Homework 5

TA: Heng-Chien Liou*

1. (Information divergence) [10]

Consider two probability density functions $\mathsf{f}(\cdot)$ and $\mathsf{g}(\cdot)$. Let μ_1 and μ_2 denote the mean of f and g respectively. Let σ_1^2 and σ_2^2 denote the variance of f and g respectively.

a) Compute $\mathrm{D}(\mathsf{f}\|\mathsf{g})$ in the following cases: (1) both f and g are Gaussian; (2) both f and g are Laplace. [6]

b) If $\mu_1 = \mu_2$, which of the above cases gives the largest/smallest KL divergence? Your answer may depend on σ_1, σ_2. [2]

c) If $\sigma_1 = \sigma_2$, which of the above cases gives the largest/smallest KL divergence? Your answer may depend on μ_1, μ_2. [2]

Solution:

a) For f and g both being Guassian, we have

$$
\begin{aligned}
\mathrm{D}(\mathsf{f}\|\mathsf{g}) &= \int_{\mathbb{R}} \mathsf{f}(x) \log \frac{\mathsf{f}(x)}{\mathsf{g}(x)} \,\mathrm{d}x \\
&= \int_{\mathbb{R}} \mathsf{f}(x) \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} \log e - \frac{(x-\mu_1)^2}{2\sigma_1^2} \log e + \log \frac{\sigma_2}{\sigma_1} \right) \mathrm{d}x \\
&= \log \frac{\sigma_2}{\sigma_1} - \frac{1}{2} \log e + \int_{\mathbb{R}} \mathsf{f}(x) \frac{(x-\mu_2)^2}{2\sigma_2^2} \log e \,\mathrm{d}x \\
&= \log \frac{\sigma_2}{\sigma_1} - \frac{1}{2} \log e + \frac{1}{2\sigma_2^2} \left(\sigma_1^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2 \right) \log e \\
&= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 - \sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} \log e
\end{aligned}
$$

For f and g both being Laplace, we have

$$
\begin{aligned}
\mathrm{D}(\mathsf{f}\|\mathsf{g}) &= \int_{\mathbb{R}} \mathsf{f}(x) \log \frac{\mathsf{f}(x)}{\mathsf{g}(x)} \,\mathrm{d}x \\
&= \int_{\mathbb{R}} \mathsf{f}(x) \left(\frac{\sqrt{2}|x-\mu_2|}{\sigma_2} \log e - \frac{\sqrt{2}|x-\mu_1|}{\sigma_1} \log e + \log \frac{\sigma_2}{\sigma_1} \right) \mathrm{d}x
\end{aligned}
$$

*with contribution by Chen-Hao Hsiao and Wen-Shao Ho

$$= \log \frac{\sigma_2}{\sigma_1} - \log e + \log e \int_{\mathbb{R}} \mathsf{f}(x) \frac{\sqrt{2}|x-\mu_2|}{\sigma_2} \, \mathrm{d}x$$

For $\mu_1 \geq \mu_2$

$$\begin{aligned}
&\int_{\mathbb{R}} \mathsf{f}(x) \frac{\sqrt{2}|x-\mu_2|}{\sigma_2} \, \mathrm{d}x \\
&= \int_{\mathbb{R}} \frac{|x-\mu_2|}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}|x-\mu_1|}{\sigma_1}} \, \mathrm{d}x \\
&= \int_{-\infty}^{\mu_2} \frac{-(x-\mu_2)}{\sigma_1\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \, \mathrm{d}x + \int_{-\mu_2}^{\mu_1} \frac{(x-\mu_2)}{\sigma_1\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \, \mathrm{d}x \\
&\quad + \int_{\mu_1}^{\infty} \frac{(x-\mu_2)}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \, \mathrm{d}x \\
&= - \left. \frac{\sqrt{2}(x-\mu_2)-\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \right|_{-\infty}^{\mu_2} + \left. \frac{\sqrt{2}(x-\mu_2)-\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \right|_{\mu_2}^{\mu_1} \\
&\quad - \left. \frac{\sqrt{2}(x-\mu_2)+\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \right|_{\mu_1}^{\infty} \\
&= \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(\mu_1-\mu_2)}{\sigma_1}} + \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(\mu_1-\mu_2)}{\sigma_1}} + \frac{\sqrt{2}(\mu_1-\mu_2)-\sigma_1}{2\sigma_2} + \frac{\sqrt{2}(\mu_1-\mu_2)+\sigma_1}{2\sigma_2} \\
&= \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}(\mu_1-\mu_2)}{\sigma_1}} + \frac{\sqrt{2}(\mu_1-\mu_2)}{\sigma_2} = \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_1}} + \frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_2}
\end{aligned}$$

