Information Theory I

Fall Semester 2017

Prof. Dr. A. Lapidoth

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Model Answers to Exercise 12 of December 6, 2017

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

Problem 1 — *Rate Distortion with Two Distortion Functions*

a) We first prove two properties of

$$\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1,\mathsf{D}_2) \triangleq \min_{\substack{P_{\hat{X}|X}:\\ \mathsf{E}[d_1(X,\hat{X})]\leq \mathsf{D}_1,\\ \mathsf{E}[d_2(X,\hat{X})]\leq \mathsf{D}_2}} I(X;\hat{X}) = \min_{\substack{P_{\hat{X}|X}:\\ \mathsf{E}[d_1(X,\hat{X})]\leq \mathsf{D}_1,\\ \mathsf{E}[d_2(X,\hat{X})]\leq \mathsf{D}_2}} I(P_X, P_{\hat{X}|X}).$$

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

$$\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1',\mathsf{D}_2') \leq \mathsf{R}^{\mathrm{I}}(\mathsf{D}_1,\mathsf{D}_2)$$

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

Proof. Fix D_1, D_2, $\mathsf{D}_1' \geq \mathsf{D}_1$, and $\mathsf{D}_2' \geq \mathsf{D}_2$. Let $P^*_{\hat{X}|X}$ be a conditional PMF that achieves the minimum in the definition of $\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1,\mathsf{D}_2)$. For this PMF,

$$\begin{aligned}\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1,\mathsf{D}_2) &= I(P_X, P^*_{\hat{X}|X}),\\ \mathsf{E}[d_1(X,\hat{X})] &\leq \mathsf{D}_1 \leq \mathsf{D}_1',\\ \mathsf{E}[d_2(X,\hat{X})] &\leq \mathsf{D}_2 \leq \mathsf{D}_2'.\end{aligned}$$

Consequently,

$$\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1',\mathsf{D}_2') = \min_{\substack{P_{\hat{X}|X}:\\ \mathsf{E}[d_1(X,\hat{X})]\leq \mathsf{D}_1',\\ \mathsf{E}[d_2(X,\hat{X})]\leq \mathsf{D}_2'}} I(P_X, P_{\hat{X}|X}) \stackrel{\text{(i)}}{\leq} I(P_X, P^*_{\hat{X}|X}) = \mathsf{R}^{\mathrm{I}}(\mathsf{D}_1,\mathsf{D}_2),$$

where (i) holds because $P^*_{\hat{X}|X}$ satisfies $\mathsf{E}[d_1(X,\hat{X})] \leq \mathsf{D}_1'$ and $\mathsf{E}[d_2(X,\hat{X})] \leq \mathsf{D}_2'$. ■

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

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

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

Proof. Let $P^{(1)}_{\hat{X}|X}$ be a conditional PMF that achieves the minimum in the definition of $\mathsf{R}^{\mathrm{I}}(\mathsf{D}^{(1)}_1, \mathsf{D}^{(1)}_2)$, and let $P^{(2)}_{\hat{X}|X}$ be a conditional PMF that achieves the minimum in the definition of $\mathsf{R}^{\mathrm{I}}(\mathsf{D}^{(2)}_1, \mathsf{D}^{(2)}_2)$. Note that the mixture $\lambda P^{(1)}_{\hat{X}|X} + \bar{\lambda} P^{(2)}_{\hat{X}|X}$ is a conditional PMF satisfying

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

Consequently,

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

where (i) holds because $I(Q,W)$ is convex in W. ■

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

$$\mathsf{E}\bigl[d_1\bigl(X^n, \hat{X}^n\bigr)\bigr] \leq \mathsf{D}_1, \tag{1}$$

$$\mathsf{E}\bigl[d_2\bigl(X^n, \hat{X}^n\bigr)\bigr] \leq \mathsf{D}_2. \tag{2}$$

(We cheat a bit and neglect δ here, i.e., we omit a continuity argument.) Indeed,

