Homework 3 (due 2/26)

DS-563 / CD-543 @ Boston University

Spring 2024

Before you start...

Collaboration policy: You may verbally collaborate on required homework problems, however, you must write your solutions independently. If you choose to collaborate on a problem, you are allowed to discuss it with **at most three** other students currently enrolled in the class.

The header of each assignment you submit must include the field "Collaborators:" with the names of the students with whom you have had discussions concerning your solutions. A failure to list collaborators may result in credit deduction.

You may use external resources such as textbooks, lecture notes, and videos to supplement your general understanding of the course topics. You may use references such as books and online resources for well known facts. However, you must always cite the source.

You may **not** look up answers to a homework assignment in the published literature or on the web. You may **not** share written work with anyone else.

Submitting: Solutions should be submitted via Gradescope (entry code: 6G4V6G). Your solutions should be typed. It is strongly suggested to use LaTeX.

Grading: Whenever we ask for an algorithm (or bound), you may receive partial credit if the algorithm is not sufficiently efficient (or the bound is not sufficiently tight).

Questions (50 points overall)

Submit solutions to **six** (6) arbitrary questions out of Questions 1–7 and answer Question 8. (If you submit answers to seven questions out of Questions 1–7, you may receive credit for an arbitrary subset of six of them.)

1. Recall the Misra–Gries algorithm. Answer the following questions. Explain your answers.

   (a) Can it handle deletions?

   (b) Can the results of computing the algorithm on disjoint subsets be combined?

   (c) Can it be (directly) represented as a linear sketch?

   (d) Is it adversarially robust? That is, can it handle adaptive streams?

2. Recall that in our algorithm for estimating the number of distinct elements, we used a random function $h : X \to \mathbb{N}$ such that for any $x \in X$ and any $i \in \mathbb{N}$, $\Pr(h(x) = i) = 2^{-(i+1)}$. We now construct a function that can play the role of h, assuming that we have a hash function $g : X \to [2^k]$, where $[2^k] = \{1, 2, 3, \ldots, 2^k\}$, with the following properties:

   - $k = \lceil 2 \log m \rceil$ and m is (an upper bound on) the length of the input stream.
   - g is pairwise independent.
   - Each $g(x)$ is distributed uniformly on $[2^k] = \{1, 2, 3, \ldots, 2^k\}$.

   Define $h'(x) = \max\{i \in \mathbb{N} : g(x) \text{ is divisible by } 2^i\}$ for all x.

   (a) Is h' pairwise independent? Explain why.

   (b) What is the probability distribution of each $h'(i)$ and how does it differ from the distribution of h that we used in class?

   (c) Argue why using h' instead of h gives essentially the same results, especially for sufficiently large m.

3. Design streaming algorithms for the following problems and analyze their space complexity:

   (a) Suppose that the input stream is a sequence of updates to an initially empty multiset $S \subseteq \mathbb{Z}$. Each update is of the form either "insert a copy of x into S" or "delete a copy of x from S." You are promised that at the end of the stream, there will be exactly one element in S. Design a small space streaming algorithm that outputs this element.

   (b) Suppose that the input stream is a sequence of integers in the range $[n] = \{1, \ldots, n\}$ and you are promised that all of them appear exactly once, except for one of them that appears twice. Design a small space streaming algorithm that outputs the number that appears twice.

4. (a) We updated the handout with useful probabilistic inequalities on the course webpage. Read the section titled "Collisions (the Birthday Paradox)."

   (b) Go to `https://www.xe.com/currencytables/` or any similar webpage. Select an arbitrary pair of currencies and check their exchange rate yesterday. Look at the two least significant digits. (Example: If the exchange rate is `3.523432`, then these digits are `32`.) Then look at the exchange rate two days ago, three days ago, and so on, always restricting your attention to the two least significant digits. How many days do you have to look back to see two days on which the exchange rate's two least significant digits are the same?

   (c) Repeat this experiment for four more pairs of currencies. Tell us what the results of your experiments are. Do you think these results match the theory, assuming that the two least significant digits are distributed uniformly?

5. Consider an experiment in which you select $n > 2$ points uniformly independently from $[0, 1]$. Prove that there is a constant $c > 0$ such that two of the points are at distance at most c/n^2 with probability at least $1/2$.

6. (a) Read Section "Hoeffding's Inequality" in the updated handout with useful probabilistic inequalities on the course webpage.

(b) Let p and q be such that $0 \leq p < q \leq 1$. Suppose that there are two kinds of coins that look exactly the same, but coins of the first kind come up heads with probability at most p and coins of the second kind come up heads with probability at least q. You know both p and q. Imagine that someone gives you a coin and you want to find out whether it is a coin of the first or second kind. Use Hoeffding's inequality to show that you can determine the kind of coin correctly with probability at least $1 - \delta$, for any $\delta \in (0, 1/2)$, by tossing it $O((q-p)^{-2} \log(1/\delta))$ times.

7. Suppose that you know how to draw samples from $\mathcal{N}(0, 1)$, i.e., the Gaussian distribution. First, select a k–dimensional vector v with each coordinate drawn independently from $\mathcal{N}(0, 1)$. Then normalize it, i.e., divide it by $\|v\|$. Why does this select uniformly at random a point on the k–dimensional unit sphere?

   Hint: Look at the probability density function of v. Recall that the probability density function of $\mathcal{N}(0, 1)$ is $\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

8. How much time (approximately) did you spend on this homework? Was is too easy/too hard?