EE/Stats 376A: Homework 7 Solutions

Due on **Friday March 17 , 5 pm**

1. **Feedback does not increase the capacity.**

   Consider a channel with feedback. We assume that all the recieved outputs are sent back immediately (and noiselessly) to the transmitter. The transmitter is allowed to use the previous outputs to decide the next input symbol. The retransmission strategy we discussed in lecture for the BEC is an example of a feedback strategy. There we showed that the retransmission strategy achieves the non-feedback capacity of the BEC, but in a much simply way. The question is whether feedback can actually *increase* the capacity of the channel.

   More precisely, a $(2^{nR}, n)$ feedback code for a DMC $p(y|x)$ is defined by a sequence of mappings $x_i(w, Y^{i-1})$ for each message $w \in \{1, \ldots, 2^{nR}\}$ where Y^{i-1} represents the previous outputs. Denote the channel capacity with feedback by C_{FB}, where C is the capacity of the channel without feedback. We will show that $C_{FB} = C$ using the following steps.

   (a) Why is $C_{FB} \geq C$.

   (b) Prove the converse: $C_{FB} \leq C$ by showing that for any sequence of schemes of rate R and vanishing probability of error:

$$\begin{aligned} nR &\overset{(a)}{\leq} I(W; Y^n) + n\epsilon_n \\ &\overset{(b)}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, W) + n\epsilon_n \\ &\overset{(c)}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, W, X_i) + n\epsilon_n \\ &\overset{(d)}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\ &\overset{(e)}{\leq} \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) + n\epsilon_n \\ &\leq nC + n\epsilon_n. \end{aligned}$$

   Provide explanations for inequalities (a)-(e) and conclude that $C_{FB} \leq C$.

Solutions

(a) (2 **points**) Since the feedback channel can use the encoding schemes of the non-feedback channel, it can always acheive the capacity acheived by the non-feedback channel. Thus $C_{FB} \geq C$.

(b) (5 **points**) Explanations:

i. (a) - Fano's inequality.

ii. (b) - Definition of mutual information and chain rule.
iii. (c) - X_i is a function of Y and W.
iv. (d) - $Y_i|X_i$ is independent of every other random variable.
v. (e) -Conditiontioning alwasys reduces the entropy.

2. **Polar Codes: Polarization II.**

In the last homework, you have "proved by simulations" the polarization phenomenon for the BEC. In this question, you will give a rigorous proof.

Let the original BEC have erasure parameter p. At stage k, there are 2^k effective BEC's. Let $p_{k,1}, p_{k,2}, \ldots, p_{k,2^k}$ be the erasure parameters of these channels. (For example, for $k = 1$, the two BEC's have erasure parameters p^2 and $1 - (1-p)^2$.).

(a) Compute
$$\frac{1}{2^k} \sum_i p_{k,i}.$$

(b) Define the second moment of the empirical distribution of the erasure parameters to be
$$v_k = \frac{1}{2^k} \sum_{i=1}^{2^k} p_{k,i}^2.$$
Compute $v_{k+1} - v_k$ in terms of the erasure parameters.

(c) What happens to v_k as $k \to \infty$? What happens to $v_{k+1} - v_k$ as $k \to \infty$?

(d) Using parts (b) and (c) or otherwise, show that for every $\epsilon > 0$, the fraction of channels with erasure parameters between ϵ and $1 - \epsilon$ goes to zero as $k \to \infty$, i.e. the BEC polarizes.

