EE376A - Information Theory
Final, Monday March 14th 2016 Solutions

Instructions:

- You have **three hours**, 3.30PM - 6.30PM
- The exam has 4 questions, totaling 120 points.
- Please start answering each question on a new page of the answer booklet.
- You are allowed to carry the textbook, your own notes and other course related material with you. Electronic reading devices [including kindles, laptops, ipads, etc.] are allowed, provided they are used solely for reading pdf files already stored on them and not for any other form of communication or information retrieval.
- You are required to provide a detailed explanation of how you arrived at your answers.
- You can use previous parts of a problem even if you did not solve them.
- As throughout the course, entropy (H) and Mutual Information (I) are specified in bits.
- log is taken in base 2.
- Throughout the exam 'prefix code' refers to a variable length code satisfying the prefix condition.
- Good Luck!

1. **Problem 1 : Exponential Noise Channel and Exponential Source** *(40 points)*

   Recall that $X \sim Exp(\lambda)$ is to say that X is a continuous non-negative random variable with density

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$$

   or, equivalently, that X is a random variable with characteristic function

$$\varphi_X(t) \stackrel{\triangle}{=} E\left[e^{itX}\right] = \frac{1}{1 - it/\lambda}.$$

   Recall also that in this case $EX = 1/\lambda$.

   (a) Find the differential entropy of $X \sim Exp(\lambda)$.
   Solution *(4 points)*:

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} f_X(x) \log(f_X(x)) dx \\ &= -\int_{0}^{\infty} \lambda e^{-\lambda x} \log(\lambda e^{-\lambda x}) dx \\ &= -\int_{0}^{\infty} \lambda e^{-\lambda x} \log(\lambda) + \lambda \int_{0}^{\infty} x \lambda e^{-\lambda x} dx \\ &= 1 - \log \lambda. \end{aligned}$$

   (b) Prove that $Exp(\lambda)$ uniquely maximizes the differential entropy among all non-negative random variables confined to $EX \leq 1/\lambda$.
   Hint: Recall our proof of an analogous fact for the Gaussian distribution.
   Solution *(5 points)*:
   Let the probability density of any such non-negative random variable be f_X, while g_X is the density of $Exp(\lambda)$ as in Part (1) above,

$$\begin{aligned} h(X) &= -\int_{0}^{\infty} f_X(x) \log(f_X(x)) dx \\ &= -\int_{0}^{\infty} f_X(x) \log(\frac{f_X(x)}{g_X(x)}) dx - \int_{0}^{\infty} f_X(x) \log(\lambda e^{-\lambda x}) dx \\ &= -D(f_X||g_X) - log(\lambda) \int_{0}^{\infty} f_X(x) dx + \lambda \int_{0}^{\infty} x f_X(x) dx \\ &= 1 - \log \lambda - D(f_X||g_X) \\ &\leq 1 - \log \lambda, \end{aligned}$$

   where the last inequality is due to the fact that $D(f_X||g_X) \geq 0$, equality holds if $X = Exp(\lambda)$.

Fix positive scalars a and b. Let $\overline{X}$ be the non-negative random variable of mean a formed by taking $\overline{X} = 0$ with probability $\frac{b}{a+b}$ and, with probability $\frac{a}{a+b}$, drawing from an exponential distribution $Exp(1/(a+b))$. Equivalently stated, $\overline{X}$ is the random variable with characteristic function

$$\varphi_{\overline{X}}(t) = \frac{b}{a+b} + \frac{a}{a+b} \cdot \frac{1}{1-it(a+b)}.$$

Let $N \sim Exp(1/b)$ and independent of $\overline{X}$.

(c) What is the distribution of $\overline{X} + N$?
Tip: simplest would be to compute the characteristic function of $\overline{X} + N$ by recalling the relation $\varphi_{\overline{X}+N}(t) = \varphi_{\overline{X}}(t) \cdot \varphi_N(t)$.
Solution *(5 points)*:

$$\begin{aligned}
\varphi_{\overline{X}+N}(t) &= \varphi_{\overline{X}}(t) \cdot \varphi_N(t) \\
&= (\frac{b}{a+b} + \frac{a}{a+b} \cdot \frac{1}{1-it(a+b)}) \frac{1}{1-itb} \\
&= \frac{1}{1-it(a+b)} \frac{a+b-itab-itb^2}{(a+b)(1-itb)} \\
&= \frac{1}{1-it(a+b)},
\end{aligned}$$

which is the characteristic function of $Exp(1/a+b)$. Thus $\overline{X} + N$ is distributed as $Exp(1/a+b)$.

