September 28 **September 2019**

Tutorial 7: Problems

Course Instructor: Himanshu Tyagi *Prepared by: Karthik*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Contents

7.1 Problems

1. Let $X \sim P$ be a discrete random variable with pmf p taking values in the set $\mathcal{X}$. For $\alpha > 0$, $\alpha \neq 1$, the Rényi entropy of order α of the distribution P is denoted by $H_\alpha(P)$ and is defined as

$$H_\alpha(P) \coloneqq \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} (p(x))^\alpha.$$

Show that for every $\alpha \in (0,1)$ and $\epsilon > 0$,

$$L_\epsilon(P) \leq H_\alpha(P) + \frac{1}{1-\alpha} \log \frac{1}{\epsilon} + 1.$$

Also show that for every $\beta > 1$ and $\delta \in (0, 1-\epsilon)$,

$$L_\epsilon(P) \geq H_\beta(P) - \frac{1}{\beta - 1} \log \frac{1}{\delta} - \log \frac{1}{1-\epsilon-\delta}.$$

Solution:
We know[1] that given $\epsilon > 0$ and a discrete random variable $X \sim P$ with pmf p, if there exists a constant λ such that $P(-\log p(X) \leq \lambda) \geq 1 - \epsilon$, then $L_\epsilon(P) \leq \lambda$. From the upper bound for $L_\epsilon(P)$ given in the question, a natural choice for λ is

$$\lambda = \lambda_\alpha = H_\alpha(P) + \frac{1}{1-\alpha} \log \frac{1}{\epsilon} + 1.$$

Therefore, it suffices to show that for this choice of λ_α, $P(-\log p(X) \leq \lambda_\alpha) \geq 1 - \epsilon$.
We shall show that $P(-\log p(X) > \lambda_\alpha) \leq \epsilon$. Towards this, we have

$$\begin{aligned}
P(-\log p(X) > \lambda_\alpha) &= P\Bigg(-\log p(X) > \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} (p(x))^\alpha + \frac{1}{1-\alpha} \log \frac{1}{\epsilon} + 1 \Bigg) \\
&\leq P\Bigg(-\log p(X) > \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} (p(x))^\alpha + \frac{1}{1-\alpha} \log \frac{1}{\epsilon} \Bigg) \\
&= P\Bigg((p(X))^{1-\alpha} < \frac{\epsilon}{\sum_{x \in \mathcal{X}} (p(x))^\alpha} \Bigg).
\end{aligned}$$

Denoting by A_α the set

$$A_\alpha \coloneqq \Bigg\{ x \in \mathcal{X} : (p(x))^{1-\alpha} < \frac{\epsilon}{\sum_{x \in \mathcal{X}} (p(x))^\alpha} \Bigg\},$$

it follows from the above set of inequalities that

$$\begin{aligned}
P(-\log p(X) > \lambda_\alpha) &\leq P(A_\alpha) \\
&= \sum_{x \in A_\alpha} p(x) \\
&= \sum_{x \in A_\alpha} (p(x))^\alpha (p(x))^{1-\alpha} \\
&\leq \sum_{x \in A_\alpha} (p(x))^\alpha \frac{\epsilon}{\sum_{x \in \mathcal{X}} (p(x))^\alpha} \\
&\leq \epsilon.
\end{aligned}$$

[1] See Lemma 3.2.1 in `https://drive.google.com/file/d/1id5BXdm9JCThZJVKgn-6vGzIPJUCrkZj/view`.

Hence, the first inequality follows.

On similar lines, we know[2] that given $\epsilon > 0$, $\delta \in (0, 1-\epsilon)$, and a discrete random variable $X \sim P$ with pmf p, if there exists a constant λ such that $P(-\log p(X) \geq \lambda) \geq 1-\delta$, then $L_\epsilon(P) \geq \lambda - \log \frac{1}{1-\epsilon-\delta}$. From the lower bound for $L_\epsilon(P)$ given in the question, a natural choice for λ is

$$\lambda = \lambda_\beta = H_\beta(P) - \frac{1}{\beta-1} \log \frac{1}{\delta}.$$

Therefore, it suffices to show that for this choice of λ_β, $P(-\log p(X) \geq \lambda_\beta) \geq 1-\delta$.

In what follows, we shall show that $P(-\log p(X) < \lambda_\beta) \leq \delta$. Towards this, we have

$$\begin{aligned} P(-\log p(X) < \lambda_\beta) &= P\bigg(-\log p(X) < H_\beta(P) - \frac{1}{\beta-1} \log \frac{1}{\delta}\bigg) \\ &= P\bigg(-\log p(X) < \frac{1}{\beta-1} \log \frac{\delta}{\sum\limits_{x \in \mathcal{X}} (p(x))^\beta}\bigg) \\ &= P\bigg((p(X))^{\beta-1} > \frac{\sum\limits_{x \in \mathcal{X}} (p(x))^\beta}{\delta}\bigg). \end{aligned}$$

Denoting by B_β the set

$$B_\beta \coloneqq \bigg\{ x \in \mathcal{X} : (p(x))^{\beta-1} > \frac{\sum\limits_{x \in \mathcal{X}} (p(x))^\beta}{\delta} \bigg\},$$

it follows from the above set of inequalities that

$$\begin{aligned} P(-\log p(X) < \lambda_\beta) &= P(B_\beta) \\ &= \sum_{x \in B_\beta} p(x) \\ &= \sum_{x \in B_\beta} \frac{(p(x))^\beta}{(p(x))^{\beta-1}} \\ &\leq \delta \cdot \frac{\sum\limits_{x \in B_\beta} (p(x))^\beta}{\sum\limits_{x \in \mathcal{X}} (p(x))^\beta} \\ &\leq \delta. \end{aligned}$$

The second inequality thus follows.

