Information Theory I

Fall Semester 2016

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 11 of November 30, 2016

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

Problem 1 — *Properties of* $\mathsf{R}(\mathsf{D})$

For any fixed $Q_{\hat{X}|X}$, we have

$$\begin{aligned} \mathsf{E}\big[d'(X,\hat{X})\big] &= \mathsf{E}\big[d(X,\hat{X}) - w(X)\big] \\ &\overset{(i)}{=} \mathsf{E}\big[d(X,\hat{X})\big] - \mathsf{E}\big[w(X)\big] \\ &\overset{(ii)}{=} \mathsf{E}\big[d(X,\hat{X})\big] - \bar{w}, \end{aligned}$$

where (i) follows from the linearity of expectation and (ii) follows from the definition of $\bar{w}$. Because $\mathsf{E}\big[d'(X,\hat{X})\big] \leq \mathsf{D}$ if and only if $\mathsf{E}\big[d(X,\hat{X})\big] \leq \mathsf{D} + \bar{w}$,

$$\mathsf{R}'(\mathsf{D}) = \min_{\substack{Q_{\hat{X}|X}: \\ \mathsf{E}[d'(X,\hat{X})] \leq \mathsf{D}}} I\big(X;\hat{X}\big) = \min_{\substack{Q_{\hat{X}|X}: \\ \mathsf{E}[d(X,\hat{X})] \leq \mathsf{D}+\bar{w}}} I\big(X;\hat{X}\big) = \mathsf{R}(\mathsf{D}+\bar{w}).$$

We now choose

$$w(x) = \min_{\hat{x}} d(x,\hat{x}) \tag{1}$$

for every $x \in \mathcal{X}$. Then, $d'(x,\hat{x}) \geq 0$ for all $x, \hat{x}$ because $d(x,\hat{x}) \geq \min_{\hat{x}'} d(x,\hat{x}')$ for all $x,\hat{x}$; and for every x, there exists at least one $\hat{x}$ such that $d'(x,\hat{x}) = 0$, for example the $\hat{x}$ that minimizes the RHS of (1). Therefore, $d'(x,\hat{x})$ satisfies the assumptions and the original rate distortion function can be recovered from $\mathsf{R}'(\cdot)$ because we have $\mathsf{R}(\mathsf{D}) = \mathsf{R}'(\mathsf{D} - \bar{w})$.

Problem 2 — *Erasure Distortion*

a) We start with the computation of the expected distortion:

$$\begin{aligned} \mathsf{E}\big[d(X,\hat{X})\big] &= \sum_{x\in\{0,1\}} \sum_{\hat{x}\in\{0,1,?\}} P(x)P(\hat{x}|x)d(x,\hat{x}) \\ &= \frac{1}{2}P_{\hat{X}|X}(0|0)\cdot 0 + \frac{1}{2}P_{\hat{X}|X}(1|0)\cdot\infty + \frac{1}{2}P_{\hat{X}|X}(?|0)\cdot 1 \\ &\quad + \frac{1}{2}P_{\hat{X}|X}(0|1)\cdot\infty + \frac{1}{2}P_{\hat{X}|X}(1|1)\cdot 0 + \frac{1}{2}P_{\hat{X}|X}(?|1)\cdot 1 \\ &= \frac{1}{2}P_{\hat{X}|X}(?|0) + \frac{1}{2}P_{\hat{X}|X}(?|1) \\ &= \Pr\big[\hat{X} = ?\big], \end{aligned}$$

where we have to set $P_{\hat{X}|X}(1|0) = P_{\hat{X}|X}(0|1) = 0$ since otherwise the expected distortion is always infinite. Hence, to find the rate distortion function we have to minimize $I(X;\hat{X})$ subject to the constraint

$$\Pr\big[\hat{X} = ?\big] \leq \mathsf{D}. \tag{2}$$

Note that for $\mathsf{D} \geq 1$, choosing $P_{\hat{X}|X}$ such that $\Pr[\hat{X} = ?] = 1 \leq \mathsf{D}$ achieves $I(X;\hat{X}) = 0$. Therefore, $R(\mathsf{D}) = 0$ for $\mathsf{D} \geq 1$.

