Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 8 of November 9, 2016

http://www.isi.ee.ethz.ch/teaching/courses/it1/

Problem 1 — *Data Processing*

Using the chain rule and Markovity we obtain:

$$\begin{aligned} I(X_1; X_2, \ldots, X_n) &= H(X_2, \ldots, X_n) - H\bigl(X_2, \ldots, X_n \big| X_1\bigr) \\ &= H(X_2) + \sum_{k=3}^{n} H\bigl(X_k \big| X_{k-1}, \ldots, X_2\bigr) - \sum_{k=2}^{n} H\bigl(X_k \big| X_{k-1}, \ldots, X_1\bigr) \\ &= H(X_2) + \sum_{k=3}^{n} H\bigl(X_k \big| X_{k-1}\bigr) - \sum_{k=2}^{n} H\bigl(X_k \big| X_{k-1}\bigr) \\ &= H(X_2) - H\bigl(X_2 \big| X_1\bigr) \\ &= I(X_1; X_2). \end{aligned}$$

Another way of seeing this is to state the Markovity property in a different form: instead of saying that $X \multimap Y \multimap Z$ if, and only if,

$$P_{Z|Y,X}(z|y,x) = P_{Z|Y}(z|y), \tag{1}$$

we can equivalently say that $X \multimap Y \multimap Z$ if, and only if, conditioned on Y we have $X \perp\!\!\!\perp Z$:

$$P_{X,Z|Y}(x,z|y) = P_{X|Y}(x|y) \cdot P_{Z|Y}(z|y).$$

To see this note that by dividing both sides by $P_{X|Y}(x|y)$ we recover (1) because

$$P_{Z|Y,X}(z|y,x) = \frac{P_{X,Z|Y}(x,z|y)}{P_{X|Y}(x|y)}.$$

The advantage of this new form is that it very clearly is symmetric, i.e., we see that Markovity does not bother about the direction of time! Applied to our situation we therefore have

$$H(X_1|X_2, X_3, \ldots, X_n) = H(X_1|X_2)$$

and

$$\begin{aligned} I(X_1; X_2, X_3, \ldots, X_n) &= H(X_1) - H(X_1|X_2, X_3, \ldots X_n) \\ &= H(X_1) - H(X_1|X_2) \\ &= I(X_1; X_2). \end{aligned}$$

© Amos Lapidoth, 2016

Problem 2 *Preprocessing the Output*

a) The statistician calculates $\tilde{Y} = g(Y)$. By the chain rule for mutual information we have

$$I(X;Y,\tilde{Y}) = I(X;\tilde{Y}) + \underbrace{I(X;Y|\tilde{Y})}_{\geq 0} \geq I(X;\tilde{Y})$$

and

$$I(X;Y,\tilde{Y}) = I(X;Y) + I(X;\tilde{Y}|Y) = I(X;Y),$$

where the last equation follows because $X \multimap Y \multimap \tilde{Y}$ form a Markov chain, i.e., given Y, $X \perp\!\!\!\perp \tilde{Y}$. Hence, we have

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

with equality if, and only if, $I(X;Y|\tilde{Y}) = 0$, i.e., $X \multimap \tilde{Y} \multimap Y$ also form a Markov chain. Note that this result is called *data processing inequality.*

Let $\tilde{Q}$ be a distribution on $\mathcal{X}$ that maximizes $I(X;\tilde{Y})$ for a given $W(\cdot|\cdot)$ and $g(\cdot)$. Then

$$\begin{aligned} \tilde{\mathsf{C}} &= \max_Q I(X;\tilde{Y}) \\ &= I(X;\tilde{Y})\big|_{Q=\tilde{Q}} \\ &\overset{(2)}{\leq} I(X;Y)\big|_{Q=\tilde{Q}} \\ &\leq \max_Q I(X;Y) \\ &= \mathsf{C}. \end{aligned}$$

Hence, the capacity $\tilde{\mathsf{C}}$ of the new channel consisting of the original channel with additional processing of the output is not larger than the capacity C of the original channel. Thus, the statistician is wrong and processing the output cannot increase capacity.

b) We have equality (no decrease in capacity) in the above sequence of inequalities only if there exists a distribution $\tilde{Q}$ that maximizes both $I(X;\tilde{Y})$ and $I(X;Y)$ and for which we have equality in the data processing inequality, i.e., $X \multimap \tilde{Y} \multimap Y$ form a Markov chain. (Note that this does not imply that g has to be injective! There exist examples where g is not injective, but nevertheless $\tilde{\mathsf{C}} = \mathsf{C}$.)

