Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 10 of November 23, 2016

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

Problem 1 *Channel Coding for a "Double"-Channel*

To prove the existence of such a codebook, we proceed along the same lines as for the achievability part of the channel coding theorem:

- Let Q and R be such that $\mathsf{R} < \min\{I(Q, W^{(1)}), I(Q, W^{(2)})\}$, and let $\epsilon' > 0$.
- For a fixed n, generate a random codebook $\mathscr{C}$ of block length n according to Q and reveal it to the encoder and both decoders.
- Use joint typicality decoding for both decoders, which fails if no codeword or more than one codeword is jointly typical with the received sequence. For a fixed message m and for n large enough, the performance analysis leads to

$$\Pr\left[\hat{M}^{(1)} \neq M \middle| M = m\right] \leq \epsilon' + 2^{n\mathsf{R}} \cdot 2^{-n(I(Q,W^{(1)})-3\epsilon')},$$
$$\Pr\left[\hat{M}^{(2)} \neq M \middle| M = m\right] \leq \epsilon' + 2^{n\mathsf{R}} \cdot 2^{-n(I(Q,W^{(2)})-3\epsilon')}.$$

  Averaging the error probabilities over all messages (which corresponds to uniform message probabilities) and applying the union bound leads to

$$\Pr\left[\hat{M}^{(1)} \neq M \text{ or } \hat{M}^{(2)} \neq M\right] \leq 2\epsilon' + 2^{n\mathsf{R}} \cdot 2^{-n(I(Q,W^{(1)})-3\epsilon')} + 2^{n\mathsf{R}} \cdot 2^{-n(I(Q,W^{(2)})-3\epsilon')}.$$

  For an appropriate choice of ϵ' and n large enough, the average probability of error satisfies

$$\Pr\left[\hat{M}^{(1)} \neq M \text{ or } \hat{M}^{(2)} \neq M\right] < \frac{\epsilon}{2} \tag{1}$$

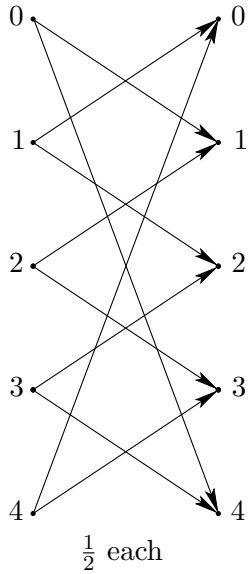
  for any $\epsilon > 0$ because $\mathsf{R} < \min\{I(Q, W^{(1)}), I(Q, W^{(2)})\}$.
- Since the codebook is generated independently of the message, there exists a (deterministic) codebook $\mathscr{C}$ for which (1) holds.
- We obtain a maximal probability of error less than ϵ by throwing away half of the codewords.

Problem 2 *Zero-Error Capacity*

The channel has the transition matrix $W(y|x)$ shown in Figure 1.

a) Since the channel is strongly symmetric, its capacity is

$$\mathsf{C} = \log|\mathcal{Y}| - H(\text{row of transition matrix}) = \log_2 5 - 1 = \log_2 2.5 \approx 1.322 \text{ bits}.$$



$$W(y|x) = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Figure 1: Transition matrix of the channel.

b) Let us construct a block code consisting of codewords of length 2. We want to achieve a capacity greater than 1 bit per channel use, i.e., greater than 2 bits per two channel uses. Thus, we need more than 4 codewords. Let's pick 5 codewords with distinct first symbols: $\{0a, 1b, 2c, 3d, 4e\}$. We must choose a, b, c, d, e so that the receiver will be able to determine which codeword was transmitted. A simple repetition code will not work, since if, say, 22 is transmitted, then 11 might be received, and the receiver could not tell whether the codeword was 00 or 22. Instead, we use the code $\{00, 13, 21, 34, 42\}$, i.e., the codewords can be described as $\mathbf{x}_i = (i, 3i \mod 5)$ for all $i = 0, \ldots, 4$. Then each codeword will be received as one of 4 possible 2-tuples which are all distinct:

$$\begin{aligned} 00 &\to \{44, 14, 41, 11\} \\ 13 &\to \{02, 22, 04, 24\} \\ 21 &\to \{10, 30, 12, 32\} \\ 34 &\to \{23, 43, 20, 40\} \\ 42 &\to \{31, 01, 33, 03\}. \end{aligned}$$

Since there are 5 possible error-free messages with 2 channel uses, the zero-error capacity of this channel is at least $\frac{1}{2}\log_2 5 \approx 1.161$ bits.

