Homework 2 Solution and Grading Policy

TA: Kai-Chun Chen

Homework Policy: (Read before You Start to Work)

- Copying from other students' solution is not allowed. If caught, all involved students get 0 point on that particular homework. Caught twice, you will be asked to drop the course.
- Collaboration is welcome. You can work together with **at most one partner** on the homework problems which you find difficult. However, you should write down your own solution, not just copying from your partner's.
- Your partner should be the same for the entire homework.
- Put your collaborator's name beside the problems that you collaborate on.
- When citing known results from the assigned references, be as clear as possible.

1. (Mixture of random processes) [14]

In this problem we look at different ways to generate mixtures of random processes, and the entropy rate of the mixture of random processes. Consider two stationary random processes $\{X_0[i] \,|\, i \in \mathbb{N}\}$ and $\{X_1[i] \,|\, i \in \mathbb{N}\}$ taking values in disjoint alphabets $\mathcal{X}_0$ and $\mathcal{X}_1$ respectively. The two processes are independent from each other, that is, $\{X_0[i]\} \perp\!\!\!\perp \{X_1[i]\}$, and they have entropy rates $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively. Let $\{\Theta_i \,|\, i \in \mathbb{N}\}$ be a **stationary** Bernoulli random process, independent of everything else.

a) Let $\Theta_i = \Theta$ for all $i \in \mathbb{N}$, where $\Theta \sim \mathrm{Ber}(q)$. Is the random process $\{X_{\Theta_i}[i]\}$ stationary? What is its entropy rate? [6]

b) Let $\{\Theta_i\}$ be Markov with a probability transition matrix

$$\mathsf{P}_{\Theta_2|\Theta_1} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}, \text{ for } \alpha, \beta \in (0,1).$$

Suppose that both $\{X_0[i]\}$ and $\{X_1[i]\}$ are i.i.d. processes in this problem. Is the random process $\{X_{\Theta_i}[i]\}$ stationary? What is its entropy rate? [8]

Solution:

a) Since $\{X_0[i]\}$ and $\{X_1[i]\}$ are stationary, we have

$$\begin{aligned}&\mathsf{P}_{X_{\Theta_1}[1],\dots,X_{\Theta_n}[n]}\\&=(1-q)\mathsf{P}_{X_0[1],\dots,X_0[n]}+q\mathsf{P}_{X_1[1],\dots,X_1[n]}\\&=(1-q)\mathsf{P}_{X_0[l+1],\dots,X_0[n]}+q\mathsf{P}_{X_1[1],\dots,X_1[l+n]}\quad\text{(by stationariness)}\\&=\mathsf{P}_{X_{\Theta_{l+1}}[l+1],\dots,X_{\Theta_{l+n}}[l+n]}\end{aligned}$$

By definition, $\{X_{\Theta_i}[i]\}$ is stationary.
Let $Y_i = X_{\Theta_i}[i]$.

$$\begin{aligned}\mathcal{H}(X_{\Theta_i}[i]) &= \lim_{n\to\infty} \mathrm{H}\left(Y_n \middle| Y^{n-1}\right)\\&=\lim_{n\to\infty} \mathrm{H}\left(Y_n,\Theta_n \middle| Y^{n-1},\Theta^{n-1}\right) && \text{(since } \mathcal{X}_0 \text{ and } \mathcal{X}_1 \text{ are disjoint)}\\&=\lim_{n\to\infty} \mathrm{H}\left(Y_n \middle| Y^{n-1},\Theta^{n}\right)+\mathrm{H}\left(\Theta_n \middle| Y^{n-1},\Theta^{n-1}\right) && \text{(by chain rule)}\\&=\lim_{n\to\infty} \mathrm{H}(Y_n|\Theta_n)+\mathrm{H}(\Theta_n|\Theta_{n-1})\\&=\lim_{n\to\infty} \mathrm{H}(Y_n|\Theta)+\mathrm{H}(\Theta|\Theta)\\&=q\mathcal{H}_1+(1-q)\mathcal{H}_0+0\end{aligned}$$

b) Since we know in advance that $\{\Theta_i\}$ is stationary, we can conclude that $\mathsf{Pr}\{\Theta_1=0\} = \frac{\beta}{\alpha+\beta}$ and $\mathsf{Pr}\{\Theta_1=1\} = \frac{\alpha}{\alpha+\beta}$.
As a result, $\{X_{\Theta_i}[i]\}$ can be shown to be stationary by decomposing $\mathsf{P}_{X_{\Theta_1}[1],\dots,X_{\Theta_n}[n]}$ and $\mathsf{P}_{X_{\Theta_{l+1}}[+1],\dots,X_{\Theta_{l+n}}[l+n]}$ simply by showing that $\mathsf{Pr}\{\Theta_n=0\} = \frac{\beta}{\alpha+\beta}$ and $\mathsf{Pr}\{\Theta_n=1\} = \frac{\alpha}{\alpha+\beta}$ for all n.