Similarly, for $\mu_1 < \mu_2$

$$\begin{aligned}
&\int_{\mathbb{R}} \mathsf{f}(x) \frac{\sqrt{2}|x-\mu_2|}{\sigma_2} \, \mathrm{d}x \\
&= \int_{-\infty}^{\mu_1} \frac{-(x-\mu_2)}{\sigma_1\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \, \mathrm{d}x + \int_{-\mu_1}^{\mu_2} \frac{-(x-\mu_2)}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \, \mathrm{d}x \\
&\quad + \int_{\mu_2}^{\infty} \frac{(x-\mu_2)}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \, \mathrm{d}x \\
&= - \left. \frac{\sqrt{2}(x-\mu_2)-\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \right|_{-\infty}^{\mu_1} + \left. \frac{\sqrt{2}(x-\mu_2)+\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \right|_{\mu_1}^{\mu_2} \\
&\quad - \left. \frac{\sqrt{2}(x-\mu_2)+\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}} \right|_{\mu_2}^{\infty} \\
&= -\frac{\sqrt{2}(\mu_1-\mu_2)-\sigma_1}{2\sigma_2} - \frac{\sqrt{2}(\mu_1-\mu_2)+\sigma_1}{2\sigma_2} + \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(\mu_2-\mu_1)}{\sigma_1}} + \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(\mu_2-\mu_1)}{\sigma_1}} \\
&= -\frac{\sqrt{2}(\mu_1-\mu_2)}{\sigma_2} + \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}(\mu_2-\mu_1)}{\sigma_1}} = \frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_2} + \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_1}}
\end{aligned}$$

Hence,

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \log\frac{\sigma_2}{\sigma_1} - \log e + \log e\left(\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_2} + \frac{\sigma_1}{\sigma_2}e^{-\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_1}}\right)$$

b) For $\mu_1 = \mu_2$

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \begin{cases} \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2-\sigma_2^2}{2\sigma_2^2}\log e & \text{Gaussian} \\ \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1}{\sigma_2}\log e - \log e & \text{Laplace} \end{cases}$$

Simple calculations based on σ_2/σ_1 and the property of quadratic forms shows that pair of Gaussian always admit a larger divergence if $\mu_1 = \mu_2$.

c) For $\sigma_1 = \sigma_2 = \sigma$,

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \begin{cases} \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}\log e & \text{Gaussian} \\ -\log e + \log e\left(\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma} + e^{-\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma}}\right) & \text{Laplace} \end{cases}$$

Let $y = |\mu_1-\mu_2|/\sigma$, then we can compare

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \begin{cases} \frac{y^2}{2}\log e & \text{Gaussian} \\ -\log \mathrm{e} + \log e\left(\sqrt{2}y + e^{-\sqrt{2}y}\right) & \text{Laplace} \end{cases}$$

The equation $\frac{y^2}{2} - (-1+\sqrt{2}y + e^{-\sqrt{2}y}) = 0$ has two real solutions, one of them being zero. We denote another one as $y_0 > 0$. For $|\mu_1-\mu_2|/\sigma \le y_0$, a pair of Laplace distribution admit a divergence greater than or equal to that of Gaussian distribution. For $|\mu_1-\mu_2|/\sigma > y_0$, a pair of Gaussian has a larger divergence.

