EE/Stats 376A - Information Theory
Midterm
February 16, 2017 Solutions

There is a total of 8 questions with a total of 84 points. You have a total of 3 hours.

All random variables in this exam are discrete and all logarithms are to the base 2, unless otherwise stated.

Please write all your answers in the exam booklet.

All answers should be justified, unless otherwise stated.

The exam is closed book but you are allowed one double-sided sheet of notes. No other materials are allowed.

Good luck!

[1.]

1. (1 point) Shannon had a good sense of humor. There is a good (but subtle) joke in the opening paragraphs of his paper. Find it!

   "The recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist and Hartley on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

   The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design."

   Solution: *'Frequently* the messages have meaning'.

2. (8 points) True or False? No justifications needed. A correct answer is worth +1 point. Incorrect answer is worth −1 points. 0 point if left blank.

   (a) $I(X;Y|Z) \leq I(X;Y)$ for any random variables X, Y, Z.

   Solution: False. Counter example: $X, Y \overset{i.i.d.}{\sim} Bern(0.5)$ and $Z = X + Y \text{mod} 2$

   (b) $I(X;Y|Z=z) \leq I(X;Y)$ for any random variables X, Y, Z and value z.

   Solution: False. Same counter example as above.

   (c) $D(p||q) = D(q||p)$ for any two distributions p and q.

   Solution: False. Relative Entropy is not symmetric w.r.t p, q. Example 2.3.1 in CT.

   (d) $H(aX) = H(X)$ for any random variable X and any $a \neq 0$.

   Solution: True. Entropy of a random variable is label **invariant.**

   (e) $H(X|aY) = H(X|Y)$ for any random variables X, Y and any $a \neq 0$.

   Solution: True. Conditional entropy is label invariant because $H(X|Y) = H(X,Y) - H(Y)$.

   (f) The entropy rate of a Markov chain which does not start at its stationary distribution is not well defined.

   Solution: False. Entropy rate of a Markov chain is defined as $H = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots X_n)$ and it doesn't depend on the initial state.

   (g) When one communicates at capacity over a memoryless channel, the input symbols $X_1, X_2, \ldots$ to the channel should be i.i.d. at the capacity-achieving input distribution p^*.

   Solution: False. For a noisy channel, X_i's must have some sort of redundancy to deal with the channel noise.

   (h) Due to the lack of knowledge of the source statistics, the Lempel-Ziv algorithm cannot compress the source to its entropy rate.

   Solution: False. LZ algorithm is a Universal scheme. For example in HW4 we showed that for i.i.d. sources LZ compresses at the the entropy rate.

3. (8 points) Let X and Y be the input and outputs of a given discrete memoryless channel with transition probabilities $p(y|x)$.

   (a) (4 points) Define what it means by the statement "$H(Y)$ is a concave function of the distribution of Y", and prove it directly from first principles, without using any facts from convex analysis.

   Solution: Concavity $H(Y)$ w.r.t to Y means

   $$H(Y) \geq \alpha H(Y_1) + (1-\alpha)H(Y_2), \quad \forall \alpha \in [0,1]$$

   where $Y_1 \sim p_1$, $Y_2 \sim p_2$ and $Y \sim \alpha p_1 + (1-\alpha)p_1$. Tp prove this, use an auxiliary random variable $Z \sim Bern(\alpha)$ and let

   $$Y = \begin{cases} Y_1 & if\ Z = 1 \\ Y_2 & if\ Z = 0 \end{cases}$$

   and we have

   $$\begin{aligned} H(Y) &\geq H(Y|Z) \\ &= H(Y|Z=1)Pr(Z=1) + H(Y|Z=0)Pr(Z=0) \\ &= \alpha H(Y_1) + (1-\alpha)H(Y_2). \end{aligned}$$

   (b) (4 points) Define what it means by the statement "$H(Y)$ is a concave function of the distribution of X", and prove it directly from first principles, without using any facts from convex analysis.

   Solution: Concavity $H(Y)$ w.r.t to X means given a fixed $p(y|x)$ (channel)

   $$H(Y) \geq \alpha H(Y_1) + (1-\alpha)H(Y_2), \quad \forall \alpha \in [0,1],$$

   where

   i. $X_1 \sim q_1$ and $X_2 \sim q_2$ and $X \sim \alpha q_1 + (1-\alpha)q_2$
   ii. $Y_1 \sim p_1$ where $p_1(y) = \sum_x p(y|x)q_1(x)$,
   iii. $Y_2 \sim p_2$ where $p_2(y) = \sum_x p(y|x)q_2(x)$,
   iv. $Y \sim p$ where $p(y) = \sum_x p(y|x)q_1(x)\alpha + p(y|x)q_2(x)(1-\alpha)$.

