Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 12 of December 7, 2016

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

Problem 1 — *Rate Distortion with Two Distortion Functions*

a) We define

$$\mathsf{R}(\mathsf{D}_1,\mathsf{D}_2) \triangleq \min_{\substack{P_{\hat{X}|X}(\cdot|\cdot):\\ \mathsf{E}[d_1(X,\hat{X})]\le \mathsf{D}_1,\\ \mathsf{E}[d_2(X,\hat{X})]\le \mathsf{D}_2}} I\big(X;\hat{X}\big) \tag{1}$$

and we will show that $\mathsf{R}(\mathsf{D}_1,\mathsf{D}_2)$ characterizes the region of achievable tuples $(\mathsf{R},\mathsf{D}_1,\mathsf{D}_2)$. We first prove two properties of $\mathsf{R}(\mathsf{D}_1,\mathsf{D}_2)$:

Lemma 1. *The function* $\mathsf{R}(\mathsf{D}_1,\mathsf{D}_2)$ *is nonincreasing in the pair* $(\mathsf{D}_1,\mathsf{D}_2)$*, i.e.,*

$$\mathsf{R}(\mathsf{D}_1',\mathsf{D}_2') \le \mathsf{R}(\mathsf{D}_1,\mathsf{D}_2)$$

if $\mathsf{D}_1' \ge \mathsf{D}_1$ *and* $\mathsf{D}_2' \ge \mathsf{D}_2$.

Proof. Let $P^*_{\hat{X}|X}$ be a conditional PMF that minimizes the RHS of (1). For this PMF,

$$\mathsf{E}[d_1(X,\hat{X})] \le \mathsf{D}_1 \le \mathsf{D}_1',$$
$$\mathsf{E}[d_2(X,\hat{X})] \le \mathsf{D}_2 \le \mathsf{D}_2',$$

therefore

$$\mathsf{R}(\mathsf{D}_1',\mathsf{D}_2') = \min_{\substack{P_{\hat{X}|X}(\cdot|\cdot):\\ \mathsf{E}[d_1(X,\hat{X})]\le \mathsf{D}_1',\\ \mathsf{E}[d_2(X,\hat{X})]\le \mathsf{D}_2'}} I\big(X;\hat{X}\big) \le I\big(X;\hat{X}\big)\Big|_{P_{\hat{X}|X}=P^*_{\hat{X}|X}} = \mathsf{R}(\mathsf{D}_1,\mathsf{D}_2).$$

□

Lemma 2. *The function* $\mathsf{R}(\mathsf{D}_1,\mathsf{D}_2)$ *is convex in the pair* $(\mathsf{D}_1,\mathsf{D}_2)$*, i.e.,*

$$\mathsf{R}(\lambda\mathsf{D}_1^{(1)}+\bar{\lambda}\mathsf{D}_1^{(2)},\lambda\mathsf{D}_2^{(1)}+\bar{\lambda}\mathsf{D}_2^{(2)}) \le \lambda\mathsf{R}(\mathsf{D}_1^{(1)},\mathsf{D}_2^{(1)})+\bar{\lambda}\mathsf{R}(\mathsf{D}_1^{(2)},\mathsf{D}_2^{(2)})$$

for $\lambda\in[0,1]$ *and* $\bar{\lambda}=1-\lambda$.

Proof. Let $P^{(1)}_{\hat{X}|X}$ be a PMF that minimizes the RHS of (1) for the pair $(\mathsf{D}^{(1)}_1, \mathsf{D}^{(1)}_2)$ and let $P^{(2)}_{\hat{X}|X}$ be a PMF that minimizes the RHS of (1) for the pair $(\mathsf{D}^{(2)}_1, \mathsf{D}^{(2)}_2)$. Note that the mixture $\lambda P^{(1)}_{\hat{X}|X} + \overline{\lambda} P^{(2)}_{\hat{X}|X}$ is a conditional PMF satisfying