2. (Differential entropy) [10]

a) Consider a Laplace random variable $X \sim \mathsf{Lap}(\mu, b)$, that is, the probability density function of X is $\mathsf{f}_X(x) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$, $x \in \mathbb{R}$. Compute its differential entropy $\mathrm{h}(X)$. [4]

b) Consider a problem of maximizing differential entropy $\mathrm{h}(X)$ subject to the constraint that $\mathsf{E}[|X|] \le \mathsf{B}$. Find the maximum differential entropy and show that a zero-mean Laplace distributed X attains the maximum value. [6]

Solution:

a)

$$\begin{aligned}\mathrm{h}(X) &= \mathsf{E}_X\left[\log\frac{1}{\mathsf{f}_X(x)}\right] \\ &= \log 2b + \log e\left(\int_{-\infty}^{\mu}\frac{\mu-x}{2b^2}e^{\frac{-(\mu-x)}{b}}\,\mathrm{d}x + \int_{\mu}^{\infty}\frac{x-\mu}{2b^2}e^{\frac{-(x-\mu)}{b}}\,\mathrm{d}x\right)\end{aligned}$$

$$
\begin{aligned}
&= \log 2b + \log e \left(\int_0^\infty \frac{t}{2} e^{-t}\,\mathrm{d}t + \int_0^\infty \frac{t}{2} e^{-t}\,\mathrm{d}t \right) \\
&= \log 2b + \log e \left(\int_0^\infty t e^{-t}\,\mathrm{d}t \right) \\
&= \log 2b + \log e \left(- t e^{-t}\Big|_0^\infty + \int_0^\infty e^{-t}\,\mathrm{d}t \right) \\
&= \log 2be
\end{aligned}
$$

We can also observe that translation doesn't change differential entropy.

b) Let $X \sim \mathrm{Lap}(0,b)$. Its differential entropy $\mathrm{h}(X) = \log(2be)$ and $\mathbf{E}[|X|] = b$.

Let Y be a random variable with $\mathsf{E}[|Y|] = b$.

Then we have $\mathrm{h}(X) = \mathsf{E}_X[-\log \mathsf{f}_X(X)] = \mathsf{E}_X\left[-\frac{1}{2b} + \left|\frac{X}{2b}\right|\right] = \mathsf{E}_Y[-\log \mathsf{f}_X(Y)]$.

Then

$$
\begin{aligned}
\mathrm{h}(Y) - \mathrm{h}(X) &= \mathsf{E}_Y\left[\log \frac{f_X(Y)}{f_Y(Y)}\right] \\
&\le \log \mathsf{E}_Y\left[\frac{\mathsf{f}_X(Y)}{\mathsf{f}_Y(Y)}\right] \quad \text{Jensen's inequality} \\
&= 0
\end{aligned}
$$

$\mathrm{Lap}(0,b)$ maximizes $\mathrm{h}(Y)$ for given $\mathbf{E}[|Y|] = b$, and the maximum value is $\log(2\mathrm{be})$.

Since $\log(2be)$ is increasing in $0 \le b \le \mathsf{B}$, the Laplace distribution $\mathrm{Lap}(0,\mathsf{B})$ also maximizes $\mathrm{h}(S)$ for given $\mathsf{E}[|S|] \le \mathsf{B}$, and the maximum differential entropy is $\log(2\mathsf{B}e)$.

3. (Channel Coding with Input-Output Cost Constraint) [10]

In this problem we explore channel coding with input and output cost constraint.

a) Consider a DMC $\left(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y}\right)$. Let $b : \mathcal{X} \times \mathcal{Y} \to [0,\infty)$ be an input-output cost function. Suppose the channel coding has to satisfy the following average cost constraint: for each codeword x^n,

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathsf{E}_{Y_i}[b(x_i, Y_i)] \le \mathsf{B}.$$

Note that Y_i follows distribution $\mathsf{P}_{Y|X}(\cdot|x_i)$.

Argue that the problem is equivalent to another channel coding problem with a properly defined input-only cost function. Show that the capacity-cost function is

$$\mathrm{C}(\mathsf{B}) = \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X \mathsf{P}_{Y|X}}[b(X,Y)] \le \mathsf{B}} \mathrm{I}(X;Y)\,. \qquad [6]$$

Hint: Consider the input-only cost function $\tilde{b}(x) := \mathsf{E}[b(x,Y)]$, and check that the steps in the proof of DMC with input cost in the lecture are still valid.

b) Using discretization techniques, the above DMC result can be extended to continuous memoryless channels. With the extension (no need to prove it here), let us consider an

AWGN channel with *average output power constraint*

$$\tfrac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[Y_i^2\right] \le \mathsf{B}.$$

where $Y = X + Z$, $Z \perp\!\!\!\perp X$, and $Z \sim \mathrm{N}(0, \sigma^2)$.