   Similar to previous part define$Z = \begin{cases} 1 & \alpha \\ 0 & 1-\alpha \end{cases}$. Now when

   i. $Z = 1$ Y is distributed according to p_1.
   ii. $Z = 0$ Y is distributed according to p_2.

   Thus from conditions 3(b)ii,3(b)iii, we have

   $$\begin{aligned} H(Y) &\geq H(Y|Z) \\ &= H(Y|Z=1)Pr(Z=1) + H(Y|Z=0)Pr(Z=0) \\ &= \alpha H(Y_1) + (1-\alpha)H(Y_2). \end{aligned}$$

4. (6 points) Consider the noisy typewriter channel, where the input and the output alphabet are both the English alphabet $\{a, b, c, \ldots, z\}$. Given an input letter, the output letter is with probability 1/2 the same as the input letter and with probability 1/2 the next letter. (The next letter for 'a' is 'b', for 'b' is 'c', etc. The next letter for 'z' is 'a'.).

   (a) (3 points) Compute the capacity of this channel.

   Solution: Since the channel is *symmetric*, the capacity of this channel is achieved when the input is uniformly distributed over $\{a, b, c, \ldots, z\}$; this implies that the output Y is also uniformly distributed over $\{a, b, c, \ldots, z\}$.

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ \big(\text{For uniform distribution}\big) &= H(Y) - H(Y|X) \\ &= \log 26 - \sum_{x \in \{a,b,c,\ldots,z\}} H(Y|X=x) Pr(X=x) \\ &= \log 26 - \sum_{x \in \{a,b,c,\ldots,z\}} H(Y|X=x) \frac{1}{26} \\ &= \log 26 - \log 2 \\ &= \log 13. \end{aligned}$$

   Capacity is thus equal to $\log 13$.

   (b) (3 points) Find a simple explicit communication scheme (not via random coding) that achieves the capacity with zero probability of error.

   Solution: Consider the scheme where the sender uses only **alternate** symbols $\mathcal{X}' = \{'a', 'c', 'e', \cdots, 'y'\}$ s.t

$$p'(x) = \begin{cases} 1/13 & x \in \mathcal{X}' \\ 0 & else \end{cases}.$$

   The decoder of this input is very trivial and has a zero probability of error. For example, it would map the output 'a', 'b' to 'a' and 'c', 'd' to $'c'$ and so on.

5. (12 points) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. Bern(p) sequence.

   (a) (2 points) Define what it means by x^n to be a typical sequence. Be as explicit as you can.

   Solution: Define a typical set by $A_\epsilon^{(n)} = \Big\{ x^n : \big| \log \frac{1}{p(x^n)} - H(X) \big| < \epsilon \Big\}$. A sequence is typical if $p(x^n) \sim 2^{-nH(X)}$.

   (b) (2 points) Suppose x^n is a typical sequence. Let $\pi(x^n)$ be a permutation of the sequence. Is $\pi(x^n)$ necessarily typical? Explain.

   Solution: Yes. Since the sequence is *i.i.d.*, for a given x^n, $p(x^n)$ depends only on the number of 1's and 0's in x^n. A permutation π doesn't change the number of $1's$ and $0's$ and thus $p(\pi(x^n)) = p(x^n)$ and hence $p(\pi(x^n))$ is also a typical sequence.

   (c) (2 points) Let $Y_1, Y_2, \ldots, Y_n$ be output of a memoryless BSC(q) with $X_1, X_2, \ldots, X_n$ as input. Define what it means by (x^n, y^n) be a jointly typical sequence. Be as explicit as you can.

   Solution: Define

$$A_\epsilon^{(n)} = \Big\{ (x^n, y^n) : \quad \big| -\tfrac{1}{n} \log p(x^n) - H(X) \big| < \epsilon, \quad \big| -\tfrac{1}{n} \log p(y^n) - H(Y) \big| < \epsilon, \quad \big| -\tfrac{1}{n} \log p(x^n, y^n) - H(X,Y) \big| < \epsilon \Big\}$$

   The sequence (x^n, y^n) is jointly typical is $p(x^n) \sim H(X)$, $p(y^n) \sim H(Y)$ and $p(y^n) \sim H(Y)$.

   (d) (2 points) Suppose (x^n, y^n) is a jointly typical sequence. Is $(\pi(x^n), y^n)$ necessarily jointly typical? Explain.