$$\begin{aligned} \mathsf{E}[d_1(X, \hat{X})] &\leq \lambda \mathsf{D}^{(1)}_1 + \overline{\lambda} \mathsf{D}^{(2)}_1, \\ \mathsf{E}[d_2(X, \hat{X})] &\leq \lambda \mathsf{D}^{(1)}_2 + \overline{\lambda} \mathsf{D}^{(2)}_2. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathsf{R}(\lambda \mathsf{D}^{(1)}_1 + \overline{\lambda} \mathsf{D}^{(2)}_1, \lambda \mathsf{D}^{(1)}_2 + \overline{\lambda} \mathsf{D}^{(2)}_2) &= \min_{\substack{P_{\hat{X}|X}(\cdot|\cdot): \\ \mathsf{E}[d_1(X,\hat{X})] \leq \lambda \mathsf{D}^{(1)}_1 + \overline{\lambda} \mathsf{D}^{(2)}_1, \\ \mathsf{E}[d_2(X,\hat{X})] \leq \lambda \mathsf{D}^{(1)}_2 + \overline{\lambda} \mathsf{D}^{(2)}_2}} I\Big(P_X, P_{\hat{X}|X}\Big) \\ &\leq I\Big(P_X, \lambda P^{(1)}_{\hat{X}|X} + \overline{\lambda} P^{(2)}_{\hat{X}|X}\Big) \\ &\stackrel{\text{(i)}}{\leq} \lambda I\Big(P_X, P^{(1)}_{\hat{X}|X}\Big) + \overline{\lambda} I\Big(P_X, P^{(2)}_{\hat{X}|X}\Big) \\ &= \lambda \mathsf{R}(\mathsf{D}^{(1)}_1, \mathsf{D}^{(1)}_2) + \overline{\lambda} \mathsf{R}(\mathsf{D}^{(2)}_1, \mathsf{D}^{(2)}_2), \end{aligned}$$

where (i) follows because $I(Q, W)$ is convex in W. □

We are now ready to prove the converse, i.e., we show that $\mathsf{R} \geq \mathsf{R}(\mathsf{D}_1, \mathsf{D}_2)$ must hold for every encoder and decoder pair satisfying

$$\mathsf{E}\big[d_1\big(X^n, \hat{X}^n\big)\big] \leq \mathsf{D}_1, \tag{2}$$

$$\mathsf{E}\big[d_2\big(X^n, \hat{X}^n\big)\big] \leq \mathsf{D}_2. \tag{3}$$

This is true because

$$\begin{aligned} \mathsf{R}\big(\mathsf{D}_1, \mathsf{D}_2\big) &\stackrel{\text{(i)}}{\leq} \mathsf{R}\big(\mathsf{E}\big[d_1\big(X^n, \hat{X}^n\big)\big], \mathsf{E}\big[d_2\big(X^n, \hat{X}^n\big)\big]\big) \\ &\stackrel{\text{(ii)}}{=} \mathsf{R}\left(\sum_{i=1}^n \frac{1}{n} \mathsf{E}\big[d_1\big(X_i, \hat{X}_i\big)\big], \sum_{i=1}^n \frac{1}{n} \mathsf{E}\big[d_2\big(X_i, \hat{X}_i\big)\big]\right) \\ &\stackrel{\text{(iii)}}{\leq} \sum_{i=1}^n \frac{1}{n} \mathsf{R}\big(\mathsf{E}\big[d_1\big(X_i, \hat{X}_i\big)\big], \mathsf{E}\big[d_2\big(X_i, \hat{X}_i\big)\big]\big) \\ &\stackrel{\text{(iv)}}{\leq} \sum_{i=1}^n \frac{1}{n} I\big(X_i; \hat{X}_i\big) \\ &= \sum_{i=1}^n \frac{1}{n} \Big[H\big(X_i\big) - H\big(X_i|\hat{X}_i\big)\Big] \\ &\stackrel{\text{(v)}}{=} \frac{1}{n} H\big(X^n\big) - \frac{1}{n} \sum_{i=1}^n H\big(X_i|\hat{X}_i\big) \\ &\stackrel{\text{(vi)}}{\leq} \frac{1}{n} H\big(X^n\big) - \frac{1}{n} \sum_{i=1}^n H\big(X_i|\hat{X}^n, X^{i-1}\big) \\ &\stackrel{\text{(vii)}}{=} \frac{1}{n} H\big(X^n\big) - \frac{1}{n} H\big(X^n|\hat{X}^n\big) \\ &= \frac{1}{n} I\big(X^n; \hat{X}^n\big) \\ &\stackrel{\text{(viii)}}{\leq} \mathsf{R}, \end{aligned}$$