Problem 3 *A Channel With Two Independent Looks at Y*

a) By the chain rule for mutual information,

$$\begin{aligned} I(X;Y_1,Y_2) &= I(X;Y_1) + I(X;Y_2|Y_1) \\ &= I(X;Y_1) + I(X,Y_1;Y_2) - I(Y_1;Y_2) \\ &= I(X;Y_1) + I(X;Y_2) + \underbrace{I(Y_1;Y_2|X)}_{=0} - I(Y_1;Y_2) \\ &= I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2), \end{aligned}$$

where we used the fact that Y_1 and Y_2 are conditionally independent given X.

© Amos Lapidoth, 2016

b) Let C_{12} denote the capacity of the first channel $X \to (Y_1, Y_2)$. Furthermore, let C_1 and C_2 denote the capacities of the channels $X \to Y_1$ and $X \to Y_2$, respectively. Then, we have

$$\begin{aligned}
\mathsf{C}_{12} &= \max_{P_X} I(X; Y_1, Y_2) \\
&\overset{a)}{=} \max_{P_X} \bigl[I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2)\bigr] \\
&\le \max_{P_X} \bigl[I(X;Y_1) + I(X;Y_2)\bigr] \\
&\le \max_{P_X} I(X;Y_1) + \max_{P_X} I(X;Y_2) \\
&= \mathsf{C}_1 + \mathsf{C}_2.
\end{aligned}$$

Thus, the capacity of the first channel is upper bounded by the sum of the capacities of the two subsequent channels.

Problem 4 — *Miscellaneous Capacities*

There are many ways to compute capacity and capacity-achieving input distributions. For a), b), c), and d), we will guess an input distribution Q and verify the Karush–Kuhn–Tucker conditions, which also provide us the capacity:

Theorem 1. *An input probability distribution Q achieves the capacity of a discrete memoryless channel with transition probabilities $W(\cdot|\cdot)$ if and only if for some number λ*

$$D\bigl(W(\cdot|x) \| (QW)(\cdot)\bigr) = \lambda \qquad \forall\, x \text{ with } Q(x) > 0, \tag{3}$$

$$D\bigl(W(\cdot|x) \| (QW)(\cdot)\bigr) \le \lambda \qquad \forall\, x \text{ with } Q(x) = 0. \tag{4}$$

Furthermore, the capacity of the channel is λ.

For e), we use Theorem 7.2.1 of Cover & Thomas, p. 191:

Theorem 2. *For a weakly symmetric channel,*

$$\mathsf{C} = \log|\mathcal{Y}| - H(\text{row of transition matrix}), \tag{5}$$

and this is achieved by a uniform distribution on the input alphabet.

a) We guess $Q(0) = Q(1) = \frac{1}{2}$, which leads to the following probability distributions:

	$y=0$	$y=1$	$y=2$
$W(y\|x=0)$	$1-\epsilon-\delta$	δ	ϵ
$W(y\|x=1)$	ϵ	δ	$1-\epsilon-\delta$
$(QW)(y)$	$\frac{1-\delta}{2}$	δ	$\frac{1-\delta}{2}$

Because

$$\begin{aligned}
D\bigl(W(\cdot|x=0)\|(QW)(\cdot)\bigr) &= (1-\epsilon-\delta)\log\frac{1-\epsilon-\delta}{(1-\delta)/2} + \delta\log\frac{\delta}{\delta} + \epsilon\log\frac{\epsilon}{(1-\delta)/2} \\
&= (1-\delta)\log 2 + (1-\epsilon-\delta)\log\frac{1-\epsilon-\delta}{1-\delta} + \epsilon\log\frac{\epsilon}{1-\delta} \\
&= (1-\delta)\left[1 + \frac{1-\epsilon-\delta}{1-\delta}\log\frac{1-\epsilon-\delta}{1-\delta} + \frac{\epsilon}{1-\delta}\log\frac{\epsilon}{1-\delta}\right] \\
&= (1-\delta)\left[1 - H_{\mathrm{b}}\left(\frac{\epsilon}{1-\delta}\right)\right]
\end{aligned}$$

and because $D\bigl(W(\cdot|x=1)\|(QW)(\cdot)\bigr) = D\bigl(W(\cdot|x=0)\|(QW)(\cdot)\bigr)$, we conclude that Q is a capacity-achieving input distribution and that the capacity is $(1-\delta)\Bigl(1 - H_{\mathrm{b}}\bigl(\frac{\epsilon}{1-\delta}\bigr)\Bigr)$ bits.

