332:542 **Information Theory and Coding** **Examination 1**

March 8, 2002

SOLUTION

This is an 80 minute exam. You may have an additional 100 minutes to answer the following five questions in the notebooks provided. You are permitted 2 double-sided sheet of notes. Make sure that you have included your name, ID number (last 4 digits only) and signature in each book used (*5 points*). Read each question carefully. All statements must be justified. Computations should be simplified as much as possible.

1. *20 points* Let $Y = g(X)$ be deterministic function of discrete random variable X.

   (a) *5 points* Give an example of a random variable X and a function $Y = g(X)$ such that $H(Y) < H(X)$.

   Suppose X is equiprobably 0 or 1, so $p_X(0) = p_X(1) = 1/2$ and $H(X) = 1$. Let $g(x) = 0$ for all x. Then $Y = g(X) = 0$ and $H(Y) = 0 < H(X)$.

   (b) *5 points* Give an example of a random variable X and a function $Y = g(X)$ such that $H(Y) = H(X)$

   A REALLY dumb (but valid) example is $g(x) = x$. Any other function $g(x)$ that is one to one also provides $H(Y) = H(X)$

   (c) *10 points* Either give an example of a random variable X and a function $Y = g(X)$ such that $H(Y) \geq H(X)$ or prove that $H(Y) \leq H(X)$. Hint: look at $H(X, Y)$.

   The question was supposed to ask whether we can have $H(Y) > H(X)$ not $H(Y) \geq H(X)$.

   The hint is that $H(X, Y) = H(X)$ because if you know X, then you know $Y = g(X)$ so Y creates no additional randomness. Expressed another way, $H(Y|X) = 0$ since Y is a deterministic function of X. As a result,

   $$H(X) = H(X, Y) = H(Y) + H(X|Y)) \geq H(Y)$$

   since $H(X|Y) \geq 0$.

2. *20 points* Let be random variables such that

   $$X \to Y \to Z, \qquad Y \to Z \to X, \qquad Z \to X \to Y$$

   If $I(X; Y) = 3$, find $I(X; Z)$ and $I(Y; Z)$. Or, if these quantities cannot be known, find tight upper and lower bounds. Make sure to justify your answers.

   This problem is just about the data processing inequality.

   $$X \to Y \to Z \implies I(X; Y) \geq I(X; Z) \tag{1}$$
   $$Y \to Z \to X \implies I(Y; Z) \geq I(Y; X) = I(X; Y) \tag{2}$$
   $$Z \to X \to Y \implies I(Z; X) \geq I(Z; Y) \tag{3}$$

   Combining (2) and (3), we have

   $$I(X; Z) = I(Z; X) \geq I(Z; Y) = I(Y; Z) \geq I(X; Y)$$

   Combining this result with (1), we obtain $I(X; Z) = I(X; Y) = 3$. Now, from (3), $I(Y; Z) \leq I(Z; X) = I(X; Y) = 3$. Combined with (2), we have $I(Y; Z) = I(X; Y) = 3$. To summarize,

   $$I(X; Y) = I(X; Z) = I(Y; Z) = 3$$

3. *15 points* Consider the code $\{0, 01\}$. Justify your answers to the following questions:

   (a) *5 points* Is the code instantaneous?
   No, since 0 *is a prefix of* 01.

   (b) *5 points* Is the code nonsingular?
   Yes, remember that noinsingular just means that two distinct inputs are mapped to distinct code symbols. Although we didn't specify the inputs, whatever they are, the output 0 *and* 01 *are distinct.*

   (c) *5 points* Is the code uniquely decodable?
   Yes, given a code sequence, the key is that each time we observe a 1*, it must the second character of the codeword* 01*. Preceding the* 1 *will be* $n \geq 1$ *zeros, corresponding to* $n-1$ *occurrences of codeword* 0 *followed by codeword* 01*. As an example,* 00101000101001 *decodes as*

$$00101000101001 \implies 0 \cdot 01 \cdot 01 \cdot 0 \cdot 0 \cdot 01 \cdot 01 \cdot 0 \cdot 01$$

4. *10 points* The source coding theorem shows that the optimal source code for random variable X has expected length $L \leq H(X) + 1$. Find an example of a random variable X for which the optimal code has $L > H(X) + 1 - \epsilon$ for any small $\epsilon > 0$. Make sure you justify your answer.

   Let X *be a Bernoulli random variable with* $p_X(0) = \delta = 1 - p_X(1)$*. So* $H(X) = H(\delta, 1-\delta)$*. For sufficiently small* δ*, we can made* $H(X) < \epsilon$*, or, equivalently,* $H(X) - \epsilon < 0$*.*

   For the source X*, the Huffman code is optimal, but the Huffman code is simply* $0 \to 0$ *and* $1 \to 1$*, which has average length* $L = 1$*. Thus*

$$L = 1 > 1 + H(X) - \epsilon$$

5. *30 points* A source has an alphabet of 4 letters, a_1, a_2, a_3, a_4 with probabilities $p_1 \geq p_2 \geq p_3 \geq p_4$.

