Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 9 of November 16, 2016

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

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

a) We use a majority decoder. So the communications system consists of a two-codeword repetition code, a BSC with $p_{\text{orig}} = 0.1$ and a majority decoder. The probability of making a decoding error given that 000 has been transmitted is thus

$$\begin{aligned}\Pr(\text{error}|000) &= \Pr(\text{received word contains two or three one's}|000)\\ &= 0.1^3 + \binom{3}{2}0.1^2 0.9^1\\ &= 0.028.\end{aligned}$$

For symmetry reasons $\Pr(\text{error}|111)$ is the same. Hence the overall channel is a BSC with crossover probability $p_{\text{new}} = 0.028$, which is smaller than $p_{\text{orig}} = 0.1$.

b) The capacity of the of the new channel is

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

However, one use of this channel corresponds to three uses of the original channel. Therefore the capacity is

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

c) The capacity of the original channel is

$$\mathsf{C}_{\text{orig}} = 1 - H_{\text{b}}(p_{\text{orig}}) \approx 0.531 \text{ bits per use of the original channel,}$$

which is more than 0.272 bits/use.

d) Think of the whole communication system (encoder, channel, decoder) as a new channel with input M and output $\hat{M}$. The encoder maps each input M onto a codeword X^n and sends it through the channel. The decoder maps each received sequence Y^n onto a decision $\hat{M}$. This new channel is also a DMC. We have

$$\begin{aligned}I(M;\hat{M}) &\le I(X^n;Y^n)\\ &\le \sum_{k=1}^{n} I(X_k;Y_k)\end{aligned}$$

$$\leq n\mathsf{C}_{\text{orig}},$$

where the first inequality follows from the data processing inequality (encoder and decoder are deterministic mappings), the second from the fact that the original channel is memoryless and used without feedback, and the last from the definition of C_{orig}. Hence,

$$\frac{I(M;\hat{M})}{n} \leq \mathsf{C}_{\text{orig}},$$

for any input M, hence also for the maximum achieving M.

Problem 2 *Nonuniqueness of Capacity-Achieving Input Distributions*

a) A trivial example is a channel with capacity zero such as the binary symmetric channel with transition probability $\epsilon = 1/2$. Obviously, any input distribution achieves the capacity of this channel.

For a more interesting example, consider a channel with law

$$\mathsf{W} = \begin{pmatrix} 1-\epsilon & \epsilon \\ 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix},$$

where the entry in row i and column j denotes $W(y_j|x_i)$.

For this channel we get for instance the following two different capacity-achieving input distributions:

$$Q_1 = \left(\frac{1}{2}, 0, \frac{1}{2}\right),$$
$$Q_2 = \left(0, \frac{1}{2}, \frac{1}{2}\right).$$

The capacity is $\mathsf{C} = 1 - H_{\text{b}}(\epsilon)$.

b) We first prove that entropy is strictly concave.

Lemma 1. *The entropy $H(P)$ is strictly concave in P, i.e.,*

$$H(\alpha P_1 + \bar{\alpha} P_2) \geq \alpha H(P_1) + \bar{\alpha} H(P_2), \quad \forall \alpha \in (0,1), \quad \bar{\alpha} \triangleq 1 - \alpha,$$

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

Proof. We introduce a chance variable Z such that $\Pr[Z=1] = \alpha$, $\Pr[Z=2] = \bar{\alpha}$. Let X_1 have distribution P_1, and let X_2 have distribution P_2. Then the chance variable X_Z has distribution $\alpha P_1 + \bar{\alpha} P_2$. We have

$$H(X_Z|Z) = \alpha H(P_1) + \bar{\alpha} H(P_2), \tag{1}$$
$$H(X_Z) = H(\alpha P_1 + \bar{\alpha} P_2). \tag{2}$$

Furthermore,

$$H(X_Z) \geq H(X_Z|Z),$$

with equality if, and only if, X_Z does not depend on Z, i.e., $P_1 = P_2$. This together with (1) and (2) concludes our proof. □