Since $P_{\hat{X}|X}(1|0) = P_{\hat{X}|X}(0|1) = 0$, we know that if $\hat{X} = 0$, then $X = 0$ with probability one, and if $\hat{X} = 1$, then $X = 1$ with probability one. Thus,

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= 1 - \Big(\Pr[\hat{X}=0]\underbrace{H(X|\hat{X}=0)}_{=0} + \Pr[\hat{X}=1]\underbrace{H(X|\hat{X}=1)}_{=0} \\ &\qquad + \underbrace{\Pr[\hat{X}=?]}_{\leq \mathsf{D} \text{ by } (2)}\underbrace{H(X|\hat{X}=?)}_{\leq 1}\Big) \\ &\geq 1 - \mathsf{D}. \end{aligned}$$

For $\mathsf{D} \in [0,1]$, the inequality is met with equality for the choice

$$P(\hat{x}|x) = \begin{cases} 1-\mathsf{D} & \text{if } \hat{x} = x, \\ \mathsf{D} & \text{if } \hat{x} = ?, \\ 0 & \text{otherwise} \end{cases}$$

because then $\Pr[\hat{X} = ?] = \mathsf{D}$ and $\Pr[X = 0|\hat{X} = ?] = \Pr[X = 1|\hat{X} = ?] = \frac{1}{2}$.

Hence, the rate distortion function is given as follows (in bits):

$$\mathsf{R}(\mathsf{D}) = \begin{cases} 1-\mathsf{D} & \text{if } 0 \leq \mathsf{D} \leq 1, \\ 0 & \text{if } \mathsf{D} > 1. \end{cases}$$

The rate distortion region, i.e., the set of all achievable pairs (R,D), is depicted in Figure 1.

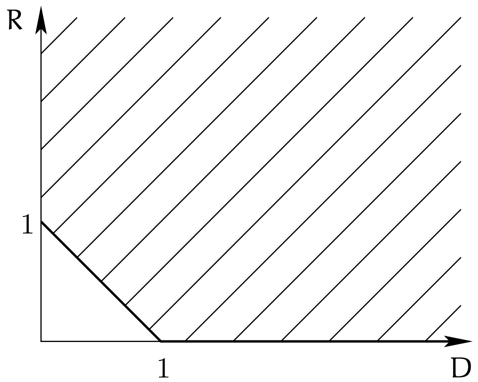

Figure 1: Rate distortion region (the rate is measured in bits).

b) We use the following scheme: for every source sequence of length n, choose a codeword that consists of the first $n(1-\mathsf{D})$ source symbols followed by $n\mathsf{D}$ question marks. This code needs $2^{n(1-\mathsf{D})}$ codewords and has therefore a rate of $(1-\mathsf{D})$ bits. The distortion achieved by this code is as follows: the first $n(1-\mathsf{D})$ digits have zero distortion, the rest has distortion 1. Hence, on average we get a distortion of D.

Note that once $\mathsf{D} \geq 1$ we will only use question marks, i.e., we only use one codeword $\hat{\mathbf{x}} = (?,\ldots,?)$. This code has rate zero and it achieves a distortion of 1.

Problem 3 ***Rate Distortion Function with Infinite Distortion***

Let's start by computing the average distortion for a certain choice of $q(\cdot|\cdot)$:

$$\begin{aligned}\mathsf{E}[d(X,\hat{X})] &= Q_X(0)q(0|0)d(0,0) + Q_X(0)q(1|0)d(0,1) + Q_X(1)q(0|1)d(1,0) + Q_X(1)q(1|1)d(1,1)\\ &= \frac{1}{2}q(1|0)\cdot\infty + \frac{1}{2}q(0|1) \overset{!}{\leq} \mathsf{D}.\end{aligned}$$

Hence, we see that $q(1|0) \overset{!}{=} 0$. Let $q(0|1) = p$ for some parameter p, $0 \leq p \leq 1$. Then, the above computation implies that

$$0 \leq p \leq 2\mathsf{D}.$$

For $\mathsf{D} \geq \frac{1}{2}$, we can take $\hat{X} = 0$ (which corresponds to $p = 1$); therefore $\mathsf{R}(\mathsf{D}) = 0$ for $\mathsf{D} \geq \frac{1}{2}$. We will prove by contradiction that $I\big(X;\hat{X}\big)$ is minimized by taking p as large as possible in the case $\mathsf{D} \in \big[0, \frac{1}{2}\big]$. Assume that $I(Q,W)$ is minimized for some $p^* < 2\mathsf{D} \leq 1$ and denote the corresponding $q(\cdot|\cdot)$ by W^*. Denote by W_0 the mapping $q(\cdot|\cdot)$ that maps every input symbol deterministically to zero. Its expected distortion is $\frac{1}{2}$ and we have $I(Q,W_0) = 0$. We now consider the mixture $\lambda W_0 + \bar{\lambda} W^*$. Its expected distortion is $\lambda \cdot \frac{1}{2} + \bar{\lambda}\cdot\frac{1}{2}p^*$, which is smaller than D for some positive value of λ because $p^* < 2\mathsf{D}$. For this value of λ, we have by the convexity of $I(Q,W)$

$$I(Q, \lambda W_0 + \bar{\lambda}W^*) \leq \lambda \underbrace{I(Q,W_0)}_{=0} + \underbrace{\bar{\lambda}}_{<1} I(Q,W^*) < I(Q,W^*),$$

