15-859 Algorithms for Big Data — Fall 2022
Problem Set 2

Due: Tuesday, October 25, 11:59pm

Please see the following link for collaboration and other homework policies: `http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall22/grading.pdf`

Problem 1: Online Leverage Scores (17 points)
The leverage scores indicate the importance of rows for sampling, as we have shown in class. In this problem, we consider an online model, where we see each row of A one at a time, and would like to find the important rows, namely those of large leverage score with respect to the previous rows that we have seen so far.

1. Given an $n \times d$ matrix A with $n \geq d$ and with rank d, recall that the i-th row leverage score ℓ_i is defined to be the i-th squared row norm of U, where U is an orthonormal basis for the column span of A. Recall from lecture that $\sum_{i=1}^n \ell_i = d$. Argue that
$$\ell_i = a_i^T (A^T A)^{-1} a_i,$$
where a_i is the i-th row of A.

2. Now suppose we see the rows of A one at a time. Define the i-th *online* leverage score to be
$$\ell_i = \min(a_i^T (A_{i-1}^T A_{i-1} + \lambda I)^{-1} a_i, 1),$$
where A_{i-1} for $i > 1$ consists of the submatrix of A consisting of its first $i-1$ rows, A_0 is just the 0 matrix and $\lambda > 0$ is a parameter. In this part we will argue that
$$\sum_{i=1}^n \ell_i = O(d \log(1 + \|A\|_2^2/\lambda)), \tag{1}$$
where $\|A\|_2^2 = \sup_x \frac{\|Ax\|_2^2}{\|x\|_2^2}$ is the supremum (maximum) over all vectors x.

   To prove this, we need the following Matrix Determinant Lemma: if S is an invertible square matrix and u is a vector, then:
$$\det(S + uu^T) = (\det S)(1 + u^T S^{-1} u).$$
   Prove (1) by showing the following:

   (a) Show that $\det(A^T A + \lambda I) \geq \det(\lambda I) \exp(\sum_i \ell_i/2)$.

   HINT: Use the Matrix Determinant Lemma applied to $S = A_i^T A_i + \lambda I$ and an appropriate vector u for each i. It may also be helpful to note $\exp(\ell_i/2) \leq 1 + \ell_i$ for each i (why?).

   (b) Show that $\det(A^T A + \lambda I) \leq (\|A\|_2^2 + \lambda)^d$ and then bound $\sum_i \ell_i$ using the bound from above part.

   HINT: Write the value of determinant in terms of eigen values of the matrix and then use an upper bound on th top eigenvalue of $A^T A + \lambda I$.

Problem 2: Gap-Dependent Bounds for Low Rank Approximation (16 points)

Suppose we are given an $n \times n$ matrix A and would like to output a $(1+\epsilon)$-approximate Frobenius norm rank-k approximation, that is, an $n \times k$ matrix U together with a $k \times n$ matrix V for which

$$\|U \cdot V - A\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2,$$

where A_k is the best rank-k approximation to A in Frobenius norm. For simplicity, we will focus on the case $k = 1$ in this problem. Let σ_i denote the i-th singular value of A. We will show that when σ_2 is a bit less than σ_1, then one can obtain fast algorithms for rank-1 approximation in terms of the approximation factor ϵ.

We will consider the following algorithm called $\mathsf{PowerMethod}(A, r)$. Suppose we choose a random $n \times 1$ Gaussian vector g (each coordinate is an independent standard normal random variable) and compute $(AA^T)^r Ag$. Let

$$u = \frac{(AA^T)^r Ag}{\|(AA^T)^r Ag\|_2}$$

be the output of $\mathsf{PowerMethod}(A, r)$.

1. Let $1 \geq \gamma \geq 1/n^2$ be a parameter. Suppose that the unit vector u is the result of running $\mathsf{PowerMethod}(A, O(\log(n)/\gamma))$. We will now argue that

$$\|u^T A\|_2^2 \geq \sigma_1^2 - \gamma\sigma_2^2, \tag{2}$$

with probability at least $1 - 1/\text{poly}(n)$. Let $A = U\Sigma V^T$ be the singular value decomposition of matrix A.

