CSE 521: Design and Analysis of Algorithms	**Fall 2023**

Problem Set 4

Deadline: Oct 30th in Canvas

1) A probability distribution D_p over $\mathbb{R}$ is said to be p-stable if for $Z, Z_1, \ldots, Z_n$ independently drawn from D_p and for any fixed $x \in \mathbb{R}^n$, the random variable $\sum_{i=1}^n x_i Z_i$ is equal in distribution to $\|x\|_p \cdot Z$. Some examples are the standard normal distribution $N(0,1)$, which is 2-stable. Another less-known example is the Cauchy distribution, which is 1-stable; it has probability density function $\Phi(x) = 1/(\pi(1+x)^2)$. It is a known theorem that such distributions exist iff $p \in (0,2]$. Note that p-stable random variables for $p \neq 2$ cannot have bounded variance, since otherwise the sum of independent copies would have to be gaussian by central limit theorem. In fact, it is known that any p-stable have **bounded and continuous** density function and they must have tail bounds $P(|Z| > \lambda) = O(1/(1+\lambda)^p)$ for all $\lambda > 0$. This implies that such distributions cannot exist for $p > 2$ (since otherwise they would have bounded variance, violating the central limit theorem).

   a) Suppose Z is p-stable; show that for any $\alpha > 0$, αZ is also p-stable.

   b) Let Z be a p-stable random variable normalized (by a constant) so that $\mathbb{P}[Z \in [-1,1]] = 1/2$ (see previous part). Fix some $\epsilon > 0$. Show that there is a constant $c > 0$ (as a function of ϵ, p such that

$$\mathbb{P}[-1+\epsilon < Z < 1-\epsilon] \leq 1/2 - c\epsilon,$$
$$\mathbb{P}[-1-\epsilon < Z < 1+\epsilon] \geq 1/2 + c\epsilon.$$

   c) Let $P \in \mathbb{R}^{m \times d}$ where $P_{i,j}$ is an independent sample of Z. Let $x \in \mathbb{R}^d$ arbitrary and $y = Px$; show that for $m = O(\log(1/\delta)/\epsilon^2)$, with probability at least $1-\delta$, the median of $|y_1|, \ldots, |y_m|$ is a $1 \pm \epsilon$ multiplicative approximation of $\|x\|_p$.

   d) Implement the algorithm in the previous part and use it to estimate the $\|x\|_1$ of the vector x given in the p4.in file in the website (will upload soon). Insert your code together with the value of m and ϵ that you use, $\|x\|_1$ and the output of your code.

   Note: Although we are not going to discuss it here, this idea can be used together with a family of k-wise independent hash functions to design streaming algorithm with poly-log memory to estimate the p-norm for $p < 2$.

2) In this problem we design an LSH for points in $\mathbb{R}^d$, with the ℓ_1 distance, i.e.

$$d(p,q) = \sum_i |p_i - q_i|.$$

   a) Let a, b be arbitrary real numbers. Fix $w > 0$ and let $s \in [0, w)$ chosen uniformly at random. Show that

$$\mathbb{P}\left[\left\lfloor \frac{a-s}{w} \right\rfloor = \left\lfloor \frac{b-s}{w} \right\rfloor\right] = \max\left\{0, 1 - \frac{|a-b|}{w}\right\}.$$

   Recall that for any real number c, $\lfloor c \rfloor$ is the largest integer which is at most c.

   Hint: Start with the case where $a = 0$.

   b) Define a class of hash functions as follows: Fix w larger than diameter of the space. Each hash function is defined via a choice of d independently selected random real numbers $s_1, s_2, \ldots, s_d$, each uniform in $[0, w)$. The hash function associated with this random set of choices is

$$h(x_1, \ldots, x_d) = \left(\left\lfloor \frac{x_1 - s_1}{w} \right\rfloor, \left\lfloor \frac{x_2 - s_2}{w} \right\rfloor, \ldots, \left\lfloor \frac{x_d - s_d}{w} \right\rfloor \right).$$

Let $\alpha_i = |p_i - q_i|$. What is the probability that $h(p) = h(q)$ in terms of the α_i values? For what values of p_1 and p_2 is this family of functions $(r, c \cdot r, p_1, p_2)$-sensitive? Do your calculations assuming that $1 - x$ is well approximated by e^{-x}.