ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Handout 5 Information Theory and Coding

Solutions to homework 2 September 27, 2011

PROBLEM 1.

(a) Code I is prefix-free, Code II is not.

(b) Both codes are uniquely decodable: Code I because it is instantaneous, Code II because the 1's at the beginning of each code word act as markers that separates the codewords and the decoding can be performed by counting the 0's between the 1's.

(c) Let X be the the indicator random variable of the event that the source letter is a_1, i.e.,

$$X = \begin{cases} 1 & \text{if the source letter is } a_1, \\ 0 & \text{otherwise} \end{cases}$$

and let Y be the indicator random variable of the event that the first letter of the codeword is a 1. The question is to find $I(X;Y)$. For Code I, we see that $X = Y$, and

$$I(X;Y) = H(X) - H(X|Y) = H(X) = -0.4 \log 0.4 - 0.6 \log 0.6.$$

For Code II, we see that $Y = 1$ with probability 1 (and thus independent of X) and thus $I(X;Y) = 0$.

(d) Since each codeword of code II begins with the letter 1 and since the letter 1 only appears at the beginning of codewords, this letter acts as an indicator of start of a codeword.

PROBLEM 2. Since the class of instantaneous codewords is a subset of the class of uniquely decodable codewords, it follows that $\bar{M}_2 \leq \bar{M}_1$. On the other hand, let $\{l_i\}$ be the codeword lengths of the uniquely decodable code for which $\bar{M} = \bar{M}_2$. Since $\{l_i\}$ satisfies the Kraft's inequality, there exists an instantaneous code with these codeword lengths. For this instantaneous code $\bar{M} = M_2$ and we see that $\bar{M}_1 \leq \bar{M} = M_2$, and we conclude that $\bar{M}_1 = \bar{M}_2$.

PROBLEM 3.

(a) $\{00, 01, 100, 101, 1100, 1101, 1110, 1111\}$.

(b) For $i > j$ observe that

$$Q_i - Q_j = \sum_{k=j}^{i-1} P(a_k) \geq P(a_j) \geq 2^{-l_j}.$$

So, the binary expansion of Q_i and Q_j must differ somewhere in the first l_j bits (if they did not the difference between Q_i and Q_j would have been less than 2^{-l_j}). Since codewords for i and j are at least l_j bits long, this implies that neither codeword can be a prefix of the other. The bound on the average codeword length follows from

$$-\log_2 P(a_i) \leq l_i < -\log_2 P(a_i) + 1.$$

This method of coding is also known as Shannon coding and predates Huffman coding.

PROBLEM 4.

(a) Consider the longest and the shortest codewords. We know that there are at least two longest codewords, suppose their length is l. Suppose the shortest codewords has length s. If s and l differ by more than 1, then we can increase the length of the shortest codeword by 1 ($s' = s + 1$) and shorten the two longest codewords by 1 ($l' = l - 1$) and still satisfy Kraft inequality:

$$[2^{-s'} + 2^{-l'} + 2^{-l'}] - [2^{-s} + 2^{-l} + 2^{-l}] = 2^{-(l-1)} - 2^{-(s+1)} \le 0.$$

But since all the codewords are equally likely, this would have decreased the average codeword length, contradicting the optimality of the Huffman code. Thus, the longest and shortest codeword lengths can differ by at most 1, and, again by Kraft inequality, their lengths must be j and $j + 1$.

(b) Let the number of codewords of length k be m_k, $k = j, j+1$. Since Huffman procedure yields a complete tree all intermediate nodes have two children. Thus, the 2^j nodes at level j of the tree are either codewords (m_j of them) or each of their two children are codewords ($m_{j+1}/2$ of them). Thus

$$m_j + m_{j+1}/2 = 2^j,$$

and also $m_j + m_{j+1} = x2^j$. From these two equations we find

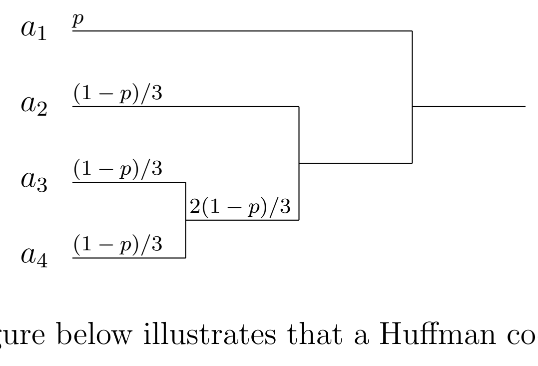
$$m_j = (2 - x)2^j \quad \text{and} \quad m_{j+1} = (x - 1)2^{j+1}.$$

(c) By the result above the average codeword length is

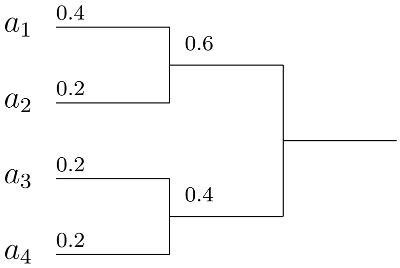
$$[jm_j + (j + 1)m_{j+1}]/(x2^j) = j + 2(x - 1)/x.$$

PROBLEM 5.

(a) Let $p = P(a_1)$, thus $P(a_2) = P(a_3) = P(a_4) = (1 - p)/3$. By the Huffman construction (see figure below) we must have $p > 2(1 - p)/3$, i.e., $q = 2/5$ in order to have $n_1 = 1$.



(b) With $P(a_1) = q$, the figure below illustrates that a Huffman code exists with $n_1 > 1$.



(c) & (d) For $K = 2$, n_1 is always 1. For $K = 3$, $n_1 = 1$ is guaranteed by $P(a_1) > P(a_2) \geq P(a_3)$. Now take $K \geq 4$ and assume $P(a_1) > 2/5$ and $P(a_1) > P(a_2) \geq \cdots \geq P(a_K)$. The Huffman procedure will combine a_{K-1} and a_K to obtain a super-symbol with probability

$$P(a_{K-1}) + P(a_K) < 2\frac{3/5}{K-1} \leq 2/5.$$

Thus, in the reduced ensemble a_1 is still the most likely element. Repeating the argument until $K = 3$, we see that $P(a_1) > q$ guarantees $n_1 = 1$ in all cases.

(e) For $K < 3$ no such q' exists. For $K \geq 3$, we claim $q' = 1/3$. Assume a_1 remains unpaired until the 2nd to last stage (otherwise there is nothing to prove). At this stage we have three nodes, and $P(a_1) < q'$ must be strictly less than one of the other two (otherwise all three would have been less than 1/3). Thus a_1 will be combined with one of them, leading to $n_1 > 1$.

Problem 6.

$$\begin{aligned} H(X) &= -\sum_{k=1}^{M} P_X(a_k) \log P_X(a_k) \\ &= -\sum_{k=1}^{M-1} (1-\alpha)P_Y(a_k) \log[(1-\alpha)P_Y(a_k)] - \alpha \log \alpha \\ &= (1-\alpha)H(Y) - (1-\alpha)\log(1-\alpha) - \alpha \log \alpha \end{aligned}$$

Since Y is a random variable that takes $M - 1$ values $H(Y) \leq \log(M - 1)$ with equality if and only if Y takes each of its possible values with equal probability.