ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Handout 7 Information Theory and Coding
Solution to homework 3 October 9, 2012

PROBLEM 1. *An* optimal set of codewords for the the two sources are as follows:

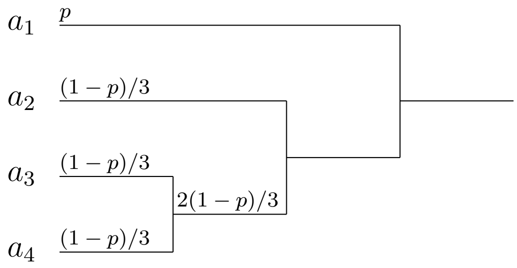
Source I		Source II	
Binary	Ternary	Binary	Ternary
00	0	00	0
01	10	01	1
100	11	100	21
101	12	101	20
110	20	110	220
111	21	1110	221
		1111	222

with average codeword lengths 2.5, 1.7, 2.55, 1.65 digits/symbol, in the order the codes appear in the table.

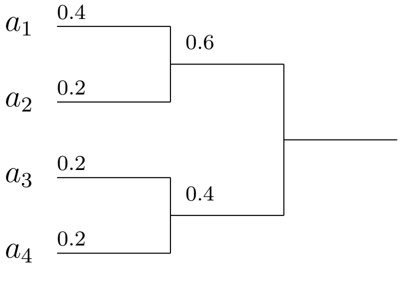
Note that for the ternary code for Source I, we need to add to the symbols of the source an extra symbol of probability zero so that the number of symbols equal 1 modulo $D-1$.

PROBLEM 2.

(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 3.

(a) Since the lengths prescribed satisfy the Kraft inequality, an instantaneous code can be used for the final stage of encoding the intermediate digits to binary codewords. In this case, each stage of the encoding is uniquely decodable, and thus the overall code is uniquely decodable.

(b) The indicated source sequences have probabilities $0.1, (0.9)(0.1), (0.9)^2(0.1), (0.9)^3(0.1), \ldots, (0.9)^7(0.1), (0.9)^8$. Thus,

$$\bar{N} = \sum_{i=1}^{8} i(0.1)(0.9)^{i-1} + 8(0.9)^8 = 5.6953.$$

(c)

$$\bar{M} = 1(0.9)^8 + 4[1 - (0.9)^8] = 2.7086.$$

(d) Let $N(i)$ be the number of source digits giving rise to the first i intermediate digits. For any $\epsilon > 0$

$$\lim_{i\to\infty} \Pr\left[\left|\frac{N(i)}{i} - \bar{N}\right| > \epsilon\right] = 0.$$

Similarly, let $M(i)$ be the number of encoded bits corresponding the the first i intermediate digits. Then

$$\lim_{i\to\infty} \Pr\left[\left|\frac{M(i)}{i} - \bar{M}\right| > \epsilon\right] = 0.$$

From this, we see that for any $\epsilon > 0$,

$$\lim_{i\to\infty} \Pr\left[\left|\frac{M(i)}{N(i)} - \frac{\bar{M}}{\bar{N}}\right| > \epsilon\right] = 0,$$

and that for a long source sequence the number of encoded bits per source digit will be $\bar{M}/\bar{N} = 0.4756$.

The average length of the Huffman code encoding 4 source digits at a time is 1.9702, yielding $1.9702/4 = 0.49255$ encoded bits per source digit.

For those of you puzzled by the fact that the 'optimum' Huffman code gives a worse result for this source than the run-length coding technique, observe that the Huffman code is the optimal solution to a mathematical problem with a given message set, but the choice of a message set can be more important than the choice of code words for a given message set.

PROBLEM 4. Let X^i denote $X_1, \dots, X_i$.

(a) By the chain rule for entropy,

$$\frac{H(X_1, X_2, \dots, X_n)}{n} = \frac{\sum_{i=1}^n H(X_i|X^{i-1})}{n} \tag{1}$$
$$= \frac{H(X_n|X^{n-1}) + \sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n} \tag{2}$$
$$= \frac{H(X_n|X^{n-1}) + H(X_1, X_2, \dots, X_{n-1})}{n}. \tag{3}$$

From stationarity it follows that for all $1 \leq i \leq n$,

$$H(X_n|X^{n-1}) \leq H(X_i|X^{i-1}),$$

which further implies, by summing both sides over $i = 1, \dots, n-1$ and dividing by $n-1$, that,

$$H(X_n|X^{n-1}) \leq \frac{\sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n-1} \tag{4}$$
$$= \frac{H(X_1, X_2, \dots, X_{n-1})}{n-1}. \tag{5}$$

Combining (3) and (5) yields,

$$\frac{H(X_1, X_2, \dots, X_n)}{n} \leq \frac{1}{n}\left[\frac{H(X_1, X_2, \dots, X_{n-1})}{n-1} + H(X_1, X_2, \dots, X_{n-1})\right] \tag{6}$$
$$= \frac{H(X_1, X_2, \dots, X_{n-1})}{n-1}. \tag{7}$$

