ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Handout 13 Introduction to Communication Systems
Homework 7 October 30, 2008

PROBLEM 1. A source has an alphabet of 4 letters, a_1, a_2, a_3, a_4, with corresponding probabilities p_1, p_2, p_3, p_4 and we have the condition $p_1 > p_2 = p_3 = p_4$. Let n_1 be the length of the codeword for a_1 in a Huffman code.

1. Find the smallest number q such that $p_1 > q$ implies that $n_1 = 1$.
2. Let $p_1 = q$ (where q is your answer in part 1.). Show by example that a Huffman code exists for $n_1 = 1$ and $n_1 = 2$.
3. Now assume the more general condition, $p_1 > p_2 \geq p_3 \geq p_4$. Does $p_1 > q$ still imply that $n_1 = 1$? Justify your answer.

PROBLEM 2.
1. Consider a source with 2^n symbols having probabilities $p_i = \frac{1}{2^n}$ for all $1 \leq i \leq 2^n$.
   (i) Let $n = 3$. Construct the Huffman code for this case and draw the corresponding tree.
   (ii) Using Huffman procedure, what is the tree for general n?
   (iii) What is the entropy of this source? Knowing just the entropy for this source, can you construct an optimal code? Note that for a code to be optimal, we first need it to be prefix-free and secondly we have to minimize the average length of the code.
2. Consider an alphabet of n letters with correponding probabilities $q_i = \frac{1}{2^i}$ if $1 \leq i \leq n-1$ and $q_n = \frac{1}{2^{n-1}}$. How does the tree of the corresponding Huffman code look like? What is the length of each codeword?

PROBLEM 3. A source S outputs symbols from an alphabet of n letters with probabilities $p_1, p_2, \ldots, p_n$. You want to construct a Huffman code for this source, but by mistake you think that the probabilities are $q_1, \ldots, q_n$ as defined in Problem 2.2, instead of $p_1, \ldots, p_n$.

1. What is the average length L of your code?
2. Consider the functional $\sum_i p_i \log(\frac{p_i}{q_i})$. This is called the Kullback-Leibler distance and it is usually denoted by $D(p_i||q_i)$. Show that $D(p_i||q_i) \geq 0$ with equality if and only if $p_i = q_i$ for all i.
3. Show that $L = H(S) + D(p_i||q_i)$. This means that we pay a penality for designing the code for the wrong distribution. This penality is given by the Kullback-Leibler distance.

PROBLEM 4. A source S_1 outputs letters from the English alphabet (which contains 26 letters) with probabilities $p_1, p_2, \ldots, p_{26}$. You construct an optimal prefix-free code C_1 for this source. Let the average length of this code be L_1. Suppose now the Canton de Vaud is taken over by Bern and hence the source becomes Swiss German. Call it S_2. Thus, instead of outputting u with probability p_{21}, S_2 outputs u with probability p'_{21} and ü with probability p''_{21}, where $p'_{21} + p''_{21} = p_{21}$. You construct an optimal code C_2 for S_2. Let the average length of C_2 be L_2.

1. Is $H(S_2)$ smaller, larger or equal to $H(S_1)$? Justify your answer.
2. Is L_2 smaller, larger or equal to L_1? Justify your answer.

Since it is inconvenient to change the whole code, assume that instead we just simply adapt the code C_1 to the source S_2. You construct the code C_1' by expanding the codeword corresponding to u with a 0 for u and a 1 for ü. Denote the average length of C_1' by L_1'.

1. Is C_1' still prefix free? Justify your answer.
2. What is the difference between the average length of your new code C_1' and the average length of the original code C_1, i.e., $L_1' - L_1$?
3. Can you bound the difference between the average length of C_2 and the average length of the original code C_1?