Solution

(a) **(5 points)** The total capacity of the original 2^k channels in the k^{th} step is $2^k(1-p)$ and the capacity of the 'effective' channels is $\sum_i \left(1 - p_{k,i}\right)$ and it is equal to $2^k(1-p)$. Thus we have
$$\begin{aligned} 2^k(1-p) &= \sum_i \left(1 - p_{k,i}\right) \\ \Longrightarrow \frac{1}{2^k} \sum_i p_{k,i} &= 1 - p. \end{aligned}$$

(b) **(10 points)** Let $p^k[j], \quad j \in \{1, 2, \cdots 2^k\}$ be erasure probabilities of the effective channels in the k^{th} step
$$\begin{aligned} v_{k+1} &= (1/2^k) \sum_{j=1}^{2^k} p^k[j]^2 + (2p^k[j] - p^k[j]^2)^2/2 \\ &= (1/2^k) \sum_{j=1}^{2^k} [p^k[j]^2 + (p^k[j](1 - p^k[j])^2]/2 \\ &= v_k + (1/2^k) \sum_{j=1}^{2^k} (p^k[j](1 - p^k[j]))^2, \end{aligned}$$
where the last step is because $(a^2 + b^2)/2 = ((a+b)/2)^2 + ((a-b)/2)^2$. Thus we have

i. $v_{k+1} > v_k$.

ii. Since $\frac{1}{2^k}\sum_{i=1}^{2^k} p_{k,i}^2 \leq v_k = \frac{1}{2^k}\sum_{i=1}^{2^k} p_{k,i} = 1-p$,we also have $v_k \leq 1-p$.

Thus $v_{k+1} - v_k = (1/2^k)\sum_{j=1}^{2^k}(p^k[j](1-p^k[j]))^2$

(c) **(5 points)** We know that $v_{k+1} > v_k$ and $v_k \leq 1-p$. Since v_k is an increasing bounded sequence, the difference $\lim_{k\to\infty} v_{k+1} - v_k = 0$.

(d) **(5 points)** Let $c_\epsilon^k = \frac{\left\{j : p^k[j] \notin (\epsilon, 1-\epsilon)\right\}}{2^k}$.

$$(1/2^k)\sum_{j=1}^{2^k}(p^k[j](1-p^k[j]))^2 \geq c_\epsilon^k \epsilon^2.$$

From part (c), we know that the LHS converges to zero, and hence $\lim_{k\to\infty} c_\epsilon^k = 0$ for all $\epsilon > 0$.

3. **Label-invariance.**

Let X and Y are real-valued random variables. In class we showed that $I(aX;Y) = I(X;Y)$ for any $a \neq 0$. In this question we will explore whether $I(X;Y)$ is label invariance in a stronger sense.

(a) Let U be a continuous real-valued random variable and let ϕ be an invertible and differentiable function from $\Re$ to $\Re$. Compute the density g of $V = \phi(U)$ in terms of the density f of U.

(b) Is it true that $I(\phi(X);Y) = I(X;Y)$ for any invertible and differentiable function ϕ? Support your answer with a proof or a counter-example.

Solutions

(a) (3 **points**) Let $Pr(U \leq u) = F(u)$. The cdf of V is

$$\begin{aligned} Pr(V \leq v) &= Pr(\phi(U) \leq v) \\ &= Pr(U \leq \phi^{-1}(v)) \\ &= F(\phi^{-1}(v)). \end{aligned}$$

The pdf of v is

$$\begin{aligned} g(v) &= \frac{dPr(V \leq v)}{dv} \\ &= \frac{dF(\phi^{-1}(v))}{dv} \\ &= \left(\phi^{-1}(v)\right)' f(\phi^{-1}(v)). \end{aligned}$$

(b) (5 **points**) $I(\phi(X);Y) = I(X;Y)$ hold true . Let's calculate $h(\phi(X))$:

$$\begin{aligned}
-h(\phi(X)) &= \int dv g(v) \log g(v) \\
&= \int dv \big(\phi^{-1}(v)\big)' f(\phi^{-1}(v)) \log \Big(\big(\phi^{-1}(v)\big)' f(\phi^{-1}(v)) \Big) \\
&= \int dv \big(\phi^{-1}(v)\big)' f(\phi^{-1}(v)) \log \big(\phi^{-1}(v)\big)' + \int dv \big(\phi^{-1}(v)\big)' f(\phi^{-1}(v)) \log f(\phi^{-1}(v)) \\
\big(\text{Let } u = \phi^{-1}(v)\big) &= \int du' f(u) \log \big(u\big)' + \int du f(u) \log f(u) \\
&= \int du' f(u) \log \big(u\big)' - h(X). \qquad (1)
\end{aligned}$$

