ECE 776 - Information theory (Fall 2008)
Midterm

Please give well-motivated answers.

Q1 (1 point). Find whether $H(X|Z) \lesseqqgtr H(X|Y)$ for X, Y, Z real variables and: (a) $Z = |Y|$; (b) $Z = Y^3$.

Sol.: For both cases (a) and (b), we have the Markov chain $X - Y - Z$, so that in general from the data processing inequality

$$I(X;Z) \leq I(X;Y)$$

and thus (by using $I(X;Z) = H(X) - H(X|Z)$ and $I(X;Y) = H(X) - H(X|Y)$)

$$H(X|Z) \geq H(X|Y).$$

However, for the case (b), the function between Y and Z is one-to-one, so that the above inequalities hold with equality (in fact, we also have $X - Z - Y$).

Q2 (1 point) Given the joint pmf $p(x, y)$ defined as below

$x \backslash y$	0	1
0	0.1	0.6
1	0.2	0.1

are the sequences $x^5 = 00010$ and $y^5 = 01111$ jointly typical (i.e., belonging to set $A_\epsilon^{(5)}$) given $\epsilon = 0.15$? Are they individually typical with respect to the marginal distributions $p(x)$ and $p(y)$?

Sol: We have the marginals $p(x) = (0.7, 0.3)$ and $p(y) = (0.3, 0.7)$ so that $H(X) = H(Y) = -0.3 \log_2 0.3 - 0.7 \log_2 0.7 = 0.8813$ bits. Moreover, the joint entropy is $H(X,Y) = -0.2 \log_2 0.1 - 0.2 \log_2 0.2 - 0.6 \log_2 0.6 = 1.571$. Now, evaluating the individual empirical entropy $-1/5 \log_2 p(x^5)$ (and similarly for x^5), we get

$$-\frac{1}{5}\log_2(0.7^4 \cdot 0.3) = 0.7591 = 0.8813 \pm 0.15,$$

so that both sequences are individually typical. To check whether they are jointly typical, we must calculate

$$-\frac{1}{5}\log_2(0.1 \cdot 0.6^3 \cdot 0.1) = 1.771 \neq 1.571 \pm 0.15.$$

Therefore, the sequences are not jointly typical with respect to the given joint distribution.

Q3 (1 point) Given the discrete memoryless channel defined by

$$p(y|x) = \begin{bmatrix} 1/4 & 3/4 & 0 \\ 3/4 & 0 & 1/4 \\ 0 & 1/4 & 3/4 \end{bmatrix},$$

calculate the capacity.

Sol.: The channel is symmetric, and therefore we have

$$C = \log_2 3 - H(1/4) = 0.774 \text{ bits/ channel use.}$$

Q4 (1 point) A radio signal X is received via two antennas, whose corresponding received signals are Y_1 and Y_2. The noises at the two antennas are independent and have the same statistics, so that $p(y_1, y_2|x) = p(y_1|x)p(y_2|x)$, with $p(y_1|x) = p(y_2|x)$ if $y_1 = y_2$ (i.e., Y_1 and Y_2 are conditionally independent and identically distributed given X). Prove that

$$I(X; Y_1, Y_2) = 2I(X; Y_1) - I(Y_1; Y_2).$$

Based on this result, argue that the capacity of the two-antenna channel is less than twice the capacity of the single-antenna channel that only measures Y_1 (or Y_2).

Sol.: We can write

$$\begin{aligned} I(X; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X) \\ &= H(Y_1, Y_2) - H(Y_1|X) - H(Y_2|X) = \\ &= H(Y_2) + H(Y_1) - I(Y_1; Y_2) - H(Y_1|X) - H(Y_2|X) = \\ &= -I(Y_1; Y_2) + 2I(X; Y_1) \end{aligned}$$

where in the second line we have used the fact that $Y_1 - X - Y_2$ (i.e., Y_1 and Y_2 are conditionally independent given X) and in the fourth, we have used the fact that Y_1 and Y_2 are conditionally and unconditionally identically distributed.
The capacity C_2 of the two-antenna system is then obtained as

$$C_2 = \max_{p(x)} I(X; Y_1, Y_2) = \max_{p(x)} 2I(X; Y_1) - I(Y_1; Y_2) \leq 2 \max_{p(x)} I(X; Y_1) = 2C_1,$$

where C_1 is the capacity of the one-antenna system.

