Information Theory I

Fall Semester 2017

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 9 of November 15, 2017

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

Problem 1 ***Encoder and Decoder as Part of the Channel***

a) The output of the majority decoder is wrong if and only if two or three bits are flipped. Thus, the crossover probability of the new BSC is

$$\tilde{\epsilon} = \binom{3}{2}\epsilon^2(1-\epsilon) + \epsilon^3 = 0.028,$$

where $\epsilon = 0.1$ denotes the crossover probability of the original BSC.

b) The capacity of the new channel is

$$\tilde{\mathsf{C}} = 1 - H_{\mathrm{b}}(\tilde{\epsilon}) \approx 0.816 \text{ bits per use of the new channel.}$$

Since one use of the new channel corresponds to three uses of the original channel,

$$\tilde{\mathsf{C}} = \frac{1 - H_{\mathrm{b}}(\tilde{\epsilon})}{3} \approx 0.272 \text{ bits per use of the original channel.}$$

c) The capacity of the original channel is

$$\mathsf{C} = 1 - H_{\mathrm{b}}(\epsilon) \approx 0.531 \text{ bits per use of the original channel.}$$

Comparing with Part b), we see that the suggested coding scheme reduces the capacity by approximately 49 %.

d) Denote the message by M, the channel inputs by X^n, the channel outputs by Y^n, the estimated message by $\hat{M}$, and the capacity of the original channel by C. Observe that $M \multimap X^n \multimap Y^n \multimap \hat{M}$ form a Markov chain, so

$$\begin{aligned}
I(M;\hat{M}) &\overset{\text{(i)}}{\leq} I(X^n;Y^n) \\
&= H(Y^n) - H(Y^n|X^n) \\
&\overset{\text{(ii)}}{=} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - \sum_{i=1}^{n} H(Y_i|X^n, Y^{i-1}) \\
&\overset{\text{(iii)}}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X^n, Y^{i-1}) \\
&\overset{\text{(iv)}}{=} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} I(X_i; Y_i) \\
&\overset{\text{(v)}}{\leq} n\mathsf{C},
\end{aligned}
$$

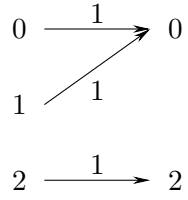
where (i) follows from the data processing inequality for mutual information; (ii) follows from the chain rule for entropy; (iii) holds because conditioning does not increase entropy; (iv) holds because $(X^n, Y^{i-1}) \multimap\!\!-\!\!\circ\!\!- X_i \multimap\!\!-\!\!\circ\!\!- Y_i$ form a Markov chain since the original channel is memoryless and used without feedback; and (v) holds because $I(X_i; Y_i) \leq \mathsf{C}$ by the definition of capacity. Denoting the capacity of the new channel from M to $\hat{M}$ by $\tilde{\mathsf{C}}$, we have

$$
\tilde{\mathsf{C}} = \max_{P_M} I(M; \hat{M}) \leq n\mathsf{C}.
$$

Thus, the capacity of the new channel per use of the original channel satisfies $\frac{\tilde{\mathsf{C}}}{n} \leq \mathsf{C}$.

Problem 2 — *Nonuniqueness of Capacity-Achieving Input Distributions*

a) One example is the BSC with crossover probability $\epsilon = \frac{1}{2}$: The capacity is zero, and all input distributions achieve it. Another example is the following channel:



The capacity is one bit; and e.g. $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$, $\left(\frac{1}{8}, \frac{3}{8}, \frac{1}{2}\right)$, and $\left(0, \frac{1}{2}, \frac{1}{2}\right)$ are capacity-achieving input distributions.

b) We first prove that entropy is strictly concave.

Lemma 1. *For all $\alpha \in (0, 1)$ and for all probability mass functions P_1 and P_2,*

$$
H(\alpha P_1 + (1-\alpha)P_2) \geq \alpha H(P_1) + (1-\alpha)H(P_2)
$$

with equality if and only if $P_1 = P_2$.

