Lecturer: Thomas Rothvoss Due date: Friday, Oct 25, 2024, 11pm, on GradeScope

Problem Set 4

CSE 521 - Design and Analysis of Algorithms

Fall 2024

Exercise 1 (10pts)
For $1 \leq p < \infty$ and $x \in \mathbb{R}^n$, $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ is the $\|\cdot\|_p$-norm which generalizes the Euclidean norm $\|\cdot\|_2$ and Manhattan norm $\|\cdot\|_1$ from the lecture. A probability distribution $\mathcal{D}_p$ over $\mathbb{R}$ is said to be p-stable if for $Z, Z_1, \dots, Z_n$ independently drawn from $\mathcal{D}_p$ 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 distributions have *bounded and continuous* density function and they must have tail bounds $\Pr(|Z| > \lambda) = O(1/(1+\lambda)^p)$ 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).

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

ii) **Optional (0 points).** Let Z be a p-stable random variable normalized (by a constant) so that $\Pr[Z \in [-1,1]] = 1/2$ (see previous part). Fix some $\varepsilon > 0$. Show that there is a constant $c > 0$ (as a function of p but not ε) such that

$$\Pr[-1+\varepsilon < Z < 1-\varepsilon] \leq 1/2 - c\varepsilon,$$
$$\Pr[-1-\varepsilon < Z < 1+\varepsilon] \geq 1/2 + c\varepsilon.$$

iii) Let Z be a p-stable random variable with $\Pr[Z \in [-1,1]] = \frac{1}{2}$. 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 $m = O(\log(1/\delta)/\varepsilon^2)$, with probability at least $1-\delta$, the median of $|y_1|, \dots, |y_m|$ is a $1 \pm \varepsilon$ multiplicative approximation of $\|x\|_p$.

Hint. It might be helpful to review the argument in Theorem 1.37.

Exercise 2 (10pts)
In this problem we design a *locally sensitive hash function (LSH)* for points in $\mathbb{R}^d$, with the ℓ_1 distance, i.e.

$$d(p,q) = \sum_{i=1}^d |p_i - q_i|.$$

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

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

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 $a = 0$.

ii) 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? Use the approximation $1 - x \approx e^{-x}$ (just ignore the error term for the sake of simplicity).