Assume that for a DMC the two input distributions P_1 and P_2 both achieve capacity C. Since $I(P,W)$ is concave in P, any convex combination of P_1 and P_2 should achieve capacity as well. Thus

$$I(\alpha P_1 + \bar{\alpha} P_2, W) = \alpha I(P_1, W) + \bar{\alpha} I(P_2, W) = \mathsf{C}, \quad \forall \alpha \in (0,1). \tag{3}$$

Let P_1W and P_2W denote the probability distributions on the channel output that result when the channel input is distributed according to P_1 and P_2, respectively, i.e.,

$$(P_1W)(y) = \sum_{x\in\mathcal{X}} P_1(x)W(y|x), \quad y \in \mathcal{Y} \qquad \text{and} \qquad (P_2W)(y) = \sum_{x\in\mathcal{X}} P_2(x)W(y|x), \quad y \in \mathcal{Y}.$$

Writing $I(X;Y)$ as

$$I(X;Y) = H(Y) - H(Y|X)$$

the left-hand side of (3) becomes

$$\begin{aligned}
&I(\alpha P_1 + \bar{\alpha} P_2, W) \\
&\quad = H(\alpha P_1 W + \bar{\alpha} P_2 W) - \alpha \sum_{x\in\mathcal{X}} P_1(x) H(Y|X=x) - \bar{\alpha} \sum_{x\in\mathcal{X}} P_2(x) H(Y|X=x) \\
&\quad \geq \alpha H(P_1W) + \bar{\alpha} H(P_2W) - \alpha \sum_{x\in\mathcal{X}} P_1(x) H(Y|X=x) - \bar{\alpha} \sum_{x\in\mathcal{X}} P_2(x) H(Y|X=x) \\
&\quad = \alpha \left(H(P_1W) - \sum_{x\in\mathcal{X}} P_1(x) H(Y|X=x) \right) + \bar{\alpha} \left(H(P_2W) - \sum_{x\in\mathcal{X}} P_2(x) H(Y|X=x) \right) \\
&\quad = \alpha I(P_1, W) + \bar{\alpha} I(P_2, W),
\end{aligned}$$

where the inequality follows from Lemma 1.

But from (3) it follows that the inequality has to hold with equality, and since $H(Y)$ is strictly concave in the distribution of Y this implies that $P_1W = P_2W$. Thus, the output distribution induced by a capacity-achieving input distribution is unique.

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) In this case, a channel blockage is equivalent to an erasure: the equivalent channel model is a binary symmetric erasure channel with erasure probability $1-q$ and crossover probability (conditioned on no erasure) equal to ϵ. See Figure 1. The output Y of the binary symmetric erasure channel takes on three possible values 0, 1, and ?. One can guess that the capacity of this channel is $q(1 - H_{\mathrm{b}}(\epsilon))$. To prove this, let B be a binary random variable denoting the state of the channel, i.e., $B = 0$ corresponds to a free channel, and $B = 1$ corresponds to a blocked channel. Note that B is known to the receiver, but not to the transmitter. Then

$$\mathsf{C} = \max I(X; Y, B) \tag{4}$$

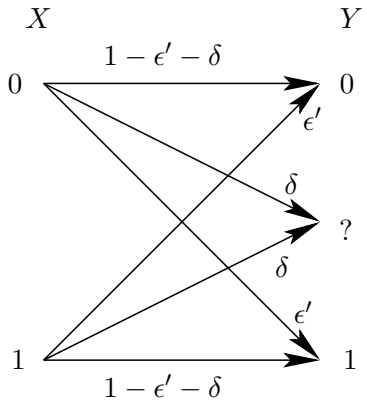


Figure 1: Binary symmetric erasure channel.

$$= \max I(X;B) + I(X;Y|B) \tag{5}$$
$$= \max I(X;Y|B) \tag{6}$$
$$= \max \Big\{(1-q)I(X;Y|B=1) + qI(X;Y|B=0)\Big\}$$
$$= \max qI(X;Y|B=0) \tag{7}$$
$$= q(1-H_{\mathrm{b}}(\epsilon)) \tag{8}$$

where (4) follows because B is a function of Y; (5) follows from the chain rule; equation (6) follows because the encoder does not know when the channel is blocked and hence X and B are independent of each other; equality in (7) holds because when the channel is blocked, then X and Y are independent; and (8) holds because the unblocked channel is a binary symmetric channel with crossover probability ϵ.