The zero-error capacity of this channel is in fact exactly $\frac{1}{2}\log_2 5$, which was proved by László Lovász in the celebrated paper "On the Shannon capacity of a graph," *IEEE Transactions on Information Theory*, vol. IT-25, no. 1, pp. 1–7, January 1979.

Here is an *alternative way* of solving this problem graphically. We introduce a graph consisting of vertices and undirected edges. In the case of block length $n = 1$ we introduce for each *input* symbol one vertex and two vertices are adjacent (i.e., they are connected by an edge) if and only if there is at least one *output* symbol which can be the possible output of both input symbols corresponding to the two vertices. In our example, 0 and 2 are adjacent because both can lead to the output symbol 1; but 0 and 1 are not adjacent because they have no common output symbol (see Figure 2).

We see that two messages of length $n = 1$ cannot be confused if they are not adjacent, so to achieve error-freeness our codebook has to consist of non-adjacent vertices. Of course we try to take the maximum number of possible non-adjacent vertices. (In graph theory this

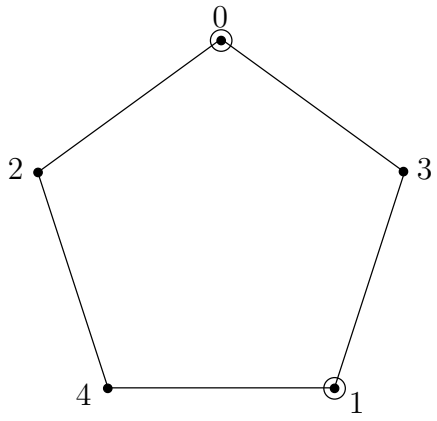


Figure 2: Graphical interpretation.

number is called the independence number.) A choice for a codebook are, e.g., the vertices 0 and 1. (By the way, because the adjacency graph for $n = 1$ for this channel looks like a pentagon, this channel is sometimes called the "pentagon channel".)

In the case of block length $n = 2$ we proceed similarly. But now, for each sequence of two input symbols we draw a vertex and there is an edge between two vertices if the two input sequences corresponding to these two vertices can be confused (i.e., if they can lead to the same output sequence). In our example, 00 and 20 are adjacent because they have the possible common output 11 (or 14). A codebook consisting of five codewords leading to error-free transmission is drawn in Figure 3.

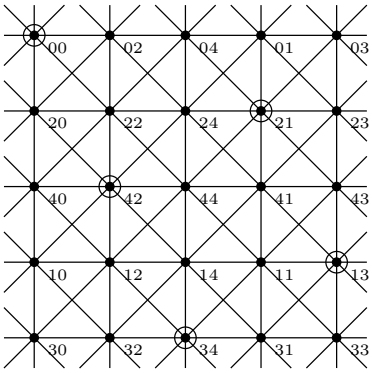


Figure 3: Codebook of five codewords (the graph has to be understood cyclically).

Problem 3 ***Symbol Error Rate vs. Message Error Rate***

a) We have

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\Pr\Big[X_k \neq \hat{X}_k\Big] &\overset{\text{(i)}}{\leq} \max_k \Pr\Big[X_k \neq \hat{X}_k\Big] \\ &\overset{\text{(ii)}}{\leq} \Pr\left(\bigcup_{k=1}^{n}\Big\{X_k \neq \hat{X}_k\Big\}\right) \\ &= \Pr\Big[(X_1, X_2, \ldots, X_n) \neq (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n)\Big],\end{aligned}$$

where (i) follows because an average is smaller than or equal to the maximum and (ii) follows because the union contains the selected event. Therefore, if $\frac{1}{n}\sum_{k=1}^{n}\Pr\big[X_k \neq \hat{X}_k\big]$ does not tend to zero as n tends to infinity, then neither does

$$\Pr\big[(X_1, X_2, \ldots, X_n) \neq (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n)\big].$$

b) We justify the steps as follows:

i) follows by the chain rule;

ii) follows because conditioning does not increase entropy;

iii) follows from $H(U_i|Z_1, \ldots, Z_n) \leq H(U_i|Z_i)$ (because conditioning does not increase entropy) and from Fano's inequality (with U_i replacing X and with Z_i replacing Y);

iv) follows from Jensen's inequality because entropy is concave; and

v) holds by the definition of P_{b}.

