Homework 3

Algorithms for Big Data: CS498 ABD/ABG, Fall 2020
Due: Wednesday at 10pm CDT, 21st October 2020

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. Frequent items and Misra-Greis Algorithm. We saw the deterministic Misra-Greis algorithm that uses k counters and outputs an estimate $\hat{f}_i$ for each f_i such that $f_i - \frac{m}{k+1} \leq \hat{f}_i \leq f_i$. Here m is the total number of elements in the stream.

- Let m' be the sum of the counters at the end of the algorithm. Show that the actual estimate provided by the algorithm is slightly stonger, namely, for each i,

$$f_i - \frac{m - m'}{k+1} \leq \hat{f}_i \leq f_i.$$

- Suppose we ran the (one-pass) Misra-Gries algorithm on two streams σ_1 and σ_2 thereby obtaining a summary for each stream consisting of k counters. Consider the following algorithm for merging these two summaries to produce a single k-counter summary.

  1. Combine the two sets of counters, adding up counts for any common items.
  2. If more than k counters remain:
     (a) $c \leftarrow$ value of $(k+1)$th counter, based on decreasing order of value.
     (b) Reduce each counter by c and delete all keys with non-positive values.

  Prove that the resulting summary is good for the combined stream $\sigma_1 \cdot \sigma_2$ (concatenation of the two streams) in the sense that frequency estimates obtained from it satisfy the bounds given in the previous part.

Problem 2. 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 $||x||_2$ is 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 $x_1 \geq x_2 \ldots \geq x_n$. For some parameter $\alpha > 1$ that characterizes the distribution we have $x_k \sim 1/k^{\alpha}$. Calculate $||y_\ell||$ for a vector x which follows the Zipfian distribution with $||x||_1 = m$.

Problem 3. JL preserves angles. 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 Π preserves the dot product between u and v to within a ϵ-additive factor with probability $1 - \delta$ with a slight increase in dimensions.
2. Show that with a slight increase in dimension, Π preserves the angle between any two vectors u, v up to a $\pm\epsilon$-additive factor. *Hint:* Taylor expansion.

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 dimenstion 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 $Q_\gamma = \{w : w \in \frac{\gamma}{\sqrt{d}}\mathbb{Z}^d, \|w\|_2 \le 1\}$ for $\gamma \in (0,1)$. Prove $|Q_\gamma| \le e^{d \cdot f(\gamma)}$ 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 $|u^T A v| \le \epsilon$ for all $u, v \in Q_\gamma$, then $|x^T A x| \le \epsilon/(1-\gamma)^2$ 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.