Proof. Introduce a chance variable S with $\Pr[S = 1] = \alpha$ and $\Pr[S = 2] = 1 - \alpha$. Let $X_1 \sim P_1$ and $X_2 \sim P_2$. Define $Z \triangleq X_S$. Then,

$$
\begin{aligned}
\alpha H(P_1) + (1-\alpha)H(P_2) &\overset{\text{(i)}}{=} \alpha H(Z|S=1) + (1-\alpha)H(Z|S=2) \\
&\overset{\text{(ii)}}{=} H(Z|S) \\
&\overset{\text{(iii)}}{\leq} H(Z) \\
&\overset{\text{(iv)}}{=} H(\alpha P_1 + (1-\alpha)P_2),
\end{aligned}
$$

where (i), (ii) and (iv) follow from the definitions of S and Z; and (iii) holds because conditioning does not increase entropy. Equality holds in (iii) if and only if S and Z are independent, i.e., if and only if $P_1 = P_2$. ■

We prove the claim by contradiction. Let Q_1 and Q_2 be capacity-achieving input distributions. Assume that they induce different output distributions, i.e., that $(Q_1W) \neq (Q_2W)$. Define the input distribution $Q \triangleq \frac{1}{2}Q_1 + \frac{1}{2}Q_2$, so $P_{XY}(x,y) = Q(x)\,W(y|x)$. Then,

$$\begin{aligned}
\mathsf{C} &\overset{\text{(i)}}{\geq} I(X;Y)\\
&= H(Y) - H(Y|X)\\
&\overset{\text{(ii)}}{=} H\big((QW)\big) - H(Y|X)\\
&\overset{\text{(iii)}}{=} H\big(\tfrac{1}{2}(Q_1W) + \tfrac{1}{2}(Q_2W)\big) - H(Y|X)\\
&\overset{\text{(iv)}}{>} \tfrac{1}{2}H\big((Q_1W)\big) + \tfrac{1}{2}H\big((Q_2W)\big) - H(Y|X)\\
&= \tfrac{1}{2}H\big((Q_1W)\big) + \tfrac{1}{2}H\big((Q_2W)\big) - \sum_x Q(x)H(Y|X=x)\\
&\overset{\text{(v)}}{=} \tfrac{1}{2}H\big((Q_1W)\big) + \tfrac{1}{2}H\big((Q_2W)\big) - \sum_x \big(\tfrac{1}{2}Q_1(x) + \tfrac{1}{2}Q_2(x)\big)H(Y|X=x)\\
&= \frac{1}{2}\left[H\big((Q_1W)\big) - \sum_x Q_1(x)H(Y|X=x)\right] + \frac{1}{2}\left[H\big((Q_2W)\big) - \sum_x Q_2(x)H(Y|X=x)\right]\\
&= \tfrac{1}{2}I(Q_1,W) + \tfrac{1}{2}I(Q_2,W)\\
&\overset{\text{(vi)}}{=} \mathsf{C},
\end{aligned}$$

where (i) follows from the definition of capacity; (ii) holds because $P_{XY}(x,y) = Q(x)\,W(y|x)$, so $P_Y = (QW)$; (iii) holds because the vector-matrix product is linear, so $\big(\big(\frac{1}{2}Q_1 + \frac{1}{2}Q_2\big)W\big) = \frac{1}{2}(Q_1W) + \frac{1}{2}(Q_2W)$; (iv) follows from Lemma 1 because $(Q_1W) \neq (Q_2W)$ by assumption; (v) holds because $H(Y|X=x)$ depends only on $W(y|x)$; and (vi) holds because Q_1 and Q_2 are capacity-achieving by assumption. But $\mathsf{C} > \mathsf{C}$ is not possible, so we conclude that all capacity-achieving input distributions induce the same output distribution.