c) Let $(X_1, \ldots, X_n) = f(M)$ and $(\hat{X}_1, \ldots, \hat{X}_n) = f(\hat{M})$, where $\hat{M}$ is the message guessed by the decoder, and let the sequence $X_1, \ldots, X_n$ take value in a finite set of cardinality $|\mathcal{X}|$. Assume that $f(\cdot)$ is injective. We prove the converse to the channel coding theorem for memoryless channels (though the converse holds more generally). Let C equal the capacity of the channel. Then,

$$\begin{aligned}
\mathsf{R} = \frac{1}{n}H(M) &\overset{\text{(i)}}{=} \frac{1}{n}H(X_1, \ldots, X_n) \\
&\overset{\text{(ii)}}{=} \frac{1}{n}H\big(X_1, \ldots, X_n\big|\hat{X}_1, \ldots, \hat{X}_n\big) + \frac{1}{n}I\big(X_1, \ldots, X_n; \hat{X}_1, \ldots, \hat{X}_n\big) \\
&\overset{\text{(iii)}}{\leq} H_{\mathrm{b}}(P_{\mathrm{b}}) + P_{\mathrm{b}}\log(|\mathcal{X}| - 1) + \frac{1}{n}I\big(X_1, \ldots, X_n; \hat{X}_1, \ldots, \hat{X}_n\big) \\
&\overset{\text{(iv)}}{\leq} H_{\mathrm{b}}(P_{\mathrm{b}}) + P_{\mathrm{b}}\log(|\mathcal{X}| - 1) + \frac{1}{n}I\big(X_1, \ldots, X_n; Y_1, \ldots, Y_n\big) \\
&\overset{\text{(v)}}{\leq} H_{\mathrm{b}}(P_{\mathrm{b}}) + P_{\mathrm{b}}\log(|\mathcal{X}| - 1) + \frac{1}{n}n\mathsf{C} \\
&= H_{\mathrm{b}}(P_{\mathrm{b}}) + P_{\mathrm{b}}\log(|\mathcal{X}| - 1) + \mathsf{C},
\end{aligned}$$

where (i) follows because f is injective; (ii) follows from the definition of mutual information; (iii) follows from Part b); (iv) follows from the data-processing inequality; and (v) follows from Lemma 7.9.2 in Cover & Thomas. Consequently,

$$\mathsf{R} - \mathsf{C} \leq H_{\mathrm{b}}(P_{\mathrm{b}}) + P_{\mathrm{b}}\log(|\mathcal{X}| - 1), \tag{2}$$

which implies that $P_{\mathrm{b}} = \frac{1}{n}\sum_{k=1}^{n}\Pr\big[X_k \neq \hat{X}_k\big]$ cannot tend to zero if $\mathsf{R} > \mathsf{C}$.

Problem 4 — *An Elementary Converse for the Binary Erasure Channel*

a) Fix an arbitrary erasure pattern $\mathbf{s} \in \mathcal{F}_{n\kappa}$ and suppose $S_1^n = \mathbf{s}$. Since exactly $n\kappa$ positions are erased, there are $n(1-\kappa)$ nonerased positions, and the number of different channel output sequences is $2^{n(1-\kappa)}$. The decoder can therefore only guess at most $2^{n(1-\kappa)}$ different messages, and at least $2^{n\mathsf{R}} - 2^{n(1-\kappa)}$ messages cannot be decoded correctly. The average probability of error conditional on $S_1^n = \mathbf{s}$ thus satisfies (the bound is trivially true if $2^{n\mathsf{R}} - 2^{n(1-\kappa)} < 0$)

$$\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\Pr\big[\phi(Y_1^n) \neq m\big|X_1^n = g(m),\ S_1^n = \mathbf{s}\big] \geq \frac{2^{n\mathsf{R}} - 2^{n(1-\kappa)}}{2^{n\mathsf{R}}}. \tag{3}$$

