15-851 Algorithms for Big Data — Spring 2025

Problem Set 2

Due: Thursday, February 27, before class

Please see the following link for collaboration and other homework policies:
http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15851-spring25/grading.pdf

Problem 1: Ridge Leverage Scores Bound Low Rank Sensitivities (16 points)
For an $n \times d$ matrix A, the i-th ridge leverage score $\tau^i(A)$ is defined to be:

$$\tau_i = a_i^T (A^T A + \lambda I)^{-1} a_i,$$

where a_i is the i-th row of A and where $\lambda = \frac{\|A - A_k\|_F^2}{k}$. Here A_k is the best rank-k approximation to A with regards to the Frobenius norm.

1. (2 points) Prove for all $i \in [n]$ that τ_i is the i-th *leverage* score of the matrix $[A; \sqrt{\lambda} I]$, where this notation means to stack A vertically on top of $\sqrt{\lambda} I$ for the $d \times d$ identity matrix I.

   HINT: Another definition for the i^{th} leverage score of a matrix M is $m_i^\intercal (M^\intercal M)^{-1} m_i$.

2. (2 points) Use the previous part and a general argument about leverage scores to show:

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

   HINT: Yet another definition for the i^{th} leverage score of a matrix M is $\sup_x \frac{(Mx)_i^2}{\|Mx\|_2^2}$.

3. (3 points) Argue that $\|A - A_{2k}\|_2^2 \leq \lambda$.

   HINT: You can use the following. Take the singular values of M to be $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$. By the Eckart–Young–Mirsky theorem, we have that $\|M - M_k\|_F^2 = \sigma_{k+1}^2 + \sigma_{k+2}^2 + \ldots + \sigma_d^2$ where M_k is the best rank-k approximation to A with regards to the Frobenius norm.

4. (9 points) Let $\mathcal{F}_k$ denote the family of all rank-k $d \times d$ projection matrices. We will show the following:

$$\tau_i \geq C \sup_{F \in \mathcal{F}_k} \frac{\|a_i^T (I - P_F)\|_2^2}{\|A(I - P_F)\|_F^2},$$

   where P_F is the projection onto a specific rank-k space F, and $C > 0$ is a constant.

(a) (2 point) Show that

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

HINT: Combine the previous two parts.

(b) (3 points) Show that there exists an x such that

$$\Pr[(Ax)_i^2 \geq \|a_i^\intercal(I - P_F)\|_2^2/C] > \frac{1}{2}$$

for some C.

HINT: Consider a specific rank-k subspace F and let $x = P_H(I - P_F)g$, where H is the at most $(3k+1)$-dimensional space spanned by the rows of A_{2k}, F, and a_i, and where g is a standard normal Gaussian vector. Here P_H denotes the projection onto H. Notice that $(Ax)_i$ is distributed as a Gaussian with a certain variance. Use this to get the desired result.

(c) (3 points) Show for the same x that

$$\Pr[\|A_{2k}x\|_2^2 + 2\lambda\|x\|_2^2 \leq C \cdot \|A(I - P_F)\|_F^2] > \frac{1}{2}$$

for some constant C.

HINT: Bound the expectations of $\|A_{2k}x\|_2^2$ and $2\lambda\|x\|_2^2$ separately. To bound the expectation of $2\lambda\|x\|_2^2$, use the fact that H is a space of at most $3k+1$ dimensions. Then use Markov's bound.

(d) (1 point) Conclude that we have

$$\tau_i \geq C \sup_{F \in \mathcal{F}_k} \frac{\|a_i^T(I - P_F)\|_2^2}{\|A(I - P_F)\|_F^2}.$$

Problem 2: Sketching for Second Order Methods (17 points)
Let $A \in \mathbb{R}^{n \times d}$ with $n \geq d$ and with rank d. We seek to solve the constrained regression problem $\min_{x \in \mathcal{C}} \|Ax - b\|_2^2$. Here $\mathcal{C}$ is a convex constraint set that the solution x must belong to (you do not need to know the notion of convexity to solve this problem). Consider the following iterative algorithm with N iterations:

1. Initialize $x^0 = 0$
2. For iterations $t = 0, 1, 2, \ldots, N-1$, generate an independent sketching matrix $S^{t+1} \in \mathbb{R}^{k \times n}$ and perform the update:

$$x^{t+1} = \text{argmin}_{x \in \mathcal{C}} \left(\frac{1}{2}\|S^{t+1}A(x - x^t)\|_2^2 - \langle A^T(b - Ax^t), x \rangle \right). \quad (1)$$

3. Return $\hat{x} = x^N$

Notice that part of (1) involves A rather than SA. Let $x^* = \text{argmin}_{x\in\mathcal{C}}\|Ax-b\|_2^2$ be the solution to the original problem. We will first focus on a specific value of t and let $S = S^{t+1}$.

1. (3 points) Argue that the minimizer to (1) is the same as that to

$$\text{argmin}_{x\in\mathcal{C}}\left(\frac{1}{2}\|SAx\|_2^2 - \langle A^Tz, x\rangle\right), \tag{2}$$

where $z = b-(I-S^TS)Ax^t$ is a fixed vector (here x^t is fixed because it was found in the previous iteration).

