Exercise Set X

These exercises are for your own benefit. Feel free to collaborate and share your answers with other students. Solve as many problems as you can and ask for help if you get stuck for too long. Problems marked * are more difficult but also more fun :).

These problems are taken from various sources at EPFL and on the Internet, too numerous to cite individually.

1 **LSH for Jaccard similarity.** Suppose we have a universe U. For non-empty sets $A, B \subseteq U$, the Jaccard index is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Design a locality sensitive hash (LSH) family $\mathcal{H}$ of functions $h : 2^U \to [0,1]$ such that for any non-empty sets $A, B \subseteq U$,

$$\Pr_{h \sim \mathcal{H}}[h(A) \neq h(B)] \begin{cases} \leq 0.01 & \text{if } J(A,B) \geq 0.99, \\ \geq 0.1 & \text{if } J(A,B) \leq 0.9. \end{cases}$$

(In this problem you are asked to explain the hash family and argue that it satisfies the above properties. Recall that you are allowed to refer to material covered in the course.)

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

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

Define a class of hash functions as follows: Fix a positive number w. 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? It may be easier to first think of the case when $w = 1$. Try to also simplify your expression if w is much larger than α_i's, using that $(1-x) \approx e^{-x}$ for small values of $x \geq 0$.

3 Consider two LSH hash families $\mathcal{H}_1$ and $\mathcal{H}_2$ designed for a distance function $\mathrm{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. For $r = 0.1$ and $c = 2$, $\mathcal{H}_1$ satisfies

$$\mathrm{dist}(p,q) \leqslant r \implies \Pr_{h \sim \mathcal{H}_1}[h(p) = h(q)] \geqslant 1/2$$
$$\mathrm{dist}(p,q) \geqslant c \cdot r \implies \Pr_{h \sim \mathcal{H}_1}[h(p) = h(q)] \leqslant 1/8$$

and $\mathcal{H}_2$ satisfies

$$\mathrm{dist}(p,q) \leqslant r \implies \Pr_{h \sim \mathcal{H}_2}[h(p) = h(q)] \geqslant 1/8$$
$$\mathrm{dist}(p,q) \geqslant c \cdot r \implies \Pr_{h \sim \mathcal{H}_2}[h(p) = h(q)] \leqslant 1/200$$

3a Which Hash family would you choose to build the data structure $\mathrm{ANNS}(r, c)$ explained in class? What would the space requirement and query time be (logs are not so important)?

3b On query $q \in \mathbb{R}^d$, asymptotically how many hash function computations are done?

4 Suppose you have a database with a set $P \subseteq \mathbb{R}^d$ of n items that are equipped with a distance function $\mathrm{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfying the following sparsity condition:

$$|\{p \in P : \mathrm{dist}(p,q) \leq 2\}| \leq 10\,.$$

Further assume that you have a $(r, c \cdot r, p_1, p_2)$-LSH hash family $\mathcal{H}$ for the considered distance function with parameters $r = 1$, $c = 2$, $p_1 = 1/2$ and $p_2 = 1/8$. That is,

$$\mathrm{dist}(p,q) \leqslant 1 \implies \Pr[h(p) = h(q)] \geqslant 1/2$$
$$\mathrm{dist}(p,q) \geqslant 2 \implies \Pr[h(p) = h(q)] \leqslant 1/8$$

where the probabilities are over $h \sim \mathcal{H}$.

Exploit the sparsity condition to modify the $\mathrm{ANNS}(c, r)$ construction seen in class so as to obtain a structure with the *same* asymptotic preprocessing and query times, but with the following improved guarantee:

> On query $q \in \mathbb{R}^d$, if $\min_{p \in P} \mathrm{dist}(p,q) \leq 1$, then we return $\arg\min_{p \in P} \mathrm{dist}(p,q)$ with probability close to 1.

(Notice that this is stronger than the guarantee seen in class as in that case one is only guaranteed to return a point p' such that $\mathrm{dist}(p', q) \leq c \cdot r$ with probability close to 1.)

What is the preprocessing time, query time, and space requirement of your solution?