Homework 4

Algorithms for Big Data: CS498 ABG/ABU, Fall 2022
Due: Thursday at 10pm, Dec 1st, 2022

Instructions and Policy:

- Unlike previous homeworks, you need only do **3** problems and the last problem which is a mini project. (Of course you're encouraged to try and welcome to submit all of them!)
- 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. JL preserves dot products and application to matrix multiplication. Recall that the distributional JL lemma implies that a projection matrix Π chosen from an appropriate distribution preserves length of any fixed vector x to within a $(1 \pm \epsilon)$-factor with probability $1 - \delta$ if the number of dimensions in the projection is $O(\log(1/\delta)/\epsilon^2)$.

1. Suppose we have two unit vectors u, v. Prove that $\mathrm{E}[|\langle \Pi u, \Pi v\rangle - \langle u, v\rangle|^2] \leq c\epsilon^2\delta$ for some fixed constant c where Π is a $d \times n$ JL matrix where $d = O(\frac{1}{\epsilon^2}\log(1/\delta))$. An easier claim that you can prove instead is that $\mathrm{P}[|\langle \Pi u, \Pi v\rangle - \langle u, v\rangle| > c\epsilon]$ is at most δ.
2. Suppose we want to multiply two matrices A, B where A is an $m \times n$ matrix and B is an $n \times h$ matrix. Argue that if one approximates AB by $D = A\Pi^T\Pi B$ then $\mathrm{E}[||D - AB||_F] \leq O(\epsilon)||A||_F||B||_F$ where Π is a JL matrix as in the preceding part.

Problem 2. Let A be an $m \times n$ matrix. Let $a_1, a_2, \ldots, a_m$ be the rows of A, each of which is viewed as vector in $\mathbb{R}^n$. Let E be the subspace of $\mathbb{R}^n$ spanned by these m vectors and assume $m \ll n$. Consider an oblivious subspace embedding Π for E that satisfies the property that for all $u \in E$, $(1-\epsilon)||u|| \leq ||\Pi u|| \leq (1+\epsilon)||u||$. Let M_k be the best rank k approximation for a matrix M in the Frobenius norm. Prove that $(1 - O(\epsilon))||B - B_k||_F \leq ||A - A_k||_F \leq (1 + O(\epsilon))||B - B_k||_F$ where $B = A\Pi^T$.

Problem 3. Let A be an $m \times n$ matrix. SVD shows that A can be written as UDV^T or equivalently as $\sum_{i=1}^{r} \sigma_i u_i v_i^T$ where r is the rank of A. Letting $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, we saw that A_k is the best rank k approximation to A in the Frobenius norm.

- Prove that σ_1 is the spectral norm of A, that is, $||A||_2 = \sigma_1$.
- Prove that A_k is also the best rank k approximation in the spectral norm. In particular, also show that $||A - A_k||_2 = \sigma_{k+1}$.

Problem 4. Describe a semi-streaming algorithm in the strict turnstile model (edges can be inserted and deleted) to check whether a graph is 2-edge connected.

Problem 5. Matchings with additional constraint. We saw an algorithm in the semi-streaming model for finding a constant factor approximation to the maximum cardinality and maximum weight matching problem. Now consider the following variant. We are given a graph $G = (V, E)$. Moreover each edge has a color from $\{1, 2, ..., k\}$ and each color i has an integer upper bound b_i. The goal is to find a maximum cardinality matching M which satisfies the additional constraint that the number of edges in M from a color class i is at most b_i. Assume that you are given the b_i values ahead of time and that each edge, when it arrives in the stream, specifies its end points and its color. Describe a constant factor approximation for this problem in the semi-streaming setting.

Problem 6. Hypergraph matching. A hypergraph $G = (V, E)$ consists of set of vertices V and a set of hyperedges E. Each hyperedge $e \in E$ is a subset of V, that is $e \subseteq V$. The rank r of G is $\max_e |e|$. Graphs are a special case with $r = 2$. $M \subseteq E$ is a matching in a hypergraph if no two hyperedges in M intersect in a vertex. Unlike matchings in graphs, finding the maximum cardinality matching in a hypergraph is NP-Hard even when $r = 3$ (the standard NP-Complete problem related to this is the 3-D matching problem). Consider the semi-streaming version of finding an approximate matching in a hypergraph where the edges arrive one by one.

- Obtain a semi-streaming algorithm for maximum cardinality matching with approximation ratio $1/r$ where r is the rank.
- Obtain an $\Omega(1/r^2)$-approximation for the weighted case.
- **Extra credit:** Obtain an $\Omega(1/r)$-approximation for the weighted case. You can skip the previous two parts if you do this.

Problem 7. In a turnstile stream updating a vector $x \in \mathbb{R}^n$ starting as the 0 vector, an *ϵ-error ℓ_1 sampler* is a streaming algorithm that when queried outputs a pair $(i, \hat{x}_i)$ such that i is output with probability $|x_i|/\|x\|_1$ and $\hat{x}_i = (1 \pm \epsilon)x_i$ (recall that in turnstile streams, each stream update is of the form $x_i \leftarrow x_i + v$ where v can be positive or negative). Pretend we have such an ℓ_1 sampler using space $S(n, \epsilon)$. Now consider the following problem: you see a stream $i_1\ i_2\ \cdots\ i_{n+1}$ with each $i_j \in [n]$. This stream must have at least one duplicate entry due to the pigeonhole principle. Show how to use a 1/2-error ℓ_1 sampler to give a one-pass streaming algorithm that reports at least one duplicate index $i \in [n]$ with probability at least $1 - \delta$. The space of your algorithm should be $O(S(n, 1/2) \cdot \log(1/\delta))$.

Problem 8. We have seen streaming algorithms for ϵ-approximate quantiles. We defined an ϵ-approximate quantile for a quantile $\phi \in (0, 1]$ as an element of rank r where $\phi n - \epsilon n \leq r \leq \phi n + \epsilon n$ where n is the number of elements. We define a stronger notion of ϵ-approximate quantiles where we wish to return an element of rank r where $(1 - \epsilon)\phi n \leq r \leq (1 + \epsilon)\phi n$. Describe how to compute an ϵ-approximate quantile summary for this stronger notion of approximation.

Problem 9. Project. Pick a paper or article or a book that addresses "big data" and "algorithms" (loosely defined) and write a two page report. It can focus on technical aspects or a reflection on certain trends in society or technology. You choose what to do but make it a sincere effort combined with good writing.