CSE 521: Design and Analysis of Algorithms **Fall 2022**

Problem Set 2

Deadline: Oct 30nd (at 11:59 PM) in gradescope

In solving these assignments and any future assignment, feel free to use these approximations:

$$1 - x \approx e^{-x}, \qquad n! \approx (n/e)^n, \qquad \left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

1) Let U be a universe. A family of hash functions $\mathcal{H} := \{h : U \to \{-1, +1\}\}$ is a sketching hash family with error $\epsilon > 0$ if for any function $f : U \to \mathbb{R}$, which is not identically zero,

$$\mathbb{P}\left[\sum_{i \in U} f(i)h(i) = 0\right] \leq \epsilon.$$

The value $sk_h(f) := \sum_{i \in U} f(i)h(i)$ is called the *sketch* of f.

a) Prove that if $f, f' : U \to \mathbb{R}$ are different functions, then

$$\mathbb{P}\left[sk_h(f) = sk_h(f')\right] \leq \epsilon.$$

b) Suppose $\mathcal{H}$ is the family of *all* functions from U to $\{-1, +1\}$. Prove that it is a sketching family with error $\epsilon = 1/2$.

c) Suppose that $\mathcal{H}$ is a family of 4-wise independent hash functions. Prove that it is a sketching family with error $\epsilon = 2/3$. In this part you can use the Paley-Zygmund inequality:

Theorem 2.1 (Paley-Zygmund Inequality)**.** *If $Z \geq 0$ is a R.V. and $0 \leq \alpha \leq 1$, then*

$$\mathbb{P}\left[Z > \alpha \mathbb{E}[Z]\right] \geq (1 - \alpha)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

2) Say we have a sequence of number $X_0, \ldots, X_{n-1} \in \{0, \ldots, n-1\}$; also let $f_i = \sum_{j=0}^{n-1} \mathbb{I}[X_j = i]$ for all $0 \leq i \leq n-1$. Given an $\epsilon > 0$, we want to output all indices i such that $f_i \geq \epsilon n$ (with high probability). You can use memory at most $O(\frac{1}{\epsilon^2} \log^C(n))$ for any constant $C > 0$, i.e., it is ok if your algorithm uses $1000 \log^{100} n/\epsilon^2$ amount of memory. The running time of your algorithm is not limited and it can depend on n. Note that this is a streaming problem and you get to read the input only once. With probability $1 - 1/n$ your algorithm should

(a) output all i such that $f_i \geq \epsilon n$, and

(b) any i in the output of your algorithm should satisfy $f_i \geq \epsilon n/2$.

Hint: First use a single hash table with a test where any i with $f_i \geq \epsilon n$ passes the test with probability 9/10 and any i where $f_i < \epsilon n/2$ fails the test with probability 9/10. Then use the median trick (and multiple hash tables) to boost these these probabilities to $1 - 1/n^2$. Finally use a union bound.

3) Let $u, v \in \mathbb{R}^d$ and $g \in \mathbb{R}^d$ be a random Gaussian vector, i.e., for each $1 \leq i \leq d$, $g_i \sim \mathcal{N}(0, 1)$.

a) What is the expected value of $\langle g, u \rangle$?

b) What is the expected value of $\langle g, u \rangle \cdot \langle g, v \rangle$?

c) What is the expected value of $|\langle g, u\rangle|$? You can use that p.d.f. of a $\mathcal{N}(0,1)$ $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

d) Consider the following hash function: $h_g(u) = \mathrm{sgn}(\langle g, u\rangle)$, where sgn is the sign function, i.e.,

$$\mathrm{sgn}(a) = \begin{cases} +1 & \text{if } a \geq 0 \\ -1 & \text{otherwise.} \end{cases}$$

Show that for a random Gaussian vector g and any two vectors u, v, $\mathbb{P}\left[h_g(u) = h_g(v)\right] = 1 - \frac{\theta(u,v)}{\pi}$ where $\theta(p, q)$ is the angle between the vector of p and q.

e) Let $P \subseteq \mathbb{R}^d$ and consider the following distance function: $\mathrm{dist}(p, q) = \frac{\theta(p,q)}{\pi}$. For what values of p_1 and p_2 is this family of functions $(r, c \cdot r, p_1, p_2)$-sensitive?

4) The permanent of an $n \times n$ matrix A is

$$per(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} A_{i,\sigma_i},$$

where the sum is over all permutations of the numbers $\{1, \ldots, n\}$. For example,

$$\mathrm{per}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad + bc.$$

a) Let $A = xx^T$ be a symmetric rank 1 matrix, for some $x \in \mathbb{R}^n$. Show that

$$\mathrm{per}(A) = n! \prod_{i=1}^{n} |x_i|^2$$

b) Let g be random Gaussian vectors in $\mathbb{R}^n$. Find $\mathbb{E}\mathrm{per}(gg^T)$.

c) Use Chebyshev's bound to prove a concentration bound on $\mathrm{per}(gg^T)$ around its expectation.

5) In this part you are supposed to implement a streaming algorithm which estimates F_2 of a given sequence of integers in the range $0, \ldots, 2^{48} - 1$ within 1 ± 0.1-multiplicative error. Implement the median trick and output your estimate. You can use that "28787381843723" as a prime in coding a 4-wise independent hash function. There will be 3 inputs to the problem; in each input you will be given a long sequence of integers in the range $0, \ldots, 2^{48} - 1$. Please upload your code together with its output.

6) **Extra Credit:** A random walker is walking on a circle with n vertices $\{1, \ldots, n\}$; it starts at vertex 1 and at each time step it moves to one of the two neighbors uniformly at random. It stops the moment that it visits every vertex at least once. What is the probability that the last vertex that it visits is $n/2$?