Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

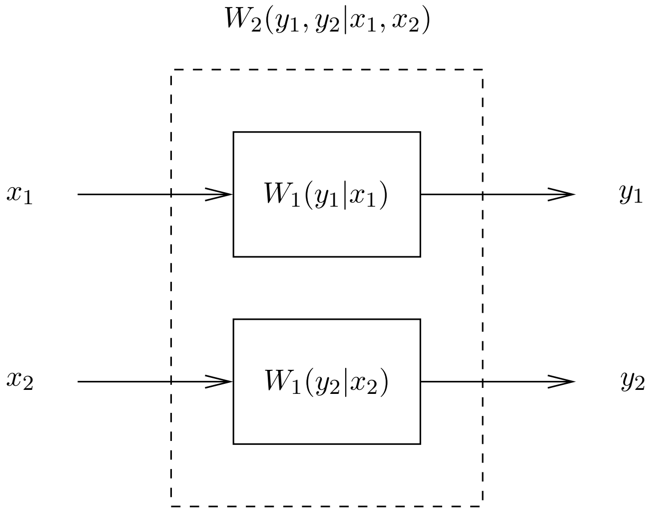
Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 7 of November 2, 2016

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

Problem 1 — *On the Achievable Rate*

The second channel, W_2, can be considered to consist of two parallel channels W_1 as shown in the figure below, i.e., one use of W_2 corresponds to two *independent* uses of W_1.



a) Consider a sequence of codes $\{\mathscr{C}_1\}_n$ of length n and achievable rate R for W_1. Now consider a sequence of codes $\{\mathscr{C}_2\}_n$ for W_2 whose codewords are of length n and consists of codewords from $\{\mathscr{C}_1\}_n$ in the following way:

$$\mathscr{C}_2 = \left\{ \mathbf{c}_2 = \begin{bmatrix} \mathbf{c}_1^{(\mathrm{u})} \\ \mathbf{c}_1^{(\mathrm{l})} \end{bmatrix} : \mathbf{c}_1^{(\mathrm{u})}, \mathbf{c}_1^{(\mathrm{l})} \in \mathscr{C}_1 \right\},$$

where $\mathbf{c}_1^{(\mathrm{u})}$ and $\mathbf{c}_1^{(\mathrm{l})}$ are row vectors. There are $2^{n\mathsf{R}}$ codewords in $\mathscr{C}_1$, and hence we can have $2^{n\mathsf{R}} \cdot 2^{n\mathsf{R}} = 2^{n2\mathsf{R}}$ codewords in $\mathscr{C}_2$ if we choose the two components independently. Thus the rate of $\mathscr{C}_2$ is $(\log 2^{n2\mathsf{R}})/n = 2\mathsf{R}$. The receiver observes the sequence $\mathbf{y}$ of length $2n$, which consists of the n-length sequences $\mathbf{y}_1$ and $\mathbf{y}_2$ of the two parallel channels with law W_1, i.e.,

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}.$$

The decoder for the channel W_2 divides $\mathbf{y}$ into $\mathbf{y}_1$ and $\mathbf{y}_2$ and decodes these sequences independently using the decoder for channel W_1. Since the rate R is achievable on W_1, both $\mathbf{y}_1$ and $\mathbf{y}_2$ can be decoded with a probability of error which tends to 0 as $n \to \infty$. By the union bound, the probability of a decoding error on W_2 also tends to 0. Thus, the rate $2\mathsf{R}$ is achievable.

b) The main idea lies in making two consecutive uses of the same channel with law W_1 instead of transmitting over two parallel independent channels of law W_1. As we are using the channel twice as often as before, the rate is going to be divided by two.

Consider a sequence of rate-R codes $\{\mathscr{C}_2\}_n$ of length n for W_2:

$$\mathscr{C}_2 = \left\{\mathbf{c}_2 = [c_{21}, c_{22}, \ldots, c_{2n}] \colon c_{2i} \in \mathcal{X} \times \mathcal{X},\ \forall i \in \{1, \ldots, n\}\right\}$$

and corresponding decoding functions

$$\phi_{2,n} \colon (\mathcal{Y} \times \mathcal{Y})^n \to \left\{1, 2, \ldots, 2^{n\mathsf{R}}\right\}$$

such that the error probability P_e of $\mathscr{C}_2$ tends to zero as $n \to \infty$. As in Part a) we can think of every codeword $\mathbf{c}_2 \in \mathscr{C}_2$ as a $2 \times n$ matrix with entries in $\mathcal{X}$ and denote the row containing the first component of each symbol as $\mathbf{c}_2^{(\mathrm{u})}$ and the row containing the second component as $\mathbf{c}_2^{(\mathrm{l})}$.

