Homework 3

CS 574: Randomized Algorithms, Fall 2025
Due: Tuesday, Oct 28th 2025 at noon

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.grainger.illinois.edu/cs374al1/sp2025/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. 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 3. We saw the simple and elegant CMV algorithm for estimating the number of distinct elements in a stream. This is based on adaptively maintaining a sample of size $O(\frac{1}{\epsilon^2} \log n)$. In a separate lecture we saw a simple sampling based algorithm for estimating the number of solutions for a DNF formula or in a similar fashion to estimate the area of the union of some shapes. The simplicity of the CMV algorithm allows one to adapt it to count the number of solutions to a DNF formulat in the streaming setting. We will instead work with estimating the area of the union of axis-parallel rectangles since the descriptive complexity of specifying a rectangle in the plan is $O(1)$ (coordinate of a corner and width and height). Formally, suppose you are given a stream of n axis-parallel rectangles in the plane where element i in the stream is a rectangle R_i specified in some natural way. Your goal is to estimate the area of the union of the rectangles $R_1, R_2, \ldots, R_n$ to within a $(1-\epsilon)$-approximation with high probability by using only poly-logarithmic space.

Problem 4. Let $G = (V, E)$ be a simple (no parallel edges), unweighted, directed, and strongly connected graph. A random walk on this graph corresponds to an irreducible finite state Markov chain and we showed that it has a uniqute stationary distribution π where $\pi_i = 1/h_{ii}$ for each $i \in V$ and this is positive. Prove that $\pi_i \geq \frac{1}{n^{(n+1)}}$ for all $i \in V$ where $n = |V|$ is the number of nodes.