(a) If $y = (AA^T)^r Ag$ show that y has the same distribution as that of $U\Sigma^{2r+1}g'$ where g' is a random vector with coordinates being independent normal random variables. From here onwards we work with the vector $y = U\Sigma^{2r+1}g'$.

HINT: Rotational invariance of Gaussian random vectors.

(b) Let $g' = (g'_1, \ldots, g'_n)$. Show that with probability $\geq 1 - 1/\text{poly}(n)$,

$$|g'_1| \geq \frac{1}{\text{poly}(n)} \quad \text{and} \quad \max_i |g'_i| \leq O(\sqrt{\log n}).$$

HINT: If $g \sim N(0, 1)$, $\Pr[|g| \geq x] \leq \exp(-x^2/2)/x$ for all $x > 0$.

(c) Show that

$$\|y\|_2^2 = \|U\Sigma^{2r+1}g'\|_2^2 = \sum_{i=1}^{n} \sigma_i^{4r+2}(g'_i)^2 \quad \text{and} \quad \|y^T A\|_2^2 = \sum_{i=1}^{n} \sigma_i^{4r+4}(g'_i)^2$$

to conclude

$$\|u^T A\|_2^2 = \frac{\|y^T A\|_2^2}{\|y\|_2^2} = \frac{\sum_{i=1}^n \sigma_i^{4r+4}(g'_i)^2}{\sum_{i=1}^n \sigma_i^{4r+2}(g'_i)^2}.$$

(d) Assuming $\sigma_2 \leq \sigma_1/2$, show that if $r = \Omega(\log(n/\gamma))$, then with probability $\geq 1 - 1/\text{poly}(n)$,

$$\|u^T A\|_2^2 \geq \sigma_1^2 - \gamma\sigma_2^2.$$

HINT: You may want to condition on the properties of g' from (b).

(e) Now assume $\sigma_2 \geq \sigma_1/2$ and let m be the largest index with $\sigma_m \geq (1-\gamma)\sigma_1$. Show that with probability $1 - 1/\text{poly}(n)$

$$\|u^T A\|_2^2 \geq \sigma_1^2 \frac{1 + (\sigma_m/\sigma_1)^2\Theta}{1 + \Theta + 1/\text{poly}(n)}$$

where

$$\Theta := (g_2'/g_1')^2(\sigma_2/\sigma_1)^{4r+2} + \cdots + (g_m'/g_1')^2(\sigma_m/\sigma_1)^{4r+2}.$$

(f) Using the above inequality, conclude that with probability $\geq 1 - 1/\text{poly}(n)$,

$$\|u^T A\|_2^2 \geq \sigma_1^2 - (C\gamma + 1/\text{poly}(n))\sigma_2^2$$

if $r = \Omega((\log n)/\gamma)$ for a large enough universal constant C and then prove (2) by rescaling γ.

2. Argue that if $\sigma_2^2 \leq \alpha\|A - A_1\|_F^2$ for some parameter $\alpha \leq 1$, then there is an algorithm to find vectors $u, v \in \mathbb{R}^n$ for which

$$\|uv^T - A\|_F^2 \leq (1+\epsilon)\|A - A_1\|_F^2,$$

with probability at least $1 - 1/\text{poly}(n)$, and which runs in $O(\text{nnz}(A)(1 + \alpha/\epsilon)\log(n))$ time. Here A_1 denotes the best rank 1 approximation of A obtained by truncating the singular value decomposition to the top singular value. Note that if $\alpha \ll 1$, then the algorithm has a significantly sub-linear dependence on the error parameter ϵ.