We construct a sequence of rate-$\frac{\mathsf{R}}{2}$ codes $\{\mathscr{C}_1\}_{2n}$ of length $2n$ and decoding functions $\phi_{1,2n} \colon \mathcal{Y}^{2n} \to \left\{1, 2, \ldots, 2^{n\mathsf{R}}\right\}$ by choosing every codeword in $\mathscr{C}_1$ to correspond to a codeword $\mathbf{c}_2$ in $\mathscr{C}_2$, i.e.,

$$\mathscr{C}_1 = \left\{\mathbf{c}_1 = \left[\mathbf{c}_2^{(u)} \| \mathbf{c}_2^{(\mathrm{l})}\right]\right\},$$

and using the same decoder as for the code $\mathscr{C}_2$

$$\phi_{1,2n}(\mathbf{y}_1 \| \mathbf{y}_2) = \phi_{2,n}(\mathbf{y}_1 \times \mathbf{y}_2).$$

Then the maximum error probability P_e of $\mathscr{C}_1$ tends to zero as $n \to \infty$, as the probability of an error is the same as for the code $\mathscr{C}_2$. Thus we achieve the rate $\mathsf{R}_1 = \frac{n\mathsf{R}}{2n} = \frac{\mathsf{R}}{2}$ on the channel W_1.

Problem 2 — *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 $Y = X$, and therefore

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

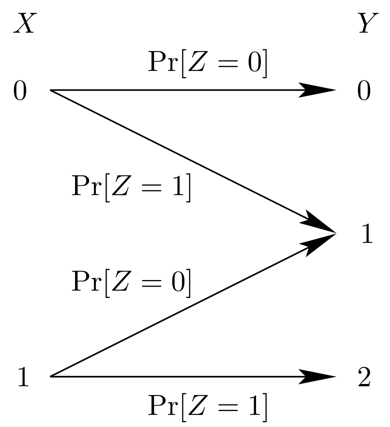
  which is achieved by a uniform distribution $P_X(\cdot)$. Hence the capacity is 1 bit per transmission.

- $a \notin \{0, \pm 1\}$: In this case Y has four possible 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

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

  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 $\mathsf{C} = 1 - \alpha = \frac{1}{2}$ bit 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 also $\mathsf{C} = \frac{1}{2}$ bit per transmission, achieved by a uniform distribution.

Problem 3 — ***Z-Channel***

Remember that the Z-channel looks as shown in Figure 1. First we express $I(X;Y)$, the mutual

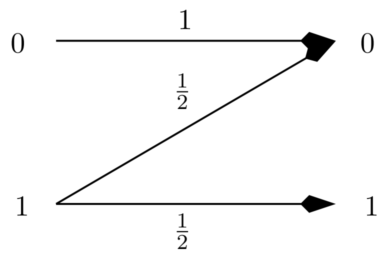


Figure 1: Z-Channel.

information between the input and output of the Z-channel, as a function of $p = \Pr[X = 1]$:

$$
\begin{aligned}
&H(Y|X = 0) = 0;\\
&H(Y|X = 1) = H_{\mathrm{b}}\left(\frac{1}{2}\right) = 1 \text{ bit};\\
&\quad\rightsquigarrow\quad H(Y|X) = \Pr[X = 0]\cdot 0 + \Pr[X = 1]\cdot 1 = p \text{ bits};\\
&\Pr[Y = 0] = \frac{1}{2}\cdot\Pr[X = 1] + 1\cdot\Pr[X = 0] = \frac{1}{2}p + 1 - p = 1 - \frac{1}{2}p;\\
&\Pr[Y = 1] = \frac{1}{2}\Pr[X = 1] = \frac{1}{2}p;\\
&\quad\rightsquigarrow\quad H(Y) = H_{\mathrm{b}}\left(\frac{p}{2}\right);\\
&\quad\rightsquigarrow\quad I(X;Y) = H(Y) - H(Y|X) = H_{\mathrm{b}}\left(\frac{p}{2}\right) - p \text{ bits}.
\end{aligned}
$$

Since $I(X;Y) = 0$ when $p = 0$ and $p = 1$, the maximum mutual information is obtained for some value of p such that $0 < p < 1$. Using elementary calculus, we determine that

$$
\frac{\mathrm{d}}{\mathrm{d}p} I(X;Y) = \frac{1}{2}\log_2\frac{2-p}{p} - 1,
$$

which is equal to zero for $p = \frac{2}{5}$. (It is reasonable that $\Pr[X = 1] < \frac{1}{2}$ because $X = 1$ is the noisy input to the channel.) So the capacity of the Z-channel in bits is $H_\text{b}(\frac{1}{5}) - \frac{2}{5} \approx 0.722 - 0.4 = 0.322$ bits per channel use.