where (i) follows from (2), (3), and the monotonicity of $\mathsf{R}(\mathsf{D}_1, \mathsf{D}_2)$; (ii) follows because we have a single-letter distortion measure (and because expectation is linear); (iii) follows from Jensen's inequality and the convexity of $\mathsf{R}(\mathsf{D}_1, \mathsf{D}_2)$ in the pair $(\mathsf{D}_1, \mathsf{D}_2)$; (iv) follows from

$$\mathsf{R}\big(\mathsf{E}\big[d_1\big(X_i, \hat{X}_i\big)\big], \mathsf{E}\big[d_2\big(X_i, \hat{X}_i\big)\big]\big) = \min_{\substack{P_{\hat{X}|X}(\cdot|\cdot):\\ \mathsf{E}[d_1(X,\hat{X})] \le \mathsf{E}[d_1(X_i,\hat{X}_i)],\\ \mathsf{E}[d_2(X,\hat{X})] \le \mathsf{E}[d_2(X_i,\hat{X}_i)]}} I\big(P_X, P_{\hat{X}|X}\big) \le I\big(P_{X_i}, P_{\hat{X}_i|X_i}\big)$$

because $P_X = P_{X_i}$ and because $P_{\hat{X}_i|X_i}$ satisfies the constraints on the expected distortion; (v) follows because $X_1, \dots, X_n$ are drawn IID; (vi) follows because conditioning does not increase entropy; (vii) follows from the chain rule; and (viii) follows from the data processing inequality and the fact that we only have $2^{n\mathsf{R}}$ codewords, so $I\big(X^n; \hat{X}^n\big) \le n\mathsf{R}$.

We prove the direct part as follows. Fix a conditional PMF $P^*_{\hat{X}|X}$ that achieves the minimum in the RHS of (1) and compute the induced marginal distribution on $\hat{X}$,

$$P_{\hat{X}}(\hat{x}) = \sum_x P_X(x) P^*_{\hat{X}|X}(\hat{x}|x), \quad \hat{x} \in \hat{\mathcal{X}}.$$

Fix $\mathsf{R} > \mathsf{R}(\mathsf{D}_1, \mathsf{D}_2)$ and generate a codebook of size $\lceil 2^{n\mathsf{R}} \rceil$ at random by drawing each symbol of each codeword IID according to the marginal $P_{\hat{X}}$. To describe a source sequence x^n, the encoder tries to find a codeword $\hat{x}^n(i)$ such that $\big(x^n, \hat{x}^n(i)\big) \in \mathcal{T}^{(n)}_{\epsilon'}(P_{X,\hat{X}})$ and then sends the smallest such index i. When observing i, the reconstructor produces $\hat{x}^n(i)$. Introduce a chance variable S that is one if the encoder was able to find at least one codeword $\hat{x}^n(i)$ such that $\big(x^n, \hat{x}^n(i)\big) \in \mathcal{T}^{(n)}_{\epsilon'}(P_{X,\hat{X}})$, and zero otherwise. The expected distortion for the random coding scheme can be bounded as

