中，每一项都至少包含两个来自 $\{E, F\}$ 的因子：

- $ED^{-1}E$ 包含两个 E;
- $ED^{-1}F$ 包含一个 E 和一个 F;
- $FD^{-1}F$ 包含两个 F;
- $(ED^{-1}E)D^{-1}F$ 包含至少三个，其中两个 E 和一个 F;
- $(ED^{-1}F)D^{-1}F$ 包含至少三个，其中一个 E 和两个 F。

因此 R 的项都是关于 E, F 的 **二阶或更高阶**项。通常在近似中我们把主要二阶项取出，把后两项视为更高阶小量。

2. 令

$$D + R \equiv D_1 - E_1 - F_1$$

为 $D + R$ 的标准分裂，D_1 为对角线，$-E_1$ 为严格下三角，$-F_1$ 为严格上三角。对其应用同样的构造:

$$(I + E_1 D_1^{-1})(D + R)(I + D_1^{-1} F_1) = D_1 + R_1.$$

按照第一部分推导的结构，R_1 由 E_1, F_1 的二阶及更高阶项构成；特别地

$$R_1 = -E_1 D_1^{-1} E_1 - E_1 D_1^{-1} F_1 - F_1 D_1^{-1} F_1 + \text{更高阶项}.$$

为给出更精确的"阶"概念，我们使用某个矩阵范数 $\|\cdot\|$，例如算子范数或 Frobenius 范数；所有范数在有限维下等价，故结论与范数选择无关。假设 D 是可逆的且 D_1 与 D 等价，即对角元无病态变化，存在常数 C_D 使得 $\|D^{-1}\|, \|D_1^{-1}\| \le C_D$。

记 $S := E + F$，则第一轮展开中主要二阶项满足形式上

$$R = \mathcal{O}(S^2),$$

即存在常数 C，依赖于 $\|D^{-1}\|$，使得当 $\|S\|$ 小时

$$\|R\| \le C\|S\|^2.$$

这是因为每一项例如 $\|ED^{-1}F\| \le \|E\|\,\|D^{-1}\|\,\|F\| \le \|D^{-1}\|\|E\|\,\|F\|$，类似地对其它项。

由定义 E_1, F_1 是 $D + R$ 的非对角部分，因此其量级与 R 同阶:

$$\|E_1\| + \|F_1\| = \mathcal{O}(\|R\|) = \mathcal{O}(\|S\|^2).$$

于是第二轮残差 R_1，它至少是 E_1, F_1 的二阶满足

$$\|R_1\| \le C'\big(\|E_1\| + \|F_1\|\big)^2 = \mathcal{O}\big(\|S\|^4\big),$$

其中常数 C' 依赖于 $\|D_1^{-1}\|$ 等但与 S 无关。换言之，第二轮的残差在范数意义下为原始非对角量 S 的 **四阶**。

因此通过对 $D + R$ 再做一次近似因式化，残差从 $\mathcal{O}(S^2)$ 提升到 $\mathcal{O}(S^4)$。

一些说明:

- 在实际数值算法里常把 $(I + ED^{-1})$ 与 $(I + D^{-1}F)$ 视为稀疏的近似因子，例如只保持某些带宽或模式，这样得到的 R 既小又稀疏结构可控，以用于构造预条件子。

- 以上阶数估计是基于小量假设，如 $\|E\| + \|F\|$ 足够小，或 D 主导矩阵，若 E, F 不小则不能用此阶次语言描述收敛性。
- 如果想要更精细的展开，可以把式 $(I + ED^{-1})A(I + D^{-1}F)$ 按幂级数或 Neumann 展开系统地截断，以得到更高阶的显式项。

□