(d) Find $I(\overline{X}; \overline{X} + N)$.
Solution *(5 points)*:

$$\begin{aligned}
I(\overline{X}; \overline{X} + N) &= h(\overline{X} + N) - h(\overline{X} + N | \overline{X}) \\
&\overset{N \perp \overline{X}}{=} h(\overline{X} + N) - h(N) \\
&= 1 + \log(a+b) - (1 + \log(b)) \\
&= \log(1 + \frac{a}{b})
\end{aligned}$$

(e) Consider the problem of communication over the additive exponential noise channel $Y = X + N$, where $N \sim Exp(1/b)$, independent of the channel input X, which is confined to being non-negative and satisfying the moment constraint $EX \leq a$. Find $C(a) = \max I(X; X+N)$, where the maximization is over all non-negative X satisfying $EX \leq a$. What is the capacity-achieving distribution?
Hint: Using findings from previous parts, show that for any non-negative random variable X, independent of N, with $EX \leq a$, we have $I(X; X+N) \leq I(\overline{X}; \overline{X}+N)$.

Solution *(7 points)*:
For any feasible X, note that $X + N$ is a non-negative random variable and $E[X+N] = E[X]+E[N] \leq a+b$, thus by the result of Part (2) above, $h(X+N) \leq 1 + \log(a + b)$. Hence,

$$\begin{aligned} I(X; X+N) &= h(X+N) - h(X+N|X) \\ &\overset{N\perp X}{=} h(X+N) - h(N) \\ &\overset{(*)}{\leq} 1 + \log(a+b) - h(N) \\ &= 1 + \log(a+b) - (1 + \log(b)) \\ &= \log(1 + \frac{a}{b}) \\ &= I(\overline{X}; \overline{X} + N), \end{aligned}$$

Thus $C(a) \leq \log(1+\frac{a}{b})$. Equality in $(*)$ holds if $X = \overline{X}$ proving $C(a) = \log(1+\frac{a}{b})$. Maximizing distribution is that of $\overline{X}$.

(f) Let now $U \sim Exp(1/m)$ and consider the rate distortion problem where the expected reconstruction error is not to exceed D and, in addition, the reconstruction is not allowed to exceed the source value. The associated rate distortion function is thus

$$R(D) = \min I(U; V),$$

where the minimum is over joint distributions respecting the given distribution of U and such that $U - V \geq 0$ while $E[U - V] \leq D$. Explain why for any U, V in the feasible set of this minimization, the following equality and inequalities hold:

$$\begin{aligned} I(U; V) &= h(U) - h(U - V|V) && (1) \\ &\geq h(U) - h(U - V) && (2) \\ &\geq \log(m/D). && (3) \end{aligned}$$

(g) Show that $R(D)$ from the previous part is given by

$$R(D) = \begin{cases} \log(m/D) & \text{if } 0 < D \leq m \\ 0 & \text{if } D > m. \end{cases}$$

Hint: Using findings from previous parts, for $0 < D \leq m$ establish existence of (U, V) in the feasible set for which the inequalities in (2) and (3) hold with equality.

2. **Problem 2: Modulo 8 Channel and a Binary Source with Erasure Distortion** *(35 points)*
Consider the memoryless channel described by

$$Y = (X + Z) \bmod 8,$$

where the channel input X, output Y, and the noise Z are real valued. For example, if $x = 6.4$ and $z = 1.8$ then $y = 0.2$. The noise is uniformly distributed on $[-B, B]$, i.e. $Z \sim \text{Uniform}[-B, B]$ and is independent of the input X. Assume that $0 < B \leq 4$ is a known channel parameter.

(a) Find the channel capacity as a function of B.

(b) A communication system is suggested where the permitted channel input values are restricted to the set $\{1, 3, 5, 7\}$. Show that, when $B = 1$, it is still possible to achieve the channel capacity. In fact, to achieve this capacity with zero probability of error and with a very simple scheme.

Consider now a Bernoulli(1/2) source under erasure distortion. I.e., the reconstruction alphabet is $\{0, 1, e\}$ and

$$d(u, v) = \begin{cases} 0 & \text{if } u = v \\ 1 & \text{if } v = e \\ \infty & \text{otherwise} \end{cases}$$

(c) Show that the rate distortion function of this source is given by $R(D) = 1 - D$ for $0 \leq D \leq 1$.

(d) Consider now a joint-source-channel-coding setting for communicating the bit source of part (c) through the modulo 8 channel defined in part (a). For every two source symbols we are allotted one channel use, i.e., the encoder translates a block of $2n$ source bits into n channel inputs.
Find the minimal achievable distortion (when $n \to \infty$) as a function of B.