Next, let $Y_i = X_{\Theta_i}[i]$. $\mathcal{H}(\{X_{\Theta_i}[i]\}) = \lim_{n\to\infty}\mathrm{H}(Y_n|Y^{n-1})$.

$$\begin{aligned}&\mathrm{H}\left(Y_n \middle| Y^{n-1}\right)\\&=\mathrm{H}\left(Y_n,\Theta_n \middle| Y^{n-1}\right) && (\mathcal{X}_0\cap\mathcal{X}_1=\varnothing)\\&=\mathrm{H}\left(Y_n \middle| \Theta_n, Y^{n-1}\right)+\mathrm{H}\left(\Theta_n \middle| Y^{n-1}\right) && \text{(by chain rule)}\\&=\mathsf{Pr}\{\Theta_n=1\}\mathrm{H}\left(X_1[n] \middle| X_1^{n-1}\right)+\mathsf{Pr}\{\Theta_n=0\}\mathrm{H}\left(X_0[n] \middle| X_0^{n-1}\right)+\mathrm{H}\left(\Theta_n \middle| Y^{n-1}\right)\\&=\mathsf{Pr}\{\Theta_n=1\}\mathrm{H}\left(X_1[n] \middle| X_1^{n-1}\right)+\mathsf{Pr}\{\Theta_n=0\}\mathrm{H}\left(X_0[n] \middle| X_0^{n-1}\right)+\mathrm{H}\left(\Theta_n \middle| \Theta^{n-1}\right) && (\mathcal{X}_0\cap\mathcal{X}_1=\varnothing)\\&=\frac{\alpha}{\alpha+\beta}\mathcal{H}_1+\frac{\beta}{\alpha+\beta}\mathcal{H}_0+\mathrm{H}(\Theta_2|\Theta_1)\\&=\frac{\alpha}{\alpha+\beta}(\mathcal{H}_1+\mathsf{H_b}(\beta))+\frac{\beta}{\alpha+\beta}(\mathcal{H}_0+\mathsf{H_b}(\alpha))\end{aligned}$$

Grading Policy:

a) Stationaryiness and reason [2] calculation of entropy rate [4]

b) Stationaryiness and reason [2] calculation of entropy rate [6]

2. (Binary hypothesis testing) [16]

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. Bernoulli p random variables, that is,

$$\Pr\{X_i = 1\} = 1 - \Pr\{X_i = 0\} = p.$$

Based on the observations so far, the goal is of a decision maker to determine which of the following two hypotheses is true:

$$\begin{aligned} \mathcal{H}_0 :& \quad p = p_0 \\ \mathcal{H}_1 :& \quad p = p_1 \end{aligned}$$

where $0 < p_0 < p_1 \leq 1/2$.

a) (Warm-up) Consider the problem of making the decision based on X_1.
Draw the optimal $(\pi_{1|0}, \pi_{0|1})$ trade-off curve. [4]

b) Suppose the decision maker waits until an 1 appears and makes the decision based on the whole observed sequence. Sketch the optimal $(\pi_{1|0}, \pi_{0|1})$ trade-off curve. [4]

c) Now suppose the decision maker waits until in total n 1's appear and makes the decision based on the whole observed sequence. Let $\varpi^*_{0|1}(n, \epsilon)$ denote the minimum type-II error probability subject to the constraint that the type-I error probability is not greater than ϵ, $0 < \epsilon < 1$. Does $\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varpi^*_{0|1}(n,\epsilon)}$ exist? If so, find it. Otherwise, show that the limit does not exist. [8]

Solution:

a) By Neyman-Pearson theorem, the optimal test is randomized LRT. So, by leveraging the parameters of the randomized LRT, that is, τ and γ, we can derive the optimal trade-off curve. Note that the likelihood ratio can only take two values: $\frac{p_1}{p_0}, \frac{1-p_1}{1-p_0}$. Therefore, discuss the range of τ we get

$$\begin{cases} \pi_{1|0} = 1, \pi_{0|1} = 0, & 0 \leq \tau < \frac{1-p_1}{1-p_0} \\ \pi_{1|0} = p_0 + \gamma(1-p_0), \pi_{0|1} = (1-\gamma)(1-p_1) = \frac{1-p_1}{1-p_0}(1-\pi_{1|0}), & \tau = \frac{1-p_1}{1-p_0} \\ \pi_{1|0} = p_0, \pi_{0|1} = 1 - p_1, & \frac{1-p_1}{1-p_0} < \tau < \frac{p_1}{p_0} \\ \pi_{1|0} = \gamma p_0, \pi_{0|1} = (1-\gamma)p_1 + (1-p_1) = 1 - \frac{p_1}{p_0}\pi_{1|0}, & \tau = \frac{p_1}{p_0} \\ \pi_{1|0} = 0, \pi_{0|1} = 1, & \tau > \frac{p_1}{p_0}. \end{cases}$$