2. Consider random variables X and Y such that Y is uniformly distributed over $\{0,1\}^b$. Let $\hat{Y}$ be an estimate of Y obtained from observing X. Show that

$$P(\hat{Y} \neq Y) \geq 1 - \frac{I(X \wedge Y) + 1}{b}.$$

Now, suppose that $X_1, \ldots, X_n, Y$ are random variables such that $X_1, \ldots, X_n$ are mutually independent and Y is as before. For each $i \in \{1, \ldots, n\}$, let $\hat{Y}_i$ denote an estimate of Y obtained from observing X_i. Show that

$$\max_{i \leq i \leq n} P(\hat{Y}_i \neq Y) \geq 1 - \frac{1}{n} - \frac{1}{b}.$$

[2] See Lemma 3.2.2 in `https://drive.google.com/file/d/1id5BXdm9JCThZJVKgn-6vGzIPJUCrkZj/view`.

Solution:

Let $p = P(\hat{Y} \neq Y)$. Then, by Fano's inequality,

$$\begin{aligned} H(Y|X) &\leq p \log(|\{0,1\}^b| - 1) + h(p) \\ &\leq bp + 1, \end{aligned}$$

where in the last line above, $h(p)$ denotes the binary entropy function, which is upper bounded by 1. Also, we have

$$\begin{aligned} H(Y|X) &= H(Y) - I(X \wedge Y) \\ &= b - I(X \wedge Y). \end{aligned}$$

Therefore, we have

$$b - I(X \wedge Y) \leq bp + 1,$$

which upon rearrangement yields the first inequality.

In order to obtain the second inequality, we note that

$$\begin{aligned} \max_{1 \leq i \leq n} P(\hat{Y}_i \neq Y) &\geq \frac{1}{n} \sum_{i=1}^{n} P(\hat{Y}_i \neq Y) \\ &\overset{(a)}{\geq} \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{I(X_i \wedge Y) + 1}{b}\right) \\ &= 1 - \frac{1}{nb} \sum_{i=1}^{n} I(X_i \wedge Y) - \frac{1}{b}. \end{aligned}$$

where the last line is from the first inequality of the question proved before. Now, we note that

$$\begin{aligned} b &= H(Y) \\ &= I((X_1, \ldots, X_n) \wedge Y) + H(Y|X_1, \ldots, X_n) \\ &\geq I((X_1, \ldots, X_n) \wedge Y) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) - H(X_i|X_1, \ldots, X_{i-1}, Y) \\ &\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_i) - H(X_i|Y) \\ &= \sum_{i=1}^{n} I(X_i \wedge Y), \end{aligned}$$

where in (b) above, we combine the fact that X_i's are mutually independent along with the fact that conditioning reduces entropy. The second inequality of the question now follows.

3. Show that

$$H(X_1, X_2, X_3) \leq \frac{1}{2} \left(H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3) \right).$$

Also show that

$$H(X_1, X_2, X_3) \geq \frac{1}{2} \left(H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_1, X_3|X_2) \right).$$

Solution:
Using the chain rule for entropy, we have

$$\begin{aligned}
2H(X_1, X_2, X_3) &= H(X_1, X_2) + \underbrace{H(X_3|X_1, X_2)}_{\leq H(X_3)} + \underbrace{H(X_1|X_2, X_3)}_{\leq H(X_1|X_3)} + H(X_2, X_3) \\
&\overset{(a)}{\leq} H(X_1, X_2) + \underbrace{H(X_3) + H(X_1|X_3)}_{=H(X_1,X_3)} + H(X_2, X_3) \\
&= H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3),
\end{aligned}$$

where (a) above follows from the fact that conditioning reduces entropy.

On similar lines, we have

$$\begin{aligned}
2H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) + H(X_2) + H(X_1, X_3|X_2) \\
&\overset{(b)}{\geq} H(X_1, X_2) + H(X_2, X_3|X_1) + H(X_1, X_3|X_2) \\
&\geq H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_1, X_3|X_2),
\end{aligned}$$

where (b) above follows from the subadditivity of entropy, and the last line above follows from the fact that conditioning reduces entropy.

4. Show that X, Y and Z constitute a Markov chain (in the same order). Then, show that $I(X \wedge Y|Z) \leq I(X \wedge Y)$. When does equality hold? Give an example where $I(X \wedge Y|Z) > I(X \wedge Y)$.