Evaluate the channel capacity $\mathrm{C}(\mathsf{B})$. [4]

Solution:

a) Follow the hint, let $\tilde{b}(x) := \mathsf{E}[b(x, Y)]$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{b}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}_{Y_i}\left[b(x_i, Y_i)\right] \le \mathsf{B}.$$

Therefore, it can be viewed as an input-only cost constraint.
Furthermore, the capacity is

$$\begin{aligned}\mathrm{C}(\mathsf{B}) &= \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X}[\tilde{b}(x)] \le \mathsf{B}} \mathrm{I}(X;Y) = \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X}\left[\mathsf{E}_{Y \sim \mathsf{P}_{Y|X}}[b(X,Y)]\right] \le \mathsf{B}} \mathrm{I}(X;Y) \\ &= \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X \mathsf{P}_{Y|X}}[b(X,Y)] \le \mathsf{B}} \mathrm{I}(X;Y)\,.\end{aligned}$$

b) The capacity can be directed calculated via a), note that for random variable Y satisfying $\mathsf{E}_{\mathsf{P}_{X,Y}}\left[Y^2\right] \le \mathsf{B}$:

$$\mathrm{h}(Y) - \mathrm{h}(Y|X) \le \frac{1}{2}\log\left(2\pi e \mathsf{B}\right) - \mathrm{h}(Z) = \frac{1}{2}\log\left(\frac{\mathsf{B}}{\sigma^2}\right).$$

The equality can be achieved by choosing $X \sim \mathrm{N}(0, \mathsf{B} - \sigma^2)$.
Hence,

$$\mathrm{C}(\mathsf{B}) = \frac{1}{2}\log\left(\frac{\mathsf{B}}{\sigma^2}\right)$$

4. (Compression with guarantee on the cross-entropy loss) [20]

Consider a discrete memoryless source $S \sim \pi$ with a finite alphabet $\mathcal{S} = \{1, 2, ..., k\}$, $|\mathcal{S}| = k < \infty$. The encoder aims to compress the source so that the decoder can give good estimates of the source sequence. In many applications, however, the decoder may not want to give a deterministic estimate. Instead, for each symbol s_i in a length-n sequence s^n, its goal is to produce a *probability vector* $\boldsymbol{q}_i$ in the k-dimensional probability simplex $\mathcal{P}_k$, where the l-th coordinate, $q_i(l)$, stands for the probability of $s_i = l$ that the decoder believes in based on what it receives from the encoder. A standard way to quantify the loss is the empirical cross entropy loss

$$\ell_{\mathrm{CE}}(s^n, \boldsymbol{q}^n) = \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{q_i(s)}.$$

Note that it can be viewed as the average distortion per symbol when the distortion function is set to be

$$d : \mathcal{S} \times \mathcal{P}_d \to [0, \infty),\ (s, \boldsymbol{q}) \mapsto d(s, \boldsymbol{q}) = \log \frac{1}{q(s)}.$$

Hence, one can study a lossy source coding problem to understand how to represent a memoryless source with the smallest rate so that the decoder can declare an estimation probability vector with the empirical cross entropy loss not greater than a prescribed level D. By the lossy source coding theorem, the rate is given by the following rate distortion function:

$$\mathrm{R}(\mathsf{D}) = \inf_{(S,\boldsymbol{Q})} \left\{ \mathrm{I}(S;\boldsymbol{Q}) \,\middle|\, \mathsf{E}\left[\log \frac{1}{Q(S)}\right] \le \mathsf{D} \text{ and } S \sim \pi \right\}$$

a) Show that for the lossy source coding problem, $\mathsf{D}_{\min} = 0$ and $\mathsf{D}_{\max} = \mathrm{H}(\pi)$.

