Homework 3

Algorithms for Big Data: CS498 ABG/ABU, Fall 2022
Due: Thursday at 10pm, October 27th, 2022

Instructions and Policy:

- Each homework can be done in a group of size at most two. Only one homework needs to be submitted per group. However, we recommend that each of you think about the problems on your own first.
- Homework needs to be submitted in pdf format on Gradescope. See `https://courses.engr.illinois.edu/cs374/fa2018/hw-policies.html` for more detailed instructions on Gradescope submissions.
- Follow academic integrity policies as laid out in student code. You can consult sources but cite all of them including discussions with other classmates. Write in your own words. See the site mentioned in the preceding item for more detailed policies.

Problem 1. Count Sketch. In the Count-Sketch analysis we showed that if we choose $w = 3/\epsilon^2$ and $d = \Omega(\log(n))$ that for each i we obtain an estimate $\tilde{x}_i$ such that with high probability $|\tilde{x}_i - x_i| \leq \epsilon ||x||_2$. This can be pessimistic in situations where the data is highly skewed with most of the $||x||_2$ concentrated in a few coordinates.

- To make this precise, for some fixed parameter $\ell \in \mathbb{Z}_+$, let $y_i \in \mathbb{R}^n$ be the vector defined by the ℓ largest coordinates (by absolute value) of x, as well as the ith coordinate of x, to 0. (All other coordinates are the same as x). Prove that for $\ell = 1/\epsilon^2$, if w is chosen to be $6/\epsilon^2$ and $d = O(\log n)$, then for all $i \in [n]$, with high probability, we have

$$|\tilde{x}_i - x_i| \leq \epsilon ||y_i||_2.$$

- The Zipfian distribution is a heavy-tailed distribution that is often used to model various forms of data. See `https://en.wikipedia.org/wiki/Zipfs_law` for more on this. In our context consider a non-negative vector $x \geq 0$ and say we sort the coordinates in absolute value and without loss of generality $x_1 \geq x_2 \geq \cdots \geq x_n$. For some parameter $\alpha > 1$ that characterizes the distribution we have $x_k \sim 1/k^\alpha$. Calculate $||y_\ell||_2$ for a vector x which follows the Zipfian distribution with $||x||_1 = m$.

Problem 2. Range Queries via Sketching. We studied CountMin and Count-Sketch, which allow for point query estimates $\tilde{x}_i$ for $i \in [n]$ over a vector x that is updated in the turnstile model. In several applications one needs to answer queries of the form $x[i, j]$ for $i < j$ where $x[i, j] = \sum_{k=i}^{j} x_k$. These are called range queries. Show how one can adapt the standard Count-Sketch data structure (with some additional logarithmic factors in space usage) to be able to answer range query estimates with additive error $\epsilon ||x||_2$. This is mainly to make you read the notes since we did not cover it in lecture.

Problem 3. We have seen algorithms for distinct element estimation and sampling in streaming using polylogarithmic space. However those algorithms are not based on linear sketching. Linear sketching allows one to handle deletions, which has important applications. This problem will help you see a way to do it via another useful idea of doing geometric search for the right value. Use the prompts below to develop a linear sketch that allows one to estimate the number of distinct elements to within a $(1-\epsilon)$-factor with high probability in the turnstile model. For simplicity you can assume the existence of ideal hash functions.

- Suppose you wish to design an algorithm that in the streaming setting decided whether the number of distinct elements d is at least T or less than $(1-\epsilon)T$ where ϵ is some fixed constant in $(0, 1/2)$ and $T > c/\epsilon^2$ for some sufficiently large constant c. Suppose one has an ideal hash function $h : [n] \to [T]$. What is the probability that $h^{-1}(0)$ is non-empty when $d \geq T$ versus when $d < (1-\epsilon)T$?
- Show how you can use the preceding and logarithmic values of T to estimate the number of distinct elements to within a $(1-\epsilon)$-factor with high probability while using space poly-logarithmic in n and polynomial in $1/\epsilon$.
- How can you extend the above to obtain a linear sketch? More formally your algorithm should generate a linear sketch for the vector $x \in \mathbb{R}^n$ such that one can estimate $||x||_0$ to within a $(1-\epsilon)$-factor.

You may want to assume that $||x||_0$ is sufficiently large as a function of $1/\epsilon$. For $||x||_0$ sufficiently small, you can use k-sparse recovery as a black box to completely recover x.

Problem 4. Fast JL. Recall the DJL lemma where we pick a random $m \times n$ matrix Π and show that for $m = O(1/\epsilon^2)$, with at least $2/3$ probability,

$$(1-\epsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1+\epsilon)\|x\|_2^2. \tag{1}$$

- Imagine picking Π as follows: for each $i \in \{1, \ldots, n\}$ we pick a uniformly random number $h_i \in \{1, \ldots, m\}$. We then set $\Pi_{h_i,i} = \pm 1$ for each $i \in \{1, \ldots, n\}$ (the sign is chosen uniformly at random from $\{-1, 1\}$), and all other entries of Π are set to 0. This Π has the advantage that in turnstile streams, we can process updates in constant time. Show that using this Π still satisfies the conditions of Equation 1 with $2/3$ probability for $m = O(1/\epsilon^2)$.
- Show that the matrix Π from the first part can be specified using $O(\log n)$ bits such that Equation 1 still holds with at least $2/3$ probability, and so that given any $i \in \{1, \ldots, n\}$, $\Pi_{h_i,i}$ and h_i can both be calculated in constant time. *Hint:* Use limited independence hash functions to generate the h_i.

Problem 5. Improved net argument for subspace embeddings. Recall that in oblivious subspace embeddings we want to preserve lengths of all vectors in a subspace of dimension d (assuming vectors are in dimension $\mathbb{R}^n$ where $n > d$). For this we showed that a JL matrix with $m = O(d/\epsilon^2)$ rows suffices via a net argument. More formally the claim is that there is a fixed set Q of $\exp(O(d))$ vectors such that preserving their lengths to a $(1 \pm \epsilon)$ factor suffices to preserve lengths of all vectors in that subspace (we then use a union bound). In lecture we describe a construction that yielded a net of size $\exp(d \log d)$ which is weaker. In this problem you will see the stronger bound via the following two parts.

- Define $Q_\gamma = \{w : w \in \frac{\gamma}{\sqrt{d}}\mathbb{Z}^d, \|w\|_2 \leq 1\}$ for $\gamma \in (0,1)$. Prove $|Q_\gamma| \leq e^{d \cdot f(\gamma)}$ for some function $f(\gamma)$ (which needn't be optimized).

  Hint: Given $z \in Q_\gamma$ define a cube C_z centered at z with side length $\gamma/\sqrt{d}$. Note these cubes are all disjoint, then use a volume argument (you may use that an ℓ_2 ball of radius r in $\mathbb{R}^d$ has volume $(C_d \cdot r/\sqrt{d})^d$ for some constant C_d which is $\Theta(1)$ as d grows).

- Show that if for some $A \in \mathbb{R}^{d \times d}$ we have $|u^T A v| \leq \epsilon$ for all $u, v \in Q_\gamma$, then $|x^T A x| \leq \epsilon/(1-\gamma)^2$ for all $x \in \mathbb{R}^d$ of unit ℓ_2 norm.

  Hint: Write $y = (1-\gamma)x$ and round down the coordinates of y to obtain $z \in Q_\gamma$. Argue that $y \in C_z$ and use that any point in a convex polytope can be written as a convex combination of the vertices of that polytope.

- Finish up the details to argue that JL matrix with $m = O(d/\epsilon^2)$ rows is yields an oblivious subspace embedding with constant probability.