We conclude with

$$\begin{aligned}
P_{\mathrm{e}}(\mathcal{F}_{n\kappa}) &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \Pr\big[\phi(Y_1^n)\neq m\big|X_1^n=g(m),\, S_1^n\in\mathcal{F}_{n\kappa}\big] \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \frac{\Pr\big[\phi(Y_1^n)\neq m,\, S_1^n\in\mathcal{F}_{n\kappa}\big|X_1^n=g(m)\big]}{\Pr\big[S_1^n\in\mathcal{F}_{n\kappa}\big|X_1^n=g(m)\big]} \\
&\overset{\text{(i)}}{=} \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\sum_{\mathbf{s}\in\mathcal{F}_{n\kappa}} \frac{\Pr\big[\phi(Y_1^n)\neq m\big|X_1^n=g(m),\, S_1^n=\mathbf{s}\big]\cdot\Pr\big[S_1^n=\mathbf{s}\big|X_1^n=g(m)\big]}{\Pr\big[S_1^n\in\mathcal{F}_{n\kappa}\big|X_1^n=g(m)\big]} \\
&\overset{\text{(ii)}}{=} \sum_{\mathbf{s}\in\mathcal{F}_{n\kappa}} \frac{\Pr\big[S_1^n=\mathbf{s}\big]}{\Pr\big[S_1^n\in\mathcal{F}_{n\kappa}\big]}\cdot\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\Pr\big[\phi(Y_1^n)\neq m\big|X_1^n=g(m),\, S_1^n=\mathbf{s}\big] \\
&\overset{\text{(iii)}}{\geq} \sum_{\mathbf{s}\in\mathcal{F}_{n\kappa}} \frac{\Pr\big[S_1^n=\mathbf{s}\big]}{\Pr\big[S_1^n\in\mathcal{F}_{n\kappa}\big]}\cdot\frac{2^{n\mathsf{R}}-2^{n(1-\kappa)}}{2^{n\mathsf{R}}} \\
&\overset{\text{(iv)}}{=} \frac{2^{n\mathsf{R}}-2^{n(1-\kappa)}}{2^{n\mathsf{R}}},
\end{aligned}$$

where (i) follows from the law of total probability; (ii) follows because S_1^n is independent of m; (iii) follows from (3); and (iv) follows because $\sum_{\mathbf{s}\in\mathcal{F}_{n\kappa}}\Pr\big[S_1^n=\mathbf{s}\big]=\Pr\big[S_1^n\in\mathcal{F}_{n\kappa}\big]$.

b) Fix a number α so that $1-\mathsf{R}<\alpha<\rho$. Then,

$$\begin{aligned}
P_{\mathrm{e}}^{(n)} &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \Pr\big[\phi(Y_1^n)\neq m\big|X_1^n=g(m)\big] \\
&\overset{\text{(i)}}{=} \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\sum_{\substack{\kappa\in[0,1]\\ n\kappa\in\mathbb{N}_0}} \Pr\big[\phi(Y_1^n)\neq m\big|X_1^n=g(m),\, S_1^n\in\mathcal{F}_{n\kappa}\big]\cdot\Pr\big[S_1^n\in\mathcal{F}_{n\kappa}\big|X_1^n=g(m)\big] \\
&\overset{\text{(ii)}}{=} \sum_{\substack{\kappa\in[0,1]\\ n\kappa\in\mathbb{N}_0}} \Pr[S_1^n\in\mathcal{F}_{n\kappa}]\,\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\Pr\big[\phi(Y_1^n)\neq m\big|X_1^n=g(m),\, S_1^n\in\mathcal{F}_{n\kappa}\big] \\
&\geq \sum_{\substack{\kappa\in[\alpha,1]\\ n\kappa\in\mathbb{N}_0}} \Pr[S_1^n\in\mathcal{F}_{n\kappa}]\,\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\Pr\big[\phi(Y_1^n)\neq m\big|X_1^n=g(m),\, S_1^n\in\mathcal{F}_{n\kappa}\big] \\
&\overset{\text{(iii)}}{\geq} \sum_{\substack{\kappa\in[\alpha,1]\\ n\kappa\in\mathbb{N}_0}} \Pr[S_1^n\in\mathcal{F}_{n\kappa}]\,\frac{2^{n\mathsf{R}}-2^{n(1-\kappa)}}{2^{n\mathsf{R}}} \\
&\overset{\text{(iv)}}{\geq} \Pr\left[\frac{1}{n}\sum_{i=1}^n S_i\geq\alpha\right]\Big(1-2^{-n(\mathsf{R}-(1-\alpha))}\Big),
\end{aligned}$$

