15-851 Algorithms for Big Data — Spring 2025

Problem Set 2 Solutions

Problem 1: Ridge Leverage Scores Bound Low Rank Sensitivities

1. We will denote $A' = [A; \sqrt{\lambda} I]$. Let us consider the i^{th} leverage score of A' for $i \in [n]$. By the definition in the hint, we have that this is

$$a_i'^{\intercal}(A'^{\intercal}A')^{-1}a_i' = a_i^{\intercal}(A'^{\intercal}A)^{-1}a_i.$$

   So, we just need to prove that $A'^{\intercal}A' = A^{\intercal}A + \lambda I$.

   We have that

$$A'^{\intercal}A' = [A; \sqrt{\lambda} I]^{\intercal}[A; \sqrt{\lambda} I] = A^{\intercal}A + (\sqrt{\lambda} I)^2 = A^{\intercal}A + \lambda I.$$

2. Using the previous part, we know that $\tau_i = \ell_i(A')$ where ℓ_i is the i^{th} leverage score. Using the hint, we can see that

$$\ell_i(A') = \sup_x \frac{(A'x)_i^2}{\|A'x\|_2^2} = \sup_x \frac{(Ax)_i^2}{\|A'x\|_2^2} = \sup_x \frac{(Ax)_i^2}{\|Ax\|_2^2 + \lambda\|x\|_2^2}.$$

3. Recall that the operator norm of a matrix is also the largest singular value of that matrix. So we have

$$\|A - A_{2k}\|_2^2 = \sigma_{2k+1}^2(A) = \sigma_{k+1}^2(A - A_k) \le \frac{1}{k}\sum_{j=1}^{k}\sigma_j^2(A - A_k) \le \frac{\|A - A_k\|_F^2}{k} \le \lambda.$$

4. (a) We follow the steps in the hint. So, we know from part 2 that we have

$$\tau_i = \sup_x \frac{(Ax)_i^2}{\|Ax\|_2^2 + \lambda\|x\|_2^2}.$$

   So, let us expand this. We have

$$\sup_x \frac{(Ax)_i^2}{\|Ax\|_2^2 + \lambda\|x\|_2^2} = \sup_x \frac{(Ax)_i^2}{\|A_{2k}x\|_2^2 + \|(A - A_{2k})x\|_2^2 + \lambda\|x\|_2^2}.$$

   Using part 3, we now get that

$$\sup_x \frac{(Ax)_i^2}{\|A_{2k}x\|_2^2 + \|(A - A_{2k})x\|_2^2 + \lambda\|x\|_2^2} \ge \sup_x \frac{(Ax)_i^2}{\|A_{2k}x\|_2^2 + 2\lambda\|x\|_2^2}.$$

   (b) Now, we let $F \in \mathcal{F}_k$ be any rank k subspace. We set H to be the span of the rows of A_{2k}, F, and a_i. Clearly H is at most a $(3k+1)$ dimensional subspace.

   Let $x = P_H(I - P_F)g$ where g is a standard normal Gaussian vector and P_H denotes the projection onto H.

   So, we have that $(Ax)_i = a_i^{\intercal}P_H(I - P_F)g = a_i^{\intercal}(I - P_F)g$. So, $(Ax)_i$ is distributed as a gaussian with variance $\|a_i^{\intercal}(I - P_F)\|_2^2$. This is a chi-squared random variable with one degree of freedom with expectation $\|a_i^{\intercal}(I - P_F)\|_2^2$. Using standard properties of the pdf of a chi-squared random variable we have

$$\mathbf{Pr}[(Ax)_i^2 \ge \|a_i^{\intercal}(I - P_F)\|_2^2/3] > 1/2.$$

   (c) We have that

$$E[\|A_{2k}x\|_2^2] = E[\|A_{2k}P_H(I - P_F)g\|_2^2] = E[\|A_{2k}(I - P_F)g\|_2^2] \le \|A(I - P_F)\|_F^2.$$

We also have

$$E[\lambda\|x\|_2^2] = E[\lambda\|P_H(I-P_F)g\|_2^2]$$
$$= E[\lambda\|HH^\intercal(I-P_F)g\|_2^2] = E[\lambda\|H^\intercal(I-P_F)g\|_2^2].$$

$H^\intercal$ has orthonormal rows, and projections to not increase norms. Therefore, each row of $H^\intercal(I - P_F)$ has length at most 1, and each row of $H^\intercal(I-P_F)g$ is a gaussian with variance at most 1. So we have

$$E[\lambda\|H^\intercal(I-P_F)g\|_2^2 \leq \lambda(3k+1) \leq 4\|A-A_k\|_F^2 \leq 4\|A(I-P_F)\|_F^2.$$

So using Markov's bound we have

$$\mathbf{Pr}[\|A_{2k}x\|_2^2 + 2\lambda\|x\|_2^2 < 20\|A(I-P_F)\|_F^2] > 1/2.$$

(d) We have with positive probability that there exists an x such that

$$\tau_i \geq \sup_x \frac{(Ax)_i^2}{\|A_{2k}x\|_2^2 + 2\lambda\|x\|_2^2} \geq \frac{1}{60}\frac{\|a_i^\intercal(I-P_F)\|_2^2}{\|A(I-P_F)\|_F^2}.$$

We proved this for arbitrary F, and so we are done.