P1 (2 point) - Generalizing the Fano inequality
We want to estimate a quantity X (taking values in a set $\mathcal{X}$) via the observation Y. Instead of producing a standard estimator $\hat{X}(Y)$ (i.e., $\hat{X}$ function of Y) and requiring that $\hat{X}(Y) = X$ (no error) with large probability, we require less from the estimate. The estimator in fact is not a single value $\hat{X}(Y)$ but rather a list of values in $\mathcal{X}$, say $L(Y)$, which depends on the observation Y. The number of elements in the list is $|L|$ (same for all Y). We define the probability of error as the probability that the real quantity X is not in the list $L(Y)$: $P_e = \Pr[X \notin L(Y)]$. Following the steps of the proof of the Fano inequality, show that

$$H(X|Y) \leq P_e \log |\mathcal{X}| + (1 - P_e) \log |L| + H(P_e).$$

Interpret this result and compare it with the standard Fano inequality (how do we get the standard Fano inequality from the relationship above?).
(Hint: As in the proof of the Fano inequality start by defining a variable E that identifies the error event).

Sol: Define the error event E

$$E = \begin{cases} 1 & \text{if } X \notin L(Y) \\ 0 & \text{if } X \in L(Y) \end{cases}$$

and notice that

$$H(X|Y) = H(X, E|Y)$$

since $H(X, E|Y) = H(X|Y) + H(E|X, Y)$ and $H(E|X, Y) = 0$. Now, we can write

$$\begin{aligned} H(X, E|Y) &= H(E|Y) + H(X|E, Y) \\ &\leq H(E) + (1 - P_e)H(X|E = 0, Y) + P_e H(X|E = 1, Y) \\ &\leq H(P_e) + (1 - P_e)\log |L| + P_e \log |\mathcal{X}|, \end{aligned}$$

since $H(X|E = 0, Y) \leq H(X|E = 0) \leq \log |L|$.
The standard Fano inequality is recovered for $|L| = 1$.

P2 (2 point)
(a) Assume that sequences x^n and y^n, taking values in sets $\mathcal{X}$ and $\mathcal{Y}$ respectively, satisfy $(x^n, y^n) \in A_\epsilon^{(n)}$ with respect to a joint distribution $p(x, y)$ (i.e., the sequences x^n and y^n are jointly typical). Show that the following is true of the conditional probability $p(y^n|x^n)$

$$2^{-n(H(Y|X)+2\epsilon)} \leq p(y^n|x^n) \leq 2^{-n(H(Y|X)-2\epsilon)}$$

(Hint: Use the definitions of typicality, joint typicality and of conditional distribution)
(b) Define $A_\epsilon^{(n)}(x^n)$ as the set of all sequences $y^n \in \mathcal{Y}^n$ that are jointly typical with a given $x^n \in A_\epsilon^{(n)}(X)$ (x^n is individually typical), that is,

$$A_\epsilon^{(n)}(x^n) = \{y^n\colon\ (x^n, y^n) \in A_\epsilon^{(n)}\}.$$

Show that $|A_\epsilon^{(n)}(x^n)| \leq 2^{n(H(Y|X)+2\epsilon)}$.
(Hint: Follow the proof of the AEP and use the result at the previous point)
(c) Fixing a given sequence $x^n \in A_\epsilon^{(n)}(X)$ (x^n is individually typical), prove the following regarding the probability that a randomly and independently generated sequence Y^n is jointly typical with x^n

$$\Pr[(Y^n, x^n) \in A_\epsilon^{(n)}] \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

