Information-Theoretic Modeling, Fall 2009

Exercises IV, due Friday 9 October.

1. E-mail your answer to this question to the teaching assistant. Mid-course questionnaire: Please give us feedback about the course. How do you find the pace: too fast or too slow? What do you find most interesting and least interesting: proofs of theorems, algorithms, whatever? How about the exercises? Feel free to give us any other comments about the course and how we could improve it in the future.

2. (Cover & Thomas, Ex5.12.) Consider a random variable X that takes on four values, $a, b, c, !$, with respective probabilities $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12}\right)$.

(a) Construct a Huffman code for this random variable.

(b) Show that there exist two different sets of optimal (Huffman) lengths for the codewords; namely, show that codeword length assignments $(1, 2, 3, 3)$ and $(2, 2, 2, 2)$ are both optimal.

(c) Conclude that there are optimal codes with code-word lengths for some symbols that exceed the Shannon code-length $\left\lceil \log_2 \frac{1}{p(x)} \right\rceil$.

3. Assume that a person x is picked at random according to distribution p from a finite population $\mathcal{X}$. You are allowed to ask an oracle a sequence of yes–no questions like "Does x belong to the subset $S_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n(i)}\}$?". You want to minimize the *expected number of questions* required to identify x.

(a) What is the optimal strategy, and what can you say about the expected number of questions that are required? (*Hint:* the sequence of answers can be regarded as the codeword of x.)

(b) What if the goal is to minimize the *worst-case* (maximum) number of questions?

4. Consider again the random variable X of Exercise 2.

(a) Construct the interval corresponding to the source sequence $acab!$ using the simple version of arithmetic coding that doesn't rescale the intervals to prevent underflows (see Fig. 2 in (Witten, Neal, and Cleary, 1987)).

(b) What is the real value, $r = 0.d_1 d_2 \cdots d_k$, with the fewest number of decimal digits (after the decimal point) such that any continuation of the value ($r' = 0.d_1 d_2 \cdots d_k d_{k+1} \cdots$) is still within the interval.

(c) *Optional (try this only if you think in binary*[1]*):* What if the real value is represented in binary digits: what is the fewest number of bits that is sufficient to guarantee that any continuation is still within the interval?

[1] There are only 10 types of people in the world — those who understand binary, and those who don't. (Recall that the binary representation of a real value $r \in [0, 1)$ is given by $0.b_1 b_2 b_3 \cdots$ such that $r = \sum_i b_i 2^{-i}$. The corresponding statement about the decimal representation is of course $r = \sum_i d_i 10^{-i}$.)

Bonus exercise. What if three possible answers instead of only two ("yes" and "no") are allowed in Exercise 2? The questions are now of the form "Which one of the (mutually exclusive) sets S_i^1, S_i^2, S_i^3 contains x?" What are the optimal strategies for minimizing the expected/worst-case number of questions? (*Hint:* Google for "ternary Huffman code".)