Homework Set #1

Due 29 September 2009

Problem 1 (Unused Code Sequence)

(a) Let C be a prefix free code that satisfies the Kraft inequality with equality. What is the probability that a random generated sequence of code alphabets begins with a codeword?

(b) Let C be a variable length code that satisfies the Kraft inequality with equality but does not satisfy the prefix condition. What is the probability that a random generated sequence of code alphabets begins with a codeword? Prove that some finite sequence of code alphabet symbols is not the prefix of any sequence of codewords.

Problem 2 (Wine Testing)

There are 6 bottles of wine, one of which you know has gone bad. You do not know which bottle contains the bad wine, but you know that the probability of each bottle being bad is $(8/23, 6/23, 4/23, 2/23, 2/23, 1/23)$. The bad wine has a distinctive taste. To find the bad wine your friend suggests you to taste a little bit of each wine until you find the bad wine.

(a) To have the least number of tastings on average, what should your strategy be? Which bottle should be tasted first?

(b) What is the average number of tastings to find the bad wine?

(c) Calculate the minimum average number of tastings if you are allowed to taste a mixture of different wines and detect a bad wine's taste inside (the distinctive taste is retained even when mixed with other good wines).

(d) Is the strategy studied in (a) optimal if you are allowed to mix wines?

Problem 3 (Huffman Codes with Costs)

Suppose that $X = i$ with probability p_i, $i = 1, 2, \cdots, m$. Let l_i be the number of binary symbols in the codeword associated with $X = i$, and let c_i denote the cost per letter of the codeword when $X = i$. Thus, the average cost C of the description of X is $C = \sum_{i=1}^{m} p_i c_i l_i$.

- Minimize C over all $l_1, l_2, \ldots, l_m$ such that $\sum 2^{-l_i} \leq 1$. Ignore any implied integer constraints on l_i. Exhibit the minimizing $l_1^*, l_2^*, \ldots, l_m^*$ and the associated minimum value C^*.

- How would you use the Huffman code procedure to minimize C over all uniquely decodable codes? Let C_{Huffman} denote this minimum.

- show that

$$C^* \leq C_{\text{Huffman}} \leq C^* + \sum_{i=1}^{m} p_i c_i.$$

Problem 4 (Huffman Codes)

Consider a source S with probabilities $p_1, \cdots, p_n$. As seen in class, in order to encode this source, the binary Huffman procedure is optimal. Let's call the average length of a Huffman code for this source l_H. In this problem we design codes for the described source so that all the codeword lengths are multiples of m ($m \geq 2$).

(a) Describe a procedure for designing a binary uniquely decodable code for the source S such that all the codeword lengths are multiples of m and the average length of the code is minimum. Call this minimum average length l_m.

(b) Derive an upper bound and a lower bound for l_m.

(c) Give an example of a source for which l_m is equal to the lower bound found in (b). Give the example for a general m.

(d) What is the minimum number of source alphabets that the example source of part (c) should have in terms of m?

(e) Show that $l_m \leq l_H + m - 1$.

(f) Give a tight example for the bound of part (e).

Problem 5 (Data Compression)

A source is to be encoded but we do not know the distribution of the source symbols exactly. What we know is that with probability λ, the source produces alphabet symbols (A, B, C, D, E, F) according to model 1 with probabilities $P_1 : (1/2, 1/4, 1/16, 1/16, 1/16, 1/16)$, and with probability $1 - \lambda$, the source produces alphabet symbols (A, B, C, D, E, F) according to model 2 with probabilities $P_2 : (1/2, 1/4, 0, 0, 1/8, 1/8)$. What is the optimal encoding strategy and the average length of the code in the following scenarios

(a) Neither the encoder, nor the decoder knows which model produces the source.
Hint: You might need to consider different regimes of $\lambda \in [0, 1]$. *You may need to consider the cases* $\lambda = 0$ *and* $\lambda = 1$, *separately as well.*

(b) Both encoder and decoder know the model producing the source.

(c) Let the symbols come from model 2, *i.e.,* $\lambda = 0$. Assume a (liar) genie tells both the encoder and the decoder that the symbols come from model 1, *i.e.,* that $\lambda = 1$, and the encoder designs an optimal code based on this information. Find the average length of the code. What is the penalty of this mis-information, *i.e.,* calculate the difference between the average length of optimal code for the true model and the designed code. How is this difference related to the Kullback-Leibler divergence, $D(\cdot||\cdot)$, between the two distributions.

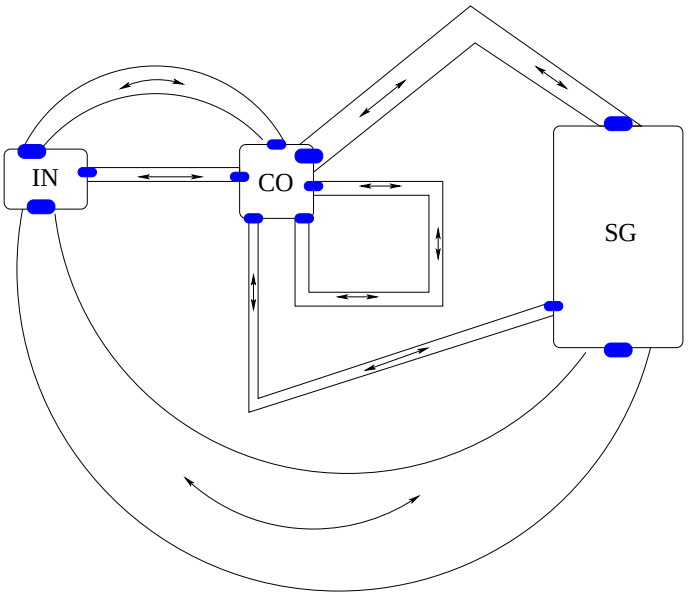

Figure 1: EPFL plan and the new student paths.

Problem 6 (Random Walk on a Graph)

One of the new students got lost at EPFL the day he arrived and for the whole day he walked around in EPFL. As he didn't know where he was going, he decided to choose one of the possible doors (illustrated in Figure 1) leading out of each building uniformly at random and follows the path out of the current building to the connecting building (regardless of the door he entered in to the current building). To make it simple, let's assume that EPFL's plan is as illustrated in Figure 1, and the points of our interest are only IN building, CO building and SG building. The sequence of the buildings he passed in his walk $(X_1, X_2, \cdots, X_i, \cdots)$ forms a stochastic process (where $X_i \in \{\text{IN}, \text{CO}, \text{SG}\}$) which we call a random walk in this problem.

(a) Show that this stochastic process is a Markov chain.

(b) Write the transition matrix of this Markov chain.

(c) With what probabilities will the student be in each of the aforementioned buildings at the end of the day; *i.e.,* what is the stationary distribution of the random walk?