Similarly we have

$$-h(\phi(X)|Y) = -h(X|Y) + \int du' f(u) \log \big(u\big)' \qquad (2)$$

Since $I(\phi(X);Y) = h(\phi(X)) - h(\phi(X) \mid Y)$, using equations (1,2), we have

$$\begin{aligned}
I(\phi(X);Y) &= h(\phi(X)) - h(\phi(X) \mid Y) \\
&= h(X). - \int du' f(u) \log \big(u\big)' - h(X|Y) + \int du' f(u) \log \big(u\big)' \\
&= h(X). - h(X|Y) \\
&= I(X;Y).
\end{aligned}$$

4. **Estimation Error and Differential Entropy**

(a) Given the variance of random variable X, $\mathrm{Var}(X) = \sigma^2$, what is the maximum value of differential entropy $h(X)$? (Note that this is different from the problem we considered in lecture, where the constraint is on the second moment $\mathbb{E}[X^2]$.)

(b) Prove that for any constant c,

$$\mathbb{E}\big[\,(X-c)^2\,\big] \geq \frac{1}{2\pi e} e^{2\,h(X)},$$

i.e. the differential entropy of a random variable yields a lower bound to the estimation error of that random variable.

(c) Given random variables X, Y and any function g prove that

$$\mathbb{E}\big[\,\big(X-g(Y)\big)^2\,\big] \geq \frac{1}{2\pi e} e^{2\,h(X|Y)}$$

Solutions

(a) (3 **points**) From problem 3, we know that $h(X) = h\big(X - \mathbb{E}[X]\big)$. Therefore, if the rv X^* attaining the maximum differential entropy for the contraint $\mathbb{E}[X^2] = \sigma^2$, then the rv $X^* - \mathbb{E}[X]$ attains the maximum differential entropy for the contraint $Var[X^2] = \sigma^2$. The maximum differential entorpy for both the cases is $\ln\big(\sigma\sqrt{2\,\pi\,e}\big)$.

(b) (**7 points**) We will prove a slightly more general result. Let $\hat{X}$ be any estimator of X; then

$$\begin{aligned}
\mathbb{E}[(X-\hat{X})^2] &\geq \min_{\hat{X}} \mathbb{E}[(X-\hat{X})^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] \\
&= Var(X) \\
&\geq \frac{1}{2\pi e} e^{2h(X)}.
\end{aligned}$$

The last inequality follows from the fact that the gaussian distribution has the maximum entropy for a given variance. We get the required result by substituting $\hat{X} = c$.

(c) (5 **points**) Similar proof as above

$$\begin{aligned}
\mathbb{E}_Y\Big[\mathbb{E}[(X-\hat{X}(Y))^2 \mid Y]\Big] &\geq \mathbb{E}_Y\Big[\min_{\hat{X}} \mathbb{E}[(X-\hat{X}(Y))^2] \mid Y\Big] \\
&= \mathbb{E}_Y\Big[\mathbb{E}[(X-\mathbb{E}[X|Y])^2] \mid Y\Big] \\
&= \mathbb{E}_Y\Big[Var(X|Y)\Big] \\
&\geq \mathbb{E}_Y\Big[\frac{1}{2\pi e} e^{2h(X|Y=y)}\Big] \\
\text{(Jensen's)} \quad &\geq \frac{1}{2\pi e} e^{2h(X|Y)}.
\end{aligned}$$

Substituting $g(Y) = \hat{X}(Y)$ gives us the required result.

5. **Exponential noise channels.**

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

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

(a) Find the mean and differential entropy of $X \sim \text{Exp}(\lambda)$.

(b) Prove that $X \sim \text{Exp}(1)$ uniquely maximizes the differential entropy among all non-negative random variables conditioned to $\mathbb{E}[X] = 1$.

