CME 307 / MS&E 311 Winter 2022

Optimization Jan 14, 2022

Prof. Yinyu Ye Optional - not graded

Homework Assignment 0 Sample Solutions

This is a diagnostic homework that covers prerequisite materials that you should be familiar with. This homework will not be graded and will not be counted towards the final grade.

Solve the following problems:

1. Consider the iterative process

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right),$$

where $a > 0$. Assuming the process converges, to what does it converge?

Solution: Taking the limit, we have

$$x^* = \frac{1}{2}(x^* + \frac{a}{x^*})$$

Solve this equation, we have $x^* = \pm\sqrt{a}$. It's obvious that the iterations don't change the signs of x_k, so we have 1) if $x_0 > 0$, then $x_k \to \sqrt{a}$; 2) if $x_0 < 0$, then $x_k \to -\sqrt{a}$. $\square$

2. Let $\{(\boldsymbol{a}_i, c_i)\}_{i=1}^m$ be a given dataset where $\boldsymbol{a}_i \in R^n$, $c_i \in \{\pm 1\}$.

(a) Compute the gradient of the following log-logistic-loss function,

$$f(\boldsymbol{x}, x_0) = \sum_{i:c_i=1} \log\left(1 + \exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)\right) + \sum_{i:c_i=-1} \log\left(1 + \exp(\boldsymbol{a}_i^T\boldsymbol{x} + x_0)\right),$$

where $\boldsymbol{x} \in R^n$ and $x_0 \in R$.

(b) Consider the following data set

$$\boldsymbol{a}_1 = (0; 0), \quad \boldsymbol{a}_2 = (1; 0), \quad \boldsymbol{a}_3 = (0; 1), \quad \boldsymbol{a}_4 = (0; 0), \quad \boldsymbol{a}_5 = (-1; 0), \quad \boldsymbol{a}_6 = (0; -1),$$

with label

$$c_1 = c_2 = c_3 = 1, \quad c_4 = c_5 = c_6 = -1,$$

show that there is no solution for $\nabla f(\boldsymbol{x}, x_0) = \boldsymbol{0}$.

Solution:

(a) (Here we treat the gradient vector as a row vector.) For $c_i = 1$,

$$\nabla \log\left(1+\exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)\right) = \frac{\exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)}{1+\exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)}(-\boldsymbol{a}_i^T,\ -1);$$

and for $c_i = -1$,

$$\nabla \log\left(1+\exp(\boldsymbol{a}_i^T\boldsymbol{x} + x_0)\right) = \frac{\exp(\boldsymbol{a}_i^T\boldsymbol{x} + x_0)}{1+\exp(\boldsymbol{a}_i^T\boldsymbol{x} + x_0)}(\boldsymbol{a}_i^T,\ 1).$$

Thus, the gradient vector $\nabla f(\boldsymbol{x}, x_0)$ is

$$\sum_{i,c_i=1} \frac{\exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)}{1+\exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)}(-\boldsymbol{a}_i^T,\ -1) + \sum_{i,c_i=-1} \frac{\exp(\boldsymbol{a}_i^T\boldsymbol{x} + x_0)}{1+\exp(\boldsymbol{a}_i^T\boldsymbol{x} + x_0)}(\boldsymbol{a}_i^T,\ 1)$$

(b) We show by contradiction that a (finite) solution does not exist. Firstly, notice that the objective is non-negative, and hence 0 is a lower bound. Then, looking at the problem data, we see that by choosing $x = (t,t)^T$ and $x_0 = 0$, taking $t \to \infty$ leads to $f(x; x_0) \to 0$. Hence 0 is the infimum of the objective function. Nevertheless, for any finite x and x_0, obviously the objective is strictly positive. Hence we conclude that the problem has no (finite) solution. □