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 $q(1-H_{\mathrm{b}}(\epsilon))$.

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

Let the BSC have crossover probability α. Let the message set be $\mathcal{M} = \{1, \ldots, 2^{n\mathrm{R}}\}$. And let $\mathbf{X}(m) \in \mathcal{X}^n$ for $m = 1, \ldots, 2^{n\mathrm{R}}$ be the codewords that the elements of $\mathcal{M}$ are mapped to.
For a BSC the optimal input distribution is the fair coin distribution. This in turn induces a fair coin distribution on the output. That is $H(X) = 1$, $H(Y) = 1$, and $H(X,Y) = H(X) + H(Y|X) = 1 + H_{\mathrm{b}}(\alpha)$.

a) Define

$$\mathcal{A}_\epsilon^{(n)} = \Bigg\{(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n \colon \left|-\frac{1}{n}\log P(\mathbf{x}) - 1\right| < \epsilon, \\ \left|-\frac{1}{n}\log P(\mathbf{y}) - 1\right| < \epsilon, \\ \left|-\frac{1}{n}\log P(\mathbf{x},\mathbf{y}) - 1 - H_{\mathrm{b}}(\alpha)\right| < \epsilon\Bigg\}.$$

Upon receiving $\mathbf{y}$, the joint typicality decoder checks for every $\tilde{m}$ whether $(\mathbf{X}(\tilde{m}), \mathbf{y})$ is in $\mathcal{A}_\epsilon^{(n)}$. If there exists one, and only one, $\tilde{m}$ such that $\mathbf{X}(\tilde{m})$ is jointly typical with $\mathbf{y}$, then the decoder declares $\hat{m} = \tilde{m}$, else it declares an error.

Now let us describe the joint typicality decoder more precisely. Note that $-\frac{1}{n}\log P(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}^n$ because $P(\mathbf{x}) = \prod_{i=1}^n P_X(x_i) = \left(\frac{1}{2}\right)^n$ is the fair coin distribution, and analogously $-\frac{1}{n}\log P(\mathbf{y}) = 1$ for all $\mathbf{y} \in \mathcal{Y}^n$. Thus, to check whether a given $(\mathbf{x},\mathbf{y})$ pair is in $\mathcal{A}_\epsilon^{(n)}$ we only need to check whether $\left| -\frac{1}{n}\log P(\mathbf{x},\mathbf{y}) - 1 - H_{\mathrm{b}}(\alpha)\right| < \epsilon$.

We can simplify this by noting that

$$\begin{aligned}-\frac{1}{n}\log P(\mathbf{x},\mathbf{y}) - 1 - H_{\mathrm{b}}(\alpha) &= -\frac{1}{n}\log P(\mathbf{x}) - \frac{1}{n}\log P(\mathbf{y}|\mathbf{x}) - 1 - H_{\mathrm{b}}(\alpha) \\ &= -\frac{1}{n}\log P(\mathbf{y}|\mathbf{x}) - H_{\mathrm{b}}(\alpha).\end{aligned}$$