which contradicts the assumption that p^* minimizes $I(Q,W)$. Consequently, the choice $p^* = 2\mathsf{D}$ is optimal in the case $\mathsf{D} \in \big[0,\frac{1}{2}\big]$. (Instead of using convexity arguments, it is also possible to use calculus to minimize the mutual information (3) as a function of p.)
Since we have

$$Q_{\hat{X}}(1) = Q_X(0)\underbrace{q(1|0)}_{=0} + \underbrace{Q_X(1)}_{=1/2}\underbrace{q(1|1)}_{=1-p} = \frac{1-p}{2},$$

we get

$$\begin{aligned} I\big(X;\hat{X}\big) &= H\big(\hat{X}\big) - H\big(\hat{X}\big|X\big)\\ &= H_{\mathrm{b}}\left(\frac{1-p}{2}\right) - Q_X(0)\underbrace{H\big(\hat{X}\big|X=0\big)}_{=0} - \underbrace{Q_X(1)}_{=1/2}\underbrace{H\big(\hat{X}\big|X=1\big)}_{=H_{\mathrm{b}}(p)}\\ &= H_{\mathrm{b}}\left(\frac{1-p}{2}\right) - \frac{1}{2}H_{\mathrm{b}}(p). \end{aligned} \tag{3}$$

Combining the results and using the choice $p^* = 2\mathsf{D}$ in the case $\mathsf{D} \in \big[0,\frac{1}{2}\big]$ leads to the following rate distortion function:

$$\mathsf{R}(\mathsf{D}) = \begin{cases} H_{\mathrm{b}}\big(\frac{1-2\mathsf{D}}{2}\big) - \frac{1}{2}H_{\mathrm{b}}(2\mathsf{D}) & \text{if } 0 \leq \mathsf{D} \leq \frac{1}{2},\\ 0 & \text{if } \mathsf{D} > \frac{1}{2}. \end{cases}$$

Problem 4 ***Rate Distortion for Uniform Source with Hamming Distortion***

First, we note that the expected distortion $\mathsf{E}\big[d(X,\hat{X})\big]$ is equal to $\Pr\big[X \neq \hat{X}\big]$:

$$\begin{aligned}\mathsf{E}\big[d(X,\hat{X})\big] &= \sum_x p(x) \sum_{\hat{x}} p(\hat{x}|x) d(x,\hat{x}) \\ &= \sum_x p(x) \sum_{\hat{x}\neq x} p(\hat{x}|x) \\ &= \Pr\big[X \neq \hat{X}\big].\end{aligned}$$

Hence, to find $R(\mathsf{D})$ we have to minimize $I(X;\hat{X})$ subject to the constraint that $\Pr\big[X \neq \hat{X}\big] \leq \mathsf{D}$. First we note that $R(\mathsf{D}) = 0$ for $\mathsf{D} \geq (m-1)/m$, because if $\hat{X}$ is independent of X, then, irrespective of $P_{\hat{X}}$, the expected distortion equals $(m-1)/m$ and the mutual information between X and $\hat{X}$ equals 0. In the following we thus consider the case where $\mathsf{D} < (m-1)/m$. Let $P_{\mathrm{e}} = \Pr\big[X \neq \hat{X}\big]$ and observe that

$$\begin{aligned}I\big(X;\hat{X}\big) &= H\big(X\big) - H\big(X|\hat{X}\big) \\ &\overset{(i)}{\geq} H\big(X\big) - H_{\mathrm{b}}(P_{\mathrm{e}}) - P_{\mathrm{e}} \log{(m-1)} \\ &\overset{(ii)}{\geq} H\big(X\big) - H_{\mathrm{b}}(\mathsf{D}) - \mathsf{D} \log{(m-1)} \\ &\overset{(iii)}{=} \log m - H_{\mathrm{b}}(\mathsf{D}) - \mathsf{D} \log{(m-1)},\end{aligned}$$

where (i) follows from Fano's inequality; (ii) follows because $P_{\mathrm{e}} \leq \mathsf{D}$ and because the function $P_{\mathrm{e}} \mapsto H_{\mathrm{b}}(P_{\mathrm{e}}) + P_{\mathrm{e}} \log{(m-1)}$ is nondecreasing in P_{e} for $P_{\mathrm{e}} \in \big[0, \frac{m-1}{m}\big]$ (we will show this later); and (iii) follows because X is distributed uniformly. Choosing $P_{\hat{X}|X}$ to be

$$P_{\hat{X}|X}(\hat{x}|x) = \begin{cases} 1-\mathsf{D} & \text{if } \hat{x} = x, \\ \frac{\mathsf{D}}{m-1} & \text{if } \hat{x} \neq x \end{cases}$$