We can then draw the trade-off curve using the equations derived above.

b) Note that our observation can only be $1, 01, 001, 0001, \cdots$, let L be the length of the observation, we have

$$\begin{aligned} \mathcal{H}_0 :& \quad L \sim \mathrm{Geo}(p_0) \\ \mathcal{H}_1 :& \quad L \sim \mathrm{Geo}(p_1) \end{aligned}$$

Similar to a), we can discuss the range of τ and get:

$$\begin{cases} \pi_{1|0} = 0, \pi_{0|1} = 1,\ \tau > \frac{p_1}{p_0} \\ \pi_{1|0} = \sum\limits_{i=1}^{n-1} (1-p_0)^{i-1} p_0 + \gamma (1-p_0)^{n-1} p_0, \\ \pi_{0|1} = \sum\limits_{i=n+1}^{\infty} (1-p_1)^{i-1} p_1 + (1-\gamma)(1-p_1)^{n-1} p_1,\ \tau = \frac{(1-p_1)^{n-1} p_1}{(1-p_0)^{n-1} p_0} \\ \pi_{1|0} = \sum_{i=1}^{n} (1-p_0)^{i-1} p_0, \pi_{0|1} = \sum_{i=n+1}^{\infty} (1-p_1)^{i-1} p_1,\ \frac{(1-p_1)^{n-1} p_1}{(1-p_0)^{n-1} p_0} > \tau > \frac{(1-p_1)^{n} p_1}{(1-p_0)^{n} p_0}. \end{cases}$$

And we can draw the trade-off curve using the equations derived above.

c) The observation can be viewed as n i.i.d. geometric random variables. To see this, for any realization of observation, insert a "—" symbol in front of the sequence, also insert a "—" right after a "1". For example, if $n = 4$ and the realization is 010001101, we write it as |01|0001|1|01|. Apperantly, the length of the subsequence between two — is a geometric random variable. Hence, in this subproblem, we are testing $\text{Geo}(p_0)^{\otimes n}$ and $\text{Geo}(p_1)^{\otimes n}$. By Chernoff-Stein lemma,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varpi^*_{0|1}(n,\epsilon)} = \mathrm{D}\left(\text{Geo}(p_0) \| \text{Geo}(p_1)\right) = \log \frac{p_0}{p_1} + \left(\frac{1-p_0}{p_0}\right) \log \frac{1-p_0}{1-p_1}.$$

Grading Policy:

a) Specify the trade-off curve [2] Argue the optimality [2]

b) Specify the trade-off curve [2] Argue the optimality [2]

c) Formulate the problem as a hypothesis testing with n instances [3], Chernoff-Stein lemma [2], and calculation [3]

3. (Mixture of information divergences) [8]

For m discrete probability distributions $\mathsf{P}_1, \mathsf{P}_2, \ldots, \mathsf{P}_m$ with the same support $\mathcal{X}$, consider the following minimization problem:

$$\min_{Q \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^{m} \lambda_i \mathrm{D}\left(P_i \| Q\right),$$

where $\mathcal{P}(\mathcal{X})$ denotes the collection of probability distributions over $\mathcal{X}$, $\sum_{i=1}^m \lambda_i = 1$, and $\lambda_i > 0$ for $i = 1, 2, \ldots, m$. Show that $\sum_{i=1}^m \lambda_i P_i$ is a minimizer to the above problem.

Solution:

Let $\overline{\mathsf{P}} = \sum\limits_{i=1}^{m} \lambda_i \mathsf{P}_i$, we have $\forall \mathsf{Q} \in \mathcal{P}(\mathcal{X})$, that

$$\sum_{i=1}^{m} \lambda_i \mathrm{D}\left(\mathsf{P}_i \| \mathsf{Q}\right) - \sum_{i=1}^{m} \lambda_i \mathrm{D}\left(\mathsf{P}_i \middle\| \overline{\mathsf{P}}\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} \sum_{x \in \mathcal{X}} \lambda_i \mathsf{P}_i(x) \log \frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
&= \sum_{x \in \mathcal{X}} \left(\sum_{i=1}^{m} \lambda_i \mathsf{P}_i(x) \right) \log \frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
&= \sum_{x \in \mathcal{X}} \overline{\mathsf{P}}(x) \log \frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
&= \mathrm{D}\left(\overline{\mathsf{P}} \middle\| \mathsf{Q}\right) \geq 0.
\end{aligned}
$$

Hence, $\overline{\mathsf{P}}$ is a minimizer.

