Information Theory I

Fall Semester 2017

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory

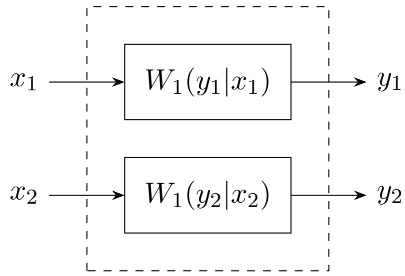
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 7 of November 1, 2017

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

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.



$W_2(y_1, y_2|x_1, x_2)$

a) Let n and R be fixed. Let $f\colon \{1, \dots, 2^{2n\mathsf{R}}\} \to \mathcal{X}^{2n}$ and $\phi\colon \mathcal{Y}^{2n} \to \{1, \dots, 2^{2n\mathsf{R}}\}$ be a length-$2n$ rate-R encoder and decoder pair for channel W_1. We construct a length-n rate-$2\mathsf{R}$ encoder and decoder pair $g\colon \{1, \dots, 2^{n2\mathsf{R}}\} \to (\mathcal{X} \times \mathcal{X})^n$ and $\psi\colon (\mathcal{Y} \times \mathcal{Y})^n \to \{1, \dots, 2^{n2\mathsf{R}}\}$ for channel W_2 by interleaving:

- Note that both encoder and decoder pairs have the same message set since $2^{2n\mathsf{R}} = 2^{n2\mathsf{R}}$.
- To encode a message m, use f to obtain the codeword $(x_1, \dots, x_{2n}) = f(m)$, and define $g(m)$ to be
$$\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}, \dots, \begin{bmatrix} x_{2n-1} \\ x_{2n} \end{bmatrix} \right).$$
(The top entries correspond to the first input of W_2, and the bottom entries correspond to the second input of W_2.)
- To decode a sequence
$$\mathbf{y} = \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \begin{bmatrix} y_3 \\ y_4 \end{bmatrix}, \dots, \begin{bmatrix} y_{2n-1} \\ y_{2n} \end{bmatrix} \right),$$
use ϕ, i.e., define $\psi(\mathbf{y})$ to be $\phi(y_1, \dots, y_{2n})$. (The top entries correspond to the first output of W_2, and the bottom entries correspond to the second output of W_2.)

Since one use of W_2 corresponds to two independent uses of W_1, the encoder and decoder pair (g, ψ) has the same probability of error as the encoder and decoder pair (f, ϕ).

If R is achievable on W_1, there must exist a sequence of encoder and decoder pairs for W_1 whose probability of error tends to zero as n tends to infinity. In this case, the above construction yields a sequence of encoder and decoder pairs for W_2 whose probability of error tends to zero as n tends to infinity, which shows that $2\mathsf{R}$ is achievable on W_2.

b) We use the same idea as in Part a). For a given length-n rate-R encoder on W_2, we construct a length-$2n$ rate-$\mathsf{R}/2$ encoder for W_1 that alternates between emitting the first input for W_2 and the second input for W_2. Similarly, for a given length-n rate-R decoder on W_2, we construct a length-$2n$ rate-$\mathsf{R}/2$ decoder for W_1 that splits the received sequence into two sequences and invokes the decoder for W_2. The encoder and decoder pair for W_1 then has the same probability of error as the encoder and decoder pair for W_2.

If R is achievable on W_2, there must exist a sequence of encoder and decoder pairs for W_2 whose probability of error tends to zero as n tends to infinity. Using the above construction and ignoring that it only yields encoder and decoder pairs for W_1 with *even* blocklength, we see that $\mathsf{R}/2$ is achievable on W_1.

Problem 3 — ***Z-Channel***

Remember that the Z-channel looks as shown in Figure 1.

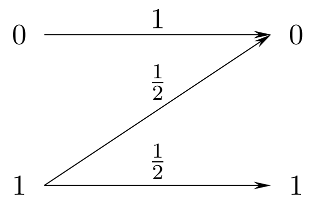


Figure 1: Z-Channel.