Fix positive scalars a and b. Let X be the non-negative random variable of mean a formed by taking $X = 0$ with probability $\frac{b}{a+b}$ and with probability $\frac{a}{a+b}$ drawing from an exponential distribution $\text{Exp}(1/(a+b))$. Let $N \sim \text{Exp}(1/b)$ and independent of X.

(c) What is the distribution of $X + N$?

(d) Find $I(X; X+N)$.

(e) Consider the problem of communication over the additive exponential noise channel $Y = X + N$, where $N \sim \text{Exp}(1/b)$, independent of the channel input X, which is confined to being non-negative and satisfying the moment constraint $\mathbb{E}[X] = a$. Find $C(a) = \max I(X; X+N)$, where the maximization is over all non-negative X satisfying $\mathbb{E}[X] = a$. What is the capacity-achieving distribution?

Solution

(a) (3 **points**) The mean of an exponential random vairable is λ^{-1}. The differential entropy is

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

(b) (5 **points**) Let $g_X(x) = \lambda e^{-\lambda x}$. Using the non-negative property of KL-divergence, we

have

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} f_X(x)\log(f_X(x))dx \\ &= -\int_0^{\infty} f_X(x)\log(f_X(x)/g_X(x))dx - -\int_0^{\infty} f_X(x)\log(\lambda e^{-\lambda x})dx \\ &= -D_{KL}(f_X||g_X) - \log\lambda\int_0^{\infty} f_X(x)dx + \lambda\int_0^{\infty} x f_X(x)dx \\ &= -D_{KL}(f_X||g_X) - \log\lambda + 1 \\ &\leq -\log\lambda + 1. \end{aligned}$$

Since $g_X(x)$ attains the upper bound, $X \sim \text{Exp}(1)$ uniquely maximizes the differential entropy among all non-negative random variables conditioned to $\mathbb{E}[X] = 1$.

(c) (7 **points**) The characteristic function of an exponential random variable X_λ, $\phi_t(X_\lambda)$ is $\frac{\lambda}{\lambda - it}$. Thus the characteristic function of X is $\frac{b}{a+b} + \frac{a}{a+b}\big(\frac{1}{1-it(a+b)}\big)$. Since N is independent rv the characteristic function of $X + N$ is

$$\begin{aligned} \phi_t(X).\phi_t(N) &= \left(\frac{b}{a+b} + \frac{a}{a+b}\big(\frac{1}{1-it(a+b)}\big)\right)\left(\frac{1}{1-itb}\right) \\ &= \frac{a+b-itab-itb^2}{\big(1-it(a+b)\big)(a+b)(1-itb)} \\ &= \frac{1}{1-it(a+b)} \end{aligned}$$

and hence $X + N$ is an exponential random variable wih parameter $\lambda = (a+b)^{-1}$.

(d) (5 **points**) Using the definition of mutual information

$$\begin{aligned} I(X+N;X) &= H(X+N) - H(X+N|X) \\ &= H(X+N) - H(N) \\ &= 1 + \log(a+b) + \big(1 - \log(b)\big) \\ &= \log\big(1 + \frac{a}{b}\big). \end{aligned}$$

(e) (5 **points**) For any feasible X, note that $X + N$ is a non-negative random variable and $\mathbb{E}[X+N] = \mathbb{E}[X]+\mathbb{E}[N] \leq a+b$, thus by previous part we have $h(X+N) \leq 1+\log(a+b)$. Hence,

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

The equality $(*)$ holds if X is a non-negative random variable of mean a formed by taking $X = 0$ with probability $\frac{b}{a+b}$ and with probability $\frac{a}{a+b}$ drawing from an exponential distribution $\text{Exp}(1/(a+b))$ (refer to previous part) and this is the capacity achieving distribution.

6. **Entropy maximization.**

Let X_1, X_2, X_3 be continuous real-valued random variables each with zero mean and unit variance. Suppose we know that $\mathbb{E}[X_1X_2] = \mathbb{E}[X_2X_3] = \rho$. Compute explicitly the differential entropy maximizing joint distribution subject to these constraints. Identify the key qualitative properties of this joint distribution. (There may be more than one!)