(b) By stationarity we have for all $1 \leq i \leq n$,

$$H(X_n|X^{n-1}) \leq H(X_i|X^{i-1}),$$

which implies that,

$$H(X_n|X^{n-1}) = \frac{\sum_{i=1}^n H(X_n|X^{n-1})}{n} \tag{8}$$
$$\leq \frac{\sum_{i=1}^n H(X_i|X^{i-1})}{n} \tag{9}$$
$$= \frac{H(X_1, X_2, \dots, X_n)}{n}. \tag{10}$$

PROBLEM 5. For a Markov chain, given X_0 and X_n are independent given X_{n-1}. Thus

$$H(X_0|X_nX_{n-1}) = H(X_0|X_{n-1})$$

But, since conditioning reduces entropy,

$$H(X_0|X_nX_{n-1}) \leq H(X_0|X_n).$$

Putting the above together we see that $H(X_0|X_{n-1}) \leq H(X_0|X_n)$.

Problem 6.

$X_1, X_2, \ldots$ are i.i.d. with distribution $p(x)$. Hence $\log p(X_i)$ are also i.i.d. and

$$\begin{aligned}
\lim(p(X_1, \ldots, X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}}} \\
&= 2^{\lim \frac{1}{n} \sum \log p(X_i)} \\
&= 2^{E(\log(p(X)))} \text{ a.e.} \\
&= 2^{-H(X)}
\end{aligned}$$

by the strong law of large numbers (assuming of course that $H(X)$ exists). Note: The abbreviation a.e. stands for ‘almost everywhere’, which is synonymous with ‘with probability 1’.

For the second part of the problem we had intended to ask a question for which taking limit and taking the expectation do not commute (i.e., the order you take them matters). This, however, is not the case here: Let G_n be the set of $(x_1, \ldots, x_n)$ for which $|p(x_1, \ldots, x_n)^{1/n} - 2^{-H(X)}| < \epsilon$. We know from the first part that $\Pr(G_n) \to 1$ as n gets large. Since $0 \le p(x_1, \ldots, x_n)^{1/n} \le 1$ for any $x_1, \ldots, x_n$, we see that

$$\begin{aligned}
|E[p(X_1, \ldots, X_n)^{1/n}] - 2^{-H(X)}| &= \left| \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \left[p(x_1, \ldots, x_n)^{1/n} - 2^{-H(X)} \right] \right| \\
&\le \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \left| p(x_1, \ldots, x_n)^{1/n} - 2^{-H(X)} \right| \\
&= \sum_{(x_1, \ldots, x_n) \in G_n} p(x_1, \ldots, x_n) \left| p(x_1, \ldots, x_n)^{1/n} - 2^{-H(X)} \right| \\
&\quad + \sum_{(x_1, \ldots, x_n) \notin G_n} p(x_1, \ldots, x_n) \left| p(x_1, \ldots, x_n)^{1/n} - 2^{-H(X)} \right| \\
&\overset{(a)}{\le} \sum_{(x_1, \ldots, x_n) \in G_n} p(x_1, \ldots, x_n) \epsilon + \sum_{(x_1, \ldots, x_n) \notin G_n} p(x_1, \ldots, x_n) \\
&= P(G_n)\epsilon + (1 - P(G_n)) \\
&\le \epsilon + (1 - P(G_n)).
\end{aligned}$$

Where (a) follows from the definition of G_n and the fact that $|p(x_1, \ldots, x_n)^{1/n} - 2^{-H}| \le 1$. Now, as n gets large $1 - P(G_n)$ approaches zero, and we see that the difference between $E[p(X_1, \ldots, X_n)^{1/n}]$ and $2^{-H(X)}$ gets smaller than any arbitrary $\epsilon > 0$, and thus

$$\lim_{n \to \infty} E[p(X_1, \ldots, X_n)^{1/n}] = 2^{-H(X)}.$$

Problem 7.

(a) Using the chain rule for conditional entropy,

$$H(X, Y \mid Z) = H(X \mid Z) + H(Y \mid X, Z) \ge H(X \mid Z),$$

with equality iff $H(Y \mid X, Z) = 0$, that is, when Y is a function of X and Z.

(b) Using the chain rule for mutual information,

$$I(X, Y; Z) = I(X; Z) + I(Y; Z \mid X) \ge I(X; Z),$$

with equality iff $I(Y; Z \mid X) = 0$, that is, when Y and Z are conditionally independent given X.

(c) Using first the chain rule for entropy and then the definition of conditional mutual information,

$$\begin{aligned} H(X,Y,Z) - H(X,Y) &= H(Z \mid X,Y) = H(Z \mid X) - I(Y;Z \mid X) \\ &\leq H(Z \mid X) = H(X,Z) - H(X)\,, \end{aligned}$$

with equality iff $I(Y;Z \mid X) = 0$, that is, when Y and Z are conditionally independent given X.

(d) Using the chain rule for mutual information,

$$I(X;Z \mid Y) + I(Z;Y) = I(X,Y;Z) = I(Z;Y \mid X) + I(X;Z)\,,$$

and therefore

$$I(X;Z \mid Y) = I(Z;Y \mid X) - I(Z;Y) + I(X;Z)\,.$$

We see that this inequality is actually an equality in all cases.