(Hint: Start by writing $\Pr[(Y^n, x^n) \in A_\epsilon^{(n)}] = \sum_{y^n \in A_\epsilon^{(n)}(x^n)} p(y^n)$, then use the definition of typicality and the result and the previous point)

Sol.: (a) We have $p(y^n|x^n) = p(x^n, y^n)/p(x^n)$ and

$$2^{-n(H(X)+\epsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\epsilon)}$$

$$2^{-n(H(X,Y)+\epsilon)} \leq p(x^n, y^n) \leq 2^{-n(H(X,Y)-\epsilon)}$$

by definition. It follows that

$$p(y^n|x^n) \leq \frac{2^{-n(H(X,Y)-\epsilon)}}{2^{-n(H(X)+\epsilon)}} = 2^{-n(H(X,Y)-H(X)-2\epsilon)} = 2^{-n(H(Y|X)-2\epsilon)}$$

and

$$p(y^n|x^n) \geq \frac{2^{-n(H(X,Y)+\epsilon)}}{2^{-n(H(X)-\epsilon)}} = 2^{-n(H(X,Y)-H(X)+2\epsilon)} = 2^{-n(H(Y|X)+2\epsilon)}.$$

(b) We have

$$1 \geq \sum_{y^n \in A_\epsilon^{(n)}(x^n)} p(y^n|x^n) \geq |A_\epsilon^{(n)}(x^n)| 2^{-n(H(Y|X)+2\epsilon)}$$

so that

$$|A_\epsilon^{(n)}(x^n)| \leq 2^{n(H(Y|X)+2\epsilon)}$$

(c) We have

$$\begin{aligned} \Pr[(Y^n, x^n) &\in A_\epsilon^{(n)}] = \sum_{y^n \in A_\epsilon^{(n)}(x^n)} p(y^n) \leq |A_\epsilon^{(n)}(x^n)| 2^{-n(H(Y)-\epsilon)} \\ &\leq 2^{n(H(Y|X)+2\epsilon)} 2^{-n(H(Y)-\epsilon)} = 2^{n(H(Y|X)-H(Y)+3\epsilon)} = \\ &= 2^{-n(I(X;Y)-3\epsilon)}. \end{aligned}$$

P3 (2 point) A random process Y_i ($i = 1, 2, ...$) is generated as shown in the figure below. Specifically, if random variable $Z = 0$, then $Y_i = X_{0i}$ for $i = 1, 2, ...$, and X_{0i} is a Markov chain with transition probabilities given by $\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$ (see figure); instead if $Z = 1$, we have that $Y_i = X_{1i}$ for $i = 1, 2, ...$, and X_{1i} is a Markov chain with transition probabilities given by $\begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix}$ (see figure). Assuming that variable Z is independent of all other variables and such that $\Pr(Z = 0) = 0.3$, and assuming that the two Markov chains are stationary (i.e., the stationary initial distribution is assumed), answer the following:
(a) Is the process Y_i stationary?
(b) Is the process Y_i a Markov chain?
(c) Are we guaranteed that the entropy rate $H(\mathcal{Y})$ exists? If so, calculate $H(\mathcal{Y})$.
(d) Is the process ergodic? Are $H(\mathcal{Y})$ bits/ symbol enough to have a lossless compression of the source?

