Mathematics of Machine Learning
Homework 3 - Solutions

Instructors: Afonso S. Bandeira & Nikita Zhivotovskiy
Course Coordinator: Pedro Abdalla Teixeira

March 19, 2021

Try to solve the questions before looking to the answers. Every item must be proved rigorously. Starred problems are harder.

Problem 1

Consider a matrix $A \in \mathbb{R}^{n\times m}$ with singular values $\sigma_1(A) \geq \sigma_2(A) \ldots \geq \sigma_{\min\{m,n\}}(A)$. Prove the following results

(a) Variational characterization for the sum of singular values:

$$\sum_{i=1}^{\min\{m,n\}} \sigma_i(A) = \sup_{\sigma_1(Q)\leq 1} \langle Q, A\rangle_F.$$

Here $\langle , \rangle$ denotes the Frobenius inner product.

(b) Schatten p-norm is indeed a norm for $p = 1, 2, \infty$.

(c) A norm $\|.\|$ is said unitarily invariant if $\|UAV\| = \|A\|$ for all $U \in \mathcal{O}(m)$ and $V \in \mathcal{O}(n)$. Are Schatten p-norm unitarily invariant for $p = 1, 2, \infty$?

Solution 1

(a) Consider the SVD decomposition of $A = U\Sigma V^T$ and choose $Q = UIV^T$ where I stands for a diagonal matrix with all diagonal elements equal to one. It is clear that $Q = UV^T$, therefore

$$\langle Q, A\rangle_F = \operatorname{Tr}(Q^T A) = \operatorname{Tr}(VU^T U\Sigma V^T) = \operatorname{Tr}\Sigma = \sum_{i=1}^{\min\{m,n\}} \sigma_i(A).$$

On the other hand,

$$\begin{aligned}
\sup_{\sigma_1(Q)\le 1} \operatorname{Tr}(Q^T A) &= \sup_{\sigma_1(Q)\le 1} \operatorname{Tr}(Q^T U\Sigma V^T) \\
&= \sup_{\sigma_1(Q)\le 1} \operatorname{Tr}(V^T Q^T U\Sigma) \\
&= \sup_{\sigma_1(Q)\le 1} \operatorname{Tr}((UQV)^T\Sigma) \\
&= \sup_{\sigma_1(Q)\le 1} \sum_i \sigma_i u_i^T Q v_i \\
&\le \sup_{\sigma_1(Q)\le 1} \sum_i \sigma_i \sigma_1(Q) \\
&\le \sum_{i=1}^{\min\{m,n\}} \sigma_i
\end{aligned}$$

This concludes the proof.

(b) It is straightforward to check positive homogeneity for all three norms. Recall that $\sigma_1(A) = \sup_{x\in\mathbb{S}^{m-1}} \|Ax\|_2$, so if $\sigma_1(A) = 0$ then A is all zeros because $\|Ax\|_2 = 0$ implies $Ax = 0$ for all $x \in \mathbb{R}^m$. It remains to prove the triangle inequality, we separate one proof for each norm. For the Schatten ∞-norm,

$$\begin{aligned}
\sigma_1(A+B) = \sup_{x\in\mathbb{S}^{m-1}} \|(A+B)x\|_2 &\le \sup_{x\in\mathbb{S}^{m-1}} \|Ax\|_2 + \sup_{x\in\mathbb{S}^{m-1}} \|Bx\|_2. \\
&= \sigma_1(A) + \sigma_1(B).
\end{aligned}$$

For the Schatten 1-norm the same trick applies, we only change the variational principle

$$\begin{aligned}\sum_{i=1}^{\min\{m,n\}} \sigma_i(A+B) &= \sup_{\sigma_1(Q)\le 1} \langle Q, A+B\rangle_F \\ &\le \sup_{\sigma_1(Q)\le 1} \langle Q, A\rangle_F + \sup_{\sigma_1(Q)\le 1} \langle Q, B\rangle_F \\ &= \sum_{i=1}^{\min\{m,n\}} \sigma_i(A) + \sigma_i(B).\end{aligned}$$

We claim that the Schatten 2-norm is equal to the Frobenius norm. If the claim is true, then the triangular inequality follows easily from

$$\begin{aligned}\|A+B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\langle A, B\rangle &\le \|A\|_F + \|B\|_F + 2(\|A\|_F\|B\|_F)^2 \\ &= (\|A\|_F + \|B\|_F)^2.\end{aligned}$$

To prove the claim observe that $\|A\|_F = \sqrt{\mathrm{Tr}(A^T A)}$ and the equality follows Proposition 3.1 in the lecture notes.

(c) Observe that the singular values of UAV and A are exactly the same because the product of two orthogonal matrices still an orthogonal matrix. Since all three Schatten norms depends only on the singular values, then all three norms are unitarily invariant.