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

where (i) follows from (1), (2), and the monotonicity of $\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1, \mathsf{D}_2)$; (ii) holds because we have a single-letter distortion measure (and because expectation is linear); (iii) follows from Jensen's inequality because $\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1, \mathsf{D}_2)$ is convex in the pair $(\mathsf{D}_1, \mathsf{D}_2)$; (iv) follows from

$$\mathsf{R}^{\mathrm{I}}\bigl(\mathsf{E}\bigl[d_1\bigl(X_i, \hat{X}_i\bigr)\bigr], \mathsf{E}\bigl[d_2\bigl(X_i, \hat{X}_i\bigr)\bigr]\bigr) = \min_{\substack{P_{\hat{X}|X}: \\ \mathsf{E}[d_1(X,\hat{X})] \leq \mathsf{E}[d_1(X_i,\hat{X}_i)], \\ \mathsf{E}[d_2(X,\hat{X})] \leq \mathsf{E}[d_2(X_i,\hat{X}_i)]}} I\bigl(P_X, P_{\hat{X}|X}\bigr) \leq I\bigl(P_{X_i}, P_{\hat{X}_i|X_i}\bigr)$$

because $P_X = P_{X_i}$ and because $P_{\hat{X}_i|X_i}$ satisfies the constraints on the expected distortion; (v) holds since $X_1, \ldots, X_n$ are IID; (vi) holds 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 have only $2^{n\mathsf{R}}$ codewords, so $I\bigl(X^n; \hat{X}^n\bigr) \leq n\mathsf{R}$.

We prove the direct part as follows. Fix D_1 and D_2. Let $P^*_{\hat{X}|X}$ be a conditional PMF that achieves the minimum in the definition of $\mathsf{R}^{\mathrm{I}}(\mathsf{D}_1, \mathsf{D}_2)$. Then, for all $\hat{x} \in \hat{\mathcal{X}}$,

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

Fix $\mathsf{R} > \mathsf{R}^{\mathrm{I}}(\mathsf{D}_1, \mathsf{D}_2)$, 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 $\bigl(x^n, \hat{x}^n(i)\bigr) \in \mathcal{T}^{(n)}_\epsilon(P_{X,\hat{X}})$ 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 $\bigl(x^n, \hat{x}^n(i)\bigr) \in \mathcal{T}^{(n)}_\epsilon(P_{X,\hat{X}})$, and zero otherwise. The expected distortion for the random coding scheme can be bounded as

$$\begin{aligned} \mathsf{E}\bigl[d_1\bigl(X^n, \hat{X}^n\bigr)\bigr] &\overset{\text{(i)}}{=} \Pr\bigl[S = 0\bigr] \underbrace{\mathsf{E}\bigl[d_1\bigl(X^n, \hat{X}^n\bigr)|S = 0\bigr]}_{\leq d_{1,\max}} + \underbrace{\Pr\bigl[S = 1\bigr]}_{\leq 1} \underbrace{\mathsf{E}\bigl[d_1\bigl(X^n, \hat{X}^n\bigr)|S = 1\bigr]}_{\leq \mathsf{D}_1(1+\epsilon)} \\ &\leq \Pr\bigl[S = 0\bigr] \cdot d_{1,\max} + \mathsf{D}_1(1 + \epsilon) \end{aligned} \tag{3}$$

where (i) follows from the law of total expectation; the first underbraced inequality holds because we assume bounded distortions; the second underbraced inequality is trivial; the third underbraced inequality holds because $S = 1$ implies that there exists a codeword $\hat{x}^n(i)$ such that $\bigl(x^n, \hat{x}^n(i)\bigr) \in \mathcal{T}^{(n)}_\epsilon(P_{X,\hat{X}})$; and for that codeword, we have $d_1\bigl(x^n, \hat{x}^n\bigr) \leq \mathsf{D}_1(1 + \epsilon)$ as a consequence of strong typicality because $P^*_{\hat{X}|X}$ satisfies $\mathsf{E}[d_1(X, \hat{X})] \leq \mathsf{D}_1$. Similarly, we obtain

$$\mathsf{E}\bigl[d_2\bigl(X^n, \hat{X}^n\bigr)\bigr] \leq \Pr\bigl[S = 0\bigr] \cdot d_{2,\max} + \mathsf{D}_2(1 + \epsilon). \tag{4}$$