Problem 4 — *Capacity of a Sum Channel*

a) Let $S \in \{1, \ldots, \nu\}$ be the chance variable representing the selected channel, i.e., $S \sim \mathbf{s}$ where $\mathbf{s}$ is the probability vector $(s_1, \ldots, s_\nu)$.

Applying the chain rule twice to $H(X, S)$ leads to

$$H(X) + \underbrace{H(S|X)}_{=0} = H(X, S) = H(S) + H(X|S),$$

where the underbraced term is zero because X determines S. Applying the chain rule twice to $H(X, S|Y)$ leads to

$$H(X|Y) + \underbrace{H(S|X, Y)}_{=0} = H(X, S|Y) = \underbrace{H(S|Y)}_{=0} + H(X|Y, S),$$

where the underbraced terms are zero because Y determines S. Therefore, we have

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(S) + H(X|S) - H(X|Y,S) \\ &= H(S) + I(X;Y|S) \\ &= H(S) + \sum_{i=1}^{\nu} s_i I(X;Y|S=i). \end{aligned}$$

By the law of total probability, we rewrite the input distribution of the sum channel as

$$P_X(x) = \sum_{i=1}^{\nu} s_i P_{X|S}(x|i).$$

Plugging these expression into the formula for the capacity leads to

$$\begin{aligned} \mathsf{C} &= \max_{P_X} I(X;Y) \\ &= \max_{\mathbf{s}} \max_{P_{X|S}} \left(H(S) + \sum_{i=1}^{\nu} s_i I(X;Y|S=i) \right) \\ &= \max_{\mathbf{s}} \left(H(S) + \sum_{i=1}^{\nu} s_i \mathsf{C}_i \right), \end{aligned}$$

where the last equality follows because $I(X;Y|S=i) \le \mathsf{C}_i$ (by the definition of the capacity) and because $I(X;Y|S=i) = \mathsf{C}_i$ if and only if, conditional on $S = i$, X is chosen to be a capacity-achieving input distribution for the i-th channel. Therefore, C is equal to the entropy of the channel selection plus a weighted average of the channel capacities, and it remains to determine the optimal channel selection.

We proceed with

$$H(S) + \sum_{i=1}^{\nu} s_i \mathsf{C}_i = \sum_{i=1}^{\nu} s_i \log \frac{1}{s_i} + \sum_{i=1}^{\nu} s_i \log 2^{\mathsf{C}_i}$$

$$
\begin{aligned}
&= -\sum_{i=1}^{\nu} s_i \log \frac{s_i}{2^{\mathsf{C}_i}} \\
&= -\sum_{i=1}^{\nu} s_i \log \frac{s_i/\alpha}{2^{\mathsf{C}_i}/\alpha} && \left(\text{where } \alpha \triangleq \sum_{i=1}^{\nu} 2^{\mathsf{C}_i}\right) \\
&= \sum_{i=1}^{\nu} s_i \log \alpha - \sum_{i=1}^{\nu} s_i \log \frac{s_i}{t_i} && \left(\text{where } t_i \triangleq \frac{2^{\mathsf{C}_i}}{\alpha}; \quad \rightsquigarrow \sum_{i=1}^{\nu} t_i = 1\right) \\
&= \log \alpha - \underbrace{D(\mathbf{s}\|\mathbf{t})}_{\geq 0}
\end{aligned}
$$

and conclude that

$$
\begin{aligned}
\mathsf{C} &= \max_{\mathbf{s}} \left(H(S) + \sum_{i=1}^{\nu} s_i \mathsf{C}_i \right) \\
&= \log \alpha = \log \sum_{i=1}^{\nu} 2^{\mathsf{C}_i},
\end{aligned}
$$

where the maximum is achieved for $\mathbf{s} = \mathbf{t}$, i.e., $s_i = \frac{2^{\mathsf{C}_i}}{\sum_{j=1}^{\nu} 2^{\mathsf{C}_j}}$ for all $i \in \{1, \ldots, \nu\}$.

b) The capacity of the BSC is $1 - H_{\mathrm{b}}(\epsilon)$ bits and the capacity of the other channel is zero. Thus by Part a) the capacity of the sum channel is

$$
\mathsf{C} = \log\left(2^{1-H_{\mathrm{b}}(\epsilon)} + 2^0\right) = \log\left(1 + 2^{1-H_{\mathrm{b}}(\epsilon)}\right).
$$