   Solution: No.

$$\begin{aligned} p(\pi(x^n), y^n) &= p(\pi(x^n)) p(y^n | \pi(x^n)) \\ &= p(x^n) \Pi_{i=1}^n p(y_i | \pi(x^n)_i) \\ \big(\text{In general}\big) &\neq p(x^n) \Pi_{i=1}^n p(y_i | x_i). \end{aligned}$$

   (e) (4 points) Repeat parts (a) and (b) if instead $X_1, X_2 \ldots X_n$ is a stationary Mickey Mouse Markov chain with symmetric crossover probability of $\alpha > 0$.

   Solution: Similar to Part (a), x^n is a typical sequence if $p(x^n) \sim 2^{-H(\alpha)}$. However, since x^n is not an i.i.d. sequence, we have

$$\begin{aligned} p(\pi(x^n)) &= p(\pi(x^n)_1) \Pi_{i=2}^N p(\pi(x^n)_{i+1} | \pi(x^n)_i) \\ \big(\text{In general}\big) &\neq p(x_1) \Pi_{i=2}^N p(x_{i+1} | x_i) \end{aligned}$$

   For example $Pr(x_n = 000111) \neq Pr(\pi(x_n) = 010101)$.

6. (11 points) The random variable X takes on k different possible values. Y is some side information we observe about X. $\hat{X}$ is an estimate of X. We assume that the random variables $X, Y, \hat{X}$ form a Markov chain.

   (a) (2 points) Define what this means.
   Solution: X and $\hat{X}$ are conditionally independent given Y. Mathematically, $p(x, \hat{x}|y) = p(\hat{x}|y)p(x|y)$, $\forall(x, \hat{x}, y)$.

   (b) (2 points) Fano's inequality states that

$$\Pr(\hat{X} \neq X) \geq \frac{H(X|\hat{X}) - 1}{\log k}.$$

   Explain the significance of Fano's inequality in the context of the proof of the converse to the noisy channel coding theorem.
   Solution: Fano's inequality lower bounds the probability of decoding error in terms of the uncertainty in decoding procedure - $H(W|\hat{W})$, where W is the message sent and $\hat{W}$ is the decoded message.

   (c) (4 points) Prove Fano's inequality. (Hint: Define the random variable E which equals 1 if $\hat{X} = X$ and 0 otherwise, and expand $H(E, X|\hat{X})$ in two ways.)
   Solution: Let $E = \mathbf{1}\{X = \hat{X}\}$. We have

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|\hat{X}, X) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}). \end{aligned}$$

   Since $H(E|\hat{X}, X) = 0$ and since $H(E|\hat{X}) \leq 1$, we have

$$\begin{aligned} H(X|\hat{X}) &\leq 1 + H(X|E = 0, \hat{X})Pr(E = 0) + H(X|E = 1, \hat{X})Pr(E = 1) \\ &\leq 1 + Pr(E = 0)H(X). \end{aligned}$$

   This implies $Pr(\hat{X} \neq X) \geq \frac{H(X|\hat{X})-1}{H(X)} \geq \frac{H(X|\hat{X})-1}{\log k}$.

   (d) (3 points) Fano's inequality gives a lower bound on the probability of error of guessing the value of X based on observing the side information Y about X. Suppose now you have no observation of any side information and you still want to guess the value of X. Give a non-trivial lower bound to the probability of error in this case.
   Solution: Without any side information, $H(X|\hat{X}) = H(X)$ and hence using Part (c), we have

$$\begin{aligned} Pr(\hat{X} \neq X) &\geq \frac{H(X) - 1}{H(X)} \\ &= 1 - \frac{1}{H(X)}. \end{aligned}$$

7. (21 points) In arithmetic coding, we represent the source symbols (assumed to be binary here) $X_1, X_2, \ldots$ as

$$X = 0.X_1X_2X_3\ldots$$

and encodes them as binary symbols $U_1, U_2, \ldots$, where

$$U = 0.U_1U_2U_3\ldots = F(X),$$

where F is the cumulative distribution function of X. You can assume F is continuous and strictly increasing.

(a) (3 points) Compute the distribution of U. Justify your answer from first principles.

Solution: The distribution of $U = F(X)$ is uniform in $[0, 1]$ (*Lemma* 1, Lecture notes 7).

(b) (1 point) What is the distribution of the sequence $U_1, U_2, \ldots$?

Solution: Since $U \sim unif(0, 1)$, $U_1, \cdots$ are i.i.d. $Bern(0.5)$.