You can also prove it by straightforward calculation: substitute the problem data and make $\nabla f(\boldsymbol{x}, x_0) = \boldsymbol{0}$. The key observation is $\boldsymbol{a}_1 = -\boldsymbol{a}_4$, $\boldsymbol{a}_2 = -\boldsymbol{a}_5$, $\boldsymbol{a}_3 = -\boldsymbol{a}_6$ so that $\nabla f(\boldsymbol{x}, x_0) = \boldsymbol{0}$ is equivalent to

$$\sum_{i\in[3]} \frac{\exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)}{1+\exp(-\boldsymbol{a}_i^T\boldsymbol{x} - x_0)}(-\boldsymbol{a}_i^T,\ -1) = \sum_{i\in[3]} \frac{\exp(-\boldsymbol{a}_i^T\boldsymbol{x} + x_0)}{1+\exp(-\boldsymbol{a}_i^T\boldsymbol{x} + x_0)}(\boldsymbol{a}_i^T,\ -1).$$

By checking the last coordinate, we must have $x_0 = 0$. Given $x_0 = 0$, the first two coordinates of the above equations is equivalent to

$$\boldsymbol{0} = 2\sum_{i\in[3]} \frac{\exp(-\boldsymbol{a}_i^T\boldsymbol{x})}{1+\exp(-\boldsymbol{a}_i^T\boldsymbol{x})}\boldsymbol{a}_i,$$

which is never true as the RHS is strictly positive. □

3. Given a symmetric matrix $A \in R^{n\times n}$ s.t. A has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, show that for every $k = 1, 2, \cdots, n$, we have:

$$\lambda_k = \max_U \left\{ \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \middle| \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0} \right\} \middle| U \text{ is a linear subspace of } R^n \text{ of dimension } k \right\} \quad (1)$$

$$= \min_U \left\{ \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \middle| \boldsymbol{x} \in U, \boldsymbol{x} \neq \boldsymbol{0} \right\} \middle| U \text{ is a linear subspace of } R^n \text{ of dimension } n-k+1 \right\} \quad (2)$$

Solution: This result is known as the *Courant-Fischer Minimax Theorem.* See Theorem 8.1.2 of [GVL13] for a sample proof.

Here we sketch the proof for (1). Let $\{v_k\}_{k=1}^n$ denote a set of orthonormal eigenbasis of A, with $Av_k = \lambda_k v_k$. Moreover, $A = \sum_{k=1}^n \lambda_k v_k v_k^T$.

When $k = 1$, the expression reduces to $\lambda_1 = \max_{x \neq 0} \frac{x^T A x}{x^T x}$, which is true for symmetric matrices, with one maximizer U^1 being spanned by v_1.

Now suppose for the sake of induction that we have shown (1) for some k and that the maximizer U^k can be taken to be the span of the first k eigenvectors, and we need to show it holds for $k+1$.

We show that a maximizer for λ_{k+1} is $U^{k+1} := U^k \cup span(v_{k+1})$. To see this, note that

$$\lambda_{k+1} = \min_{x \in U^{k+1}} \frac{x^T A x}{x^T x}$$

so that $\lambda_{k+1} \leq RHS$ of (1).

On the other hand, for any subspace U of dimension $k+1$ that is not spanned by the first $k+1$ eigenvectors of A, minimization in RHS will choose an eigenvector corresponding to an eigenvalue smaller than λ_{k+1}.

4. Given symmetric matrices $A, B, C \in R^{n \times n}$ s.t. A has eigenvalues $a_1 \geq a_2 \geq \cdots \geq a_n$, B has eigenvalues $b_1 \geq b_2 \geq \cdots \geq b_n$ and C has eigenvalues $c_1 \geq c_2 \geq \cdots \geq c_n$, if $A = B + C$, show that for every $k = 1, 2, \cdots, n$, we have:

$$b_k + c_n \leq a_k \leq b_k + c_1. \tag{3}$$

Solution: We show that $a_k \leq b_k + c_1$. The other inequality is similar. According to (1), define U_k to be the dim-k linear subspace such that

$$a_k = \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\} \tag{4}$$

and let $\boldsymbol{x}^*$ be the minimizer of $\min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\}$. It follows that