Problem 5 — *Using Two Channels At Once*

Suppose we are given two channels, $(\mathcal{X}_1, W_1(\cdot|\cdot), \mathcal{Y}_1)$ and $(\mathcal{X}_2, W_2(\cdot|\cdot), \mathcal{Y}_2)$, which we can use at the same time. We can define the product channel as the channel $(\mathcal{X}_1 \times \mathcal{X}_2, W(\cdot,\cdot|\cdot,\cdot) = W_1(\cdot|\cdot)W_2(\cdot|\cdot), \mathcal{Y}_1 \times \mathcal{Y}_2)$. To find the capacity of the product channel, we must find the distribution $Q(x_1, x_2)$ on the input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$ that maximizes $I(X_1, X_2; Y_1, Y_2)$. Because the channels are independent, we have

$$
H(Y_1, Y_2|X_1, X_2) = \underbrace{H(Y_1|X_1, X_2)}_{= H(Y_1|X_1)} + \underbrace{H(Y_2|X_1, X_2, Y_1)}_{= H(Y_2|X_2)}.
$$

Consequently,

$$
\begin{aligned}
I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \\
&= H(Y_1) + H(Y_2|Y_1) - H(Y_1, Y_2|X_1, X_2) \\
&= H(Y_1) + H(Y_2|Y_1) - H(Y_1|X_1) - H(Y_2|X_2) \\
&\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\
&= I(X_1; Y_1) + I(X_2; Y_2), \qquad (1)
\end{aligned}
$$

where the inequality follows from the fact that conditioning cannot increase entropy, and where we have equality if and only if Y_1 and Y_2 are independent. Hence

$$
\begin{aligned}
\mathsf{C} &= \max_{Q(x_1,x_2)} I(X_1, X_2; Y_1, Y_2) \\
&\leq \max_{Q(x_1,x_2)} \big(I(X_1; Y_1) + I(X_2; Y_2) \big)
\end{aligned}
$$

$$
\begin{aligned}
&\le \max_{Q_1(x_1,x_2)} I(X_1;Y_1) + \max_{Q_2(x_1,x_2)} I(X_2;Y_2) \\
&= \max_{Q_1(x_1)} I(X_1;Y_1) + \max_{Q_2(x_2)} I(X_2;Y_2) \\
&= \mathsf{C}_1 + \mathsf{C}_2,
\end{aligned}
$$

where the first inequality follows from (1), thus equality can be reached if Y_1 and Y_2 are independent. In the second inequality, equality can be achieved if X_1 and X_2 are independent, in which case also Y_1 and Y_2 are independent. Therefore,

$$\mathsf{C} = \mathsf{C}_1 + \mathsf{C}_2.$$

Note: We can use the above result inductively to generalize to a combination of many channels. By induction,

$$\mathsf{C} = \sum_{k=1}^{n} \mathsf{C}_k.$$

Alternatively, we can view the n discrete memoryless channels as n channels connected in parallel. The input $\mathbf{x} = (x_1, \dots, x_n)$ and output $\mathbf{y} = (y_1, \dots, y_n)$ are n-tuples of inputs and outputs to the individual channels. The output from each channel depends only on the input of this channel. That is,

$$W(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{n} W(y_k|x_k).$$

Thus,

$$H(Y_1, \dots, Y_n|X_1, \dots, X_n) = \sum_{k=1}^{n} H(Y_k|X_k).$$

Notice that because Y_i depends on X_i, Y_j depends on X_j, and X_i and X_j can depend on each other, Y_i and Y_j are not necessarily independent. Therefore,

$$H(Y_1, \dots, Y_N) \le \sum_{k=1}^{n} H(Y_k),$$

with equality if the inputs are statistically independent. Using the above equations,

$$
\begin{aligned}
I(X_1, \dots, X_n; Y_1, \dots, Y_n) &= H(Y_1, \dots, Y_n) - H(Y_1, \dots, Y_n|X_1, \dots, X_n) \\
&\le \sum_{k=1}^{n} (H(Y_k) - H(Y_k|X_k)) \\
&= \sum_{k=1}^{n} I(X_k; Y_k) \\
&\le \sum_{k=1}^{n} \mathsf{C}_k,
\end{aligned}
$$

with equality if the inputs are statistically independent and if the individual input probabilities are chosen to achieve capacity on the individual channels. Thus the capacity of the parallel combination is

$$\mathsf{C} = \max_{Q} I(X_1, \dots, X_n; Y_1, \dots, Y_n) = \sum_{k=1}^{n} \mathsf{C}_k.$$

Note: Do not confuse $X_1, \dots, X_n$ with the usual time-dependent chance variables. Here they denote the several inputs of the many channels that are combined.