First we express $I(X;Y)$, the mutual information between the input and output of the Z-channel, as a function of $p = \Pr[X = 1]$:

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

Since $I(X;Y) = 0$ if $p = 0$ or $p = 1$, the maximum mutual information is obtained for some value of p such that $0 < p < 1$. Using 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 is

$$
H_{\mathrm{b}}\left(\frac{1}{5}\right) - \frac{2}{5} \approx 0.722 - 0.4 = 0.322 \quad \text{bits per channel use.}
$$

The capacity of the parallel channel can be computed as

$$\mathsf{C} = \max_{Q} I(X_1, X_2; Y_1, Y_2),$$

where the maximum is over all input distributions on the alphabet $\mathcal{X}_1 \times \mathcal{X}_2$. Because the channels are independent, we have

$$H(Y_1, Y_2|X_1, X_2) \overset{\text{(i)}}{=} H(Y_1|X_1) + H(Y_2|X_2).$$

(Since

$$\Pr[Y_1 = y_1, Y_2 = y_2|X_1 = x_1, X_2 = x_2] = \Pr[Y_1 = y_1|X_1 = x_1] \cdot \Pr[Y_2 = y_2|X_2 = x_2],$$

we have

$$H(Y_1, Y_2|X_1 = x_1, X_2 = x_2) = H(Y_1|X_1 = x_1) + H(Y_2|X_2 = x_2),$$

which implies (i).) Therefore,

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

where (ii) holds because conditioning does not increase entropy. Thus,

$$\begin{aligned} \mathsf{C} &= \max_{Q} I(X_1, X_2; Y_1, Y_2) \\ &\leq \max_{Q} \big\{ I(X_1; Y_1) + I(X_2; Y_2) \big\} \\ &\overset{\text{(iii)}}{\leq} \mathsf{C}_1 + \mathsf{C}_2, \end{aligned} \tag{1}$$

where (iii) holds because $I(X_1; Y_1) \leq \mathsf{C}_1$ and $I(X_2; Y_2) \leq \mathsf{C}_2$ by the definition of C_1 and C_2. Choosing $Q(x_1, x_2) = Q_1(x_1)\, Q_2(x_2)$, where Q_1 and Q_2 are capacity-achieving input distributions for W_1 and W_2, respectively, leads to

$$\begin{aligned} I(X_1, X_2; Y_1, Y_2) &\overset{\text{(iv)}}{=} I(X_1; Y_1) + I(X_2; Y_2) \\ &\overset{\text{(v)}}{=} \mathsf{C}_1 + \mathsf{C}_2, \end{aligned} \tag{2}$$

where (iv) holds because X_1 and X_2 are independent and because the channels are independent, so Y_1 and Y_2 independent, which implies that (ii) holds with equality; and (v) holds because X_1 and X_2 are distributed according to Q_1 and Q_2, respectively, which are assumed to be capacity-achieving input distributions. Combining (1) and (2), we conclude that

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

Note: The above result can be generalized to n independent parallel channels. By induction,

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

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

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 holds 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

$$\begin{aligned} H(S) + \sum_{i=1}^{\nu} s_i \mathsf{C}_i &= \sum_{i=1}^{\nu} s_i \log \frac{1}{s_i} + \sum_{i=1}^{\nu} s_i \log 2^{\mathsf{C}_i} \\ &= -\sum_{i=1}^{\nu} s_i \log \frac{s_i}{2^{\mathsf{C}_i}} \\ &= -\sum_{i=1}^{\nu} s_i \log \frac{s_i/\alpha}{2^{\mathsf{C}_i}/\alpha} \qquad \left(\text{where } \alpha \triangleq \sum_{i=1}^{\nu} 2^{\mathsf{C}_i} \right) \end{aligned}$$

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

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

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