CSE 521: Design and Analysis of Algorithms **Winter 2017**

Midterm

Deadline: Feb 10th (at 12:00PM) in Canvas

In this assignment you are not allowed to collaborate. But, you may contact the instructor or TA for hints. We will study locally sensitive hash functions in more depth. We also have one question related to the Shwartz-Zippel Lemma.

1) 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 $a = 0$.

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

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

Define a class of hash functions as follows: Fix $w \gg r$. 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? You can do your calculations assuming that you are in a regime where $1 - x$ is well approximated by e^{-x}.

3) In this problem we design an LSH for points in $\mathbb{R}^d$ with the ℓ_2 distance function,

$$d(p, q) = \|p - q\|_2 = \sqrt{\sum_i (p_i - q_i)^2}.$$

Let $w \gg r$, and let s be uniformly distributed in $[0, w)$. Let g be a d-dimensional Gaussian vector, i.e., for all $1 \le i \le d$, $g_i \sim \mathcal{N}(0, 1)$ and all coordinates of g are chosen independently. Consider the hash function

$$h(p) = \left\lfloor \frac{\langle g, p \rangle - s}{w} \right\rfloor$$

a) Show that for any two points p, q, $\langle g, p \rangle - \langle g, q \rangle$ is distributed as a normal random variable. What is the mean and variance of this random variable? In this part you can use the fact that any linear combination of independent normal random variables is also a normal random variable.

b) Use Problem (1) to estimate the probability that $h(p) = h(q)$. Note that this probability is over the randomness of g and s. In this part you can use the fact that for a random variable $X \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}[|X|] = \sigma\sqrt{2/\pi}$. To make calculations simple, assume that w is large enough such that $\mathbb{P}[|\langle g, p \rangle - \langle g, q \rangle| > w] = 0$.

c) Use the statement of part (b) to determine for what values of p_1, p_2, is this family $(r, c \cdot r, p_1, p_2)$ sensitive?

4) In this problem we design an algorithm to estimate the size of the largest matching of a bipartite graph. Recall that for a matrix A, $\text{rank}(A)$ is the number of linearly independent columns of A; it is also the same as the number of linearly independent rows of A. It turns out that for any matrix $A \in \mathbb{R}^{n \times n}$, $\det(A) \neq 0$ if and only if $\text{rank}(A) = n$. Let $G = (X, Y, E)$ be a given bipartite graph. Using the above terminology, we can rewrite the algorithm that tests whether G has a perfect matching as follows: For each edge (x_i, y_j) of G, choose $A_{i,j}$ uniformly and independently from the set $\{0, 1, \ldots, n^2\}$, and let the rest of entries of A be 0. Return yes if $\text{rank}(A) = n$ and no otherwise. Since $\det(A)$ is a polynomial of degree n, you will use the Shwartz-Zippel Lemma to show this algorithm succeeds with probability $1 - 1/n$.

a) Let A be the following matrix: For each nonadjacent pair x_i, y_j, let $A_{i,j} = 0$; choose the rest of the entries of A arbitrarily. Show that if $\text{rank}(A) = k$, then G has a matching of size at least k.

b) Now, suppose for any edge (x_i, y_j) we choose $A_{i,j}$ uniformly and independently from $\{0, 1, \ldots, n^2\}$, and we let the rest of the entries be 0. Show that with high probability, $\text{rank}(A)$ is equal to the size of the largest matching of G.

5) **Extra Credit.** In this problem you are supposed to implement the NNS algorithm for the hamming distance. You are given n points $P \subseteq \{0, 1\}^d$ that you are supposed to preprocess and store based on the algorithm that we discussed in class. Then, you will be given t query points; for each query point you need to find a point at distance no more than twice the closest point.

In the input file lsh.in you are given n, d, t in this order. The input is followed by points of P, the $i+1$-st row of the input is contains the i-th point of P. Then, the input is followed by query points (so the $n + 1 + i$-th row of the input has the i-th query point). In the i-th line of the output, write the index of the point P that is closest to the i-th query point. Please submit your code together with the output to Canvas.