From the lecture we know that for $\mathsf{R} > \mathsf{R}^{\mathrm{I}}(\mathsf{D}_1, \mathsf{D}_2) = I\bigl(P_X, P^*_{\hat{X}|X}\bigr)$ 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 $\mathsf{E}\bigl[d_1\bigl(X^n, \hat{X}^n\bigr)\bigr] \leq \mathsf{D}_1 + \delta$ and $\mathsf{E}\bigl[d_2\bigl(X^n, \hat{X}^n\bigr)\bigr] \leq \mathsf{D}_2 + \delta$ from (3) and (4), respectively. Finally, we argue that we can get the same performance with a deterministic codebook. Note that the RHS of (3) and (4) depend only 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 $\Pr[S = 0|\mathcal{C}^*] \leq \Pr[S = 0]$, and this codebook also achieves $\mathsf{E}\bigl[d_1\bigl(X^n, \hat{X}^n\bigr)\bigr] \leq \mathsf{D}_1 + \delta$ and $\mathsf{E}\bigl[d_2\bigl(X^n, \hat{X}^n\bigr)\bigr] \leq \mathsf{D}_2 + \delta$. (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 $g^{(n)} \colon \{1,\ldots,2^{n\mathsf{R}}\} \to \hat{\mathcal{X}}_1 \times \hat{\mathcal{X}}_2$. Then,

$$\begin{aligned} \mathsf{E}[d_1(X^n, \hat{X}_1^n)] &\leq \mathsf{D}_1 + \delta, \\ \mathsf{E}[d_2(X^n, \hat{X}_2^n)] &\leq \mathsf{D}_2 + \delta \end{aligned}$$

if and only if

$$\begin{aligned} \mathsf{E}[\tilde{d}_1(X^n, \hat{X}^n)] &\leq \mathsf{D}_1 + \delta, \\ \mathsf{E}[\tilde{d}_2(X^n, \hat{X}^n)] &\leq \mathsf{D}_2 + \delta. \end{aligned}$$

Using Part a), we see that rates $\mathsf{R} > \mathsf{R}(\mathsf{D}_1,\mathsf{D}_2)$ are achievable and rates $\mathsf{R} < \mathsf{R}(\mathsf{D}_1,\mathsf{D}_2)$ are not achievable, where

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

c) Let $P^*_{\hat{X}|X}$ be a conditional PMF that achieves the minimum in the definition of $\mathsf{R}^{\mathsf{I}}(\mathsf{D}_1,\mathsf{D}_2)$. Let the conditional PMF $\tilde{P}_{\hat{X}_1,\hat{X}_2|X}$ be given by

$$\tilde{P}_{\hat{X}_1,\hat{X}_2|X}(\hat{x}_1,\hat{x}_2|x) = P^*_{\hat{X}|X}(\hat{x}_1|x)\,\mathrm{I}\{\hat{x}_2 = \hat{x}_1\}.$$

Observe that $\tilde{P}_{\hat{X}_1,\hat{X}_2|X}$ satisfies

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

Consequently,

$$\begin{aligned} \mathsf{R}^{\mathsf{J}}(\mathsf{D}_1,\mathsf{D}_2) &= \min_{\substack{P_{\hat{X}_1,\hat{X}_2|X}: \\ \mathsf{E}[d_1(X,\hat{X}_1)]\leq \mathsf{D}_1, \\ \mathsf{E}[d_2(X,\hat{X}_2)]\leq \mathsf{D}_2}} I\bigl(P_X, P_{\hat{X}_1,\hat{X}_2|X}\bigr) \\ &\leq I\bigl(P_X, \tilde{P}_{\hat{X}_1,\hat{X}_2|X}\bigr) \\ &= I\bigl(P_X, P^*_{\hat{X}|X}\bigr) \\ &= \mathsf{R}^{\mathsf{I}}(\mathsf{D}_1,\mathsf{D}_2). \end{aligned}$$