$$\begin{aligned}\mathsf{E}\big[d_1\big(X^n, \hat{X}^n\big)\big] &\overset{\text{(i)}}{=} \Pr\big[S = 0\big] \underbrace{\mathsf{E}\big[d_1\big(X^n, \hat{X}^n\big)|S = 0\big]}_{\le d_{1,\max}} + \underbrace{\Pr\big[S = 1\big]}_{\le 1} \underbrace{\mathsf{E}\big[d_1\big(X^n, \hat{X}^n\big)|S = 1\big]}_{\le \mathsf{D}_1(1+\epsilon')} \\ &\le \Pr\big[S = 0\big] \cdot d_{1,\max} + \mathsf{D}_1(1 + \epsilon') \end{aligned} \tag{4}$$

where (i) follows from the law of total expectation; the first underbraced inequality follows because we assume bounded distortions; the second underbraced inequality is trivial; the third underbraced inequality follows because $S = 1$ implies that there exists a codeword $\hat{x}^n(i)$ such that $\big(x^n, \hat{x}^n(i)\big) \in \mathcal{T}^{(n)}_{\epsilon'}(P_{X,\hat{X}})$; and for that codeword, we have $d_1\big(x^n, \hat{x}^n\big) \le \mathsf{D}_1(1 + \epsilon')$ as a consequence of strong typicality because $P^*_{\hat{X}|X}$ satisfies $\mathsf{E}[d_1(X, \hat{X})] \le \mathsf{D}_1$. Similarly, we obtain

$$\mathsf{E}\big[d_2\big(X^n, \hat{X}^n\big)\big] \le \Pr\big[S = 0\big] \cdot d_{2,\max} + \mathsf{D}_2(1 + \epsilon'). \tag{5}$$

From the lecture we know that for $\mathsf{R} > \mathsf{R}(\mathsf{D}_1, \mathsf{D}_2) = I\big(P_X, P^*_{\hat{X}|X}\big)$ and ϵ' small enough, $\Pr[S = 0]$ tends to zero as n tends to infinity. Therefore, for an appropriate choice of ϵ' and n large enough, we obtain $\mathsf{E}\big[d_1\big(X^n, \hat{X}^n\big)\big] \le \mathsf{D}_1 + \epsilon$ and $\mathsf{E}\big[d_2\big(X^n, \hat{X}^n\big)\big] \le \mathsf{D}_2 + \epsilon$ from (4) and (5), respectively. Finally, we argue that we can get the same performance with a deterministic codebook. Note that (4) and (5) only depend on $\Pr[S = 0]$. By the law of total probability, we have

$$\Pr[S = 0] = \sum_{\mathcal{C}} P(\mathcal{C}) \Pr[S = 0|\mathcal{C}],$$

so there must exist a deterministic codebook $\mathcal{C}^*$ with $\Pr[S = 0|\mathcal{C}^*] \le \Pr[S = 0]$, and this codebook also achieves $\mathsf{E}\big[d_1\big(X^n, \hat{X}^n\big)\big] \le \mathsf{D}_1 + \epsilon$ and $\mathsf{E}\big[d_2\big(X^n, \hat{X}^n\big)\big] \le \mathsf{D}_2 + \epsilon$. (Since the codebook is generated independently of the source sequence, X^n is also distributed IID when conditioned on $\mathcal{C}^*$.)

b) This result follows from the first part by defining $\hat{\mathcal{X}} = \hat{\mathcal{X}}_1 \times \hat{\mathcal{X}}_2$, $\hat{X} = (\hat{X}_1, \hat{X}_2)$ and setting $\tilde{d}_1(x,\hat{x}) = d_1(x,\hat{x}_1)$ and $\tilde{d}_2(x,\hat{x}) = d_2(x,\hat{x}_2)$. The two reconstructors $g_1^{(n)}$ and $g_2^{(n)}$ can be viewed as the coordinates of a single reconstructor $g^{(n)}\colon \{1,\dots,2^{n\mathsf{R}}\} \to \hat{\mathcal{X}}_1 \times \hat{\mathcal{X}}_2$. The result from the first part can then be applied to show that $(\mathsf{R}, \mathsf{D}_1, \mathsf{D}_2)$ is achievable if, and only if,

