Homework 3

Algorithms for Big Data
CS498ABD Spring 2019
Due: 10am, Wednesday, March 27th

Instructions:

- Each home work can be done in a group of size at most two. Only one home work 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 class mates. Write in your own words. See the site mentioned in the preceding item for more detailed policies.

Exercise 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 have run 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 increasing 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.

Exercise 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 $\ell \in \mathbb{N}$, 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

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

Exercise 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 constant probability if the number of dimensions in the projection is $\Omega\big(1/\epsilon^2\big)$.

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 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 $(1 \pm \epsilon)$-multiplicative factor.

Hint: Taylor expansion...

Exercise 4. TBD.