New York University Tandon School of Engineering
Computer Science and Engineering

CS-GY 9223D: Homework 1.
Due Friday, September 18th, 2020, 11:59pm ET.

Collaboration is allowed on this problem set, but solutions must be written-up individually. Please list collaborators for each problem separately, or write "No Collaborators" if you worked alone.

For just this first problem set, 10% extra credit will be given if solutions are typewritten (using LaTeX, Markdown, or some other mathematical formatting program).

Problem 1: Short answers.

(20 pts) Do these first!

1. For any given $k > 0$, give an example of a random variable for which Chebyshev's inequality is tight up to constant factors. Specifically, for any given k, describe a random variable X with variance σ^2 such that $\Pr[|X - \mathbb{E}X| \geq k\sigma] \geq \frac{1}{10k^2}$.

2. A biased random coin comes up heads with probability $1/n$ for some $n > 1$. Show that, after n random flips, the probability that you never see heads is $\leq .3679$. Show that after $n \log n$ flips, the probability that you never see heads is $\leq 1/n$. **Hint:** You might need to use a little calculus! I used the Taylor series for $\log(1 - x)$.

3. In class, we saw that, when hashing m items into a hash table of size $O(m^2)$, the expected number of collisions was < 1. In particular, this meant that with probability $> 9/10$) we could easily find a "perfect" hash function into the table that had no collisions. Use (2) above to to give an alternative proof of this fact. Specifically, to have no collisions we much have the following events all happen in sequence: the second item inserted into the hash table doesn't collide with an existing item, the third item inserted doesn't collide with an existing item, ..., the m^{th} item inserted doesn't collide with an existing item. Analyze the probability these events all happen.

4. Since I was a bit informal in lecture, prove the Union Bound using... Markovs Inequality + Linearity of Expectation. That is, prove that for any random events $A_1, \ldots, A_k$,

$$\Pr[A_1 \cup A_2 \cup, \ldots, \cup A_k] \leq \sum_{i=1}^{k} \Pr[A_i].$$

   Hint: Markov's inequality applies to a *random variable*. You're going to need to define a new random variable in a clever way to get the proof.

Problem 2: If at first you don't success, try again.

(10 pts) In class, we saw that, when hashing m items into a hash table of size $O(m^2)$, the expected number of collisions was < 1. In particular, this meant we could easily find a "perfect" hash function into the table that had no collisions.

Consider the following alternative scheme: build two tables, each of size $O(m^{1.5})$ and choose a separate random hash function (independently at random) for each table. To insert an item, hash it to one bucket in each table and place it in the emptier bucket.

1. Show that, if we're hashing m items, with probability $1/2$, there will be no collisions in either table. You may assume a uniformly random hash functions.

2. Modify the above scheme to use $O(\log m)$ tables. Prove that this approach yields a collision-free hashing scheme with space $O(m \log m)$. Again, you may assume a fully random hash function.

Problem 3: Pinning down the median.

(15 pts) A very common objective in statistical analysis is to estimate the *median* (not the mean) of a dataset from uniformly random samples. For example, a census might poll random citizens in a city to request information about their income. From this sample, the goal is to estimate the city's *median income*.

1. Suppose we have a list S of n numbers with median M. We sample k numbers $X_1, \ldots, X_k$ uniformly at random (with replacement) from S. Show that as long as $k \geq O\left(\frac{1}{\epsilon^2}\right)$, then $\tilde{M} = \text{median}(X_1, \ldots, X_k)$ is a good approximate median in the following sense: with probability $9/10$, at least a $\frac{1}{2} - \epsilon$ fraction of numbers in S are $\leq \tilde{M}$ and at least a $\frac{1}{2} - \epsilon$ fraction of numbers in S are $\geq \tilde{M}$.

2. **Extra Credit – optional!** Show that it is *impossible* to estimate the *value* of the true median M with $o(n)$ random samples from S, even if we just want to get within a constant approximation factor, and succeed with constant probability. For example, we can't even guarantee that $.5M \leq \tilde{M} \leq 2M$ with probability $\geq 2/3$ unless we take nearly n samples from S.

Problem 4: Randomized methods for COVID-19 group testing.

(15 pts) One of the most important factors in controlling the COVID-19 outbreak has been testing. Unfortunately, testing can be expensive and slow. A popular proposal to make it cheaper is by testing patients in *groups*. In particular, the biological samples from multiple patients (e.g., multiple nose swabs) are combined into single test tube and tested for COVID-19 all at once. If the test comes back negative, we know everyone in the group is negative. If the test comes back positive, we do not know which patients in the group actually had COVID-19, so further testing would be necessary. There's a trade-off here, but it turns out that, overall, group testing can save on the total number of tests run.

1. Consider the following deterministic "two-level" testing scheme. We divide a population of n individuals to be tested into C arbitrary groups. We then test each of these groups in aggregate. For any group that comes back positive, we retest all members of the group individually. Show that there is a choice for C such that, if k individuals in the population have COVID-19, we can find all of those individuals with $\leq 2\sqrt{nk}$ tests. You can assume k is known in advance (often it can be estimated accurately from the positive rate of prior tests). This is already an improvement on the naive n tests when $k < 25\% \cdot n$.

2. We can use randomness to do even better. Consider the following scheme: Collect $q = O(\log n)$ nose swabs from each individual (I know... not pleasant). Then, repeat the following process q times: randomly partition our set of n individuals into C groups, and test each group in aggregate. Once this process is complete, report that an individual "has COVID" if the group they were part of tested positive all q times. Report that an individual "is clear" if any of the groups they were part of tested negative. Show that for $C = O(k)$, with probability $9/10$, this scheme finds all truly positive patients and reports no false positives. Thus, we only require $O(k \log n)$ tests!

3. **(Hard) Extra Credit – optional.** Show that no scheme can use $o(k \log(n/k))$ tests and succeed with probability $> 2/3$. So, for small k, the approach above is essentially optimal up to constant factors!

4. **Optional qeustions to think about for a possible final project or research project:**

   - Clearly administering $O(\log n)$ nasal swabs to every individual is not ideal, even if it reduces the total number of tests that need to be run in a lab. Are there different schemes which trade-off between total number of tests, number of swabs that need to be run up-front, and number of patient revisits to a testing center (as required e.g., in part 1)?
   - Could you improve the above method if you had some prior knowledge on the probability a patient would test positive (e.g., from a survey on if they had symptoms, if they had been exposed to someone with COVID, if they traveled recently, etc.)? Can you surpass the lower bound? I have some ideas on this, and would be happy to point students to some relevant papers.
   - What if instead of finding all individuals, I just wanted to estimate the COVID-positive rate in my population. How many tests are needed, and can you reduce that number with clever group testing schemes?