COS 330: Great Ideas in Theoretical Computer Science

Fall 2025

Problem Set 8

Module: Information and Codes

Below is a reminder of key aspects of the PSet:

- The only goal of this PSet is to help you develop your problem-solving skills in preparation for the exams. Your performance on this PSet will not directly contribute to your grade, but will indirectly improve your ability to do well on the exams.
- Because your performance does not directly impact your grade, you may use any resources you like (collaboration, AI, etc.) to help you complete the PSet.
- We strongly suggest taking a serious stab at the PSet alone, to help self-evaluate where you're at. But, we also suggest collaborating with friends, visiting office hours, asking on Ed, and/or using AI tools to help when stuck. Even when able to complete the entire PSet on your own, you may still find any of these methods useful to discuss the PSet afterwards.
- Throughout the PSet, we've included some general tips to help put these into broader context. Exams will not have these, and future PSets may have fewer.
- We strongly suggest treating this like any other PSet, and writing up your solutions as if you were handing them in for a grade. At minimum, we very strongly suggest writing up sufficiently many solutions to discuss with your Coach.

Aligning Expectations

Recall that the 🗎 symbol implies that the following problem is an "exam-style" problem We highly recommend that you write-up a full solution to this problem.

Problem 1 : Entropy and Concentration of Probability 🗎

Let X be a discrete random variable taking values in a finite set $\{x_1, \ldots, x_m\}$ with probabilities $p_i = \Pr[X = x_i]$, sorted so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Recall that the entropy of X is given by:

$$H(X) = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i}.$$

In this problem you will show how a small entropy implies that a large fraction of the probability mass is concentrated on a small set of possible values.

(a) Given a probability distribution on m elements, the maximum entropy is $\log_2 m$, achieved by the uniform distribution. In other words, $H(X) \leq \log_2 m$ and equality holds if and only if $p_i = 1/m$ for all i. You may assume this fact without proof. (Optional: you may find it instructive to prove it if you'd like. There are multiple ways to do so, the simplest one uses a probability inequality known as Jensen's inequality.)

Prove the following generalization of the maximum entropy fact. Let $q_1, \ldots, q_t \geq 0$ with $q_1 + \cdots + q_t = s$, where $0 < s \leq 1$. Show that

$$\sum_{i=1}^{t} q_i \log_2 \frac{1}{q_i} \;\leq\; s \log_2 \frac{t}{s}.$$

(Hint. Try to relate the q_i's to a probability distribution so you can apply the maximum entropy fact.)

(b) Suppose that

$$H(X) \;\leq\; \frac{1}{2} \log_2 \log_2 m. \qquad (*)$$

Partition the indices $\{1, \ldots, m\}$ into two sets:

$$S = \left\{ i : p_i < \frac{1}{\log_2 m} \right\} \qquad \text{and} \qquad L = \left\{ i : p_i \geq \frac{1}{\log_2 m} \right\}.$$

Use part (a) to show that $\sum_{i \in S} p_i \leq \frac{1}{2}$.

(c) Prove that $|L| \leq \log_2 m$. Conclude that under assumption $(*)$, there exists a set of at most $\log_2 m$ values of X whose total probability is at least $1/2$.

Problem 2 : Pooled Testing with One Faulty Test

A lab is testing blood samples for a rare disease. There are n patients labeled $\{1, 2, \ldots, n\}$, and exactly one patient is infected. The lab can run pooled tests: for each test, you choose a subset $S \subseteq \{1, \ldots, n\}$ of patients, mix their samples, and receive a result that is positive if and only if the infected patient is in S. However, the testing machine is unreliable: at most one test result can be wrong (either a false positive or false negative).

You must design all your tests in advance (non-adaptively) before seeing any results. After receiving the (possibly corrupted) test outcomes, you must correctly identify the infected patient. Let t denote the total number of tests. A testing strategy is robust if it always correctly identifies the infected patient, even if one test result is flipped.

(a) Let $C \subseteq \{0,1\}^t$ be a binary code with minimum Hamming distance at least 3 and $|C| = 2^k$ codewords. Assume $n \leq 2^k$ and fix an injective mapping $x \mapsto c_x$ from $\{1, \ldots, n\}$ into C, in other words, label each codeword by a distinct integer.

Design a non-adaptive testing scheme with t tests using the codewords $\{c_x\}$ and prove that your scheme is robust.

(b) For a robust testing strategy with t tests, let $v_x \in \{0,1\}^t$ be the ideal outcome vector when patient x is infected (assuming all tests are correct). Define the Hamming ball

$$B_1(v_x) = \{w \in \{0,1\}^t : \text{dist}(w, v_x) \leq 1\},$$

where $\text{dist}(w, v_x)$ denotes the Hamming distance (the number of coordinates in which w and v_x differ).

1. Explain why any observed outcome when patient x is infected must lie in $B_1(v_x)$.
2. Show that $|B_1(v_x)| = 1 + t$.

(c) Prove that for the strategy to be robust, the balls $B_1(v_x)$ and $B_1(v_{x'})$ must be disjoint for all $x \neq x'$. Use this to show that any robust strategy with t tests can handle at most

$$N_{\max}(t) \leq \frac{2^t}{t+1}$$

patients.

(d) For each integer $r \geq 2$, there exists a binary code with length $t = 2^r - 1$, 2^k codewords where $k = 2^r - 1 - r$, and minimum distance 3 (one example of such codes are Hamming codes, see the precept for how they are defined). Using this fact, show the following:

1. For every $r \geq 2$, there is a robust testing strategy using $t = 2^r - 1$ tests that can handle up to $N_r = 2^{2^r - 1 - r}$ patients.
2. This strategy achieves $N_r = \frac{2^t}{t+1}$, matching the upper bound from part (c).
3. Conclude that for any n, there exists a robust testing scheme using $O(\log n)$ tests.