Problem 2

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with positive eigenvalues $\lambda_1 > \lambda_2 \geq \dots \lambda_n > 0$. Consider the following iteration for a given initial vector θ^0,

$$\theta^t := \frac{A\theta^{t-1}}{\|A\theta^{t-1}\|_2}.$$

(a) Prove that the iteration converges to a normalized eigenvector v_1 associated with the largest eigenvalue λ_1 if the initial vector is not orthogonal to v_1.

(b) Give examples in which the iteration above do not converge to v_1.

Solution 2

(a) Let $v_1, \dots, v_n$ be the standard orthonormal basis of normalized eigenvectors of A and write vector as $\theta^0 = \sum_{i=1}^n c_i v_i$. Without loss of generality we may assume that θ^0 is a unit vector. It is easy to see by induction that,

$$\theta^t = \frac{\sum_{i=1}^n c_i \lambda_i^t v_i}{\sqrt{\sum_{i=1}^n c_i^2 \lambda_i^{2t}}} = \frac{\lambda_1^t \sum_{i=1}^n (\frac{\lambda_i}{\lambda_1})^t c_i v_i}{\lambda_1^t \sqrt{\sum_{i=1}^n c_i \frac{\lambda_i^{2t}}{\lambda_1^{2t}}}} = \frac{\sum_{i=1}^n (\frac{\lambda_i}{\lambda_1})^t c_i v_i}{\sqrt{\sum_{i=1}^n c_i \frac{\lambda_i^{2t}}{\lambda_1^{2t}}}}.$$

Since, for every $i \geq 2$ the ratio $\frac{\lambda_i}{\lambda_1}$ is strictly less than one, the numerator converges to $c_1 v_1$ as t goes to infinity and the denominator converges to $|c_1|$. Observe here we need the assumption on $c_1 = \langle \theta^0, v_1 \rangle$ is non-zero. We conclude that θ^t converges to v_1 or $-v_1$.

(b) Observe that if the initial vector θ^0 is equal to v_2, then $\theta^1 = \theta^0$ because v_2 is an eigenvector. Therefore, for all $t > 0$, $\theta^t = v_2$.

The algorithm does not converge to v_1. Analogously, if we take the initial vector θ^0 to be any eigenvector linearly independent from v_1, the algorithm fails to converge to v_1.

Problem 3

* Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix and let B be a principal submatrix of A of dimension $n - 1$ (B is obtained by deleting the same row and column of A). Prove the following facts:

(a) If $\alpha_1 \geq \ldots \geq \alpha_n$ are the eigenvalues of A and $\beta_1 \geq \ldots \beta_n$ are the eigenvalues of B, then

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \ldots \geq \alpha_n.$$

Hint: Use the Courant-Fischer principle for symmetric matrices.

(b) Let $G = (V, E)$ be a graph. If $G(S)$ is the graph induced by a subset $S \subset V$, then the average degree of $G(S)$ is at most the largest eigenvalue of the adjacency matrix A of G.

Hint: Find an upper bound for the average degree of $G(S)$ in terms of a submatrix of A.

Solution 3

(a) Without loss of generality, assume that B is obtained by deleting the first row and the first column of A. By the Courant-Fischer principle to A we can write

$$\alpha_k = \sup_{S \subset \mathbb{R}^n; \dim(S) = k} \inf_{x \in S} \frac{x^T A x}{x^T x}.$$

Now we apply for B,

$$\beta_k = \sup_{S \subset \mathbb{R}^{n-1}; \dim(S)=k} \inf_{y \in S} \frac{y^T B y}{y^T y} = \sup_{S \subset \mathbb{R}^{n-1}; \dim(S)=k} \inf_{y \in S} \frac{[0\ x] B [0\ x]^T}{[0\ x][0\ x]^T}.$$

Here $[0\ x]$ stands for a vector that has the first entry equal to 0 and the others entries are the same as the vector x. Observe that the supremum for β_k is taken over a special family of subspaces (the ones with first coordinate equal to zero) and therefore it is at most the supremum of the same quantity over a more general family of subspaces. It implies that $\beta_k \leq \alpha_k$. We repeat the argument for $-A$ and $-B$ to conclude that $\alpha_{k+1} \leq \beta_k$.

(b) Let B be a matrix adjacency matrix induced by S, obtained by deleting all rows and columns corresponding to vertices not in S and let β_1 be its largest eigenvalue. The average degree d_a can be bounded via Courant-Fischer principle as follows

$$\beta_1 = \sup_x \frac{x^T B x}{x^T x} \geq \frac{1^T B 1}{1^T 1} = \frac{\sum_{i,j} B(i.j)}{n} = d_a$$

Here 1^T denotes a vector with all entries equal to one. By letter "a", for every principal submatrix obtained the largest eigenvalue cannot increase, so β_1 is at most the largest eigenvalue of A.