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

$$
\begin{aligned}
k\mathsf{R}(\mathsf{D}) &\overset{\text{(a)}}{\leq} k\mathsf{R}\big(\mathsf{E}\big[d\big(U^k,\hat{U}^k\big)\big]\big) \\
&\overset{\text{(b)}}{=} k\mathsf{R}\left(\sum_{i=1}^{k}\frac{1}{k}\,\mathsf{E}\big[d\big(U_i,\hat{U}_i\big)\big]\right) \\
&\overset{\text{(c)}}{\leq} \sum_{i=1}^{k}\mathsf{R}\big(\mathsf{E}\big[d\big(U_i,\hat{U}_i\big)\big]\big) \\
&\overset{\text{(d)}}{\leq} \sum_{i=1}^{k} I\big(U_i;\hat{U}_i\big) \\
&= \sum_{i=1}^{k}\Big[H\big(U_i\big)-H\big(U_i|\hat{U}_i\big)\Big] \\
&\overset{\text{(e)}}{=} H\big(U^k\big)-\sum_{i=1}^{k}H\big(U_i|\hat{U}_i\big) \\
&\overset{\text{(f)}}{\leq} H\big(U^k\big)-\sum_{i=1}^{k}H\big(U_i|\hat{U}^k,U^{i-1}\big) \\
&\overset{\text{(g)}}{=} H\big(U^k\big)-H\big(U^k|\hat{U}^k\big) \\
&= I\big(U^k;\hat{U}^k\big) \\
&\overset{\text{(h)}}{\leq} I\big(U^k;Y^n\big) \\
&= H\big(Y^n\big)-H\big(Y^n|U^k\big) \\
&\overset{\text{(i)}}{=} \sum_{i=1}^{n}\Big[H\big(Y_i|Y^{i-1}\big)-H\big(Y_i|U^k,Y^{i-1}\big)\Big] \\
&\overset{\text{(j)}}{\leq} \sum_{i=1}^{n}\Big[H\big(Y_i\big)-H\big(Y_i|U^k,Y^{i-1}\big)\Big] \\
&\overset{\text{(k)}}{=} \sum_{i=1}^{n}\Big[H\big(Y_i\big)-H\big(Y_i|U^k,Y^{i-1},X_i\big)\Big] \\
&\overset{\text{(l)}}{=} \sum_{i=1}^{n}\big[H\big(Y_i\big)-H\big(Y_i|X_i\big)\big] \\
&= \sum_{i=1}^{n} I\big(X_i;Y_i\big) \\
&\overset{\text{(m)}}{\leq} n\mathsf{C},
\end{aligned}
$$

where (a) follows from the monotonicity of $\mathsf{R}(\mathsf{D})$; (b) holds 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) holds because $U_1,\ldots,U_k$ are independent; (f) holds because conditioning does not increase entropy; (g) follows from the chain rule; (h) follows from the data-processing inequality; (i) follows from the chain rule; (j) holds because conditioning does not increase entropy; (k) holds because X_i is a function of U^k and Y^{i-1}; (l) holds because $(U^k,Y^{i-1}) \multimap\!\!- X_i \multimap\!\!- Y_i$ form a Markov chain since we have a DMC; and (m) follows from the definition of C.

Problem 3 ***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

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

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

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

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

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

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

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

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 rate* is bounded away from zero, i.e., that on average a constant fraction of the bits will not be decoded correctly.)