$$\mathsf{R} \geq \min_{\substack{P_{\hat{X}|X}(\cdot|\cdot):\\ \mathsf{E}[\tilde{d}_1(X,\hat{X})]\leq \mathsf{D}_1,\\ \mathsf{E}[\tilde{d}_2(X,\hat{X})]\leq \mathsf{D}_2}} I\bigl(X;\hat{X}\bigr) = \min_{\substack{P_{\hat{X}_1,\hat{X}_2|X}(\cdot,\cdot|\cdot):\\ \mathsf{E}[d_1(X,\hat{X}_1)]\leq \mathsf{D}_1,\\ \mathsf{E}[d_2(X,\hat{X}_2)]\leq \mathsf{D}_2}} I\bigl(X;\hat{X}_1,\hat{X}_2\bigr)$$

c) This follows by noting that if $P_{\hat{X}|X}$ achieves

$$\min_{\substack{P_{\hat{X}|X}(\cdot|\cdot):\\ \mathsf{E}[d_1(X,\hat{X})]\leq \mathsf{D}_1,\\ \mathsf{E}[d_2(X,\hat{X})]\leq \mathsf{D}_2}} I\bigl(X;\hat{X}\bigr),$$

then

$$P_{\hat{X}_1,\hat{X}_2|X}(\hat{x}_1,\hat{x}_2|x) = P_{\hat{X}|X}(\hat{x}_1|x)\,\mathrm{I}\{\hat{x}_2 = \hat{x}_1\}$$

satisfies $\mathsf{E}[d_1(X,\hat{X}_1)] \leq \mathsf{D}_1$ and $\mathsf{E}[d_2(X,\hat{X}_2)] \leq \mathsf{D}_2$, so that the minimum over all permissible $P_{\hat{X}_1,\hat{X}_2|X}(\cdot,\cdot|\cdot)$ can only be smaller.

Problem 2 — *Source–Channel Separation with Feedback*

The first half of the proof proceeds along the same lines as the proof for the converse of the rate-distortion theorem; the second half of the proof is similar to the proof that feedback does not increase the capacity of a DMC. The claim is true because

$$\begin{aligned}
k\mathsf{R}(\mathsf{D}) &\overset{(a)}{\leq} k\mathsf{R}\bigl(\mathsf{E}\bigl[d\bigl(U^k,\hat{U}^k\bigr)\bigr]\bigr)\\
&\overset{(b)}{=} k\mathsf{R}\left(\sum_{i=1}^{k}\frac{1}{k}\,\mathsf{E}\bigl[d\bigl(U_i,\hat{U}_i\bigr)\bigr]\right)\\
&\overset{(c)}{\leq} \sum_{i=1}^{k}\mathsf{R}\bigl(\mathsf{E}\bigl[d\bigl(U_i,\hat{U}_i\bigr)\bigr]\bigr)\\
&\overset{(d)}{\leq} \sum_{i=1}^{k} I\bigl(U_i;\hat{U}_i\bigr)\\
&= \sum_{i=1}^{k}\Bigl[H\bigl(U_i\bigr) - H\bigl(U_i|\hat{U}_i\bigr)\Bigr]\\
&\overset{(e)}{=} H\bigl(U^k\bigr) - \sum_{i=1}^{k} H\bigl(U_i|\hat{U}_i\bigr)\\
&\overset{(f)}{\leq} H\bigl(U^k\bigr) - \sum_{i=1}^{k} H\bigl(U_i|\hat{U}^k,U^{i-1}\bigr)\\
&\overset{(g)}{=} H\bigl(U^k\bigr) - H\bigl(U^k|\hat{U}^k\bigr)\\
&= I\bigl(U^k;\hat{U}^k\bigr)\\
&\overset{(h)}{\leq} I\bigl(U^k;Y^n\bigr)
\end{aligned}$$

