ECE 776 - Information theory (Spring 2011)
Midterm

Please give well-motivated answers.

Q1 (1 point) A binary i.i.d. sequence $X^4 = (X_1, ..., X_4)$ is passed through a memoryless binary symmetric channel with cross-over probability p to produce the output sequence $Y^4 = (Y_1, ..., Y_4)$. In other words, $p_{Y|X}(1|0) = p_{Y|X}(0|1) = p$ and $p(y^4|x^4) = \prod_{i=1}^4 p_{Y|X}(y_i|x_i)$. Set $p_X(1) = q$ and calculate $p_Y(1)$. For $p = 0.2$ and $q = 0.2$, is $x^4 = (0, 0, 1, 0)$ a typical sequence in $A_{0.1}^{(4)}(p_X(x))$ (i.e., $\epsilon = 0.1$) and is $y^4 = (1, 0, 0, 0)$ in $A_{0.1}^{(4)}(p_Y(y))$?

Sol.: We have

$$H(X) = H(p_X(x)) = -0.2 \log_2 0.2 - 0.8 \log_2 0.8 = 0.722 \text{ bit.}$$

Moreover,

$$p_Y(1) = q(1-p) + (1-q)p = 2 \cdot 0.2 \cdot 0.8 = 0.32,$$

so that

$$H(Y) = H(p_Y(y)) = -0.32 \log_2 0.32 - 0.68 \log_2 0.68 = 0.904 \text{ bit.}$$

For sequence $x^4 = (0, 0, 1, 0)$, we calculate

$$\begin{aligned} -\frac{1}{4} \log_2 p(x^4) &= -\frac{1}{4} \sum_{i=1}^4 \log_2 p(x_i) = -\frac{1}{4}(3 \cdot \log_2 0.8 + \log_2 0.2) \\ &= 0.823 \leq H(X) + \epsilon, \end{aligned}$$

and $|0.823 - H(X)| > 0.1$, so we can conclude that $x^4 \notin A_{0.1}^{(4)}(p_X(x))$. For $y^4 = (1, 0, 0, 0)$,

$$-\frac{1}{4} \log_2 p(y^4) = -\frac{1}{4}(3 \cdot \log_2 0.68 + \log_2 0.32) = 0.828$$

and $|0.828 - H(Y)| \leq 0.1$, so $y^4 \in A_{0.1}^{(4)}(p_Y(y))$.

Q2 (1 point) For the problem at the point above, with any $0 \leq p,q \leq 1$, argue using the AEP that the number of typical sequences at the output, $A_\epsilon^{(n)}(p_Y(y))$, is larger than, or equal to, the number of typical sequences at the input, $A_\epsilon^{(n)}(p_X(x))$, for any n. For which values of p and q do the two sets have the same number of sequences? (you can neglect ϵ in your arguments).

Sol.: The set of typical sequences at the input has around $2^{nH(q)}$ sequences, whereas the set of typical output sequences is about $2^{nH(p*q)}$, where we denoted $p * q = q(1-p) + (1-q)p$. We thus need to argue that $H(p * q) \geq H(q)$. To see this you can use the strict concavity of the entropy. In fact,

$$H(q) = (1-p)H(q) + pH(1-q)$$

for any $0 \leq p, q \leq 1$ since $H(q) = H(1-q)$. Now using Jensen's inequality

$$(1-p)H(q) + pH(1-q) \leq H(p * q),$$

with equality if and only if $p = 0, 1$ or $q = 1/2$. Under the latter conditions, the sets of typical sequences at input and output have the same number of sequences.
Alternatively (ed equivalently), you can argue that $p * q$ is strictly closer to 0.5 than q for any $p \neq 0$ or 1 and $q \neq 0.5$.

Q3 (1 point) Consider now an asymmetric binary channel with $p_{Y|X}(1|0) = 0.9$ and $p_{Y|X}(0|1) = 0.1$. Moreover, the input is $X \sim Ber(0.2)$. Calculate the joint distribution of X and Y and find the best estimate $\hat{X}$ of X given Y as a function $\hat{X} = g(Y)$ (in the sense of minimizing the probability of error). Calculate the probability of error of such an estimator and check that it satisfies the Fano's inequality. Interpret your results.