b) From symmetry one could mistakingly guess $Q(0) = Q(1) = Q(2) = \frac{1}{3}$. Because input 1 does not look useful, it is actually more intuitive to guess $Q(1) = 0$ and $Q(0) = Q(2) = \frac{1}{2}$. This leads to the following probability distributions:

	$y=0$	$y=1$	$y=2$
$W(y\|x=0)$	$\frac{3}{4}$	$\frac{1}{4}$	0
$W(y\|x=1)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$W(y\|x=2)$	0	$\frac{1}{4}$	$\frac{3}{4}$
$(QW)(y)$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$

Because

$$\begin{aligned} D\big(W(\cdot|x=0)||(QW)(\cdot)\big) &= \frac{3}{4}\log\frac{3/4}{3/8} + \frac{1}{4}\log\frac{1/4}{1/4} \\ &= \frac{3}{4}, \end{aligned}$$

because $D\big(W(\cdot|x=2)||(QW)(\cdot)\big) = D\big(W(\cdot|x=0)||(QW)(\cdot)\big)$, and because

$$\begin{aligned} D\big(W(\cdot|x=1)||(QW)(\cdot)\big) &= \frac{1}{3}\log\frac{1/3}{3/8} + \frac{1}{3}\log\frac{1/3}{1/4} + \frac{1}{3}\log\frac{1/3}{3/8} \\ &= \frac{1}{3}\log\frac{8}{9} + \frac{1}{3}\log\frac{4}{3} + \frac{1}{3}\log\frac{8}{9} \\ &= \frac{1}{3}\log\frac{256}{243} \\ &\approx 0.025 \leq \frac{3}{4}, \end{aligned}$$

we conclude that Q is a capacity-achieving input distribution and that the capacity is $\frac{3}{4}$ bits.

c) Setting $Q(1)$ to zero separates the output values (i.e., $y = 0$ only if $x = 0$ and $y = 1$ or $y = 2$ only if $x = 2$), so we guess $Q(1) = 0$ and $Q(0) = Q(2) = \frac{1}{2}$. This leads to the following probability distributions:

	$y=0$	$y=1$	$y=2$
$W(y\|x=0)$	1	0	0
$W(y\|x=1)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
$W(y\|x=2)$	0	$\frac{1}{2}$	$\frac{1}{2}$
$(QW)(y)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Because

$$\begin{aligned} D\big(W(\cdot|x=0)||(QW)(\cdot)\big) &= 1\log\frac{1}{1/2} = 1, \\ D\big(W(\cdot|x=2)||(QW)(\cdot)\big) &= \frac{1}{2}\log\frac{1/2}{1/4} + \frac{1}{2}\log\frac{1/2}{1/4} = 1, \\ D\big(W(\cdot|x=1)||(QW)(\cdot)\big) &= \frac{1}{2}\log\frac{1/2}{1/2} + \frac{1}{4}\log\frac{1/4}{1/4} + \frac{1}{4}\log\frac{1/4}{1/4} = 0 \leq 1, \end{aligned}$$

we conclude that Q is a capacity-achieving input distribution and that the capacity is 1 bit.

d) Note that this is not a weakly symmetric channel. We guess $Q(0) = Q(1) = Q(2) = \frac{1}{3}$, which leads to the following probability distributions:

	$y=0$	$y=1$	$y=2$	$y=3$
$W(y\|x=0)$	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{3}$
$W(y\|x=1)$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$W(y\|x=2)$	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$
$(QW)(y)$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{1}{3}$

Because

$$\begin{aligned} D\big(W(\cdot|x=0)||(QW)(\cdot)\big) &= \frac{1}{3}\log\frac{1/3}{2/9} + \frac{1}{3}\log\frac{1/3}{2/9} + \frac{1}{3}\log\frac{1/3}{1/3} \\ &= \frac{2}{3}\log\frac{3}{2} \end{aligned}$$

and $D\big(W(\cdot|x=1)||(QW)(\cdot)\big) = D\big(W(\cdot|x=2)||(QW)(\cdot)\big) = D\big(W(\cdot|x=0)||(QW)(\cdot)\big)$, we conclude that Q is a capacity-achieving input distribution and that the capacity is $\frac{2}{3}\log\frac{3}{2}$ bits.

e) The transition matrix is

$$W = \begin{bmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & 1-\epsilon & \epsilon \\ \epsilon & 0 & 1-\epsilon \end{bmatrix},$$

thus W is weakly symmetric (it is even strongly symmetric). By Theorem 2, the uniform distribution achieves capacity and the capacity is $\log 3 - H_{\mathrm{b}}(\epsilon)$.