Sol.: (a) Yes. In fact, the distribution $p_Y(y_{k_1}, y_{k_2}, ..., y_{k_m})$ for any given set of time instants $k_1, k_2, ..., k_m$ reads

$$p_Y(y_{k_1}, y_{k_2}, ..., y_{k_m}) = 0.3p_0(y_{k_1}, y_{k_2}, ..., y_{k_m}) + 0.7p_1(y_{k_1}, y_{k_2}, ..., y_{k_m}),$$

and $p_0(y_{k_1}, y_{k_2}, ..., y_{k_m})$ and $p_1(y_{k_1}, y_{k_2}, ..., y_{k_m})$ are the joint distributions for the two stationary Markov chains X_{0i} and X_{1i}.
(b) From the reasoning above, we can write:

$$\begin{aligned} p_Y(y_1, y_2, ..., y_n) &= 0.3p_0(y_1, y_2, ..., y_n) + 0.7p_0(y_1, y_2, ..., y_n) = \\ &= 0.3p_0(y_1)p_0(y_2|y_1)p_0(y_3|y_2)\cdots p_0(y_n|y_{n-1}) \\ &\quad +0.7p_1(y_1)p_1(y_2|y_1)p_1(y_3|y_2)\cdots p_1(y_n|y_{n-1}) \end{aligned}$$

with $p_0(y_n|y_{n-1})$ and $p_1(y_n|y_{n-1})$ denoting the transition probabilities for the two Markov chains. As such, we have that Y_i is not a Markov chain.

(c) Yes, since the process is stationary.

$$H(\mathcal{Y}) = \lim_{n\to\infty} \frac{H(Y^n)}{n}$$

and $H(Y^n) = H(Y^n|Z) + I(Y^n;Z)$ so that

$$H(\mathcal{Y}) = \lim_{n\to\infty} \frac{H(Y^n|Z)}{n}$$

where we have used the fact that $I(Y^n;Z)/n \leq H(Z)/n \to 0$. Now,

$$H(Y^n|Z) = 0.3H(X_0^n) + 0.7H(X_1^n)$$

so that

$$H(\mathcal{Y}) = 0.3H(\mathcal{X}_0) + 0.7H(\mathcal{X}_1).$$

The entropy rates of the two Markov chains are easily calculated

$$\begin{aligned} H(\mathcal{X}_0) &= H(X_{02}|X_{01}) = 2\frac{0.1}{0.2}H(0.1) = H(0.1) = 0.469 \\ H(\mathcal{X}_1) &= H(X_{12}|X_{11}) = \frac{0.8}{1.4}H(0.2) + \frac{0.6}{1.4}H(0.4) = 0.829 \end{aligned}$$

and finally,

$$H(\mathcal{Y}) = 0.3 \cdot 0.469 + 0.7 \cdot 0.829 = 0.721$$

(d) The process is not ergodic. This can be seen by, e.g., calculating the temporal average

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \to \begin{cases} E[X_{0i}] = 0.5 & \text{with prob. } 0.3 \\ E[X_{1i}] = 0.8/1.4 & \text{with prob. } 0.7 \end{cases},$$

while the ensemble average is $E[X_i] = 0.3 \cdot E[X_{0i}] + 0.7 \cdot E[X_{1i}]$. Thefore, the AEP does not apply and we cannot conclude that $H(\mathcal{Y})$ bits/ symbol are enough to have a lossless compression of the source.

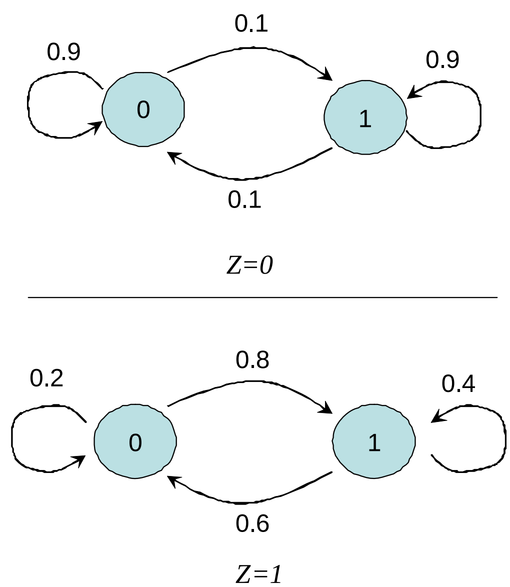

Figure 1: