Homework 1 (due 2/5)

DS-563 / CD-543 @ Boston University

Spring 2025

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. You may **not** use LLMs (such as ChatGPT or Claude) to find solutions for homework problems or to write them up.

Submitting: Solutions should be submitted via Gradescope. 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

1. **(10 points)** Read again the handout with useful probabilistic inequalities shared in class (early version). Which one is your favorite and why?

   Note: For the rest of this homework, please be very explicit what probabilistic inequalities you are using when you apply them.

2. **(20 points)** We asked in class whether Markov's inequality still holds if we remove the assumption that the variable is non-negative. We showed that the answer was no, but in our example, we could bound the probability of X, our variable, being greater than any fixed $a > 0$ by $1/2$.

   Your task is to construct an even worse example. More specifically, show that for any $a > 0$ and $\delta \in (0, 1/2)$, there is a random variable X such that $E[X] = 1$ and $\Pr[X \leq a] \leq \delta$.

3. **(20 points)** In minimization problems, the goal is to output a solution that minimizes some value as much as possible. Examples include finding the shortest path from x to y, finding the shortest tour that passes by all mailboxes for your friend who is a mail carrier, or finding the most time efficient way of assembling cars in a factory.

   Suppose that you have a randomized algorithm for a minimization problem. For a given input instance x, let $\mathrm{OPT}(x) \geq 0$ be the optimal value. The algorithm outputs a solution that *in expectation* has value at most $4\,\mathrm{OPT}(x)$, where the expectation is taken over its internal randomness. How much can you narrow the following probabilities, compared to the obvious range $[0, 1]$? Explain.

   (a) The probability of outputting a solution of value at most $3\,\mathrm{OPT}(x)$.

   (b) The probability of outputting a solution of value at most $4\,\mathrm{OPT}(x)$.

   (c) The probability of outputting a solution of value at most $7\,\mathrm{OPT}(x)$.

   Note: When we ask "Explain" in this question we mean that whatever range you are able to come up with, an ideal solution would prove both

   - that values outside of the range are not possible
   - and that the range cannot be narrowed further.

   For instance, if the correct solution is $[1/7, 5/9)$, you should show that

   - the corresponding probability is at least $1/7$ and strictly less than $5/9$,
   - there is a distribution on possible outputs of such an algorithm in which the probability is $1/7$,
   - for any sufficiently small $\epsilon > 0$, there is a distribution on possible outputs of such an algorithm in which the probability is $5/9 - \epsilon$.

   If your bounds do not match (i.e., there is a gap between your upper and lower bounds and the values of the probability that you can construct) but are correct, that is okay. You will receive partial credit based on how much you were able to narrow the answer.

4. **(20 points)** Consider an experiment in which an event $\mathcal{E}$ occurs with probability p. Why does it suffice to run the experiment independently $\Omega(1/p)$ times to see $\mathcal{E}$ occur at least once with probability $1/2$?

   Hint: What is the probability that you repeat the experiment k times and it does not occur? You may find the following inequality useful: $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$.

5. **(20 points)** Count-Min Sketch, which we introduced in class, can be used to provide *point* frequency estimates. For instance, if we consider a database of tax payers, it could be used to provide an estimate of the number of people who paid exactly \$1,234 in taxes last year. In many settings, however, *range* queries may be more useful than *point* queries. In the example that we mentioned, we may be more interested in approximately learning the number of people who paid between \$1,234 and \$2,341 in taxes.

   Our task is to build on top of Count-Min Sketch to provide range queries. For simplicity, we assume that the domain of values that we consider is $[n] = \{1, 2, 3, \ldots, n\}$, where $n = 2^t$ for a positive integer t.

(a) First, construct a data structure that can provide frequency queries for a specific subset of ranges of the form

$$\left\{i \cdot 2^j + 1, i \cdot 2^j + 2, \ldots, (i+1) \cdot 2^j\right\},$$

where i and j are non-negative integers such that $j \leq t$ and $(i+1) \cdot 2^j \leq n$. The estimates should have similar guarantees to those for Count-Min Sketch, i.e., the estimate should never be an underestimate and with probability $1 - \delta$, the provided estimate should not be greater than the actual value by more than an ϵ fraction of all items.

Example: If $n = 4 = 2^2$, we have to be able to answer frequency queries for the following ranges/sets: $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{1, 2\}$, $\{3, 4\}$, and $\{1, 2, 3, 4\}$.

(b) Prove that any range $\{a, \ldots, b\}$ with $a, b \in [n]$ such that $a \leq b$ can be decomposed into $O(t) = O(\log n)$ ranges as above.

(c) Explain how to put the above two ideas together in order to provide range queries for an arbitrary range?

(d) How much space do you need to provide estimates with guarantees as in Count-Min Sketch, ignoring the space needed to store hash functions? More precisely, we want no estimate to be an underestimate, and with probability at least $1 - \delta$, we can overestimate by at most an ϵ-fraction of all items.

Note: More efficient solutions may receive more credit, but up to a constant factor, any solution should be fine, as long as it is correct and readable.

6. **(10 points)** Please fill out the Homework 1 survey.

   (To receive full credit for this question, please state here that you filled out the survey. Thanks!)