Problem 3 *The Binary Jammer Channel*

a) If the encoder and decoder do not know when the channel is blocked, the interference simply increases the channel crossover probability. If the channel is free, which occurs with probability q, a transmitted 0 is received as a 1 with probability ϵ. If the channel is blocked, which occurs with probability $1-q$, a transmitted 0 is received as a 1 with probability 1/2. The combined probability that a 0 is received as a 1 is therefore $q\epsilon + \frac{1-q}{2}$, and the channel capacity is

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

b) The channel is equivalent to the binary symmetric erasure channel depicted in Figure 1, where the output Y is equal to ? if the channel was blocked. (If the channel is blocked, the output bit is independent of the input bit, so we can ignore the output bit in that case.) The capacity of the channel is

$$\mathsf{C} = q\big(1 - H_{\mathrm{b}}(\epsilon)\big),$$

which can be obtained for example from Exercise 8, Problem 5, Part a).

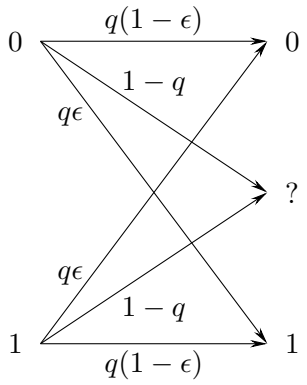


Figure 1: Binary symmetric erasure channel.

c) We first construct a good codebook of length qn for the original BSC (without interference). We know that for large enough n, this codebook can be constructed to transmit up to $2^{nq(1-H_\mathrm{b}(\epsilon))}$ messages. Next, for any $\delta > 0$, consider using the jammed BSC $(1+\delta)n$ times. For large n, with high probability the number of unblocked channel uses is close to $(1+\delta)qn$ and is therefore larger than qn. Our coding scheme is to keep sending each letter of the codeword for the original BSC until an interference free channel use occurs, and then move on to the next letter of the codeword. If more than nq interference-free channel uses occur, the encoder pads with an arbitrary input symbol. Accordingly, the decoder ignores the output when the channel is jammed and uses the decoder for the original code on the first qn remaining output symbols. (Of course, if there are less than nq channel uses without interference, an error occurs. But this happens with small probability.) This is equivalent to decoding on the original BSC without interference, and therefore with high probability the decoder will make the correct decision. Thus, with high probability, we can reliably transmit information at any rate up to $q(1-H_\mathrm{b}(\epsilon))/(1+\delta)$, for any $\delta > 0$. This implies that we can transmit at any rate up to $q(1-H_\mathrm{b}(\epsilon))$.

It is also possible to use Part b) to show that any rate smaller than $q(1-H_\mathrm{b}(\epsilon))$ is achievable: the encoder can simply ignore whether the channel is blocked or not and use a coding scheme for the situation in Part b); this works because the capacity in Part b) is $q(1-H_\mathrm{b}(\epsilon))$ bits per channel use.

Problem 4 *Typical Decoding vs. Maximum-Likelihood Decoding*

Denote the message set by $\mathcal{M} = \{1, \ldots, 2^{n\mathsf{R}}\}$, and denote the codeword for message $m \in \mathcal{M}$ by $x^n(m)$.

a) Remember that $\mathcal{A}_\epsilon^{(n)}(P_{XY})$ is the set of all $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying:

(A) $2^{-n(H(P_X)+\epsilon)} \le \prod_{i=1}^{n} P_X(x_i) \le 2^{-n(H(P_X)-\epsilon)}$;

(B) $2^{-n(H(P_Y)+\epsilon)} \le \prod_{i=1}^{n} P_Y(y_i) \le 2^{-n(H(P_Y)-\epsilon)}$; and

(C) $2^{-n(H(P_{XY})+\epsilon)} \le \prod_{i=1}^{n} P_{XY}(x_i, y_i) \le 2^{-n(H(P_{XY})-\epsilon)}$.