$$a_k = \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T (B+C) \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\} \leq \frac{\boldsymbol{x}^{*T} (B+C) \boldsymbol{x}^*}{\boldsymbol{x}^{*T} \boldsymbol{x}^*} \tag{5}$$

$$\leq \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\} + \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T C \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in R^n, \boldsymbol{x} \neq \mathbf{0} \right\} \tag{6}$$

$$\leq \max_{U} \left\{ \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U, \boldsymbol{x} \neq \mathbf{0} \right\} |\dim(U) = k \right\} + \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T C \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in R^n, \boldsymbol{x} \neq \mathbf{0} \right\} \tag{7}$$

$$= b_k + c_1, \tag{8}$$

completing the proof.

Similarly, According to (2), define U_k to be the $(n-k+1)$-dimensional linear subspace such that

$$a_k = \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\} \tag{9}$$

and let $\boldsymbol{x}^*$ be the maximizer of $\max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\}$. It follows that

$$a_k = \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T (B + C) \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\} \geq \frac{\boldsymbol{x}^{*T} (B + C) \boldsymbol{x}^*}{\boldsymbol{x}^{*T} \boldsymbol{x}^*} \tag{10}$$

$$\geq \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U_k, \boldsymbol{x} \neq \mathbf{0} \right\} + \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T C \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in R^n, \boldsymbol{x} \neq \mathbf{0} \right\} \tag{11}$$

$$\geq \min_{U} \left\{ \max_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in U, \boldsymbol{x} \neq \mathbf{0} \right\} | \dim(U) = n - k + 1 \right\} + \min_{\boldsymbol{x}} \left\{ \frac{\boldsymbol{x}^T C \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} | \boldsymbol{x} \in R^n, \boldsymbol{x} \neq \mathbf{0} \right\} \tag{12}$$

$$= b_k + c_n. \tag{13}$$

□

5. Let $A \in R^{n \times n}$ be a positive-semidefinite matrix with Schur decomposition $A = Q \Lambda Q^T$, where $Q = [\boldsymbol{q}_1 | \cdots | \boldsymbol{q}_n]$ is an orthogonal matrix, $\Lambda = \mathbf{diag}\{\lambda_1, \ldots, \lambda_n\}$ satisfies $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \geq 0$. Show that for any $k = 1, \ldots, n$,

$$\min_{\mathbf{rank}(B) = k} \|A - B\|_2 = \|A - A_k\|_2 = \lambda_{k+1}, \tag{14}$$

and

$$\min_{\mathbf{rank}(B) = k} \|A - B\|_F = \|A - A_k\|_F = \sqrt{\sum_{j=k+1}^{n} \lambda_j^2}, \tag{15}$$

where A_k is defined as

$$A_k := \sum_{j=1}^{k} \lambda_j \boldsymbol{q}_j \boldsymbol{q}_j^T. \tag{16}$$

Here $\| \cdot \|_2$ stands for the spectrum (L_2) norm and $\| \cdot \|_F$ stands for the Frobenius norm.

Solution: This result is a special case of the *Eckhart-Young Theorem.* See Theorem 2.4.8 of [GVL13] for a sample proof for the general case. This is also well-known as the best low-rank matrix approximation and the results generalize to general matrices (using SVD decomposition).

We give a sketch of the special case here. We first show (14). Let B be any rank k matrix. By rank-nullity theorem we can find orthonormal vectors $x_1, \ldots, x_{n-k}$ that span the null space of B. In dimension n, the null space of B which is $n - k$ dimensional, and the span of $\{\boldsymbol{q}_i\}_{i=1}^{k+1}$, which is $k + 1$ dimensional, have non-empty intersection.