where (i) follows from the law of total probability; (ii) follows because S_1^n is independent of m; (iii) follows from Part a); and (iv) follows because $2^{n\mathsf{R}}-2^{n(1-\kappa)}\geq 2^{n\mathsf{R}}-2^{n(1-\alpha)}$ as $\kappa\geq\alpha$. The assumption $1-\mathsf{R}<\alpha$ implies $\mathsf{R}>1-\alpha$; therefore,

$$1-2^{-n(\mathsf{R}-(1-\alpha))}\to 1 \quad \text{as } n\to\infty.$$

By the weak law of large numbers and since $\rho-\alpha>0$,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n S_i\geq\alpha\right]=\Pr\left[\frac{1}{n}\sum_{i=1}^n S_i\geq\rho-(\rho-\alpha)\right]\to 1 \quad \text{as } n\to\infty.$$

Combining the results shows that $P_{\mathrm{e}}^{(n)}$ tends to one as n tends to infinity.

Problem 5 ***Source-Channel Separation and Average Bit-Error Probability***

a) We upperbound $I\big(U_1^k;\hat{U}_1^k\big)$ with the data processing inequality:

$$I\big(U_1^k;\hat{U}_1^k\big) \le I(X_1^n;Y_1^n) \le n\mathsf{C},$$

where the second inequality follows from Lemma 7.9.2 in Cover & Thomas.

b) In the computation, (a) follows from the chain rule; (b) holds because

$$I\big(U_i;\hat{U}^k\big|U^{i-1}\big) = I\big(U_i;\hat{U}^k,U^{i-1}\big) - I\big(U_i;U^{i-1}\big)$$

by the chain rule and since U_i is independent of U^{i-1}; (c) is true because

$$H\big(U_i|\hat{U}^k,U^{i-1}\big) \le H\big(U_i|\hat{U}_i\big) \le H_{\mathrm{b}}\big(\Pr\big[U_i\neq\hat{U}_i\big]\big) + \Pr\big[U_i\neq\hat{U}_i\big]\underbrace{\log(2-1)}_{=0},$$

where the first inequality follows because conditioning does not increase entropy and the second inequality follows from Fano's inequality, and

$$I\big(U_i;\hat{U}^k,U^{i-1}\big) = \underbrace{H\big(U_i\big)}_{=1} - \underbrace{H\big(U_i|\hat{U}^k,U^{i-1}\big)}_{\le H_{\mathrm{b}}\left(\Pr\left[U_i\neq\hat{U}_i\right]\right)} \ge 1 - H_{\mathrm{b}}\big(\Pr\big[U_i\neq\hat{U}_i\big]\big);$$

and (d) holds since entropy is concave.

c) Combining the previous parts, we find that

$$\begin{aligned} n\mathsf{C} &\ge I\big(U^k;\hat{U}^k\big) \\ &\ge k\left(1 - H_{\mathrm{b}}\left(\frac{1}{k}\sum_{i=1}^{k}\Pr\big[U_i\neq\hat{U}_i\big]\right)\right). \end{aligned}$$

Since $H_{\mathrm{b}}(\cdot)$ is invertible on $[0,\,1/2]$, we have

$$\frac{1}{k}\sum_{i=1}^{k}\Pr\big[U_i\neq\hat{U}_i\big] \ge H_{\mathrm{b}}^{-1}\Big(1-\frac{n}{k}\mathsf{C}\Big) > 0,$$

where the last inequality holds by the assumption that $\frac{k}{n} > \mathsf{C}$.