Problem 1 (10 Points) - Amortized Analysis

Design a data structure to support the following two operations for a dynamic multiset S of integers, which allows duplicate values:

1. INSERT(S, x) inserts x into S
2. DELETE-LARGER-HALF(S) deletes the largest $\lceil |S|/2 \rceil$ elements from S

Show how to implement this data structure so that any sequence of m INSERT and DELETE-LARGER-HALF operations runs in $O(m)$ time. Your implementation should also include a way to output the elements of S in $O(|S|)$ time. You can assume that your initial multiset is empty.

Hint: to delete larger half, need to find the median. You can use the result that there is a linear time algorithm for finding the median in an unordered array (if you don't know this result, you should learn about it - it is pretty cool).

Problem 2 (10 Points) - Amortized Analysis

An **ordered stack** is a data structure that stores a sequence of items and supports the following operations:

1. ORDEREDPUSH(x) removes all items smaller than x from the beginning of the sequence and then adds x to the beginning of the sequence.
2. POP() deletes and returns the first item in the sequence, or NULL if the sequence is empty.

Suppose we implement an ordered stack with a simple linked list, using the obvious ORDEREDPUSH and POP algorithms. Prove that if we start with an empty data structure, the amortized cost of each ORDERED-PUSH or POP operation is $O(1)$.

Problem 3 (15 Points) - Concentration Inequalities

The simplest model for a random graph consists of n vertices, and tossing a fair coin for each pair $\{i, j\})$ to decide whether this edge should be present in the graph. Call this model $G(n, 1/2)$. A triangle is a set of 3 vertices with an edge between each pair.

1. What is the expected number of triangles?
2. What is the variance?
3. Give an expression for the bound in the decay of probability, and show that the number of triangles is concentrated around the expectation. That is, if μ is the expectation of your random variable X, show that $\Pr[|X - \mu| \geq \delta\mu] \leq \Phi(\delta, n)$ for some function Φ. To show that it is concentrated around the expectation, for what value of δ does $\Phi(\delta, n) \leq 1/4$?

PROBLEM 4 (15 Points) - Concentration Inequalities

In this problem, we are in the setting where given a set S (which is not known to us - and this set does not have repeated elements), we only have access to S by querying a random element from S uniformly at random. Thus, if we want to sample s elements from S, we will obtain a sequence of elements $a_1, \ldots, a_s \in S$ where each a_k was drawn from S uniformly at random. Thus it could be the case where $a_i = a_j$ for some $i \neq j$.

1. Show that given n distinct numbers in $[0, 1]$ it is impossible to estimate the value of the median within say 1.1 multiplicative approximation factor with $o(n)$ samples.

   Hint: to show an impossibility result you show two different sets of n numbers that have very different medians but which generate, with high probability, identical samples of size $o(n)$.

2. Now calculate the number of samples needed (as a function of t) so that the following is true: with high probability, the median of the sample has at least $n/2 - t$ numbers (from the given set of n numbers) less than it and at least $n/2 - t$ numbers (from the given set of n numbers) more than it.

PROBLEM 5 (10 Points) - Balls and Bins

Consider again the experiment in which we toss m labeled balls at random into n labeled bins, and let the random variable X be the number of empty bins. We have seen that $\mathbb{E}[X] = n \cdot \left(1 - \frac{1}{n}\right)^m$.

(a) By writing $X = \sum_i X_i$ for suitable random variables X_i, show how to derive the following exact formula for the variance of X:

$$Var[X] = n \cdot \left(1 - \frac{1}{n}\right)^m + n(n-1) \cdot \left(1 - \frac{2}{n}\right)^m - n^2 \cdot \left(1 - \frac{1}{n}\right)^{2m}$$

(b) What is the asymptotic value (as $n \to \infty$) of $Var[X]$ in the cases $m = n$ and $m = n \ln(n)$?

Hint: you may use the approximations $(1 - x/n)^n \sim e^{-x} \cdot (1 - x^2/2n)$ and $(1 - 1/n)^{xn} \sim e^{-x} \cdot (1 - x/2n)$

(c) When throwing n balls into n bins, what is the expected number of bins with *exactly* one ball? Compute an exact formula and its asymptotic approximation.

Problem 6 (10 Points) - Coupon Collector

1. Let us consider the coupon collector problem: we toss $m = n\log n + cn$ balls into n bins, where c is a constant, and we are interested in the probability that there is no empty bin. We saw in class that

$$Pr[\text{some bin is empty}] \leq n \cdot \left(1 - \frac{1}{n}\right)^m \sim n \cdot \frac{1}{e^{m/n}} = \frac{1}{e^c}.$$

   Prove that

$$Pr[\text{some bin is empty}] = \Omega\left(\frac{1}{e^c} - \frac{1}{2e^{2c}}\right).$$

   Hint: inclusion-exclusion

2. Consider the following variation of the coupon collector's problem. Each box of cereal contains one of $2n$ different coupons. The coupons are organized into n pairs, so that coupons 1 and 2 are a pair, coupons 3 and 4 are a pair, and so on. Once you obtain one coupon from every pair, you can obtain a prize. Assuming that each coupon in each box is chosen independently and uniformly at random from the $2n$ possibilities, what is the expected number of boxes you must buy before you can claim the prize?

Problem 7 (15 Points) - More Hash Functions

Consider the following examples of hash families. For each one, prove that it is 2-universal or give a counterexample. Unless stated otherwise, the universe is $U = \mathbb{F}_p = \{0, 1, \ldots, p-1\}$ and the codomain of the hash functions is $\{0, 1, \ldots, n-1\}$.

1. Let p be a prime number and $n \leq p$ be an integer. Let

$$\mathcal{H} := \{h_a(x) = (ax \bmod p) \bmod n \ \mid \ a \in \{1, \ldots, p-1\}\}$$

2. Let m be an integer multiple of n, where $m \geq n$ and the universe $U = \{0, \ldots, m-1\}$. Let

$$\mathcal{H} := \{h_{a,b}(x) = (ax + b \bmod m) \bmod n \ \mid \ a \in \{1, \ldots, m-1\}, b \in \{0, 1, \ldots, m-1\}\}$$

3. Let p be a prime number and $n \leq p$ be an integer. Let

$$\mathcal{H} := \{h_{a,b}(x) = (ax + b \bmod p) \bmod n \ \mid \ a, b \in \{0, 1, \ldots, p-1\}, a \neq 0\}$$

Problem 8 (15 Points) - Karger Strikes Back

To improve the probability of success of the randomized min-cut algorithm, it can be run multiple times.

1. Consider running the algorithm twice. Determine the number of edge contractions and bound the probability of finding a min-cut.

2. Consider the following variation. Starting with a graph with n vertices, first contract the graph down to k vertices using the randomized min-cut algorithm. Make ℓ copies of the graph with k vertices, and now run the randomized algorithm on these reduced graphs independently. Determine the number of edge contractions and bound the probability of finding a min-cut.

3. Find optimal (or at least near-optimal) values of k and ℓ for the variation in the previous part that maximizes the probability of finding a min-cut while using the same number of edge contractions as running the original algorithm twice.