Telecom Paris ACCQ202, Information Theory

ASSIGNMENT 6

Exercise 1. (Convexity)

a. For distributions p and q on a finite alphabet, show that $D(p||q)$ is convex in the pair (p, q).*i.e.,* if (p_1, q_1) and (p_2, q_2) are two pairs of pmfs, then,

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2),$$

for all $0 \le \lambda \le 1$.
Hint – Suppose that the alphabet size is m. Then, the left-side is a sum of m terms. Apply log-sum inequality to a particular term and sum over all m terms.

Solution. Consider a symbol x.

$$\begin{aligned}\lambda p_1(x)\log\frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x)\log\frac{p_2(x)}{q_2(x)} &= \lambda p_1(x)\log\frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x)\log\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \\ &\ge \Big\{\lambda p_1(x) + (1-\lambda)p_2(x)\Big\}\log\frac{\lambda p_1(x)+(1-\lambda)p_2(x)}{\lambda q_1(x)+(1-\lambda)q_2(x)}\end{aligned}$$

by log-sum inequality. Sum over all x to obtain the desired inequality. □

b. For $(X, Y) \sim p(x)p(y|x)$, show that $I(X; Y)$ is a convex function of $p(y|x)$ for fixed $p(x)$.
Hint –

i. Consider $p_1(y|x)$ and $p_2(y|x)$ and their convex combination $p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x)$.
ii. Write out the joint distribution $p_\lambda(x, y)$ and the marginal $p_\lambda(y)$.
iii. Consider the KL divergence between $p_\lambda(x, y)$ and $p(x)p_\lambda(y)$.

Solution. Fix the distribution of X to be p^X. Let

$$\begin{aligned}p_\lambda^{XY}(x,y) &= p^X(x)p_\lambda(y|x) \\ &= p^X(x)\left[\lambda p_1(y|x) + (1-\lambda)p_2(y|x)\right] \\ &= \lambda p^X(x)p_1(y|x) + (1-\lambda)p^X(x)p_2(y|x),\end{aligned}$$

and thus its marginal

$$p_\lambda^Y(y) = \sum_x \lambda p^X(x)p_1(y|x) + (1-\lambda)p^X(x)p_2(y|x).$$

Now,

$$\begin{aligned}p^X(x)p_\lambda^Y(y) &= p^X(x)\sum_{x'}\lambda p^X(x')p_1(y|x') + (1-\lambda)p^X(x')p_2(y|x') \\ &= \lambda p^X(x)\sum_{x'}p^X(x')p_1(y|x') + (1-\lambda)p^X(x)\sum_{x'}p^X(x')p_2(y|x')\end{aligned}$$

Since $I(X;Y)$ is the KL divergence between p_λ^{XY} and $p^X p_\lambda^Y$, by convexity of KL divergence in $(p_{XY}, p_X p_Y)$ proved in part a, we have

$$D(p_\lambda^{XY}||p^X p_\lambda^Y) \leq \lambda D\left(p^X p_1 \middle|\middle| p^X \sum_{x'} p^X(x')p_1(\cdot|x')\right) + (1-\lambda) D\left(p^X p_2 \middle|\middle| p^X \sum_{x'} p^X(x')p_2(\cdot|x')\right).$$

The first term is λ times the KL divergence between the joint distribution $p^X p_1$ and the product of their marginals. Similarly, the second term is $(1-\lambda)$ times the KL divergence between the joint distribution $p^X p_2$ and the product of their marginals. Thus, we have completed the proof. □

Exercise 2. (Converse to the rate distortion theorem) Recall that the rate distortion function is given by

$$R(D) = \min_{p(\widehat{x}|x):\sum_{x,\widehat{x}} p(x)p(\widehat{x}|x)d(x,\widehat{x}) \leq D} I(X;\widehat{X}).$$

a. Prove that $R(D)$ is a non-increasing convex function of D.
Hint – To prove that $R(D)$ is convex, consider two rate distortion pairs, (R_1, D_1) and (R_2, D_2), which lie on the rate distortion curve. Let the joint distributions that achieve these pairs be $p_1(x,\widehat{x}) = p(x)p_1(\widehat{x}|x)$ and $p_2(x,\widehat{x}) = p(x)p_2(\widehat{x}|x)$. Consider the distribution

$$p_\lambda = \lambda p_1 + (1-\lambda)p_2.$$

Since the distortion is a linear function of the distribution, we have

$$D(p_\lambda) = \lambda D_1 + (1-\lambda)D_2.$$

Also,

$$I_{p_\lambda}(X;\widehat{X}) \leq \lambda I_{p_1}(X;\widehat{X}) + (1-\lambda) I_{p_2}(X;\widehat{X}).$$

b. Consider any $(2^{nR}, n)$ rate distortion code defined by functions f_n and g_n. Let $\widehat{X}^n = g_n(f_n(X^n))$ be the reproduced sequence corresponding to X^n. Justify the steps with labels on the equality or the inequality signs.

$$\begin{aligned}
I(X^n;\widehat{X}^n) &= H(X^n) - H(X^n|\widehat{X}^n) \\
&\overset{(a)}{=} \sum_{i=1}^n H(X_i) - H(X^n|\widehat{X}^n) \\
&\overset{(b)}{=} \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\widehat{X}^n, X_{i-1}, \ldots, X_1) \\
&\overset{(c)}{\geq} \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\widehat{X}^i) \\
&= \sum_{i=1}^n I(X_i;\widehat{X}_i).
\end{aligned}$$

c. Assume that the expected distortion $\mathbb{E}d(X^n, \widehat{X}^n) \leq D$ for this code. Justify the steps with labels on the equality or the inequality signs.

$$\begin{aligned}\sum_{i=1}^{n} R(\mathbb{E}d(X_i, \widehat{X}_i)) &= n \sum_{i=1}^{n} \frac{1}{n} R(\mathbb{E}d(X_i, \widehat{X}_i)) \\ &\overset{(a)}{\geq} nR\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}d(X_i, \widehat{X}_i)\right) \\ &\overset{(b)}{=} nR(\mathbb{E}d(X^n, \widehat{X}^n)).\end{aligned}$$

d. Using the results above, show that $R \geq R(D)$. *Hint* – Start with $nR \geq H(\widehat{X}^n)$.