Grading Policy

Reasonable Procedure[4] Correctness [4] Per mistake [-1]

4. (Rényi's divergence) [12]

Alfréd Rényi introduced the following generalization of information divergence called *Rényi's divergence of order* α (for simplicity, only deal with the discrete case):

$$
\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) := \frac{1}{\alpha - 1} \log \left(\sum_{a \in \mathcal{X}} \mathsf{P}(a)^\alpha \mathsf{Q}(a)^{1-\alpha} \right), \quad \alpha \in (0,1) \cup (1, \infty),
$$

where P, Q are both probability distributions over a finite alphabet $\mathcal{X}$, and $\operatorname{supp} \mathsf{P} \subseteq \operatorname{supp} \mathsf{Q}$.

a) (Non-negativity) Show that $\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) \geq 0$, with equality if and only if $\mathsf{P} = \mathsf{Q}$. [4]

b) (Relation with KL divergence) Show that $\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) \geq \mathrm{D}(\mathsf{P}\|\mathsf{Q})$ for $\alpha > 1$ and $\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) \leq \mathrm{D}(\mathsf{P}\|\mathsf{Q})$ for $\alpha < 1$. Furthermore, $\lim_{\alpha \to 1} \mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) = \mathrm{D}(\mathsf{P}\|\mathsf{Q})$. [4]

c) (Data processing) Show that $\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q})$ satisfies the data processing inequality. [4]

Solution:

a) We divide the value of α into two cases:
If $\alpha \in (0,1)$, we can lower bound $\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q})$ by Hölder's inequality, which states that for any $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, we have

$$
\sum_{i=1}^{n} |x_i y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |y_i|^q \right)^{\frac{1}{q}},
$$

the equality holds iff $(|x_1|^p, ..., |x_n|^p)$ and $(|y_1|^q, ..., |y_n|^q)$ are linearly dependent.

$$
\begin{aligned}
\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) &= \frac{1}{\alpha - 1} \log \left(\sum_{a \in \mathcal{X}} \mathsf{P}(a)^\alpha \mathsf{Q}(a)^{1-\alpha} \right) \\
&\geq \frac{1}{\alpha - 1} \log \left(\sum_{a \in \mathcal{X}} (\mathsf{P}(a)^\alpha)^{\frac{1}{\alpha}} \right)^{\alpha} \left(\sum_{a \in \mathcal{X}} (\mathsf{Q}(a)^{1-\alpha})^{\frac{1}{1-\alpha}} \right)^{1-\alpha}
\end{aligned}
$$

$$(\text{by Hölder's inequality with } p = \frac{1}{\alpha}, q = \frac{1}{1-\alpha})$$
$$= 0$$

If $\alpha \in (1, \infty)$, we can bound it by Jensen's inequality,

$$\begin{aligned}
\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) &= \frac{1}{\alpha-1}\log\left(\sum_{a\in\mathcal{X}}\mathsf{P}(a)^\alpha\mathsf{Q}(a)^{1-\alpha}\right) \\
&= \frac{1}{\alpha-1}\log\left(\sum_{a\in\mathcal{X}}\mathsf{P}(a)\left(\frac{\mathsf{P}(a)}{\mathsf{Q}(a)}\right)^{\alpha-1}\right) \\
&= \frac{1}{\alpha-1}\log\mathsf{E}_{X\sim\mathsf{P}}\left[\left(\frac{\mathsf{P}(X)}{\mathsf{Q}(X)}\right)^{\alpha-1}\right] \\
&\geq \frac{1}{\alpha-1}\mathsf{E}_{X\sim\mathsf{P}}\left[\log\left(\frac{\mathsf{P}(X)}{\mathsf{Q}(X)}\right)^{\alpha-1}\right] \\
&= \mathsf{E}_{X\sim\mathsf{P}}\left[\log\frac{\mathsf{P}(X)}{\mathsf{Q}(X)}\right] = \mathrm{D}(\mathsf{P}\|\mathsf{Q}) \geq 0
\end{aligned}$$

Furthermore, both equality holds iff $\mathsf{P}(a) = \mathsf{Q}(a)$ for all $a \in \mathcal{X}$, which is $\mathsf{P} = \mathsf{Q}$.