2. (7 points) When minimizing a convex function f over a convex set $\mathcal{C}$, if p is the optimum solution and q is any feasible solution, then

$$\langle\nabla f(p), q-p\rangle \geq 0,$$

where $\nabla f(p)$ is the gradient of function f evaluated at p. We will not prove or formally define these terms, but for our problem (2) they imply

$$\langle(SA)^TSAx^{t+1} - A^Tz, x^* - x^{t+1}\rangle \geq 0, \tag{3}$$

where x^{t+1} is the optimal to (2) and x^* is the minimizer to $\frac{1}{2}\min_{x\in\mathcal{C}}\|Ax-b\|_2^2$. Applying similar reasoning to the program $\frac{1}{2}\min_{x\in\mathcal{C}}\|Ax-b\|_2^2$ also implies:

$$\langle A^TAx^* - A^Tb, x^{t+1} - x^*\rangle \geq 0. \tag{4}$$

You do not need to prove (3) or (4) and can take both as given. Show how to combine (3) and (4) to prove:

$$\|SA\Delta\|_2^2 \leq |(x^*-x^t)A^T(I-S^TS)A\Delta|. \tag{5}$$

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

3. (7 points) Suppose that S^t is a $(1+\epsilon)$-approximate subspace embedding of A, for each $t = 0, 1, 2, \ldots, N-1$. Argue that

$$\|A\hat{x} - Ax^*\|_2^2 \leq \varepsilon^{\Theta(N)}\cdot\|Ax^*\|_2^2$$

HINT: Again for notational convenience take $S = S^{t+1}$. It suffices to show $\|Ax^* - Ax^{t+1}\|_2 \leq O(\epsilon)\|Ax^*-Ax^t\|_2$ for each t and then apply induction. For a single iteration, use (5) and lower bound the left hand side and upper bound the right hand side using properties of a subspace embedding.

Problem 3: Block Leverage Scores (17 points)

Suppose we are given an $n \times d$ matrix A with full column rank and would like to sample and reweight a subset of its rows to obtain a subspace embedding. We learned how to do this using leverage score sampling in class. Suppose that the rows of A are partitioned into t groups, denoted $A^1, \ldots, A^t$, where A^i is an $n_i \times d$ matrix, and now we would like to either include or exclude entire groups in our sample. This is useful if, e.g., each group has multiple rows that collectively mean something and one would like to preserve this meaning by sampling the entire group.

We define the i-th block leverage score

$$\mathcal{L}_i(A) = \mathrm{Tr}(A^i(A^TA)^{-1}(A^i)^T),$$

where Tr denotes the trace, which is the sum of diagonal elements of a square matrix. Notice that if A^i has a single row, then this coincides with the usual definition of leverage scores.

1. (2 points) Prove that if T is the set of rows of A in A^i, then $\mathcal{L}_i(A) = \sum_{j \in T} \ell_j(A)$, where $\ell_j(A)$ is the leverage score of the j-th row of A.

2. (2 points) For any fixed constant $\epsilon > 0$, show how to compute estimates $\mathcal{L}_i(A)$, for all $i \in \{1, 2, \ldots, t\}$, in total time $O(\mathrm{nnz}(A) + d^2) \log n$ such that with constant probability, simultaneously for all $i \in \{1, 2, \ldots, t\}$ the estimate $\tilde{\mathcal{L}}_i$ is a $(1 \pm \varepsilon)$ approximation to $\mathcal{L}_i(A)$.

3. (5 points) This part is meant to give you an understanding of the meaning of a block leverage score and might or might not be needed in the next part. Prove that

$$\mathcal{L}_i(A) = \sup_X \frac{\|A^iX\|_F^2}{\|AX\|_2^2}.$$

   HINT: Start by taking the SVD of A and then perform a change of variables.

4. (3 points) Write A^i in its SVD as $U^i\Sigma V^\intercal$. Prove that

$$\mathcal{L}_i(A) \geq \|U^i\|_2^2.$$

5. (5 points) Now suppose that for each $i \in \{1, 2, \ldots, t\}$ we have an estimate $\tilde{\mathcal{L}}_i$ satisfying

$$\beta\mathcal{L}_i \leq \tilde{\mathcal{L}}_i \leq \mathcal{L}_i.$$

   Let $(q_1, \ldots, q_t)$ be a probability distribution with $q_i \geq \frac{\tilde{\mathcal{L}}_i}{d}$ for all $i \in \{1, 2, \ldots, t\}$.

   Let $k = O(d(\log d)/(\beta\epsilon^2))$. Define a sampling and rescaling matrix $S = D \cdot \Omega^T$. For each $j \in [k]$, independently and with replacement pick a block index $i \in [t]$ with probability q_i. So, we have sampled indices $i_1, \ldots, i_k$. Let K be the total number of rows that are in block matrices $A^{i_1}, \ldots, A^{i_k}$. We have that D is $K \times K$ and Ω is $n \times K$.

Take r_i to be the number of rows in A^i. For some $j \in [k]$, we have sampled block index i_j, corresponding to block matrix A^{i_j}. Take index ℓ to be $r_{i_1} + r_{i_2} + \ldots + r_{i_{j-1}} + 1$. For each $z \in [r_{i_j}]$, take x_z to be the row index (of A) corresponding to the z-th row of A^{i_j} and set $\Omega_{x_z,\ell+z-1} = 1$. Also, set $D_{\ell+z-1,\ell+z-1} = 1/(q_{i_j}k)^{1/2}$.

Show that with probability at least $9/10$, SA is a subspace embedding, meaning that simultaneously for all x, we have $\|SAx\|_2^2 = (1 \pm \epsilon)\|Ax\|_2^2$.

HINT: Follow the leverage score matrix Chernoff analysis from class, but you will need to change sampled rows to sampled blocks in a few places, and make modifications in the analysis to deal with blocks.