Upon receiving y^n, the weak typicality decoder checks for every $\tilde{m} \in \mathcal{M}$ whether $(x^n(\tilde{m}), y^n)$ is in $\mathcal{A}_\epsilon^{(n)}(P_{XY})$. If there exists exactly one $\tilde{m}$ such that $x^n(\tilde{m})$ is jointly typical with y^n, then the decoder declares $\hat{m} = \tilde{m}$, otherwise it declares an error.

We know analyze $\mathcal{A}_\epsilon^{(n)}(P_{XY})$. We have $P_{XY}(x,y) = P_X(x)\,W(y|x)$, where $W(y|x)$ denotes the channel transitions of a BSC with crossover probability α, so

$P_{XY}(x,y)$	$y=0$	$y=1$
$x=0$	$\frac{1}{2}(1-\alpha)$	$\frac{1}{2}\alpha$
$x=1$	$\frac{1}{2}\alpha$	$\frac{1}{2}(1-\alpha)$

We have $P_X(0) = P_X(1) = \frac{1}{2}$, so $H(P_X) = 1$ bit and (A) is satisfied for all $x^n \in \mathcal{X}^n$. We have $P_Y(0) = P_Y(1) = \frac{1}{2}$, so $H(P_Y) = 1$ bit and (B) is satisfied for all $y^n \in \mathcal{Y}^n$. We also have $H(P_{XY}) = H(X) + H(Y|X) = 1 + H_\mathrm{b}(\alpha)$. Taking logarithms and dividing by $-n$ shows that (C) is satisfied if and only if

$$H(P_{XY}) - \epsilon \le -\frac{1}{n}\log\prod_{i=1}^n P_{XY}(x_i,y_i) \le H(P_{XY}) + \epsilon,$$

which is equivalent to

$$\left| -\frac{1}{n}\log\prod_{i=1}^n P_{XY}(x_i,y_i) - H(P_{XY}) \right| \le \epsilon. \tag{1}$$

Fix x^n and y^n, and let d denote the Hamming distance between x^n and y^n (the number of discrepancies between x^n and y^n). Then,

$$\begin{aligned}
\prod_{i=1}^n P_{XY}(x_i,y_i) &= \prod_{i=1}^n \big(P_X(x_i) P_{Y|X}(y_i|x_i)\big) \\
&\overset{\text{(i)}}{=} 2^{-n} \prod_{i=1}^n P_{Y|X}(y_i|x_i) \\
&\overset{\text{(ii)}}{=} 2^{-n}\,\alpha^d (1-\alpha)^{n-d},
\end{aligned}$$

where (i) holds because $P_X(0) = P_X(1) = \frac{1}{2}$; and (ii) follows from the definition of d because the crossover probability is α. Observe that

$$\begin{aligned}
-\frac{1}{n}\log\prod_{i=1}^n P_{XY}(x_i,y_i) - H(P_{XY}) &= 1 + \frac{d}{n}\log\frac{1}{\alpha} + \frac{n-d}{n}\log\frac{1}{1-\alpha} - 1 - H_\mathrm{b}(\alpha) \\
&= \frac{d}{n}\log\frac{1}{\alpha} + \frac{n-d}{n}\log\frac{1}{1-\alpha} - \alpha\log\frac{1}{\alpha} - (1-\alpha)\log\frac{1}{1-\alpha} \\
&= \left(\frac{d}{n} - \alpha\right)\log\frac{1}{\alpha} - \left(\frac{d}{n} - \alpha\right)\log\frac{1}{1-\alpha} \\
&= \left(\frac{d}{n} - \alpha\right)\log\frac{1-\alpha}{\alpha}.
\end{aligned}$$

Combining this with (1), we obtain that (x^n, y^n) is in $\mathcal{A}_\epsilon^{(n)}(P_{XY})$ if and only if

$$\left| \left(\frac{d}{n} - \alpha\right)\log\frac{1-\alpha}{\alpha} \right| \le \epsilon,$$

which is equivalent to

$$\alpha - \frac{\epsilon}{\left|\log \frac{1-\alpha}{\alpha}\right|} \le \frac{d}{n} \le \alpha + \frac{\epsilon}{\left|\log \frac{1-\alpha}{\alpha}\right|}. \tag{2}$$

(This makes sense because the expected number of channel crossovers divided by n is α.) Feasible values of $\frac{d}{n}$ are depicted in Figure 2.

