Problem Set

Due: July 3rd at 9:30 AM

1 Instructions

Please cite all the references (including web-sites, names of friends, etc.) which you have used/consulted as the source of information for each of the questions. BTW, when a question asks you to design an algorithm - it **requires** you to (1) Clearly spell out the **steps** of your algorithm in pseudo code (2) **Prove** that your algorithm is correct and (3) **Analyze** the running time. You can submit a handwritten document.
Caution: Start to work on this problem set as soon as possible and please start asking questions if a problem is unclear.

2 Problems

1. Show that for any two random variables X and Y, $E[X+Y] = E[X] + E[Y]$.

2. The l be a 'random' line which has been segmented into intervals of length 2. Show that when the distance between two points in the plane is at most 1, then with probability at least $1/2$, they will project within the same interval on l.

3. Let A be an array of length n consisting of the majority element, i.e. an element that occurs more than $\lceil \frac{n}{2} \rceil + 1$ times. Design an algorithm to find this element that uses only a constant amount of memory space.

4. Assume that a cereal box company is running a promotion. Each cereal box contains one of the cars (with equal chances) among a collection of 100 different types of toy cars. What is the expected number of different types of cars we would have collected if we buy only 150 boxes. Compute this by hand or via (computer/R) simulation.

5. (Refer to Section on Fingerprint matching in the MMDS textbook.) Recall the algorithm for the fingerprint matching using the minutia and the normalized grid. Assume that for any fingerprint there is a 20% chance that a minutia appears in a grid cell. Moreover, assume that there is a 90% chance that if a fingerprint has a minutia in a grid cell, then another copy of that fingerprint will have minutia in the same grid cell. To construct the sensitive family of functions F, each function $f \in F$ is constructed as follows. For each hash function f, we associate a bucket B_f. We select four random grid squares, and define a function f on these grid squares. A fingerprint is mapped to B_f by f if and only if it has minutia in those four specific grid squares. Two fingerprints are said to be similar with respect to f if both of them are mapped to B_f. Answer the following:

(a) Suppose we use 1000 functions from the family F and construct a new 1000-way OR family F_1. What is the probability that the two fingerprints of the same finger will be hashed together in at least one of the buckets among these 1000 functions. What is the probability that the fingerprints from two different fingers will be hashed together in at least one of the buckets? What are the false-positive and false-negative rates of F_1.

(b) Do the same analysis when we use 2000-way OR family instead of 1000-way OR family.

(c) Suppose we have two 1000-way OR families F_1 and F_2. Construct a 2-way AND family F_3 of F_1 and F_2. Compute the false-negative and the false-positive rates of F_3.

(d) Compare the false-positive and the false-negative rates and the running times of the three scenarios (a), (b), and (c). What will be your recommendation to a user of this system?

6. Let us assume that we have a large collection B of binary vectors in dimension $d = 10,000$. We are asked to compute a data structure so that the following queries can be answered efficiently. Given any query binary vector q in dimension d, we are interested to report all the binary vectors in B that are approximately 95% similar to q. We say that two vectors $a = a_1a_2 \dots a_d$ and $b = b_1b_2 \dots b_d$ are 95% similar if $a_i = b_i$ for at least 95% of indices i, $1 \leq i \leq d$. Design an algorithm that computes such a data structure and show how each query can be answered efficiently. The time to answer the query q should not exceed $O((k+1)d)$, where k is the number of vectors in B that are at least 95% similar to q. It is fine if you have some false positives and negatives, but their percentage shouldn't be large.

7. Assume that a very large data stream S consists of elements from a universe U. Each element in S has the property that it may occur at most twice. Let s_1 be the count of the number of elements in S that occur exactly once. Similarly, let s_2 is the total count of the elements that occur exactly twice in S. To count the number of distinct elements in the stream S you may take a sample $S' \subset S$, say each element of S is chosen uniformly at random with probability $0 < p < 1$. Let the count of the number of elements in S' that occur exactly once be s'_1 and the number of elements that occur exactly twice be s'_2. Prove or disprove each of the following statements. In case you think the statement is wrong, provide the correct analysis.

(a) $E[s'_1] = ps_1$

(b) $E[s'_1] = ps_1 + 2p(1-p)s_2$

(c) $E[s'_2] = ps_2$

(d) $E[s'_2] = p^2 s_2$

(e) $\frac{s_2}{s_1+s_2} = \frac{s'_2}{s'_1+s'_2}$

8. (Bonus Question) Consider the LSH sensitive family for similarity of sets. Assume that we have S-curves corresponding to the following two scenarios:

(a) First an r-way AND construction followed by a b-way OR construction. (This is what we did in the class.)

(b) A b-way OR construction followed by an r-way AND construction.

We are interested to know where is the steepest rise in the S-curve. Note that the steepest rise is where the slope of the tangent to the S-curve is largest. Can we express the point where the steepest rise occurs as a function of r and b. Consult Exercises 3.6.3 and 3.6.4 in the MMDS textbook.