CSE 521: Design and Analysis of Algorithms **Fall 2018**

Problem Set 2

Deadline: Oct 29th in Canvas

1) Suppose we have a universe U of elements. For $A, B \subseteq U$, the Jaccard distance of A, B is defined as

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

This definition is used practice to calculate a notion of similarity of documents, webpages, etc. For example, suppose U is the set of English words, and any set A represents a document considered as a bag of words. Note that for any two $A, B \subseteq U$, $0 \leq J(A,B) \leq 1$. If $J(A,B)$ is close to 1, then we can say $A \approx B$.

(a) Let $h : U \to [0,1]$ where for each $i \in U$, $h(i)$ is chosen uniformly and independently at random. For a set $S \subseteq U$, let $h_S := \min_{i \in S} h(i)$. Show that

$$\mathbb{P}[h_A = h_B] = J(A,B).$$

(b) Now, suppose we have sets $A_1, A_2, \ldots, A_n$, we can use the above idea to output the Jaccard similarity of all pairs of sets. In the input files j1.in, j2.in, j3.in, j4.in you are given the description of n sets. The first line of the the input contains n followed by $|U|$. The elements in each set are a subset of $\{1, \ldots, |U|\}$. In the next n lines, each line has the list of numbers in one of the sets. For all $1 \leq i, j \leq n$, in the $n(i-1)+j$ line of the output you should write the Jaccard similarity of the i-th and j-th set within 1 ± 0.1 multiplicative error, except for j4.in for which it is enough to write down the Jaccard similarity within 0.25-*additive* error. The input file h4.in has only 10 percent of the grade. Below you can see a sample input and output files. Upload your code together with all output files to Canvas. You will receive full grade of each test case as long as you get 90 percent of the numbers correct.

```
j0.in
3 6
1 6 4
3 2 6
1 2 4
```

```
j0.out
0.21
0.49
0.2
```

Note that the correct Jaccard distances are $0.2, 0.5, 0.2$ but it is enough to estimate the distance within 1 ± 0.1 multiplicative error, so you may output 0.21 instead of the correct distance of 0.2. Note that the naive algorithm would take $O(n^2|U|)$ to calculate all pairwise similarities.

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}.

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: $h_g(u) = \text{sgn}(\langle g, u \rangle)$, where sgn is the sign function, i.e., $\text{sgn}(a) = 1$ if $a \geq 0$ and $\text{sgn}(a) = -1$ otherwise. Show that for a random Gaussian vector g and any two vectors u, v, $\mathbb{P}[h_g(u) = h_g(v)] = 1 - \frac{\theta(u,v)}{\pi}$ 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: $\text{dist}(p, q) = \frac{\theta(p,q)}{\pi}$. For what values of p_1 and p_2 is this family of functions $(r, c \cdot r, p_1, p_2)$-sensitive? You can do your calculations assuming that $1 - x$ is well approximated by e^{-x}.

4) Describe an example (i.e., an appropriate set of points in $\mathbb{R}^n$) that shows that the Johnson-Lindenstrauss dimension reduction method, the linear transformation obtained by projecting on Gaussian vectors scaled properly, does not preserve ℓ_1 distances within even factor 2.