b) Show that for any jointly distributed $(S, \boldsymbol{Q}) \sim P$,

$$\mathrm{H}(S|\boldsymbol{Q}) \le \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log \frac{1}{Q(S)}\right].$$

Then, argue that $\mathrm{R}(\mathsf{D}) \ge \mathrm{H}(\pi) - \mathsf{D}$, for $0 \le \mathsf{D} \le \mathrm{H}(\pi)$.

c) Show that for $0 \le \mathsf{D} \le \mathrm{H}(\pi)$,

$$\mathrm{R}(\mathsf{D}) \le \min_{(S,\hat{S}),\ \hat{S}\in\mathcal{S}} \left\{ \mathrm{I}\left(S;\hat{S}\right) \,\middle|\, \mathrm{H}\left(S\middle|\hat{S}\right) \le \mathsf{D} \text{ and } S \sim \pi \right\}.$$

d) Show that for $0 \le \mathsf{D} \le \mathrm{H}(\pi)$,

$$\mathrm{R}(\mathsf{D}) = \min_{(S,\hat{S})} \left\{ \mathrm{I}\left(S;\hat{S}\right) \,\middle|\, \mathrm{H}\left(S\middle|\hat{S}\right) \le \mathsf{D} \text{ and } S \sim \pi \right\} = \mathrm{H}(\pi) - \mathsf{D}.$$

Hence, $\mathrm{R}(\mathsf{D}) = \max\{0, \mathrm{H}(\pi) - \mathsf{D}\}$.

Solution:

a) For $\mathsf{D}_{\min}$, one could choose $q_l(s) = \begin{cases} 1 & l = s \\ 0 & otherwise \end{cases}$

$\mathsf{E}_S[d(S, q(S))] = \mathsf{E}_S[\log 1] = 0$, so $\mathsf{D}_{\min} = 0$.

For $\mathsf{D}_{\max}$, since

$$\begin{aligned} \mathsf{E}_S\left[\log \frac{1}{\pi(S)}\right] - \mathsf{E}_S\left[\log \frac{1}{q(S)}\right] &= \mathsf{E}_S\left[\log \frac{q(S)}{\pi(S)}\right] \\ &\le \log \mathsf{E}_S\left[\frac{q(S)}{\pi(S)}\right] = 0 \quad \forall q. \end{aligned}$$

$\mathsf{D}_{\max} = \min_q \mathsf{E}_S[d(S, q)] = \mathsf{E}_S\left[\log \frac{1}{\pi(S)}\right] = \mathrm{H}(\pi)$.

b) Denote the law of P as $\mathsf{P}_{S,\boldsymbol{Q}}$ and its marginal as $\boldsymbol{Q}$ as $\mathsf{P}_{\boldsymbol{Q}}$

$$
\begin{aligned}
\mathrm{H}(S|\boldsymbol{Q}) - \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] &= \mathrm{h}(S,\boldsymbol{Q}) - \mathrm{h}(\boldsymbol{Q}) - \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] \\
&= \mathsf{E}_{(S,\mathbf{Q})\sim P}\left[\log\frac{Q(S)P_{\mathbf{Q}}(\mathbf{Q})}{\mathsf{P}_{S,\mathbf{Q}}(S,\mathbf{Q})}\right] \\
&\le \log \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\frac{Q(S)\mathsf{P}_{\boldsymbol{Q}}(\boldsymbol{Q})}{\mathsf{P}_{S,\boldsymbol{Q}}(S,\boldsymbol{Q})}\right] \\
&= \log\left(\sum_{s\in\mathcal{S}}\int_{\boldsymbol{q}\in\mathcal{P}_k} q(s)\mathsf{P}_{\boldsymbol{Q}}(\boldsymbol{q})\right) = 0
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{R}(\mathsf{D}) &= \inf_{(S,\boldsymbol{Q})}\left\{\mathrm{I}(S;\boldsymbol{Q})\middle|\mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right]\le \mathsf{D} \text{ and } S\sim\pi\right\} \\
&= \mathrm{H}(\pi) - \sup_{(S,\boldsymbol{Q})}\left\{\mathrm{H}(S|\mathbf{Q})\middle|\mathsf{E}_{(S,\mathbf{Q})\sim P}\left[\log\frac{1}{Q(S)}\right]\le \mathsf{D} \text{ and } S\sim\pi\right\} \\
&\ge \mathrm{H}(\pi) - \mathsf{D}, \quad \forall \mathsf{D}_{\min} = 0 \le \mathsf{D} \le \mathsf{D}_{\max} = \mathrm{H}(\pi)
\end{aligned}
$$