Problem 2: Sketching for Second Order Methods

1. We expand the initial expression. So we have that

$$\begin{aligned}
&\text{argmin}_{x\in\mathcal{C}}(\frac{1}{2}\|SA(x-x^t)\|_2^2 - \langle A^\intercal(b-Ax^t), x\rangle)\\
=&\text{argmin}_{x\in\mathcal{C}}(\frac{1}{2}x^\intercal A^\intercal S^\intercal SAx - \frac{1}{2}x^\intercal A^\intercal S^\intercal SAx^t - \frac{1}{2}(x^t)^\intercal A^\intercal S^\intercal SAx - x^\intercal A^\intercal b + x^\intercal A^\intercal Ax^t)\\
=&\text{argmin}_{x\in\mathcal{C}}(\frac{1}{2}\|SAx\|_2^2 - x^\intercal A^\intercal S^\intercal SAx^t - x^\intercal A^\intercal b + x^\intercal A^\intercal Ax^t)\\
=&\text{argmin}_{x\in\mathcal{C}}(\frac{1}{2}\|SAx\|_2^2 - x^\intercal A^\intercal(S^\intercal SAx^t + b - Ax^t))\\
=&\text{argmin}_{x\in\mathcal{C}}(\frac{1}{2}\|SAx\|_2^2 - x^\intercal A^\intercal(b - (I - S^\intercal S)Ax^t))\\
=&\text{argmin}_{x\in\mathcal{C}}(\frac{1}{2}\|SAx\|_2^2 - \langle A^\intercal(b-(I-S^\intercal S)Ax^t), x\rangle)
\end{aligned}$$

2. Let $\Delta = x^* - x^{t+1}$.

$$\langle (SA)^\intercal SAx^{t+1} - A^\intercal z, \Delta\rangle \geq 0$$

and

$$\langle A^\intercal Ax^* - A^\intercal b, (-\Delta)\rangle \geq 0.$$

Expanding the former,

$$(x^{t+1})^T(SA)^T(SA)\Delta - b^TA\Delta + (x^t)^TA^TA\Delta - (x^t)^TA^TS^TSA\Delta \geq 0,$$

and expanding the latter,

$$-(x^*)^TA^TA\Delta + b^TA\Delta \geq 0.$$

Adding them together, we have

$$(x^{t+1} - x^t)^T(A^TS^TSA)\Delta \geq (x^* - x^t)^TA^TA\Delta.$$

Now we add $(x^t - x^*)^T A^T S^T SA\Delta$ to both sides, and we get

$$(x^{t+1} - x^*)^T A^T S^T SA\Delta \geq (x^* - x^t)^T A^T A\Delta + (x^t - x^*)^T A^T S^T SA\Delta$$

which is the same as

$$-\Delta^T A^T S^T SA\Delta \geq (x^* - x^t)^T A^T (I - S^T S)A\Delta,$$

and rearranging gives

$$|(x^* - x^t)^T A^T (I - S^T S)A\Delta| \geq \|SA\Delta\|_2^2.$$

3. Let $S = S^{t+1}$. Applying Cauchy-Schwarz to the upper bound in the previous part, we have

$$|(x^* - x^t)^T A^T (I - S^T S)A\Delta| \leq \|\Sigma V^T (x^* - x^t)\|_2 \cdot \|\Sigma V^T \Delta\|_2 \cdot \epsilon,$$

where $A = U\Sigma V^T$ in it SVD, and we have used $\|I - U^T S^T SU\|_2 \leq \epsilon$ since S is a subspace embedding for A.

Also, we have

$$\|SA\Delta\|_2^2 \geq (1-\epsilon)\|A\Delta\|_2^2,$$

also because S is a subspace embedding.

Combining these two bounds and the previous part, we obtain

$$\|A\Delta\|_2 = O(\epsilon)\|A(x^* - x^t)\|_2.$$

The claim now follows inductively across the N iterations, applying the same analysis as above with t.