Sol.: The joint distribution $p_{XY}(x, y)$ is

$$\begin{array}{cc|c} X\backslash Y & 0 & 1 \\ \hline 0 & 0.8 \cdot 0.1 = 0.08 & 0.8 \cdot 0.9 = 0.72 \\ 1 & 0.2 \cdot 0.1 = 0.02 & 0.2 \cdot 0.9 = 0.18 \end{array}.$$

Note that X and Y are independent. In fact, the probability of receiving Y is independent of which X is sent. We have $p_{XY}(x, y) = p_X(x)p_Y(y)$ with $p_Y(1) = 0.9$.
The best estimator is thus $\hat{x} = g(y) = 0$ for both $y = 0$ and $y = 1$. The error probability is thus

$$\begin{aligned} P_e &= \Pr[\hat{X} \neq X] = \Pr[X \neq 0] \\ &= 0.02 + 0.18 = 0.2. \end{aligned}$$

From Fano inequality, we have

$$H(X|Y) \leq H(X|\hat{X}) \leq H(P_e) + P_e(\log_2(2) - 1) = H(P_e),$$

but

$$\begin{aligned} H(X|Y) &= H(X|\hat{X}) = H(X) \\ &= H(0.8) = 0.72 \end{aligned}$$

and $H(P_e) = H(X) = H(0.8)$. Fano inequality is thus satisfied with equality.

Q4 (1 point) Consider an arbitrary binary random vector X^n, not necessarily i.i.d. or even stationary, such that $H(X^n) \geq n\alpha$ for some $0 \leq \alpha \leq 1$. Assume that this sequence goes through a memoryless binary symmetric channel with cross over probability p, producing Y^n. Then, the following inequality is well known:

$$H(Y^n) \geq nH(p * H^{-1}(\alpha)),$$

where $H^{-1}(\alpha)$ is the inverse of the binary entropy function in the interval $[0, 1/2]$ (i.e., $q = H^{-1}(\alpha)$ is the value of $0 \leq q \leq 1/2$ such that $H(q) = \alpha$) and $a * b = a(1 - b) + b(1 - a)$. Provide an example of a random vector $H(X^n)$ with $H(X^n) = n\alpha$ for which this inequality is strict (i.e., $H(Y^n) > nH(p * H^{-1}(\alpha))$). You can build your example with $n = 2$, $p = 0.1$ and $\alpha = \frac{1}{2}$. (Hint: The vector X^n should not be i.i.d.! Moreover, $H^{-1}(1/2) = 0.11$).

Sol.: First, we note that if X^n is i.i.d., then it must be $Ber(H^{-1}(\alpha))$, since we want $H(X^n) = nH(X) = nH(H^{-1}(\alpha)) = n\alpha$. In this case, we have $H(Y^n) = nH(Y) = nH(p * H^{-1}(\alpha))$, so that the inequality at hand holds with equality.
We then consider a constant vector $X_1 = X_2$. To impose $H(X^2) = H(X_i) = 2\frac{1}{2} = 1$, we select X_i as $Ber(1/2)$. To calculate, $H(Y^2) = H(Y_1, Y_2)$ we evaluate the joint distribution $p_{Y_1Y_2}(y_1, y_2)$:

$Y_1 \backslash Y_2$	0	1
0	$1/2 \cdot 0.9^2 + 1/2 \cdot 0.1^2 = 0.41$	$0.1 \cdot 0.9 = 0.09$
1	$0.1 \cdot 0.9 = 0.09$	$1/2 \cdot 0.1^2 + 1/2 \cdot 0.9^2 = 0.41$

.

We then obtain

$$\begin{aligned} H(Y_1, Y_2) &= -2 \cdot 0.41 \cdot \log_2(0.41) - 2 \cdot 0.09 \cdot \log_2(0.09) \\ &= 1.68. \end{aligned}$$

Which we must compare to

$$\begin{aligned} nH(p * H^{-1}(\alpha)) &= 2H(0.1 * H^{-1}(1/2)) \\ &= 2H(0.1 * 0.11) \\ &= 2H(0.188) = 1.39. \end{aligned}$$

We clearly have $H(Y^2) > 2H(0.1 * H^{-1}(1/2))$.