   (a) Suppose $p_1 > p_2 = p_3 = p_4$. Find the smallest number q such that $p_1 > q$ implies $n_1 = 1$, where n_1 is the length of the code word for a_1 in a binary Huffman code for the source.
   In this case, $p_2 = p_3 = p_4 = (1 - p_1)/3$*. Here, the Huffman code combines symbols 3 and 4 into a super-letter with probability* $2(1-p_1)/3$*. As long the superletter probability is strictly less than* p_1*, then the superletter will be combined with* a_2*, yielding* $n_1 = 1$*. That is, we must have* $p_1 > 2(1 - p_1)/3$ *to ensure* $n_1 = 1$*. Equivalently,* $p_1 > 0.4 = q$ *implies* $N_1 = 1$*.*

   (b) Show by example that if $p_1 = q$ (your answer in part (a)), then a Huffman code exists with $n_1 > 1$.
   If $p_1 = 0.4$ *and* $p_2 = p_3 = p_4 = 0.2$*, then we combine* a_3 *and* a_4 *into a super-letter* a' *with probability* 0.4*. Since* a_1 *and* a' *both have probability* 0.4*, at the second step, we can choose to combine* a_1 *and* a_2*. The result is that* $n_1 = 2$*.*

   (c) Now assume the more general condition $p_1 > p_2 \geq p_3 \geq p_4$. Does $p_1 > q$ still imply that $n_1 = 1$? Why or why not?

Yes, $p_1 > q = 0.4$ implies $n_1 = 1$. For a proof by contradiction, suppose $p_1 > 0.4$ and there is a Huffman procedure yielding $n_1 > 1$. In this case, we must have $p_3 + p_4 \geq p_1 > 0.4$. Since $p_2 \geq p_3$ and $p_2 \geq p_4$, we have that

$$p_2 \geq \frac{p_3 + p_4}{2} > \frac{0.4}{2} = 0.2$$

It follows that

$$p_1 + p_2 + (p_3 + p_4) > 0.4 + 0.2 + 0.4 = 1$$

which is a contradiction since $p_1 + p_2 + p_3 + p_4 = 1$.

(d) Now assume that the source has an arbitrary number K of letters, with $p_1 > p_2 \geq \cdots \geq p_K$. Does $p_1 > q$ now imply that $n_1 = 1$? Explain.

Yes, $p_1 > q = 0.4$ is sufficient to ensure $n_1 = 1$. This can be proven by induction. It is trivially true for $K = 1, 2, 3$ and we have shown it is true for $K = 4$. Suppose it is true that $n_1 = 1$ for any collection of K letters with $p_1 > 0.4$ and $p_1 > p_2 \geq p_3 \geq \cdots p_K$. Now suppose we have a source with $K + 1$ letters with probabilities $p_1 > p_2 \geq p_3 \geq \cdots p_{K+1}$ such that the Huffman procedure produces $n_1 > 1$. At the first step, the Huffman procedure combines a_K and a_{K+1} into a super-letter with probability $p_K + p_{K+1}$. If $p_K + p_{K+1} < p_1$, then the we have remaining an encoding problem with only K letters, $p_1 > 0.4$ and a_1 strictly the most probable symbol. By our induction hypothesis, the Huffman procedure will yield $n_1 = 1$. Otherwise, if $p_K + p_{K+1} \geq p_1 > 0.4$, then we must have

$$p_j \geq \frac{p_K + p_{K+1}}{2} > 0.2 \qquad j = 2, \ldots, K-1$$

Since $K \geq 4$, it follows that

$$p_1 + (p_2 + \cdots + p_{K-1}) + (p_K + p_{K+1}) > 0.4 + (K-2)(0.2) + 0.4 \geq 1.2$$

which a contradiction.

(e) *20 points* Now assume the source has K letters $a_1, \ldots, a_K$, with $p_1 \geq p_2 \geq \cdots \geq p_K$. Find the largest number q' such that $p_1 < q'$ implies that $n_1 > 1$.

First, we should not that this problem only makes sense for $K \geq 4$ since $K \leq 3$ and $p_1 \geq p_2 \geq p_3$ always implies there exist a Huffman code with $n_1 = 1$.

Examining the $K = 4$ case makes it easy to "guess" q'. When $p_1 = q'$, we need to guarantee that $p_3 + p_4 = q'$. That is, we need to ensure that the smallest possible $p_3 + p_4$ is $p_3 + p_4 = q'$ given $p_1 = q'$. Given $p_1 = q'$, $p_3 + p_4$ is minimized when p_2 is as large as possible. This occurs when $p_2 = p_1 = q'$. Thus we obtain

$$1 = p_1 + p_2 + (p_3 + p_4) = 3q' \quad \Longrightarrow \quad q' = 1/3$$

A more careful analysis goes something like this. First, we can conclude that $q' \leq 1/3$ since if $q' > 1/3$, then we can choose $p_1 = 1/3$ and a probability distribution for the other letters $p_2, \ldots, p_K$ such that after $K - 3$ Huffman steps, we have probabilities $\{p_1, p_2', p_3'\} = \{1/3, 1/3, 1/3\}$. In this case, the Huffman procedure can choose to encode a_1 with length $n_1 = 1$. To see that $q' = 1/3$. For pruposes of contradiction, suppose $p < 1/3$ and the Huffman procedure allows $n_1 = 1$. In this case, after $K - 3$ Huffman steps, we must have probabilities $\{p_1, p_2', p_3'\}$ such that $p_1 \geq p_2'$ and $p_1 \geq p_3'$. Hence

$$p_2' \leq p_1 < 1/3 \qquad p_3' \leq p_1 < 1/3$$

It follows that $p_1 + p_2' + p_3' < 1$, which is a contradiction.