Information Theory I

Fall Semester 2017

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory

Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 8 of November 8, 2017

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

Problem 1 — *An Additive Noise Channel*

The channel output is $Y = X + Z$, where $X \in \{0, 1\}$ and $Z \in \{0, a\}$. We must distinguish various cases depending on the value of a:

- $a = 0$: In this case $Y = X$, therefore $H(X|Y) = 0$, and

$$\mathsf{C} = \max_{P_X} I(X;Y) = \max_{P_X} \{H(X) - H(X|Y)\} = \max_{P_X} H(X) = 1 \text{ bit},$$

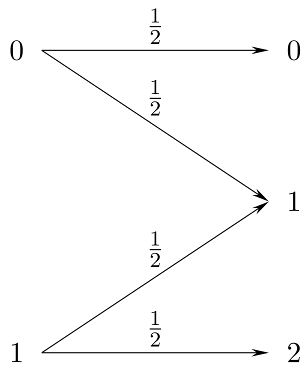
  which is achieved by a uniform distribution on the input X.

- $a \notin \{0, \pm 1\}$: In this case Y has four different values. If Y is 0 or a, we know that $X = 0$. If Y is 1 or $1 + a$, we know that $X = 1$. Hence $H(X|Y) = 0$, and therefore

$$\mathsf{C} = \max_{P_X} I(X;Y) = \max_{P_X} \{H(X) - H(X|Y)\} = \max_{P_X} H(X) = 1 \text{ bit},$$

  which is achieved by a uniform distribution on the input X.

- $a = 1$: In this case Y has three possible output values: 0, 1, and 2. The channel looks as follows:



  One sees that the channel is equivalent to a binary erasure channel with erasure probability $\alpha = \frac{1}{2}$. The capacity of the binary erasure channel is $\mathsf{C} = 1 - \alpha = \frac{1}{2}$ bits per transmission, which is achieved by a uniform distribution on the input X.

- $a = -1$: This is similar to the case when $a = 1$: Y also can take on three different values: -1, 0, and 1, where now 0 is the "erasure" output. We again have a BEC and the capacity is $\mathsf{C} = \frac{1}{2}$ bits per transmission, achieved by a uniform distribution.

Problem 2 ***Data Processing***

Using the chain rule, we obtain:

$$
\begin{aligned}
I(X_1; X_2, \dots, X_n) &= H(X_2, \dots, X_n) - H\big(X_2, \dots, X_n \big| X_1\big) \\
&= H(X_2) + \sum_{k=3}^{n} H\big(X_k \big| X_{k-1}, \dots, X_2\big) - \sum_{k=2}^{n} H\big(X_k \big| X_{k-1}, \dots, X_1\big) \\
&\overset{\text{(i)}}{=} H(X_2) + \sum_{k=3}^{n} H\big(X_k \big| X_{k-1}\big) - \sum_{k=2}^{n} H\big(X_k \big| X_{k-1}\big) \\
&= H(X_2) - H\big(X_2 \big| X_1\big) \\
&= I(X_1; X_2),
\end{aligned}
$$

where (i) holds because $X_1 \multimap X_2 \multimap X_3 \multimap \cdots \multimap X_n$ form a Markov chain.
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

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

and

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

Problem 3 ***Preprocessing the Output***

a) Observe that $X \multimap Y \multimap \tilde{Y}$ form a Markov chain, so for every fixed distribution P_X,

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

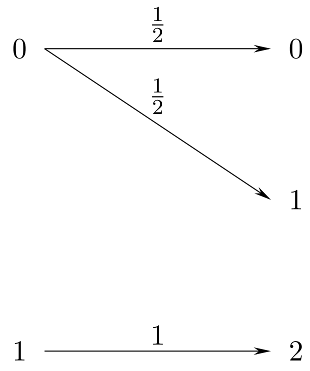
by the *data processing inequality.* Denoting the capacity of the new channel with additional processing of the output by $\tilde{\mathsf{C}}$, we have

$$
\tilde{\mathsf{C}} = \max_{P_X} I(X; \tilde{Y}) \overset{\text{(i)}}{\leq} \max_{P_X} I(X; Y) = \mathsf{C},
$$

where (i) follows from (2). This shows that processing the output of a channel cannot increase its capacity.

b) We have equality (no decrease in capacity) in (i) only if there exists a distribution P_X^* that maximizes $I(X;Y)$ and for which we have equality in the data processing inequality, i.e., for which $X \multimap \tilde{Y} \multimap Y$ form a Markov chain.

(Note that this does not imply that g has to be injective! A counterexample is the channel



together with $g\colon \{0,1,2\} \to \{0,2\}$, $g(0) = g(1) = 0$ and $g(2) = 2$. Clearly, g is not injective, but $\mathsf{C} = \tilde{\mathsf{C}} = 1$ bit.)

Problem 4 — *A Channel With Two Independent Looks at Y*

a) Using the chain rule for mutual information three times, we obtain

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

where the underbraced term is zero since Y_1 and Y_2 are conditionally independent given X.

b) Let C_{12} denote the capacity of the 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

$$\begin{aligned} \mathsf{C}_{12} &= \max_{P_X} I(X;Y_1,Y_2) \\ &\overset{\text{(i)}}{=} \max_{P_X} \bigl\{ I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2) \bigr\} \\ &\overset{\text{(ii)}}{\leq} \max_{P_X} \bigl\{ I(X;Y_1) + I(X;Y_2) \bigr\} \\ &\leq \max_{P_X} I(X;Y_1) + \max_{P_X} I(X;Y_2) \\ &= \mathsf{C}_1 + \mathsf{C}_2, \end{aligned}$$

where (i) follows from Part a); and (ii) holds because mutual information is nonnegative. Thus, the capacity of the first channel is upper bounded by the sum of the capacities of the two other channels.

Problem 5 ***Miscellaneous Capacities***

There are many ways to find the capacity and a capacity-achieving input distribution. For a), b), c), and d), we will guess an input distribution and verify the Karush–Kuhn–Tucker conditions:

Theorem 1. *If an input probability distribution Q and $\lambda \in \mathbb{R}$ satisfy*

$$D\big(W(\cdot|x)||(QW)(\cdot)\big) = \lambda \qquad \forall x \text{ with } Q(x) > 0,$$
$$D\big(W(\cdot|x)||(QW)(\cdot)\big) \le \lambda \qquad \forall x \text{ with } Q(x) = 0,$$

then Q achieves capacity and the capacity is λ.

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

Theorem 2. *For a weakly symmetric channel,*

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

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

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

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

b) From symmetry one could mistakenly guess $Q(0) = Q(1) = Q(2) = \frac{1}{3}$. Because input 1 does not look useful, it is actually better 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

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

$$= \frac{2}{3} \log \frac{3}{2}$$

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 $\log 3 - H_{\mathrm{b}}(\epsilon)$.