P1 (2 points) Your friend picks uniformly at random a card X from a deck of 40 cards, keeping it hidden from you. You can ask one question Q, out of a set of predetermined questions, about the card X, and receive an answer $A = f(X, Q)$, where $f(\cdot, \cdot)$ is a given deterministic function. When selecting the question Q, the object X is unknown, and thus Q and X must be independent.
a. You can choose a distribution $p(q)$ on the questions Q. In order to maximize the usefulness of such question, you want to maximize $I(Q, A; X)$. Why? Show that $I(Q, A; X) = H(A|Q)$ and interpret this equality (Hint: Use the chain rule).
b. Suppose that a second friend can select a question Q knowing X in order to increase $I(Q, A; X)$ and help you out. In this case, Q is not independent of X. Quantify the increase in mutual information $I(Q, A; X)$ with respect to point a. and interpret your result.
c. Now suppose that two independent questions Q_1, Q_2 bot distributed according to $p(q)$ are asked, independent of X, eliciting answers A_1 and A_2, where $A_i = f(X, Q_i)$ for $i = 1, 2$. Show that two questions are less valuable than twice a single question in the sense that $I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1)$ (Hint: The date processing inequality may be useful).

Sol.: a. Maximizing $I(Q, A; X)$ maximizes the amount of information that the pair (Q, A) brings about the card X. Using the fact that Q and X are independent, we have

$$\begin{aligned} I(Q, A; X) &= I(Q; X) + I(A; X|Q) \\ &= I(A; X|Q) \\ &= H(A|Q) - H(A|X, Q) \\ &= H(A|Q), \end{aligned}$$

where the last line follows from the fact that $A = f(X, Q)$. In other words, the amoung of information that (Q, A) brings about the card X is equal to the uncertainty about the answer A when asking question Q.
b. We now have

$$\begin{aligned} I(Q, A; X) &= I(Q; X) + I(A; X|Q) \\ &= I(Q; X) + H(A|Q), \end{aligned}$$

so that the increase in mutual information is $I(Q; X)$. This can be easily interpreted – The choice of the question itself brings information about X.
c. We have

$$I(X; Q_1, A_1, Q_2, A_2) = I(X; Q_1, A_1) + I(X; Q_2, A_2|Q_1, A_1).$$

We then need to prove that

$$I(X; Q_2, A_2|Q_1, A_1) \leq I(X; Q_2, A_2) = I(X; Q_1, A_1),$$

where the last equality is clear. This is equivalent to proving

$$H(Q_2, A_2|Q_1, A_1) \geq H(Q_2, A_2|X).$$

But this is true by the data processing inequality since $(Q_1, A_1) - X - (Q_2, A_2)$.

P2 (2 points) A constant process $X_1 = X_2 = ... = X_n$ with $X_i \sim Ber(0.1)$ (i.e., $\Pr[X_i = 1] = 0.1$) is passed through a memoryless binary symmetric channel with cross-over probability $p = 0.3$, producing output $Y^n = (Y_1, ..., Y_n)$.
a. Is Y^n stationary? Is it i.i.d.? Is it ergodic?
b. Demonstrate a lossless compression scheme that achieves compression rates arbitrarily close to the entropy rate $H(\mathcal{Y})$.
c. Answer the two questions above for the case where the input X^n is i.i.d. $Ber(0.1)$.