satisfies $\Pr\big[X \neq \hat{X}\big] \leq \mathsf{D}$ and achieves equality because then

$$P_{\hat{X}}(\hat{x}) = \sum_x P_{X,\hat{X}}(x,\hat{x}) = \sum_x P_X(x) P_{\hat{X}|X}(\hat{x}|x) = \frac{1}{m}(1-\mathsf{D}) + (m-1)\frac{1}{m}\frac{\mathsf{D}}{m-1} = \frac{1}{m}$$

and

$$P_{X|\hat{X}}(x|\hat{x}) = \frac{P_{X,\hat{X}}(x,\hat{x})}{P_{\hat{X}}(\hat{x})} = \frac{P_{\hat{X}|X}(\hat{x}|x)P_X(x)}{P_{\hat{X}}(\hat{x})} = \frac{P_{\hat{X}|X}(\hat{x}|x)m^{-1}}{m^{-1}} = P_{\hat{X}|X}(\hat{x}|x).$$

Consequently,

$$\begin{aligned}H\big(X|\hat{X}\big) &= \sum_{\hat{x}} P_{\hat{X}}(\hat{x}) H\big(X|\hat{X} = \hat{x}\big) \\ &= \sum_{\hat{x}} P_{\hat{X}}(\hat{x}) \left[(1-\mathsf{D}) \log \frac{1}{1-\mathsf{D}} + (m-1)\frac{\mathsf{D}}{m-1} \log \frac{m-1}{\mathsf{D}}\right] \\ &= (1-\mathsf{D}) \log \frac{1}{1-\mathsf{D}} + (m-1)\frac{\mathsf{D}}{m-1} \log \frac{m-1}{\mathsf{D}} \\ &= (1-\mathsf{D}) \log \frac{1}{1-\mathsf{D}} + \mathsf{D} \log \frac{1}{\mathsf{D}} + \mathsf{D} \log{(m-1)} \\ &= H_{\mathrm{b}}(\mathsf{D}) + \mathsf{D} \log{(m-1)}.\end{aligned}$$

Hence, the rate distortion function is

$$\mathsf{R}(\mathsf{D}) = \begin{cases} \log m - H_{\mathrm{b}}(\mathsf{D}) - \mathsf{D}\log(m-1) & 0 \le \mathsf{D} \le \frac{m-1}{m}, \\ 0 & \mathsf{D} > \frac{m-1}{m}. \end{cases}$$

It remains to show that the function $P_{\mathrm{e}} \mapsto H_{\mathrm{b}}(P_{\mathrm{e}}) + P_{\mathrm{e}} \log{(m-1)}$ is nondecreasing in P_{e} for $P_{\mathrm{e}} \in \bigl[0, \frac{m-1}{m}\bigr]$. This can be done for example by computing its derivative. An alternative way is the following concavity argument. Observe that

$$\begin{aligned} H_{\mathrm{b}}(P_{\mathrm{e}}) + P_{\mathrm{e}} \log{(m-1)} &= P_{\mathrm{e}} \log \frac{1}{P_{\mathrm{e}}} + (1-P_{\mathrm{e}}) \log \frac{1}{1-P_{\mathrm{e}}} + P_{\mathrm{e}} \log{(m-1)} \\ &= P_{\mathrm{e}} \log \frac{m-1}{P_{\mathrm{e}}} + (1-P_{\mathrm{e}}) \log \frac{1}{1-P_{\mathrm{e}}} \\ &= \log m + P_{\mathrm{e}} \log \frac{(m-1)/m}{P_{\mathrm{e}}} + (1-P_{\mathrm{e}}) \log \frac{1/m}{1-P_{\mathrm{e}}} \\ &= \log m - D\Bigl(\bigl(P_{\mathrm{e}}, 1-P_{\mathrm{e}}\bigr) \big|\big| \bigl(\tfrac{m-1}{m}, \tfrac{1}{m}\bigr)\Bigr). \end{aligned}$$

Because relative entropy is nonnegative and equal to zero if and only if both arguments are equal, $P_{\mathrm{e}} \mapsto H_{\mathrm{b}}(P_{\mathrm{e}}) + P_{\mathrm{e}} \log{(m-1)}$ is maximized for $P_{\mathrm{e}} = \frac{m-1}{m}$. Since $P_{\mathrm{e}} \mapsto H_{\mathrm{b}}(P_{\mathrm{e}}) + P_{\mathrm{e}} \log{(m-1)}$ is concave in P_{e} (entropy and linear functions are concave; and the sum of concave functions is concave), it is nondecreasing in P_{e} for $P_{\mathrm{e}} \in \bigl[0, \frac{m-1}{m}\bigr]$.