(c) (1 point) Justify why this is an optimal compression scheme.

Solution: The sequence $U_1U_2\cdots$ is i.i.d. $Bern(1/2)$; since it cannot be further compressed, the above scheme has (asymptotically) achieved optimal compression.

(d) (1 point) After observing the first k source symbols $x_1, x_2, \ldots, x_k$, compute the interval where $u = F(x)$ lies. Which bits of the encoded symbols can be computed at this point?

Solution: Let $A_k = F(0.x_1x_2\cdots x_k)$ and $B_k = F(0.x_1x_2\cdots x_k + 2^{-k})$ and let $U \in [A_k, B_k] = 0.u_1u_2\cdots u_\infty$. Define l as

$$l = \max_i \Big(0.u_1u_2\cdots u_i0000..\ > A_k, \text{ and } B_k > 0.u_1u_2\cdots u_i11111..\ \Big).$$

In other words, return the first l bits common between A_k, B_k

(e) (3 points) Suppose the source symbols are i.i.d. Bern(p). Give recursive formulas for updating the interval in part (d) as you observe more source symbols. How does the computational complexity of the coder scales with the number of source symbols observed?

Solution: $A_0 = 0, B_0 = 1$ and

$$\begin{aligned} A_{n+1} &= A_n + \mathbf{1}\{x_{n+1} = 1\}(1-p)(B_n - A_n) \\ B_{n+1} &= B_n - \mathbf{1}\{x_{n+1} = 0\}p(B_n - A_n). \end{aligned}$$

Time Complexity: $O(n)$.

(f) (4 points) Suppose the source symbols are now not i.i.d. but follows a stationary Mickey mouse Markov chain with symmetric cross over probability α. Give recursive formulas for updating the interval in

part (d) as you observe more source symbols. How does the computational complexity of the coder scale with the number of source symbols observed?

Solution: The iniitilization step is

$$A_1 = \begin{cases} 0 & x_1 = 0 \\ 1/2 & x_1 = 1 \end{cases}$$

and for A_2 we have

$$\begin{aligned} A_2 &= Pr(X < 0.x_1x_2) \\ &= Pr(X_1 < x_1) + Pr(X_1 = x_1, X_2 < x_2) \\ &= A_1 + Pr(X_1 = x_1)Pr(X_2 < x_2|X_1 = x_1) \\ &= A_1 + (B_1 - A_1)Pr(X_2 < x_2|X_1 = x_1) \end{aligned} \quad (1)$$

Thus for A_{n+1} we have

$$A_{n+1} = A_n + (B_n - A_n)Pr(X_{n+1} < x_{n+1}|X_n = x_n)$$

and similarly for B_{n+1} , we have

$$B_{n+1} = B_n - (B_n - A_n)Pr(X_{n+1} > x_{n+1}|X_n = x_n).$$

Time Complexity: $O(n)$.

(g) (5 points) Suppose you have your disposal a sequence of i.i.d. coin flips with biased Heads probability p and you want to use them to simulate the source in part (f). Explain how you can do that. Draw a system diagram to illustrate your scheme.(You can assume you have at your disposal the arithmetic encoders and decoders for the sources in part (e) and (f).)

Solution:

i. Pass the i.i.d. sequence generated by coin flips with $Bern(p)$ into encoder of Part (e) to generate an i.i.d sequence $U_1, U_2, \cdots \sim Bern(0.5)$.

ii. Pass $U_1, U_2, \cdots$ into decoder of Part (f) whose output will simulate the MMMC.

(h) (3 points) On the average, how many coin flips do you need to generate a Markov source symbol in part (g)?

Solution: The compression rate of MMMC is $H(\alpha)$ and the compression rate of encoder in Part (e) $H(p)$. Thus we need $\frac{1}{H(p)}$ $Bern(p)$ bits to generate 1 $Bern(0.5)$ bit and we need $H(\alpha)$ $Bern(0.5)$ bits to generate 1 bit of MMMC. Therefore we need $\frac{H(\alpha)}{H(p)}$ bits to generate 1 bit of MMMC.

8. (17 points)

(a) (3 points) State the data processing theorem. Under what condition will the inequality in the data processing theorem becomes equality?

Solution: For random variables X, Y, Z such that $X \to Y \to Z$, we have

$$I(X;Y) \geq I(X;Z). \tag{2}$$

We get an equality$I(X;Y) - I(X;Z) = I(X;Y,Z) - I(X;Z) = I(X;Y|Z) = 0$. So having equality in equation (2)means that both $X \to Z \to Y$ and $X \to Y \to Z$ should hold true.

