15-859 Algorithms for Big Data — Fall 2020

Problem Set 3

Due: Thursday, November 5, 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: Sampling for ℓ_1-Subspace Embeddings and ℓ_1-Regression (25 points)
In class we saw that there was an oblivious matrix R with $O(d\log(d))$ rows for which with probability $\geq 9/10$,

$$\|Ax\|_1 \leq \|RAx\|_1 \leq O(d\log(d))\|Ax\|_1$$

for all $x \in \mathbb{R}^d$. Certain applications may need a stronger guarantee, i.e., we may want to compute a matrix S with a small number of rows such that

$$(1-\varepsilon)\|Ax\|_1 \leq \|SAx\|_1 \leq (1+\varepsilon)\|Ax\|_1 \tag{1}$$

for all $x \in \mathbb{R}^d$. We show in this problem that given a well-conditioned basis for A, we can compute a sampling matrix S that samples a few rows of A and for which (1) holds. Suppose you are given a matrix U that has the same column space as A and satisfies

1. $\|U\|_1 \leq \alpha$ where $\|U\|_1 = \sum_{i,j} |U_{i,j}|$, and
2. for all $x \in \mathbb{R}^d$, $\|x\|_\infty \leq \beta\|Ux\|_1$

For each $i = 1, 2, \ldots, n$, let

$$p_i = \min\left(1, s \cdot \frac{\|U_{i*}\|_1}{\|U\|_1}\right).$$

Here for a matrix A, $\|A\|_1 = \sum_{i,j} |A_{i,j}|$. Let S be an $n \times n$ diagonal matrix, where $S_{i,i} = \frac{1}{p_i}$ with probability p_i, where all diagonal entries of S are chosen independently.

Show that if $\alpha, \beta \leq \text{poly}(d)$, then for a large enough $s = \text{poly}(d/\epsilon)$, the matrix S satisfies (1) with probability $\geq 9/10$. What is the expected number of nonzero entries in S? This gives a bound on the expected number of rows sampled by S.

HINT: For each $x \in \mathbb{R}^d$, compute $E_S[\|SUx\|_1]$ and $E_S[\|SUx\|_1^2]$. Then apply Bernstein's inequality:

Fact 1 *(Bernstein's Inequality) Let $Z_i \geq 0$ be independent random variables with $\sum_i \mathbf{E}[Z_i^2] < \infty$, and define $Z = \sum_i Z_i$. Suppose $Z_i - \mathbf{E}[Z_i] \leq \Delta$ for all i. Then for any $t > 0$,*

$$\Pr[|Z - \mathbf{E}[Z]| > t] \leq 2 \cdot e^{\left(\frac{-t^2}{2\sum_i \mathbf{E}[Z_i^2] + 2t\Delta/3}\right)}.$$

This gives a high probability bound for a fixed x. Use an ε-net and argue that the bound holds for all x.

Note that to solve the regression problem $\min_x \|Ax - b\|_1$, we can first compute a well-conditioned basis for the matrix $[A, b]$ and then find a sampling matrix as above such that $\|S[A, b][x; y]\|_1 = (1 \pm \epsilon)\|[A, b][x; y]\|_1$ for all $x \in \mathbb{R}^d, y \in \mathbb{R}$. Now one can see that a solution to the problem $\min_x \|SAx - Sb\|_1$ is a $1 + O(\epsilon)$ approximation to $\min_x \|Ax - b\|_1$.

Problem 2: Streaming: What Can and Cannot Be Done (25 points)
In this problem we will explore the efficiency of different but related tasks in the one-way communication model. In these problems, Alice has two vectors $a \in \{-n, -n+1, \ldots, n\}^n$ and $x \in \{-n, -n+1, \ldots, n\}^n$, while Bob has two vectors $b \in \{-n, -n+1, \ldots, n\}^n$ and $y \in \{-n, -n+1, \ldots, n\}^n$. We assume there is an infinitely long shared random string r that both Alice and Bob have access to.

Alice looks at her input and the random string r, and generates a message $M(a, x)$ which she sends to Bob. Bob should then output an approximation to a function with probability at least $9/10$, based on M, r, and his inputs b and y.

For each of the functions below, determine the 1-way communication complexity of the desired approximation.

1. Bob should output a number Z with
$$(1 - \epsilon)[\|x + y\|_1 + \|a + b\|_1] \leq Z \leq (1 + \epsilon)[\|x + y\|_1 + \|a + b\|_1].$$
2. Bob should output a number Z with
$$(1 - \epsilon)[\|x + y\|_1 - \|a + b\|_1] \leq Z \leq (1 + \epsilon)[\|x + y\|_1 - \|a + b\|_1].$$
3. Bob should output a number Z with
$$(1 - \epsilon)[\|x + y\|_1 \cdot |\|a + b\|_2^2 - 1|] \leq Z \leq (1 + \epsilon)[\|x + y\|_1 \cdot |\|a + b\|_2^2 - 1|].$$
4. Let D_a be an $n \times n$ diagonal matrix with the entries of a on the diagonal. Similarly, define D_b. Bob should output a number Z with
$$(1 - \epsilon)\|(D_a - D_b) \cdot (x - y)\|_1 \leq Z \leq (1 + \epsilon)\|(D_a - D_b) \cdot (x - y)\|_1.$$

HINT: Some of these problems have efficient protocols. Try to use the efficient sketching algorithms we covered in class and think carefully what the approximation guarantee implies. For some parts it will be helpful to remember that a and b have integer coordinates.

Some of these problems have lower bounds. To prove your lower bounds you can use the communication lower bound for the following problem:

Definition 2 (Indexing) *Suppose Alice has a set $A \subseteq [n]$ and Bob has an index $i \in [n]$. Bob wants to know if $i \in A$. Any one-way randomized protocol to solve this problem with probability $\geq 9/10$ requires $\Omega(n)$ bits of communication.*