Solution (10 **points**)**:**

We have to solve the following problem

$$\begin{aligned} \underset{f}{\text{argmax}} \quad & h(X_1, X_2, X_3) \\ s.t \quad & \mathbb{E}[X_1X_2] = \rho \\ & \mathbb{E}[X_2X_3] = \rho. \end{aligned} \tag{3}$$

Consider

$$\begin{aligned} h(X_1, X_2, X_3) &= h(X_1, X_2) + h(X_3 \mid X_1, X_2) \\ &\leq h(X_1, X_2) + h(X_3 \mid X_2) \end{aligned} \tag{4}$$

- Under the constraints (3), the maximum value of $h(X_1, X_2)$ is acheived by X_1, X_2 being zero-mean gaussian with $\mathbb{E}[X_1X_2] = \rho, \mathbb{E}[X_1^2] = \rho, \mathbb{E}[X_2^2] = 1$.
- Similarly under the constraints (3), the maximum value of $h(X_3|X_2)$ is attained by X_2, X_3 being being zero-mean gaussian with $\mathbb{E}[X_2X_3] = \rho, \mathbb{E}[X_2^2] = \rho, \mathbb{E}[X_3^2] = 1$.

For the **zero-mean** gaussian distribution f^* on X_1, X_2, X_3 satisfying the above two constraints, and satisfying the markov chain $X_1 \leftrightarrow X_2 \leftrightarrow X_3$, inequality (4) will be an equality and hence f^* is be optimal. How do we calculate f^*? f^* is completely characterized by covariance matrix Σ. We have

$$\begin{aligned} \mathbb{E}\big[X_1X_3\big] &= \mathbb{E}\Big[\mathbb{E}\big[X_1X_3\big]\big|X_2\Big] \\ \text{(By markov chain property and gaussian property)} \quad &= \mathbb{E}\Big[\mathbb{E}\big[X_1\big]\big|X_2\mathbb{E}\big[X_3\big]\big|X_2\Big] \\ &= \mathbb{E}\Big[\rho X_2\rho X_2\Big] \\ &= \rho^2\mathbb{E}\Big[X_2X_2\Big] \\ &= \rho. \end{aligned}$$

We also know that $\mathbb{E}[X_i^2] = 1 \quad i = \{1, 2, 3\}$ and $\mathbb{E}\big[X_1X_2\big] = \mathbb{E}\big[X_2X_3\big] = \rho$,and thus we know all the entries in Σ which gives us f^*.

7. **Entropy maximization.**

Let X_1, X_2, X_3 be binary random variables taking values in $\{0, 1\}$. Suppose we know that the probability that $X_i = 1$ is α_i, $i = 1, 2, 3$, and probability that $X_i = 1$ and $X_{i+1} = 1$ is β_i, for $i = 1, 2$.

(a) Is there always a joint distribution on X_1, X_2, X_3 which satisfies these constraints? Explain. If not, give conditions on the α_i's and β_i's such that such a joint distribution exists.

(b) Compute the parametric form of the entropy maximizing joint distribution subject to these constraints when such a distribution exists.

(c) By using the parametric form, show that under the entropy maximizing joint distribution, X_1, X_2, X_3 forms a Markov chain. (This gives an alternate proof as the one we gave in the lecture.)

Solution

(a) (7 **points**) No, some obvious necessary sufficient conditions are all $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ should be in interval $[0, 1]$. Also we should have $\beta_1 \leq \min\{\alpha_1, \alpha_2\}$ and $\beta_2 \leq \min\{\alpha_2, \alpha_3\}$. However, these necessary conditions are not sufficient. Note that

$$\alpha_2 - \beta_2 = Pr(X_2 = 1, X_3 = 0) \leq 1 - \alpha_3, \quad \alpha_2 - \beta_1 = Pr(X_1 = 0, X_2 = 1) \leq 1 - \alpha_1.$$