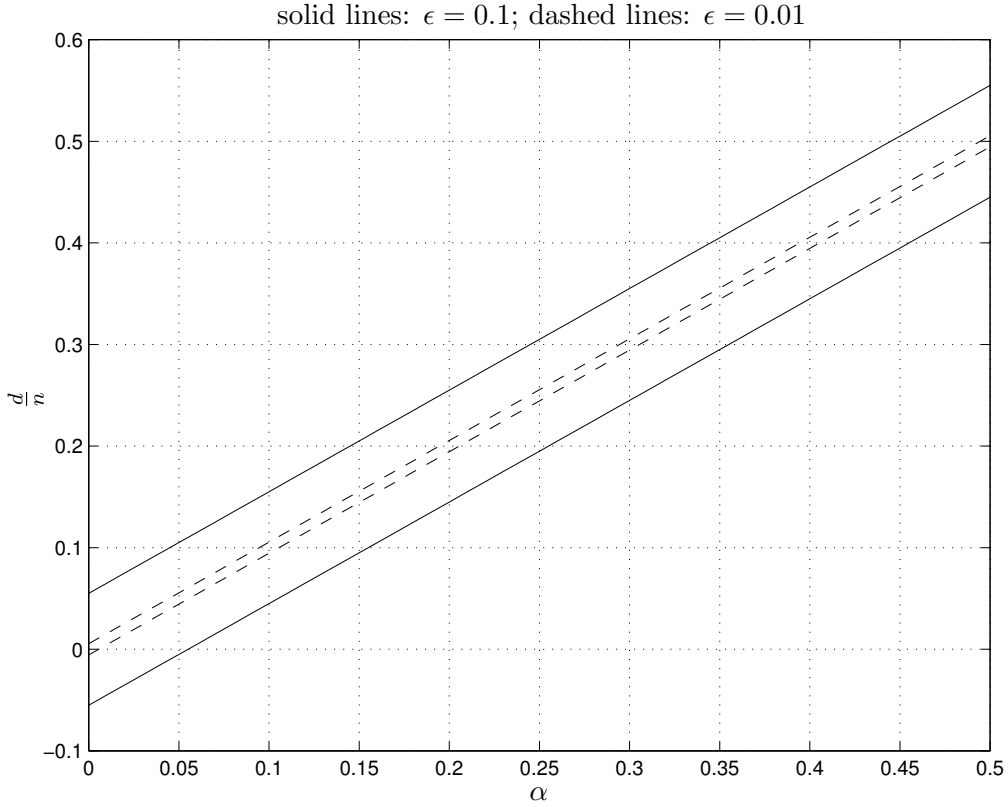
Let d equal the number of discrepancies between $\mathbf{x}$ and $\mathbf{y}$ (Hamming distance). Then

$$\begin{aligned}-\frac{1}{n}\log P(\mathbf{y}|\mathbf{x}) - H_{\mathrm{b}}(\alpha) &= -\frac{1}{n}\log\Big(\alpha^d(1-\alpha)^{n-d}\Big) - H_{\mathrm{b}}(\alpha) \\ &= -\frac{d}{n}\log\alpha - \frac{n-d}{n}\log(1-\alpha) + \alpha\log\alpha + (1-\alpha)\log(1-\alpha) \\ &= \left(\alpha - \frac{d}{n}\right)\log\alpha + \left(1 - \alpha - \frac{n-d}{n}\right)\log(1-\alpha) \\ &= \left(\alpha - \frac{d}{n}\right)\log\frac{\alpha}{1-\alpha}.\end{aligned}$$

Hence, to determine whether $(\mathbf{x},\mathbf{y}) \in \mathcal{A}_\epsilon^{(n)}$ we only need to check that the number of discrepancies d is in the open interval

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

The interval is shown in the following figure for two different values of ϵ:



Upon receiving $\mathbf{y}$ we compute the discrepancies with respect to $\mathbf{X}(1), \ldots, \mathbf{X}(2^{n\mathsf{R}})$. If there is one, and only one, $\mathbf{X}(\tilde{m})$ with discrepancy in the above interval then we declare that the transmitted message was $\tilde{m}$, else we declare an error. Note that if $\alpha = \frac{1}{2}$ we will only declare errors.

b) The maximum likelihood decoder also has a simple 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 $\mathbf{y}$ and the codeword $\mathbf{X}(m)$.

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

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 $\mathbf{X}(\tilde{m})$ with the smallest discrepancy $d_{\tilde{m}}$, and if $\alpha > \frac{1}{2}$ the maximum is attained by the $\mathbf{X}(i)$ 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.