$$
\begin{aligned}
&= H(Y^n) - H(Y^n|U^k) \\
&\overset{\text{(i)}}{=} \sum_{i=1}^{n} \Big[H(Y_i|Y^{i-1}) - H(Y_i|U^k, Y^{i-1}) \Big] \\
&\overset{\text{(j)}}{\le} \sum_{i=1}^{n} \Big[H(Y_i) - H(Y_i|U^k, Y^{i-1}) \Big] \\
&\overset{\text{(k)}}{=} \sum_{i=1}^{n} \Big[H(Y_i) - H(Y_i|U^k, Y^{i-1}, X_i) \Big] \\
&\overset{\text{(l)}}{=} \sum_{i=1}^{n} \big[H(Y_i) - H(Y_i|X_i) \big] \\
&= \sum_{i=1}^{n} I(X_i; Y_i) \\
&\overset{\text{(m)}}{\le} n\mathsf{C},
\end{aligned}
$$

where (a) follows from the monotonicity of $\mathsf{R}(\mathsf{D})$; (b) follows because we use a single-letter distortion measure; (c) follows from Jensen's inequality and the convexity of $\mathsf{R}(\mathsf{D})$; (d) follows from the definition of $\mathsf{R}(\mathsf{D})$; (e) follows because $U_1, \ldots, U_k$ are independent; (f) follows because conditioning does not increase entropy; (g) follows from the chain rule; (h) follows from the data-processing inequality as $\hat{U}^k$ is a function of Y^n; (i) follows from the chain rule; (j) follows because conditioning does not increase entropy; (k) follows because X_i is a function of U^k and Y^{i-1}; (l) follows because the output Y_i of a DMC depends only on its input X_i; and (m) follows from the definition of C.

Problem 3 — *Source-Channel Separation and Average Bit-Error Probability*

Using the distortion function

$$
d(u, \hat{u}) \triangleq I\{u \neq \hat{u}\} = \begin{cases} 0 & \text{if } u = \hat{u}, \\ 1 & \text{if } u \neq \hat{u}, \end{cases}
$$

we see that

$$
\frac{1}{k} \sum_{i=1}^{k} \Pr\big[U_i \neq \hat{U}_i\big] = \frac{1}{k} \sum_{i=1}^{k} \mathsf{E}\big[d(U_i, \hat{U}_i)\big] = \mathsf{E}\big[d(U^k, \hat{U}^k)\big] \ge \mathsf{D}_{\min},
$$

where $\mathsf{D}_{\min}$ denotes the smallest expected distortion that can be achieved. From Problem 2 we know that $\mathsf{D}_{\min}$ has to satisfy

$$
\mathsf{R}(\mathsf{D}_{\min}) \le \frac{n}{k}\mathsf{C}.
$$

Since we can always achieve an expected distortion of $\frac{1}{2}$ (by setting $\hat{U}_i$ to zero for example), we have $\mathsf{D}_{\min} \in \big[0, \frac{1}{2}\big]$, and $\mathsf{R}(\mathsf{D}_{\min}) = 1 - H_{\mathrm{b}}(\mathsf{D}_{\min})$ follows from the example given in the lecture (Ber(1/2) source with Hamming distortion). Combining these results, we obtain

$$
1 - H_{\mathrm{b}}(\mathsf{D}_{\min}) \le \frac{n}{k}\mathsf{C}.
$$

By the assumption that $\frac{k}{n} > \mathsf{C}$ and by the monotonicity of the binary entropy function, we get

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

which shows that reliable communication above channel capacity is not possible (even with feedback, since the result from Problem 2 also holds if feedback is provided to the encoder). (Without going into the details, note that this result is stronger than the converse for channel coding given in the lecture: There it was shown that the *block error probability*, i.e., the probability that at least one bit will not be decoded correctly, is bounded away from zero. Here you have shown that the *bit error probability* is bounded away from zero, i.e., that on average a constant fraction of the bits will not be decoded correctly.)