(b) (1 point) We are given a symmetric memoryless channel $p(y|x)$ with binary input alphabet $\{0,1\}$. What is the capacity-achieving optimal input distribution?

Solution: Since the conditional distribution is symmetric, $p(x) \sim Unif(\{0,1\})$ is the capacity-achieving optimal input distribution.

(c) (2 points) Let C be the capacity of the channel. The maximum mutual information over two uses of the channel is:

$$\max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2),$$

where the optimization is over all joint distributions of X_1, X_2. Solve this optimization problem and express the answer in terms of C. What is the optimizing joint input distribution?

Solution: Using the definition of mutual informations,we have

$$\begin{aligned}
\max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2) &= \max_{p(x_1,x_2)} \Big(H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \Big) \\
\big(\text{Using the memoryless property}\big) &= \max_{p(x_1,x_2)} \Big(H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \Big) \\
&\leq \max_{p(x_1,x_2)} \Big(H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \Big) \\
&= \max_{p(x_1,x_2)} \Big(I(X_1;Y_1) + I(X_2;Y_2) \Big) \\
\implies \max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2) &\leq 2C.
\end{aligned}$$

Equality can be achieved if $H(Y_1, Y_2) = H(Y_1) + H(Y_2)$; which is true when $X_1 \perp X_2$ and $X_1, X_2 \overset{i.i.d.}{\sim} Unif(\{0,1\})$.

(d) (4 points) Now suppose we want to send two binary symbols U, $V \in \{0,1\}$ on the two uses of the channel, such that

$$\begin{aligned}
X_1 &= U + V \pmod 2 \\
X_2 &= U.
\end{aligned}$$

The maximum mutual information between $((U, V)$ and the channel outputs is given by

$$\max_{p(u_1,u_2)} I(U_1, U_2; Y_1, Y_2),$$

where the optimization is over all joint distributions of U, V. Solve this optimization problem and express the answer in terms of C. What is the optimizing joint distribution on U and V? (Hint: what is the relationship between $I(U_1, U_2; Y_1, Y_2)$ and $I(X_1, X_2; Y_1, Y_2)$?)

Solution: From the construction U, V we have

$$\begin{aligned} U &= X_1 \\ V &= X_1 + X_2 \text{ (mod 2)}. \end{aligned}$$

and hence using the condition for equality for the data processing theorem from Part (a), $I(U_1, U_2; Y_1, Y_2) = I(X_1, X_2; Y_1, Y_2)$.

From Part (c), we know that the capacity achieving distribution of X_1 and X_2 is $X_1, X_2 \overset{i.i.d.}{\sim} Unif(\{0, 1\})$, which can be achieved by $U, V \overset{i.i.d.}{\sim} Unif(\{0, 1\})$. Hence we have

$$\max_{p(u_1,u_2)} I(U_1, U_2; Y_1, Y_2) = \max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2) = 2C.$$

(e) (2 point) Using part (d) or otherwise, show that under the optimal joint distribution on U and V,

$$I(V; Y_1, Y_2) + I(U; Y_1, Y_2|V) = 2C.$$

Solution: Using chain rule for mutual information, we have

$$I(U, V : Y_1, Y_2) = I(V; Y_1, Y_2) + I(U; Y_1, Y_2|V)$$

and for the optimal distribution, it is equal to $2C$.

(f) (5 points) Can one say that one of the two terms $I(V; Y_1, Y_2)$ and $I(U; Y_1, Y_2|V)$ is *definitely* greater than or equal to C, while the other term is *definitely* less than or equal to C? Explain. (Hint: can you give an interpretation to the term $I(U; Y_1, Y_2|V)$?)

Solution: Expanding mutual information in terms of entropy, we have

$$I(V; Y_1, Y_2) = H(V) - H(V|Y_1, Y_2)$$

and

$$\begin{aligned} I(U; Y_1, Y_2|V) &= I(X_1, X_2; Y_1; Y_1|V) \\ &= I(X_2; Y_1, Y_2|V) + I(X_1; Y_1, Y_2|V, X_2) \\ &\geq I(X_2; Y_1, Y_2|V) \\ &\geq I(X_2; Y_2|V) \\ &= I(X_2; Y_2). \end{aligned}$$

because X_2 and Y_2 are independent of of V. Thus we have $I(U; Y_1, Y_2|V) \geq C$ and therefore $I(V; Y_1, Y_2) \leq C$.