NYU | TANDON SCHOOL OF ENGINEERING

New York University Tandon School of Engineering
Computer Science and Engineering

CS-GY 6763 INET: Homework 2.
Due Sunday, Feb. 1st, 2026, 11:59pm ET.

## Problem 1: Universal Hash Function That is Better than Uniform.

Consider the following implementation of a hash function, $h$, that maps integers from $1, \ldots, n$ to integers from $1, \ldots, m$ (where $m < n$, and for simplicity, assume $n$ is divisible by $m$). Randomly permute the integers $1, \ldots, n$ to obtain a list $\sigma_1, \ldots, \sigma_n$. Then, for all $i \in 1, \ldots, m$, set

$$h(\sigma_{m\cdot(i-1)+j}) = i \quad \text{for all } j = 1, \ldots, n/m.$$

To give and example, suppose $n = 9$ and $m = 3$. We might have that:

$$\sigma_1, \ldots, \sigma_9 = 5, 2, 8, 1, 6, 3, 9, 4, 7.$$

We would choose our hash function so that:

$$\begin{aligned} h(5) = h(2) = h(8) = 1 \\ h(1) = h(6) = h(3) = 2 \\ h(9) = h(4) = h(7) = 3. \end{aligned}$$

Prove that a hash function chosen in this way is *universal*, meaning that, for any two distinct integers $x, y \in 1, \ldots, n$,

$$\Pr[h(x) = h(y)] \leq \frac{1}{m}.$$

Morever, show that inequality is *strict*. I.e., for any two distinct integers $x, y \in 1, \ldots, n$,

$$\Pr[h(x) = h(y)] < \frac{1}{m}.$$

Notably, this means that the probability of collision for any two inputs is strictly better than for a uniformaly random hash function, which always gives collision probability $1/m$.

## Problem 2: Mark-and-Recapture Full Analysis Analysis.

Now that you have learned Cheybshev's inequality, you have the tools to prove the claim on Slide 31 of Lecture 1. Specifically, if we collect $O(\sqrt{n}/\epsilon)$ samples uniformly from a set of unknown size $n$, show that the mark-and-recapture estimate, $\tilde{n} = \frac{m(m-1)}{2D}$ satisfies

$$(1-\epsilon)n \leq \tilde{n} \leq (1+\epsilon)n,$$

with probability $9/10$.

**Hint:** $\tilde{n}$ is a difficult random variable to work with because it involves the inverse of another random variable. Instead of trying to analyze it directly, prove this claim indirectly by showing that $D$ concentrates around its expectation, and this is enough to imply that $\tilde{n}$ is accurate.

## Problem 3: Randomized Disease Group Testing (Optional).

*This problem only requires basic probability calculations – it does not use any material specifically from Lecture 2.*

One of the most important factors in controlling diseases like bird flu or, some years ago, COVID-19, is testing. However, testing often requires processing in a lab, so can be expensive and slow. One way to make it cheaper is to test patients/livestock/etc. in *groups*. The biological samples from multiple individuals (e.g., multiple nose swabs) are combined into a single test tube and tested for the disease 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 have the disease, 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 $C$ groups of the same size. We then test each of these groups. 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 a disease (would test positive), 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 better. Consider the following scheme: Collect $q = O(\log n)$ biological samples from each individual (in reality, divide one sample into $q$ parts). 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 "is positive" if the group they were part of tested positive all $q$ times. Report that an individual "is negative" if *any* of the groups they were part of tested negative. Prove 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!

   **Hint:** If your proof would also work for $q = O(1)$ then it has a bug! See EC below.

3. **Challenge** 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!