Problem 3: Block Leverage Scores

1. We have that

$$\mathcal{L}_i(A) = Tr(A^i (A^\intercal A)^{-1} (A^i)^\intercal) = \sum_j A^i_j (A^\intercal A)^{-1} (A^i_j)^\intercal = \sum_j \ell_j(A).$$

2. We saw in class how to estimate ℓ_i to within a $(1 \pm \varepsilon)$ approximation by taking $\tilde{\ell}_i = |e_i ARG|_2^2$. So the algorithm $\mathcal{L}_i(A)$ is to compute $\tilde{\ell}_i$ for each $i \in T$ where T is the set of rows of A in A^i and then add them up.

   So using the previous part, we have that

$$\tilde{\mathcal{L}}_i(A) = \sum_{i\in T} \tilde{\ell}_i = \sum_{i \in T} \ell_i(1 \pm \varepsilon) = (1 \pm \varepsilon)\mathcal{L}_i(A).$$

   In terms of runtime, note that we only have to compute product ARG once, which takes $(\mathsf{nnz}(A) + d^2)\log n$ time. Computing $|e_i ARG|_2^2$ for each $i \in [n]$ takes $O(\mathsf{nnz}(A)\log n)$ time.

3. Let $U\Sigma V^\intercal$ be the SVD decomposition of A. So, we can also rewrite A^i as $U^i \Sigma V^\intercal$ where U^i consists of the first i rows of U. Since our quantity is scale invariant, finding

$$\sup_X \frac{\|A^i X\|_F^2}{\|AX\|_2^2}$$

   is the same as maximizing $\|U^i \Sigma V^\intercal X\|_F^2$ subject to the constraint that $\|U\Sigma V^\intercal X\|_2^2 = 1$. Note that U has orthonormal columns and therefore the constraint becomes $\|\Sigma V^\intercal X\|_2^2 = 1$. Now, let us say that $Y = \Sigma V^\intercal X$. So, we want to maximize $\|U^i Y\|_F^2$ subject to $\|Y\|_2^2 = 1$. This is maximized when we have $Y = I$. So we have that

$$\sup_X \frac{\|A^i X\|_F^2}{\|AX\|_2^2} = \sum_{r=1}^{i} \|U_r\|_2^2,$$

   giving us the result.

4. We have that
$$\mathcal{L}_i(A) = \sum_j \ell_j(A) = \sum_j |U_j|_2^2 = \|U^i\|_F^2 \geq \|U^i\|_2^2.$$

5. We want to show that $|SAx|_2^2 = (1 \pm \varepsilon)|Ax|_2^2$ for all x. Doing the standard change of variable, this is equivalent to showing that
$$|SUy|_2^2 = (1 \pm \varepsilon)|y|_2^2$$
for all y and U with orthonormal columns. Like in class we will show with high probability that
$$\|U^\intercal S^\intercal SU - I\|_2 \leq \varepsilon.$$
We will use Matrix Chernoff. Let us set up our random variables and bound the required terms.

For $j \in [t]$, let i_j be the index that was picked in the j-th trial. Let
$$X_j = I_d - \frac{U_{i_j}^\intercal U_{i_j}}{q_{i_j}}$$
where U_{i_j} is the block matrix corresponding to the sampled rows in the j-th trial.

We can see that all the X_j are independent copies of a symmetric matrix. Now, let us verify the expectation. So we have
$$E[X_j] = I_d - \sum_i q_i \cdot \frac{U_i^\intercal U_i}{q_i} = I_d - \sum_i U_i^\intercal U_i = I_d - U^\intercal U = I_d - I_d = 0^d.$$
We also have
$$\|X_j\|_2 \leq \|I_d\|_2 + \frac{\|U_{i_j}^\intercal U_{i_j}\|_2}{q_{i_j}} \leq 1 + \max_i \frac{\|U_i\|_2^2}{q_i} = 1 + \max_i \frac{|U_i|_2^2 \cdot d}{\beta \mathcal{L}_i} \leq 1 + \frac{d}{\beta}$$
where the last inequality follows from the previous part.

Finally we have
$$\begin{aligned} E[X^\intercal X] &= I_d - 2E\left[\frac{U_{i_j}^\intercal U_{i_j}}{q_{i_j}}\right] + E\left[\frac{U_{i_j}^\intercal U_{i_j} U_{i_j}^\intercal U_{i_j}}{q_{i_j}^2}\right] \\ &= \sum_i \frac{U_i^\intercal U_i U_i^\intercal U_i}{q_i} - I_d \preceq \sum_i \frac{\|U_i^\intercal U_i\|_2 U_i^\intercal U_i}{q_i} - I_d \preceq \frac{d}{\beta} \sum_i U_i^\intercal U_i - I_d \leq (\frac{d}{\beta} - 1) I_d. \end{aligned}$$
So we therefore have that
$$\|E[X^\intercal X]\|_2 \leq \frac{d}{\beta} - 1.$$
The rest follows from the lecture.

References