Math 345 - Problem Set 2 - due Wednesday, March 7

These will be graded on both correctness and clarity, so write careful, organized solutions (a good idea to keep in mind is that another person from the class should be able to read and understand your solution). All of your answers must be justified. **Honor Code:** For this and all problem sets, you are encouraged to work on solving these problems in groups. However, you MUST write up your solutions individually; in particular, you may not look at someone else's write-up. In addition, you must indicate who you worked with.

1. Consider the following symbol code:

$$\begin{aligned} a &\to 00 \\ b &\to 10 \\ c &\to 11 \\ d &\to 101 \end{aligned}$$

   (a) Show that there are arbitrarily long sequences of bits $\mathbf{b} = b_1 b_2 \cdots b_m$ such that there exist strings of symbols $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_M)$ with $x_1 \neq y_1$ such that $\mathbf{b}$ is a prefix for the encodings of both $\mathbf{x}$ and $\mathbf{y}$. (In other words, you may have to go arbitrarily far through the sequence of bits before you can even decode the first symbol!)

   (b) Show that this code is uniquely decodable. (Hint: as happens in math and life, there is a trick. Play around with examples, and hopefully you'll discover it. If you don't you can always come talk to me.)

2. Suppose C is a binary prefix symbol code with codelengths l_i such that

$$\sum_{i=1}^{n} 2^{-l_i} = 1.$$

   Prove that any binary string $b_1 b_2 \cdots b_m$ is a prefix of the encoding of some $(x_1, x_2, \ldots, x_N)$ (in other words, decoding any string would never run into an error – hitting something that's not a codeword – though it might unexpectedly hit the end of the string).

3. (from Cover and Thomas) Let $m = 2^n$, and let X be uniformly distributed over $\{1, 2, \ldots, m\}$. If we ask the right questions, we can determine the value of X in n yes/no questions. Suppose we ask random questions, though. To be precise, we choose S from among the 2^m subsets of $\{1, 2, \ldots, m\}$, each with equal probability, and ask, "Is $X \in S$?"

   (a) Without loss of generality, suppose that $X = 1$ is the actual value. What is the probability that $X = 2$ yields the same answers as $X = 1$ for k such (independent) random questions?

   (b) What is the expected number of objects in $\{2, 3, \ldots, m\}$ that have the same answers to the questions as does the correct object 1? (Remember that $E[X + Y] = E[X] + E[Y]$)

   (c) Suppose we ask $k = n + \sqrt{n}$ random questions. What is the expected number of wrong objects agreeing with the correct one?

   (d) Use the Markov/Chebyshev inequality

   $$P(t \geq \alpha) \leq \frac{\bar{t}}{\alpha}$$

   to show that the probability of error (one or more wrong objects remaining after $k = n + \sqrt{n}$ random questions) goes to zero as $n \to \infty$.

   (e) (Optional) So with random questions, we only need a few ($\sqrt{n}$) more questions than with the best possible deterministic questions. What (if anything) was special about $\sqrt{n}$?

4. **NOT required!** I promised to put this on the homework, but decided that we're doing a better proof of the Lossless Source Coding Theorem in class.

   Given a positive integer N and given $\beta > 0$, create the lossless code $C_{N,\beta}$ for an i.i.d. string $X^N = (X_1, \ldots, X_n)$ as follows:

   - Encode strings from $T_{N,\beta}$ as a 0 followed by $\lceil \log_2 |T_{N,\beta}| \rceil$ bits.
   - Encode strings from $X^N \backslash T_{N,\beta}$ as a 1 followed by $\lceil \log_2 |X^N \setminus T_{N,\beta}| \rceil$ bits.

   Prove that, given ϵ, there exists a β such that, for sufficiently large N, the expected length of an encoded string (using $C_{N,\beta}$) is at most

   $$N(H(X) + \epsilon).$$