We claim that the set of above necessary conditions are sufficient to guarantee the existence of a joint distribution which agrees with the above marginals. To show this, without loss of generality suppose that $\beta_2 \leq \beta_1$. Then, let

$$\begin{aligned}
&Pr(X_1 = 1, X_2 = 1, X_3 = 1) = \beta_2 \\
&Pr(X_1 = 1, X_2 = 1, X_3 = 0) = \beta_1 - \beta_2 \\
&Pr(X_1 = 0, X_2 = 1, X_3 = 1) = 0 \\
&Pr(X_1 = 0, X_2 = 1, X_3 = 0) = \alpha_2 - \beta_1 \\
&Pr(X_1 = 1, X_2 = 0, X_3 = 1) = \min\{\alpha_3 - \beta_2, \alpha_1 - \beta_1\}.
\end{aligned}$$

Note that based on the given conditions the sum of above non-negative numbers is upper-bounded by 1 and we can set the rest of probabilities such that we meet the constraints and the sum add up to 1.

(b) (3 **points**) The above problem falls in to the case of Ising model and the parametrix form is

$$p(x_1, x_2, x_3 \mid \theta) = \frac{1}{Z(\theta)} \exp\big\{ -\theta_1 x_1 - \theta_2 x_2 - \theta_3 x_3 - \theta_{12} x_1 x_2 - \theta_{23} x_2 x_3 \big\},$$

where $Z(\theta)$ is the normalizing constant.

(c) (5 **points**) From part (b) we have

$$\begin{aligned}
p(x_1, x_2, x_3 \mid \theta) &= \frac{1}{Z(\theta)} \exp\big\{ -\theta_1 x_1 - \theta_2 x_2 - \theta_3 x_3 - \theta_{12} x_1 x_2 - \theta_{23} x_2 x_3 \big\} \\
&= \left(\frac{1}{\sqrt{Z(\theta)}} \exp\big\{ -\theta_1 x_1 - 0.5\theta_2 x_2 - \theta_{12} x_1 x_2 \big\} \right) \left(\frac{1}{\sqrt{Z(\theta)}} \exp\big\{ -0.5\theta_2 x_2 - \theta_3 x_3 - \theta_{23} x_2 x \right. \\
&:= g(x_1, x_2) g(x_2, x_3).
\end{aligned}$$

We have

$$\begin{aligned} p(x_1, x_3|x_2) &= \frac{p(x_1, x_3, x_2)}{p(x_2)} \\ &= \frac{g(x_1, x_2)g(x_2, x_3)}{p(x_2)} \\ &= \frac{g(x_1, x_2)g(x_2, x_3)}{\sum_{x_1=\{0,1\}, x_3=\{0,1\}} g(x_1, x_2)g(x_2, x_3)} \\ &= \frac{g(x_1, x_2)}{\sum_{x_1=\{0,1\}} g(x_1, x_2)} \frac{g(x_2, x_3)}{\sum_{x_3=\{0,1\}} g(x_2, x_3)} \\ &= p(x_1|x_2)p(x_3|x_2). \end{aligned} \tag{5}$$

The last inequality follows because

$$\begin{aligned} p(x_1|x_2) &= \frac{p(x_1, x_2)}{p(x_2)} \\ &= \frac{\sum_{x_3=\{0,1\}} g(x_1, x_2)g(x_2, x_3)}{\sum_{x_1=\{0,1\}, x_3=\{0,1\}} g(x_1, x_2)g(x_2, x_3)} \\ &= \frac{g(x_1, x_2)}{\sum_{x_1=\{0,1\}} g(x_1, x_2)} \frac{\sum_{x_3=\{0,1\}} g(x_2, x_3)}{\sum_{x_3=\{0,1\}} g(x_2, x_3)} \\ &= \frac{g(x_1, x_2)}{\sum_{x_1=\{0,1\}} g(x_1, x_2)}. \end{aligned}$$

Equation (5) proves that X_1, X_2, X_3 forms a Markov chain.