(e) The following communication system is suggested for the setting of the previous part: every two source bits are mapped into a channel input in the following way:

$$\begin{aligned} 00 &\to 1 \\ 01 &\to 3 \\ 11 &\to 5 \\ 10 &\to 7 \end{aligned}$$

This mapping is known as a Gray Code. Show that this system (with the corresponding optimal decoder) achieves the minimal distortion in the case where $B = 2$.

3. **Problem 3: Removing bias from the world** *(25 points)*
Computer simulations require unbiased and independent random samples. However, in most cases, the data obtained contains biased and dependent samples. In this problem we will try to understand aspects of this problem.

Consider a biased coin with unknown bias $0 < p < 1$. The biased coin generates independent coin tosses $X_1, X_2, ..., X_n$. We wish to obtain a sequence of unbiased and independent bits $Z_1, Z_2, Z_3, ..., Z_K$ from the biased coin toss experiment. Specifically, we want a mapping $f(X_1, X_2, ..., X_n) = (Z_1, Z_2, ...Z_K)$, where K may depend upon $(X_1, X_2, ...X_n)$ and, conditioned on $K = k$, $Z_1, Z_2, \ldots, Z_k$ are fair coin flips.
We define the rate of a scheme to be:

$$R = \frac{E[K]}{n} \tag{4}$$

(a) Von Neumann gave a simple procedure of generating an unbiased coin, where he considered 2 coin tosses, and mapped the output as follows:
$f(01) = 0$
$f(10) = 1$
$f(00) = f(11) = \lambda$
Here, λ corresponds to a null output (i.e., a string of length 0 where we don't output anything). What is the rate achieved by Von-Neumann's scheme?

(b) In case of Von-Neumann's method, we assign a bit $\{0, 1\}$ to an outcome in the 'equiprobable set' of $\{01, 10\}$, and assign λ otherwise. Suggest a simple generalization of Von-Neumann's method where you consider n biased coin tosses at a time. (Hint: All sequences within a type have the same probability, regardless of the bias of the coin)

(c) Show that for any scheme $R \leq H_b(p)$. Can you suggest a sequence of schemes, for increasing n, with rates approaching $H_b(p)$?

(d) We now consider the opposite problem: generating random samples using the unbiased and independent bits. Specifically, we want to generate independent samples $(W_1, W_2, ..., W_K)$ with a specific probability mass function, using unbiased and independent bits $(Z_1, Z_2, ...Z_n)$.

i. Give a procedure for generating i.i.d samples of a random variable W with probability mass function $\{0.5, 0.25, 0.25\}$
(Hint: A technique similar to Huffman coding might be useful here)

ii. Is your procedure optimal?

Solution:

(a) *5 points*
Using $n = 2$, the average number of unbiased bits produced is given by:

$$E[K] = 1P(K = 1) + 0P(K = 0)$$

$$= \quad 2p(1-p)$$

This gives $R = p(1-p)$

(b) *5 points*

Consider a type $T(\frac{m}{n})$ $(0 < m < n)$ which corresponds to m number of H from the n coin tosses. Let, $n_m = \lfloor log\left(|T(\frac{m}{n})|\right) \rfloor$. Consider some 2^{n_m} sequences in the type $T(\frac{m}{n})$, and assign an unique n_m length binary code for each of these 2^{n_m} sequences.

One simple scheme of generating unbiased bits would be, to consider a type corresponding to a fixed $m = \hat{m}$, and output the $n_{\hat{m}}$ length identifier if any one of the $2^{n_{\hat{m}}}$ sequences occur, and output λ otherwise. Note that this scheme would generate $K = 0$ or $K = n_{\hat{m}}$ number of unbiased bits using the n coin tosses.

We can in fact do a better job by not restricting to a single type. Thus, if any of the 2^{n_m} sequences from any type-sets occur, then we output the corresponding n_m length identifier bits, otherwise output λ. Notice that, in this scheme, K can take any of the values n_m.