Let $\boldsymbol{z}$ be a unit norm vector in this intersection. We then have

$$\begin{aligned} \|A - B\|_2^2 &\geq \|(A - B)\boldsymbol{z}\|_2^2 = \|A\boldsymbol{z}\|_2^2 \\ &= \sum_{i=1}^{k+1} \lambda_i^2 (\boldsymbol{q}_i^T \boldsymbol{z})^2 \geq \lambda_{k+1}^2 \end{aligned}$$

where in the last inequality we have used that $\sum_{i=1}^{k+1}(\boldsymbol{q}_i^T\boldsymbol{z})^2 = \|\boldsymbol{z}\|^2 = 1$, since $\boldsymbol{z}$ is in the span of $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{k+1}$.

For (15), we use the identity that

$$\begin{aligned}
\|C\|_F^2 &= Tr(C^T C) \\
&= Tr(C^T C \sum_{j=1}^{n} \boldsymbol{v}_j \boldsymbol{v}_j^T) \\
&= \sum_{j=1}^{n} (\boldsymbol{v}_j^T C^T C \boldsymbol{v}_j) = \sum_{j=1}^{n} \|C\boldsymbol{v}_j\|^2
\end{aligned}$$

for any orthonormal basis $\{\boldsymbol{v}_j\}_{j=1}^n$ and write

$$\begin{aligned}
\|A - B\|_F^2 &= \sum_{j=1}^{n} \|(A - B)\boldsymbol{x}_j\|^2 \\
&= \sum_{j=1}^{n-k} \|A\boldsymbol{x}_j\|^2 + \sum_{j=n-k+1}^{n} \|(A - B)\boldsymbol{x}_j\|^2 \\
&\geq \sum_{j=1}^{n-k} \|A\boldsymbol{x}_j\|^2
\end{aligned}$$

where again we assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-k}$ span the null space of B.

Finally, $\sum_{j=1}^{n-k} \|A\boldsymbol{x}_j\|^2 \geq \sum_{j=k+1}^{n} \|A\boldsymbol{q}_j\|^2 = \sum_{j=k+1}^{n} \lambda_j^2$. This identity says that projections onto any $n-k$ dimensional subspace (LHS) is bounded below by the projection onto the $n-k$ dimensional subspace spanned by $\{\boldsymbol{q}_j\}_{j=k+1}^n$ (RHS). Equivalently, $\{\boldsymbol{q}_j\}_{j=1}^k$ span the best fit k-dimensional subspace for A, in the sense that

$$\sum_{j=1}^{k} \|A\boldsymbol{x}_j\|^2 \leq \sum_{j=1}^{k} \|A\boldsymbol{q}_j\|^2$$

for any orthonormal system $\{\boldsymbol{x}_j\}_{j=1}^k$.

To prove $\sum_{j=1}^{k} \|A\boldsymbol{x}_j\|^2 \leq \sum_{j=1}^{k} \|A\boldsymbol{q}_j\|^2$, we use the important fact that

$$\boldsymbol{q}_j \in \arg \max_{v \perp span(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{j-1})} \|A\boldsymbol{v}\|^2,$$

that is the j-th unit eigenvector of A maximizes $\|A\boldsymbol{v}\|^2$ among all unit vectors that are not in the span of the first $j-1$ eigenvectors. Clearly the inequality holds for $k = 1$. Suppose for the sake of induction we have shown it for some k. Let $\{\boldsymbol{y}_j\}_{j=1}^{k+1}$ be a solution to

$$\max_{\text{orthonormal } \{\boldsymbol{x}_j\}} \sum_{j=1}^{k+1} \|A\boldsymbol{x}_j\|^2$$

Without loss of generality we can let $\boldsymbol{y}_{k+1}$ be orthogonal to the span of $\{\boldsymbol{q}_j\}_{j=1}^k$. Then $\|A\boldsymbol{y}_{k+1}\|^2 \leq \|A\boldsymbol{q}_{k+1}\|^2$, so that

$$\sum_{j=1}^{k+1} \|A\boldsymbol{y}_j\|^2 \leq \sum_{j=1}^{k+1} \|A\boldsymbol{q}_j\|^2$$

completing the induction step.

References

[GVL13] Gene H Golub and Charles F Van Loan. Matrix computations. edition, 2013.