Sol.:
a. Given the generation mechanism, we have that with probability 0.1, the process Y^n is i.i.d. $Ber(0.7)$, while with probability 0.9 process Y^n is i.i.d. $Ber(0.3)$. The output process Y^n is stationary since the mechanism used for generation does not depend on time. However, it is neither i.i.d. nor ergodic. To show this, it is enough to notice that the empirical average

$$\frac{1}{n} \sum_{i=1}^{n} Y_i$$

does not converge to $E[Y] = 0.1 \cdot 0.7 + 0.9 \cdot 0.3 = 0.34$, but to a random variable equal to 0.7 with probability 0.1 and equal to 0.3 with probability 0.9.
b. A lossless compression scheme that achieves compression rates arbitrarily close to the entropy rate $H(\mathcal{Y})$ can be devised by applying Shannon codes on blocks of sufficient length as discussed in class. We recall that the scheme works for any stationary process.
c. In this case, the process Y^n is i.i.d. $Ber(0.1 \cdot 0.7 + 0.9 \cdot 0.3 = 0.34)$. Therefore, lossless compression schemes that achieves compression rates arbitrarily close to the entropy $H(\mathcal{Y}) = H(Y)$ can be devised based on typicality or using the approach mentioned above.

P3 (2 points) Consider a Markov chain $X^n = X_1, ..., X_n$ over the alphabet $\{0, 1, 2\}$ with transition probability $\Pr[X_{m+1} = i | X_m = j]$ defined by the following table

$i \backslash j$	0	1	2
0	1/2	1/8	3/8
1	3/8	1/2	1/8
2	1/8	3/8	1/2

Assume that the initial distribution is $\Pr[X_1 = i] = 1/3$ for $i = 0, 1, 2$.
a. Is X^n stationary? (Hint: Check whether the initial distribution is the same as the stationary distribution)
b. Find the entropy $H(X)$ and the entropy rate $H(\mathcal{X})$. Which one is larger? Why?
c. Consider the process $Z^n = Z_1, ..., Z_n$, where $Z_1 = X_1$ and $Z_m = (X_m - X_{m-1}) \bmod 3$ for $m = 2, ..., n$ (recall that $(0 + 3k) \bmod 3 = 0$, $(1 + 3k) \bmod 3 = 1$, $(2 + 3k) \bmod 3 = 2$ for $k = ..., -2, -1, 0, 1, 2, ...$). Focus on $Z_2, ..., Z_n$ (that is, excluding Z_1). Is this process stationary? Is it i.i.d.?
d. Focusing on $Z_2, ..., Z_n$, calculate the entropy $H(Z)$ and the entropy rate $H(\mathcal{Z})$.

Sol.: a. The stationary distribution (μ_0, μ_1, μ_2) satisfies the "equilibrium" conditions

$$\begin{aligned} \mu_0 &= \frac{1}{2}\mu_0 + \frac{3}{8}\mu_1 + \frac{1}{8}\mu_2 \\ \mu_1 &= \frac{1}{8}\mu_0 + \frac{1}{2}\mu_1 + \frac{3}{8}\mu_2 \\ \mu_2 &= \frac{3}{8}\mu_0 + \frac{1}{8}\mu_1 + \frac{1}{2}\mu_2, \end{aligned}$$

which are clearly satisfied by the initial distribution. So, the process is stationary.
b. Since the process is a stationary Markov chain, we can calculate

$$H(X) = H((1/3, 1/3, 1/3)) = \log_2 3 = 1.58 \text{ bit},$$

and

$$\begin{aligned} H(\mathcal{X}) &= H(X_2|X_1) = H((1/2, 1/8, 3/8)) \\ &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{8}\log_2\frac{1}{8} - \frac{3}{8}\log_2\frac{3}{8} \\ &= 1.4 \text{ bit}. \end{aligned}$$

As we expect, $H(\mathcal{X}) < H(X)$, since the process has memory.
c. The process is i.i.d. (and thus also stationary). In fact, Z_m for $m \geq 2$ is distributed as

$$p(z) = \begin{cases} 1/2 & z = 0 \\ 1/8 & z = 1 \\ 3/8 & z = 2 \end{cases}$$

independently of all previous Z_{m-i} for $i \geq 1$.
d. The entropy is

$$\begin{aligned} H(Z) &= H((1/2, 1/8, 3/8)) \\ &= 1.4 \text{ bit}, \end{aligned}$$

and the entropy rate is $H(\mathcal{Z}) = H(Z)$.