Solution:
We have

$$\begin{aligned}
I(X \wedge Y|Z) &= H(X|Z) - H(X|Y, Z) \\
&\overset{(a)}{=} H(X|Z) - H(X|Y) \\
&\overset{(b)}{\leq} H(X) - H(X|Y) \\
&= I(X \wedge Y),
\end{aligned}$$

where (a) above follows from the fact that X, Y, Z constitute a Markov chain (in the same order), and (b) follows from the fact that conditioning reduces entropy. Equality holds if and only if $H(X|Z) = H(X)$, which happens if and only if $I(X \wedge Z) = 0$, i.e., X and Z are independent.

An example for the case $I(X \wedge Y|Z) > I(X \wedge Y)$ is as follows. Let X and Y be independent Bernoulli random variables with parameter $p = 0.5$. Let $Z = X \bigoplus Y$, where the operator $\bigoplus$ denotes the 'XOR' operation between X and Y. Then, it can be easily verified that $Z \sim \text{Ber}(0.5)$.

Since X and Y are independent, $I(X \wedge Y) = 0$, while

$$\begin{aligned}
I(X \wedge Y|Z) &= H(X|Z) - H(X|Y, Z) \\
&\overset{(a)}{=} H(X|Z) \\
&= 1,
\end{aligned}$$

where (a) above follows by noting that given Y, Z, the value of X is completely determined. The value of $H(X|Z)$ in the last line is easy to arrive at.

7.2 Exercises

1. It is of interest to determine the bias of a coin by observing n independent tosses from it. Towards this, consider the following M-ary hypothesis testing problem.

$$\begin{aligned}
\mathcal{H}_1 : &\quad X_1 \ldots, X_n \stackrel{iid}{\sim} \mathrm{Ber}(p_1) \\
\mathcal{H}_2 : &\quad X_1 \ldots, X_n \stackrel{iid}{\sim} \mathrm{Ber}(p_2) \\
&\vdots \\
\mathcal{H}_M : &\quad X_1 \ldots, X_n \stackrel{iid}{\sim} \mathrm{Ber}(p_M).
\end{aligned}$$

   (a) Describe the maximum likelihood rule for this problem.

   (b) Fix $\epsilon > 0$. Assuming that all the p_i's are distinct, use the union bound to determine the largest value of M such that the average probability of error for the ML rule above under a uniform prior on the hypotheses remains below ϵ.

   Hint: For each $i \neq j$, define the set B_{ij} as $B_{ij} = \{x : P_i(x) \geq P_j(x)\}$. Suppose $A_1, \ldots, A_M$ denotes the partition of $\mathcal{X} = \{0,1\}^n$ corresponding to the ML test. Show that $A_i \subseteq B_{ij}$ for all $j \neq i$. Then, we have

$$\begin{aligned}
P_e(\text{ML test}) &= \frac{1}{M} \sum_{i=1}^{M} P_i(A_i^c) \\
&= \sum_{i=1}^{M} P_i\Bigg(\bigcup_{j \neq i} A_j\Bigg).
\end{aligned}$$

   Use the above relation and union bound. Get an upper bound in terms of $\max_{j \neq i} P_i(B_{ji})$ and use the results from binary hypothesis testing for P_i vs P_j to upper bound $P_i(B_{ji})$.

2. Given $\alpha > 0$, show that among all positive integer-valued random variables with mean equal to α, the entropy is maximised for a Geometric random variable.

   Hint: Let $Z \sim P = \mathrm{Geo}(\mu)$, where μ is such that $E[Z] = \alpha$. Let $X \sim Q$ be any other random variable satisfying $E[X] = \alpha$. Write down the formula for $D(Q||P)$ and conclude that $H(P) \geq H(Q)$.

3. Can you find random variables X, Y and Z such that **all** of the following conditions hold.

   (a) X is independent of Y.

   (b) X is a function of (Y, Z).

   (c) $H(X) = 2$ and $H(Z) = 1$.

   If yes, give an example. If no, justify why.

   Hint: Since X and Y are independent, $I(X \wedge Y) = 0$. Manipulate this mutual information (try to bring in the random variable Z somehow) and show eventually that $H(X) \leq H(Z)$, thereby yielding that there cannot be such random variables as specified in the question.

4. Let W be an $m \times m$ matrix with nonnegative entries w_{ij}, $1 \leq i, j \leq m$. Suppose that W is a stochastic matrix. That is, for all $i \in \{1, \ldots, m\}$, $\sum_{j=1}^{m} w_{ij} = 1$. Let $p = (p_1, \ldots, p_m)$ be a probability mass function. Define $\hat{p}$ as

$$\hat{p} = pW.$$

   (a) Show that $\hat{p}$ is a valid probability mass function.

(b) Suppose $q = (q_1, \ldots, q_m)$ is any other probability mass function. Then, show that $D(pW||qW) \leq D(p||q)$. Identify the condition for equality.

(c) Suppose now that W is doubly stochastic, i.e.,

$$\sum_{j=1}^{m} w_{ij} = 1 \text{ for all } i \in \{1, \ldots, m\}, \quad \sum_{i=1}^{m} w_{ij} = 1 \text{ for all } j \in \{1, \ldots, m\}.$$

Then show that $H(pW) \geq H(p)$.