Problem 3: CUR Decompositions (17 points)

In this problem we will show that for any matrix $A \in \mathbb{R}^{n \times n}$, there is a product of matrices $C \cdot U \cdot R$, where $C \in \mathbb{R}^{n \times O(k \log^2 k)}$, $U \in \mathbb{R}^{O(k \log^2 k) \times O(k \log k)}$, and $R \in \mathbb{R}^{O(k \log k) \times n}$, where C is a subset of columns of A, R is a subset of rows of A, and U is an arbitrary matrix, for which

$$\|C \cdot U \cdot R - A\|_F^2 = O(1)\|A - A_k\|_F^2. \tag{3}$$

Here A_k is the optimal rank k approximation to A obtained by truncating the singular value decomposition of A to the top k singular values.

The above decomposition is useful since $C \cdot U \cdot R$ has rank at most $O(k \log k)$, and so is low rank, and since it consists of an actual subset of columns and rows of A, if A has sparse rows or columns, then so does the low rank factorization. Note also U is a very small matrix and is easy to store.

1. To show (3), we first generalize the affine embedding proof from class for leverage score sampling. Namely, given an $n \times d$ matrix A and an $n \times m$ matrix B, let S be an $O(d \log d) \times n$ leverage score sampling matrix, as described in class, from the row leverage scores of A. Note that S does not depend on B. We would like to show with probability at least 9/10: if $\hat{X}$ is the minimizer of $\|SAX - SB\|_F^2$, then

$$\|A\hat{X} - B\|_F^2 = O(1) \min_{X \in \mathbb{R}^{d \times m}} \|AX - B\|_F^2. \tag{4}$$

Here, we will only have two properties of S, which you may freely use without proof:

(a) With probability at least 19/20, $(1/2)\|Ax\|_2 \leq \|SAx\|_2 = 2\|Ax\|_2$ simultaneously for all x. This is the **subspace embedding property** we proved in class when proving properties of leverage score sampling.

(b) For any fixed matrix M that does not depend on S,

$$E_S[\|SM\|_F^2] = \|M\|_F^2.$$

Use the above two properties and try to copy the argument in class for affine embeddings to show (4). Also write an explicit expression for $\hat{X}$ and argue that rows of $\hat{X}$ are spanned by a $O(d \log d)$ size subset of rows of B.

Note that the guarantee of (4) is weaker than the final guarantee for affine embeddings we showed in class, as (4) is only about the minimizers of the original and sampled problems (this is necessary, as we do not have all three conditions on S that we had for CountSketch for affine embeddings in lecture).

HINT: You may want to use normal equations for least squares regression multiple times to upper bound $\|A\hat{X} - B\|_F^2$ in terms of $\|AX^* - B\|_F^2$ where X^* is the optimal solution to

$$\min_X \|AX - B\|_F^2.$$

You may also apply Markov inequality to bound $\|SM\|_F^2$ for some matrix M that appears in the analysis.

2. Now suppose we use the low rank approximation algorithm from class to find matrices $\hat{U} \in \mathbb{R}^{n \times k}$ and $\hat{V} \in \mathbb{R}^{k \times n}$ for which $\|\hat{U} \cdot \hat{V} - A\|_F^2 = O(1)\|A - A_k\|_F^2$ in $\text{nnz}(A) + n \cdot \text{poly}(k)$ time. Show how to use the previous part to find a subset of $O(k \log k)$ rows of A, denoted as an $O(k \log k) \times n$ matrix R, for which $\min_{U \in \mathbb{R}^{n \times O(k \log k)}} \|U \cdot R - A\|_F^2 = O(1)\|A - A_k\|_F^2$.

HINT: Apply the result from above part with $A = \hat{U}$ and $B = A$.

3. Given R, show how to use the previous part to find a subset of $O(k \log^2 k)$ columns of A, denoted as an $n \times O(k \log^2 k)$ matrix C, for which

$$\min_{U \in O(k \log^2 k) \times O(k \log k)} \|CUR - A\|_F^2 = O(1)\|A - A_k\|_F^2.$$

It turns out the optimal U to the above expression is given by $U = C^{-}(PAQ)R^{-}$, where P is the projection onto C and Q is the projection onto R, and C^{-} and R^{-} are the Moore Penrose pseudoinverses. You do not need to prove this. Thus, one has found a CUR-decomposition with the guarantees of (3)!