15-851 Algorithms for Big Data — Spring 2024

Problem Set 1

Due: Thursday, Februrary 8, 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-spring24/grading.pdf`

Problem 1: Sparse Regression (12 points)
Given an $n \times d$ matrix A as well as an $n \times 1$ vector b, in the sparse regression problem we would like to find a vector $x \in \mathbb{R}^d$ with at most k-non zero entries so that $\|Ax - b\|_2$ is approximately minimized. Find an upper bound s on the number of rows of a Gaussian matrix S, so that with failure probability at most $1/10$, we have that if x' is the vector with at most k-non zero entries minimizing $\|SAx' - Sb\|_2$, then

$$\|Ax' - b\|_2 \leq (1+\epsilon) \min_x \|Ax - b\|_2.$$

Your upper bound on s may be a function of n, d, k, and $1/\epsilon$, and you should give a proof of its correctness. The smaller your upper bound, the more credit you will receive.

HINT: It may help to note that the proof for the subspace embedding result covered in class actually proves that a Gaussian matrix with $O((d + \log(1/\delta))/\varepsilon^2)$ rows is a $(1+\varepsilon)$-subspace embedding for a d -dimensional subspace with probability at least $1 - \delta$.

Problem 2: Gaussian Subspace Embeddings with Exactly d Rows (13 points)

Given an $n \times d$ matrix A with $n > d$ and $\text{rank}(A) = d$, we would like to choose a matrix S of i.i.d. $N(0,1)$ Gaussian random variables with as few rows as possible so that with positive probability S is a subspace embedding, that is,

$$\|SAx\|_2 = (1 \pm \epsilon)\|Ax\|_2$$

simultaneously for all vectors x.

1. Show that if S has fewer than d rows, this task is impossible.

2. Now show that if S has d rows, then with probability at least $e^{-\Theta(d^2 \log d)}$, it holds that S is a $(1 + 1/\text{poly}(d))$-subspace embedding.

   HINT: as we did in class, we may assume that the columns of A are orthonormal, then try to argue that with this probability $SA = I + T$, where I is a $d \times d$ identity matrix and all the entries in T are at most $1/\text{poly}(d)$. Then use this form of SA to argue that it is a subspace embedding.

HINT: for a standard Gaussian random variable and any $0 < \delta < 1$, we have

$$\mathbf{Pr}\big[|g| \leq \delta\big] = \Omega(\delta) \quad \text{and} \quad \mathbf{Pr}\big[1 - \delta \leq |g| \leq 1 + \delta\big] = \Omega(\delta)\ .$$

This can be derived by looking at its probability density function.

Note that even though the success probability is extremely small, this gives the best compression possible and is useful in certain applications. If d is small, one could repeat the above experiment $e^{\Theta(d^2 \log d)}$ times, and with good probability, one of the repetitions will provide the desired compression. Feel free to ask me more if you are interested in the applications!

Problem 3: Active Regression (13 points)
You are given an $n \times d$ matrix A, $n > d$, and would like to solve *active regression*, that is, you would like to find x' so that

$$\|Ax' - b\|_2 \leq C \min_x \|Ax - b\|_2, \tag{1}$$

where $C > 0$ is some constant, and where you would like to succeed with probability at least $9/10$. The catch is that you are not directly given the vector b. Instead, you have *query access* to the vector b, and would like to minimize the number of entries in b that you read in order to solve the above problem.

Suppose that A has orthonormal columns and each of its rows A_i is such that $\|A_i\|_2^2 = O(d/n)$. Show that there is a scheme which reads only $O(d \log d)$ entries of b and produces an x' which satisfies (1) for some constant $C > 0$ (the precise constant that you get does not matter).

HINT: Consider using the Matrix Chernoff bound we studied in class and applying it to obtain a subspace embedding for the matrix A which is based on sampling rows of A. Then try to involve b in your analysis with the triangle inequality. You may also need to use Markov's inequality to bound the dilation of a fixed vector, and combine your analysis with a union bound.

Problem 4: Fast High Probability Matrix Product (12 points)
Design a distribution on sketching matrices S which have $O(\log n)$ rows, so that for any fixed matrices $A, B \in \mathbb{R}^{n \times n}$, we have that

1. $S \cdot A$ and $S \cdot B$ can each be computed in $O(n^2)$ time, and
2. with probability at least $1 - 1/n^{.99}$, it holds that

$$\|A^T S^T S B - A^T B\|_F \leq \frac{1}{10} \|A\|_F \cdot \|B\|_F.$$

HINT: It may be helpful to combine various sketching matrices from class. You may reduce the dimension multiple times.

HINT: For a random Gaussian matrix, how many rows do we need to approximate the matrix product? You may not need to use the JL-moment property for this.

HINT: To make the problem easier, we allow you to refer to the proof of problem 3 here:

`https://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall17/ps1sol.pdf`

for part of your argument. You may need to change some parameters in this argument (you don't need to reprove everything but can say how the above proof changes) and combine it with other arguments to solve the problem.