(c) *10 points*

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &\geq H(Z_1, Z_2, \ldots, Z_K) \\ &= H(Z_1, Z_2, \ldots, Z_K, K) \\ &= H(Z_1, Z_2, \ldots, Z_K|K) + H(K) \\ &\geq H(Z_1, Z_2, \ldots, Z_K|K) \\ &= \sum_{k=1}^{n} P(K=k)k \\ &= E[K] \end{aligned}$$

Thus, $H(X_1, X_2, \ldots, X_n) \geq E[K]$. As $H(X_1, X_2, \ldots, X_n) = nH_b(p)$, this results in $R \leq H_b(p)$

The scheme in part b) results in rates approaching $H_b(p)$ as n increases. To see this, note that for any m, $n_m = n(H_b(\frac{m}{n}) - \epsilon_1))$, where $\epsilon_1 \to 0$ as $n \to \infty$. This can be obtained from the bounds on type sizes derived in the class. Now, consider the types which have sequences in the strongly typical set $T_\delta(p)$. For all these types, the empirical entropy of the sequences is $H_b(p) - \epsilon_2(\delta)$, where ϵ_2 is small. Thus, all the types in the strongly typical set output $\approx nH_b(p)$ bits, as $n \to \infty, \delta \to 0$. Also, as $n \to \infty$, the probability of a sequence being present in the strongly typical set goes to 1. This, results in $E[K] \approx nH_b(p)$, which gives the required result.

(d) *5 points*

i. Let the alphabet set of W be $\{w_1, w_2, w_3\}$. Consider the mapping:

$$\begin{aligned} 0 &\to w_1 \\ 10 &\to w_2 \\ 11 &\to w_3 \end{aligned}$$

This mapping generates the alphabets with the required probability mass function of $0.5, 0.25, 0.25$. This procedure requires on an average 1.5 coin tosses to produce W of the given distribution.

ii. Using similar arguments to the proof in part c), we can show that:
$E[K]H(W) \leq n$
As, $H(W) = 1.5$ in our case, this implies that our procedure is in fact optimal.

4. **Problem 4 : Modulo Channel** *(20 points)*

(a) Consider the DMC defined as follows: Output $Y = X \oplus_2 Z$ where X, taking values in $\{0, 1\}$, is the channel input, $\oplus_2$ is the modulo-2 summation operation, and Z is binary channel noise uniformly distributed over $\{0, 1\}$ and independent of X. What is the capacity of this channel?

Solution *(4 points)*:
The channel is a BSC with crossover probability 1/2, so the capacity is zero.

(b) Consider the channel of the previous part, but suppose that instead of modulo-2 addition $Y = X \oplus_2 Z$, we perform modulo-3 addition $Y = X \oplus_3 Z$. Now what is the capacity?

Solution *(8 points)*:
Note that now $Y \in \{0, 1, 2\}$. Since Z is Bernoulli($\frac{1}{2}$), the channel output being 1 indicates that X could be 0 or 1 with equal probability. Thus, it gives no information about the input. This simply becomes a Binary Erasure channel with erasure probability $\frac{1}{2}$, and the capacity is $1 - \frac{1}{2} = \frac{1}{2}$.

(c) Now suppose the noise Z is no longer independent of the input X, but is instead described by the following conditional distribution:

$$p(Z = z|X = 0) = \begin{cases} 1/4 \text{ if } z = 0 \\ 3/4 \text{ if } z = 1, \end{cases}$$

and

$$p(Z = z|X = 1) = 1/2 \quad \text{both for } z = 0 \text{ and } z = 1.$$

A random code of size 2^{nR} is generated uniformly (that is all codewords are drawn i.i.d. $X \sim Bern(0.5)$). Find the value V such that if $R < V$ then the average probability of decoding error (average both across the messages and the randomness in the codebook) vanishes with increasing blocklength while if $R > V$ then it does not.

Solution *(8 points)*:
In this problem, several students wrongly interpreted V to be the capacity of the channel. This is incorrect, as the problem gives a random codebook generation according to i.i.d. Bern(0.5) codewords. Recall from the direct and converse theorems in the lecture that the maximum supported rate under a given random codebook is simply the Mutual Information between X and Y.
If X is uniform and the noise Z is distributed as described, one may determine the information between X and Y as follows:

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(Z|X).$$

For X chosen Bernoulli(1/2), we have that

$$H(Z|X) = \frac{1}{2} h_b\left(\frac{1}{4}\right) + \frac{1}{2}.$$

To compute $H(Y)$, we first observe that the distribution of Y when X is Bernoulli$(1/2)$ is given by $(p_0, p_1, p_2) = (\frac{1}{8}, \frac{5}{8}, \frac{1}{4})$. The entropy of the this distribution may then be calculated.

Performing both computations, we find that $I(X;Y) = H(Y) - H(Z|X) = 0.393$. This is the maximum supported rate by the channel under a uniformly generated codebook. Note that this is not the capacity of the channel which is achieved for a different input distribution.