c) The key to this problem is to make the observation that, for all pair of random variable $(S,\hat{S})$ over $\mathcal{S}\times\mathcal{S}$, we can associate a random vector $\boldsymbol{Q}_{S,\hat{S}}$ over $\mathcal{P}_k$ such that

$$
\begin{aligned}
\boldsymbol{Q}_{S,\hat{S}} &= \boldsymbol{q}_i = \mathsf{P}_{S|\hat{S}}(\cdot|i) \text{ if } \hat{S} = i \\
\text{equivalently } \boldsymbol{Q}_{S,\hat{S}} &= \sum_{i=1}^{k}\mathsf{P}_{S|\hat{S}}(\cdot|i)\mathbb{1}\left\{\hat{S}=i\right\}
\end{aligned}
$$

a mixture of k vectors $\boldsymbol{q}_i$ determined by the outcome of $\hat{S}$. Note that this specify a Markov chain $S-\hat{S}-\boldsymbol{Q}_{S,\hat{S}}$. For this Markov chain, by the data processing inequality, $\mathrm{I}\left(S;\boldsymbol{Q}_{S,\hat{S}}\right)\le\mathrm{I}\left(S;\hat{S}\right)$. And

$$
\mathsf{E}_{(S,\boldsymbol{Q}_{S,\hat{S}})}\left[\log\frac{1}{Q_{S,\hat{S}}(S)}\right] = \mathsf{E}_{(S,\hat{S})}\left[\log\frac{1}{\mathsf{P}_{S|\hat{S}}(S|\hat{S})}\right]
$$

Hence,

$$
\begin{aligned}
\mathrm{R}(\mathsf{D}) &= \inf_{(S,\boldsymbol{Q})}\left\{\mathrm{I}(S;\boldsymbol{Q})\middle|\mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right]\le\mathsf{D} \text{ and } S\sim\pi\right\} \\
&= \inf_{(S,\hat{S},\boldsymbol{Q})}\left\{\mathrm{I}(S;\boldsymbol{Q})\middle|\mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right]\le D \text{ and } S\sim\pi\right\} \\
&\le \min_{(S,\hat{S})}\left\{\mathrm{I}(S;\boldsymbol{Q})\middle|\mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right]\le D \text{ and } S\sim\pi \text{ and } \boldsymbol{Q}=\boldsymbol{Q}_{S,\hat{S}}\right\}
\end{aligned}
$$

$$\leq \min_{(S,\hat{S})} \left\{ \mathrm{I}\left(S;\hat{S}\right) \middle| \mathsf{E}_{(S,\hat{S})}\left[\log \frac{1}{\mathsf{P}_{S|\hat{S}}(S|\hat{S})}\right] \leq \mathsf{D} \text{ and } S \sim \pi \right\}$$
$$= \min_{(S,\hat{S})} \left\{ \mathrm{I}\left(S;\hat{S}\right) \middle| \mathrm{H}\left(S \middle| \hat{S}\right) \leq \mathsf{D} \text{ and } S \sim \pi \right\}$$

d) $\mathrm{R}(\mathsf{D}) = \mathrm{H}(\pi) - \mathsf{D}$ since

$$\begin{aligned} \mathrm{I}\left(S;\hat{S}\right) &= \mathrm{H}(S) - \mathrm{H}\left(S \middle| \hat{S}\right) \\ &= \mathrm{H}(\pi) - \mathrm{H}\left(S \middle| \hat{S}\right) \\ &\geq \mathrm{H}(\pi) - \mathsf{D}. \end{aligned}$$

The equality holds since $0 \leq \mathsf{D} \leq \mathrm{H}(\pi)$, $\exists \mathsf{P}_{S,\hat{S}}$ s.t. $\mathrm{H}\left(S \middle| \hat{S}\right) = \mathsf{D}$.