Homework 2

CS 574: Randomized Algorithms, Fall 2025
Due: Thursday, Oct 2nd 2025 at noon

Instructions and Policy:

- Each homework can be done in a group of size at most two. Only one homework needs to be submitted per group. However, we recommend that each of you think about the problems on your own first.
- Homework needs to be submitted in pdf format on Gradescope. See `https://courses.grainger.illinois.edu/cs374al1/sp2025/hw-policies.html` for more detailed instructions on Gradescope submissions.
- Follow academic integrity policies as laid out in student code. You can consult sources but cite all of them including discussions with other classmates. Write in your own words. See the site mentioned in the preceding item for more detailed policies.

Optional Problem. Recall the variant of Quick Sort you saw in the previous home work. Prove that its running time is $O(n \log n)$ with high probability.

Optional Problem. Perfect hashing. Exercise 3.3 in Kent Quanrud's notes.

Problem 1. Sampling, Chebyshev vs Chernoff. Suppose you want to estimate the average of n numbers via sampling, for example the average wealth of people in a town. The average can be very skewed by outliers — perhaps there are a few billionaires that will not make it to the sample but will clearly affect the average. However, we can obtain an accurate estimate if we assume that the numbers are within some limited range. Assume the input numbers $z_1, z_2, \ldots, z_n$ are from $[a, b]$ where $a, b \in \mathbb{R}$ with $a \leq b$. Suppose you sample k input numbers (with replacement) and output their average as the estimate for the true average $\alpha = (\sum_i z_i)/n$. Let X be the random variable denoting the output value.

- Using Chebyshev's inequality, show that for $k \geq \frac{(b-a)^2}{\delta \epsilon^2}$, we have
$$\mathrm{P}[|X - \alpha| \geq \epsilon] \leq \delta.$$
- Using the Chernoff inequality, show that there exists a constant $c > 0$ such that for $k \geq \frac{c(b-a)^2 \log(2/\delta)}{\epsilon^2}$, we have
$$\mathrm{P}[|X - \alpha| \geq \epsilon] \leq \delta.$$

Problem 2. Chernoff bounds.

- Suppose we throw m balls into n bins. Via Chernoff bounds we saw that when $m = n$ the maximum bin load is $O(\log n / \log \log n)$ with high probability. For general m the bound we can prove is $O(\frac{m}{n} \log n / \log \log n)$. When m/n is large this is not such a good bound. Prove via Chernoff bounds that for any fixed $\epsilon \in (0, 1)$ the maximum load is $(1+\epsilon)m/n + O(\log n/\epsilon^2)$ with high probability.

- Let $X = \sum_{i=1}^{n} X_i$ where the X_i are independent geometrically distributed random variables with parameters $p_1, \ldots, p_n$. Recall that a geometric distribution is defined over non-negative integers $\{1, 2 \ldots, \}$; if Y is distributed according to geometric distribution with parameter p then $\mathrm{P}[Y = i] = (1-p)^{i-1}p$ which is the probability that number of independent tosses of a coin before one sees a head is i; here p is the probability of a head in each toss of the coin. Prove Chernoff bounds for X using the moment generating function method.

Problem 3. Importance sampling. Importance sampling is a fundamental technique in statistics that is also used in several algorithms. Here we illustrate it in two simple scenarios.

- Recall the problem of estimating the mean via uniform sampling. The issue is the variance which can be high due to outliers. Suppose we want to estimate the mean of n non-negative numbers $a_1, a_2, \ldots, a_n$ and we also have some crude estimates $w_1, w_2, \ldots, w_n$ (we will not worry about where these estimates came from) such that for each i, $w_i/\alpha \leq a_i \leq \alpha w_i$ for some constant $\alpha > 1$ (say $\alpha = 5$). Consider estimating the mean via weighted sampling with $w_1, w_2, \ldots, w_n$: pick an $i \in [n]$ where the probability of picking i is w_i/W with $W = \sum_{i=1}^{n} w_i$. The estimator is $\frac{1}{n} a_i W / w_i$. You take k such samples and average them. Let Z be this average. Argue that Z is an exact estimator for the mean of $a_1, \ldots, a_n$. Upper bound the variance of Z as a function of α, k, and μ where $\mu = \frac{1}{n} \sum_{i=1}^{n} a_i$. Using Chebyshev's inequality, what is the number of samples k you need to guarantee that $\mathrm{P}[|Z - \mu| \geq \epsilon\mu] \leq \delta$ for given ϵ and δ in $[0, 1]$?

- Suppose you are using uniform sampling to estimate the mean of n non-negative numbers $a_1, \ldots, a_n$. This requires generating a random integer between 1 and n but what if we had only access to random bits? Let $N = 2^k$ where $2^{k-1} < n \leq 2^k$. If $n = N$ then we can pick k bits at random and generate perfectly uniform numbers between 1 and n. Otherwise, a standard technique is rejection sampling; pick a random number r in $[N]$ and reject it if $r > n$, otherwise we have a uniformly random number in $[n]$. However this has the problem that we can lose almost half the samples in expectation (if $n = 2^{k-1} + 1$). Further, it also has the disadvantage that there is no apriori bound on the number of useful samples from a given set of h samples (all of them could be rejected). One way to use all the generated samples is via importance sampling. Suppose we set up an onto mapping $\phi : [N] \to [n]$. We generate $i \in [N]$ and would like to use $j = \phi(i)$ where $j \in [n]$. Clearly this can create a non-uniform distribution over $[n]$. Show how you can adjust for this non-uniformity (knowing ϕ) so that you can still get an unbiased estimator for the mean. Suppose you use k such samples and take their average. Let Z be this random variable. Upper bound variance of Z when compared to the variance of the estimator based on uniform samples from $[n]$.

Problem 4. We saw JL lemma as a technique to dimensionality reduction of a set of n vectors. The same ideas lead to what are called *oblivious subspace embeddings* which are quite useful. The goal of this problem is for you to learn what they are. Do Exercise 9.4 in Kent Quanrud's notes.

Problem 5. Max k-Cut is the following problem: given a graph $G = (V, E)$ we wish to partition V into k parts $V_1, V_2, \ldots, V_k$ so as to maximize the number of edges that cross the partition (have end points in two different parts). Obtain an efficient randomized algorithm that yields a $(1 - 1/k)$-approximation. Suppose $k = 5$. Obtain a deterministic algorithm by derandomizing the randomized algorithm using limited independence.

Problem 6. Extra Credit: Consider the following process with balls and bins. We start with n balls and n bins. In each iteration the balls are thrown into bins independently. Call a bin *lonely* if it falls into a bin by itself. We remove all the lonely balls and repeat (note that the number of bins remains the same). The process ends when there are no balls left. Let X be the random variable for the number of iterations of this process. Give an upper bound on the expectation of X.