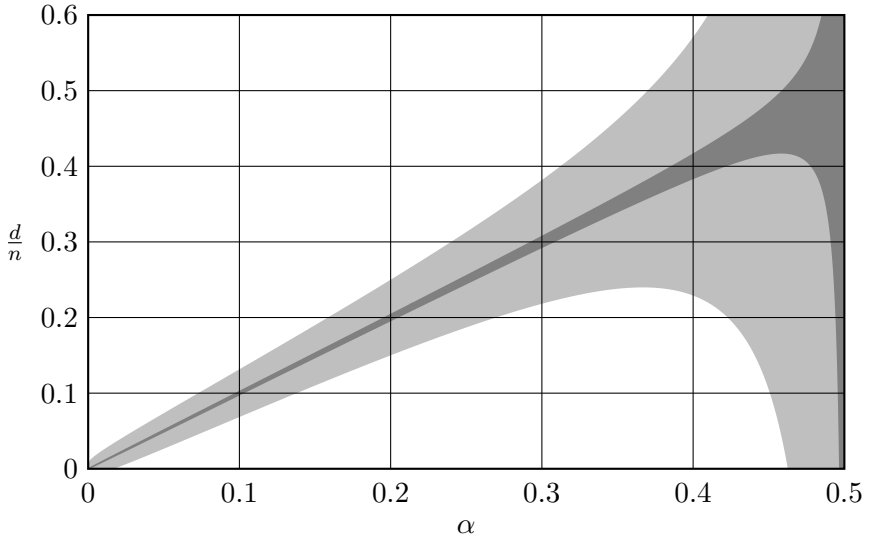


Figure 2: Values of $\frac{d}{n}$ satisfying (2) as a function of α for $\epsilon = 0.1$ (light gray) and $\epsilon = 0.01$ (gray).

b) The maximum likelihood decoder also has a description in terms of d, the number of discrepancies between a codeword and the received signal. Let d_m equal the number of discrepancies between the received signal y^n and the codeword $x^n(m)$.

$$\begin{aligned}
\text{ML-decoder}(y^n) &\triangleq \operatorname{carg\,max}_{\tilde{m}} P(y^n|x^n(\tilde{m})) \\
&= \operatorname{carg\,max}_{\tilde{m}} \alpha^{d_{\tilde{m}}}(1-\alpha)^{n-d_{\tilde{m}}} \\
&= \operatorname{carg\,max}_{\tilde{m}} (1-\alpha)^n \left(\frac{\alpha}{1-\alpha}\right)^{d_{\tilde{m}}},
\end{aligned}$$

where carg max stands for "choosing-arg max", i.e., if there are several $\tilde{m}$ that achieve the maximum, then one of the possible solutions is chosen randomly. If $\alpha < \frac{1}{2}$ then the maximum is attained by the $x^n(\tilde{m})$ with the smallest discrepancy $d_{\tilde{m}}$, and if $\alpha > \frac{1}{2}$ the maximum is attained by the $x^n(\tilde{m})$ with the largest discrepancy $d_{\tilde{m}}$. Note that when $\alpha = \frac{1}{2}$, then all $\tilde{m}$ achieve the maximum and the ML decoder will pick a decoding randomly among all $\tilde{m}$.

c) Since the priors are uniform, the maximum likelihood decoder minimizes the probability of error and thus always leads to a lower average probability of error.