b) In a), we have proved the case when $\alpha > 1$. For the case when $\alpha < 1$, we can directly obtain the result by substituting the inequality since $\frac{1}{\alpha-1} < 0$ here. Thus, it suffices to show that $\lim_{\alpha\to1}\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) = \mathrm{D}(\mathsf{P}\|\mathsf{Q})$.

$$\begin{aligned}
\lim_{\alpha\to1}\mathrm{D}_\alpha(\mathsf{P}\|\mathsf{Q}) &= \lim_{\alpha\to1}\frac{\log\left(\sum_{a\in\mathcal{X}}\mathsf{P}(a)^\alpha\mathsf{Q}(a)^{1-\alpha}\right)}{\alpha-1} \\
&\overset{H}{=} \lim_{\alpha\to1}\frac{\sum_{a\in\mathcal{X}}\mathsf{P}(a)^\alpha\mathsf{Q}(a)^{1-\alpha}\log\left(\frac{\mathsf{P}(a)}{\mathsf{Q}(a)}\right)}{\sum_{a\in\mathcal{X}}\mathsf{P}(a)^\alpha\mathsf{Q}(a)^{1-\alpha}} \\
&= \sum_{a\in\mathcal{X}}\mathsf{P}(a)\frac{\mathsf{P}(a)}{\mathsf{Q}(a)} = \mathrm{D}(\mathsf{P}\|\mathsf{Q})
\end{aligned}$$

c) Follow the notation proving data processing inequality used in lecture, we have $\forall x, y$,

$$\frac{\mathsf{P}_{X,Y}(x,y)}{\mathsf{Q}_{X,Y}(x,y)} = \frac{\mathsf{P}_X(x)W_{Y|X}(y|x)}{\mathsf{Q}_X(x)W_{Y|X}(y|x)} = \frac{\mathsf{P}_X(x)}{\mathsf{Q}_X(x)}$$

As a result,

$$\begin{aligned}
\mathrm{D}_\alpha(\mathsf{P}_{X,Y}\|\mathsf{Q}_{X,Y}) &= \frac{1}{\alpha-1}\sum_{a\in\mathcal{X}}\sum_{b\in\mathcal{Y}}\mathsf{P}_{X,Y}(a,b)\left(\frac{\mathsf{P}_{X,Y}(a,b)}{\mathsf{Q}_{X,Y}(a,b)}\right)^{\alpha-1} \\
&= \frac{1}{\alpha-1}\sum_{a\in\mathcal{X}}\mathsf{P}_X(a)\left(\frac{\mathsf{P}_X(a)}{\mathsf{Q}_X(a)}\right)^{\alpha-1} = \mathrm{D}_\alpha(\mathsf{P}_X\|\mathsf{Q}_X)
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\mathrm{D}_\alpha(\mathsf{P}_X\|\mathsf{Q}_X) &= \mathrm{D}_\alpha(\mathsf{P}_{X,Y}\|\mathsf{Q}_{X,Y}) \\
&= \frac{1}{\alpha-1}\mathsf{E}_{X,Y\sim\mathsf{Q}_{X,Y}}\left[\left(\frac{\mathsf{P}_{X,Y}(X,Y)}{\mathsf{Q}_{X,Y}(X,Y)}\right)^\alpha\right] \\
&= \frac{1}{\alpha-1}\mathsf{E}_{Y\sim\mathsf{Q}_Y}\left[\mathsf{E}_{X\sim\mathsf{Q}_{X|Y=Y}}\left[\left(\frac{\mathsf{P}_{X,Y}(X,Y)}{\mathsf{Q}_{X,Y}(X,Y)}\right)^\alpha \middle| Y\right]\right] \\
&\geq \frac{1}{\alpha-1}\mathsf{E}_{Y\sim\mathsf{Q}_Y}\left[\mathsf{E}_{X\sim\mathsf{Q}_{X|Y=Y}}\left[\left(\frac{\mathsf{P}_{X,Y}(X,Y)}{\mathsf{Q}_{X,Y}(X,Y)}\right) \middle| Y\right]^\alpha\right] \\
&= \frac{1}{\alpha-1}\mathsf{E}_{Y\sim\mathsf{Q}_Y}\left[\left(\frac{\mathsf{P}_Y(Y)}{\mathsf{Q}_Y(Y)}\right)^\alpha\right] = \mathrm{D}_\alpha(\mathsf{P}_Y\|\mathsf{Q}_Y)
\end{aligned}
$$

Grading Policy

a) Proof of inequality [3] Equivalent condition of equality [1]

b) Proof of inequalities [2] Calculation of limit [2]

c) Correct proof [4] Wrong for some α [-1]