15-859 Algorithms for Big Data — Fall 2020

Problem Set 1

Due: Thursday, October 1, before class

Please see the following link for collaboration and other homework policies:
http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall20/grading.pdf

Problem 1: Generalization Error for Regression (14 points)
In class we looked at the sketch-and-solve paradigm for finding a solution $\hat{x} \in \mathbb{R}^d$ for which

$$\|A\hat{x} - b\|_2 \leq (1+\epsilon) \min_x \|Ax - b\|_2,$$

where $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^{n \times 1}$ and $n \geq d$. In this problem, we will study different error guarantees for least squares regression.

Suppose in both parts that $S \in \mathbb{R}^{s \times n}$, where $s = O(d/\epsilon^2)$ is a matrix of i.i.d. standard normal random variables, scaled by $1/\sqrt{s}$. Here we assume the constant in the $O(\cdot)$ notation is sufficiently large.

Let $\hat{x}$ be the solution to $\text{argmin}_x \|SAx - Sb\|_2$, and let x_{OPT} be the solution to $\min_x \|Ax - b\|_2$. Suppose you are given a single new vector $v \in \mathbb{R}^d$, independent of S and would like to use your solution $\hat{x}$ to "predict" the corresponding entry of b you would get if v were an additional row appended to A. More precisely, you would like $\langle \hat{x}, v\rangle$ to be close to $\langle x_{OPT}, v\rangle$.

1. (6 points) Show with probability at least 2/3, $\|\hat{x} - x_{OPT}\|_2 \leq O(\epsilon)\|Ax_{OPT} - b\|_2 \cdot \|A^\dagger\|_2$.
2. (1 point) Using the above, show $|\langle \hat{x}, v\rangle - \langle x_{OPT}, v\rangle| = O(\epsilon)\|v\|_2\|Ax_{OPT} - b\|_2\|A^\dagger\|_2$.
3. (2 points) The previous part shows a "generalization error" bound, i.e., how well we can predict the "labels" (entries of b) given only the "examples" (rows of A). However, we can obtain a better error bound with the same random matrix S.

   First, show that

   $$|\langle \hat{x}, v\rangle - \langle x_{OPT}, v\rangle| = |\langle (SA)^\dagger S(b - AA^\dagger b), v\rangle|.$$

4. (5 points) Use the previous part to show that with probability at least 2/3,

   $$\|\hat{x} - x_{OPT}\|_\infty \leq O\left(\frac{\epsilon \cdot \sqrt{\log d}}{\sqrt{d}}\right) \|Ax_{opt} - b\|_2 \cdot \|A^\dagger\|_2.$$

   Here fore a vector y, we have that $\|y\|_\infty$ is the maximum absolute value of any entry of y. For this problem it will be helpful to recall properties about dot products of vectors with Gaussian vectors that we did in class.

Problem 2: Ridge Regression (12 points)
Consider the ridge regression problem: given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^{n \times 1}$, find an $x \in \mathbb{R}^d$ so as to minimize $\|Ax - b\|_2^2 + \|x\|_2^2$. Here the $\|x\|_2^2$ term is known as a regularizer or "ridge", which encourages the solution x to not have very large norm, which could be an indication of overfitting the data in A. We will again consider the overconstrained case $n > d$.

We would like to design an $s \times n$ sketching matrix S so that

$$\|SAx - Sb\|_2^2 + \|x\|_2^2 = (1 \pm \epsilon)(\|Ax - b\|_2^2 + \|x\|_2^2) \tag{1}$$

simultaneously for all vectors x. Intuitively, the number s of rows of S should be smaller than without the ridge since if $\|Ax - b\|_2$ is small compared to $\|x\|_2$, we do not need to approximate $\|Ax - b\|_2$ very well. This problem will make this intuition formal.

1. (6 points) Consider the matrix

$$\hat{A} = \begin{bmatrix} A & b \\ I_d & 0 \end{bmatrix},$$

where I_d is the $d \times d$ identity matrix. Let us write an orthonormal basis for the columns of $\hat{A}$ in the following form, for some U_1 and U_2, where U_1 has n rows.

$$\begin{bmatrix} U_1 & b^1 \\ U_2 & b^2 \end{bmatrix},$$

where (b^1, b^2) is a unit vector in the direction of $(b, 0)$ projected orthogonally to the column span of

$$\begin{bmatrix} A \\ I_d \end{bmatrix}.$$

Prove that $\|U_1, b^1\|_F^2 \leq 1 + \sum_{i=1}^d \frac{1}{1 + \frac{1}{\sigma_i^2}}$, where σ_i is the i-th singular value of A.

HINT: it may be useful to write $A = U\Sigma V^T$ in its singular value decomposition (SVD). Then consider the matrix

$$\begin{bmatrix} U\Sigma(\Sigma^T\Sigma + I_d)^{-1/2} & b^1 \\ V(\Sigma^T\Sigma + I_d)^{-1/2} & b^2 \end{bmatrix}$$

2. (6 points) The quantity $1 + \sum_{i=1}^d \frac{1}{1 + \frac{1}{\sigma_i^2}}$ is known as the statistical dimension t, and is always at most $d + 1$. Suppose that S is a CountSketch matrix with $O(t^2/\epsilon^2)$ rows and n columns. Argue that (1) holds for this choice of S with constant probability.

HINT: It may be helpful to consider the matrix:

$$\hat{S} = \begin{bmatrix} S & 0 \\ 0 & I_d \end{bmatrix}$$

and to consider the product $\hat{S} \cdot \hat{A}$.

Problem 3: Finding a Column Basis of a Matrix (12 points) Suppose you are given an $n \times n$ matrix C and promised it has at most k linearly independent columns, but you do not know where they are. Show how to find a maximal subset of linearly independent columns of C in $O(\text{nnz}(C)) + n \cdot \text{poly}(k)$ time. Your algorithm should succeed with probability at least 2/3.

Problem 4: Sublinear Approximate Matrix Product (12 points)
Let $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ be given. In class we saw that if S is a Countsketch with $O(1/\epsilon^2)$ rows, then with probability at least 2/3,

$$\|AS^T SB - AB\|_F \leq \epsilon \|A\|_F \|B\|_F. \tag{2}$$

The time to compute AS^T and SB is $O(\text{nnz}(A) + \text{nnz}(B))$, which is not ideal if say, A is a sparse matrix but B is a dense matrix. Show how to define a different random matrix S so that one can compute AS^T and SB in total time $\min(\text{nnz}(A), \text{nnz}(B)) + O(n/\epsilon^2)$, and so that AS^T and SB achieve the guarantee in (2) with probability at least 2/3.

HINT: think about different kinds of sampling and rescaling matrices S. In particular, think of a way of sampling to approximate the matrix product, based on a distribution defined only in terms of the matrix which has fewer non-zero entries. Try to think about distributions defined in terms of rows or columns of a matrix.