For solution, see Section 10.4 *in "Elements of Information Theory, Cover & Thomas, 2nd edition".*

Exercise 3. (Uniquely decodable codes) Given an alphabet $\mathcal{X} = \{1, \ldots, m\}$ and a probability distribution $P = (p_1, \ldots, p_m)$ on $\mathcal{X}$, solve (using Lagrange multipliers) the following convex optimization problem:

$$\min_{\ell_1,\ldots,\ell_m \in \mathbb{R}} \sum_{i=1}^{m} p_i \ell_i \text{ subject to } \sum_{i=1}^{m} 2^{-\ell_i} \leq 1.$$

Conclude that for a uniquely decodable code, the minimum expected codeword length is greater than or equal to $H(P)$. Why is it *greater than or equal to* and not *equal to*?

Solution. In order to find the optimal $\ell = (\ell_1, \ldots, \ell_m)$, we define the Lagrangian

$$L(\ell, \lambda) = \sum_{i=1}^{m} p_i \ell_i - \lambda \left(\sum_{i=1}^{m} 2^{-\ell_i} - 1\right).$$

Taking derivative with respect to ℓ_i and λ and setting them equal to 0 yields

$$\frac{dL}{d\ell_i} = p_i - \lambda 2^{-\ell_i} \ln 2 = 0$$

and $\lambda = 1/\ln 2$ since $\sum_{i=1}^{m} p_i = 1$, whereby

$$\ell_i = -\log_2 p_i.$$

Thus, the minimum is $L^* = \sum_{i=1}^{m} p_i \ell_i = H(P)$. The minimum expected codeword length for a uniquely decodable code, $\overline{L}$ is obtained by solving the same optimization problem with an additional constraint, namely, $\ell_i \in \mathbb{N}$; therefore, the minimum can only be larger. Hence, $\overline{L} \geq L^* = H(P)$. □

Exercise 4. (Rényi entropy) For a distribution P on a finite alphabet $\mathcal{X}$, the Rényi entropy of order α, $H_\alpha(P)$ is given by

$$H_\alpha(P) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{m} p_i^\alpha\right).$$

for $\alpha \geq 0$, $\alpha \neq 1$.

a. Show that the Shannon entropy $H(P)$ satisfies

$$H(P) = \lim_{\alpha \to 1} H_\alpha(P).$$

Hint – Use L'Hôpital's rule.

Solution. Let $\mathcal{X} = \{1, \ldots, m\}$ and $P = (p_1, \ldots, p_m)$. By L'Hôpital's rule,

$$\begin{aligned}\lim_{\alpha \to 1} H_\alpha(P) &= \lim_{\alpha \to 1} \frac{\ln \sum_{i=1}^m p_i^\alpha}{1-\alpha} \\ &= \left. \frac{\frac{1}{\sum_{i=1}^m p_i^\alpha} \sum_{i=1}^m p_i^\alpha \ln p_i}{-1} \right|_{\alpha=1} \\ &= -\sum_{i=1}^m p_i \ln p_i \\ &= H(P).\end{aligned}$$

□

b. For i.i.d. random variables X and Y on $\mathcal{X}$, what is $\mathbb{P}[X = Y]$ in terms of $H_2(P)$?

Solution. We have

$$H_2(P) = -\log \sum_{i=1}^m p_i^2.$$

Then,

$$\begin{aligned}\mathbb{P}[X = Y] &= \sum_{i=1}^m \mathbb{P}[X = i, Y = i] \\ &= \sum_{i=1}^m \mathbb{P}[X = i] \cdot \mathbb{P}[Y = i] \\ &= \sum_{i=1}^m p_i^2 \\ &= 2^{-H_2(P)}.\end{aligned}$$

□

c. In the limit as $\alpha \to \infty$, H_α converges to H_∞ defined by

$$H_\infty(P) = -\log \max_i P_i.$$

Show that $H_2 \leq 2H_\infty$.

Solution. Since "sum $\geq \max$", we have

$$\begin{aligned}H_2(P) &= -\log \sum_{i=1}^m p_i^2 \\ &\leq -\log \left\{ \max_i p_i^2 \right\} \\ &= 2H_\infty(P).\end{aligned}$$

□

d. Show that for a fixed P,
$$H_0 \geq H_1 \geq H_2 \geq H_\infty,$$
where H_1 denotes the Shannon entropy.

Solution. We have
$$H_0(P) = \log \sum_{i=1}^{m} p_i^0 = \log m \geq H_1(P).$$
By Jensen's inequality, we have
$$\begin{aligned} H_2(P) &= -\log \left\{ \sum_{i=1}^{m} p_i \cdot p_i \right\} \\ &\leq \sum_{i=1}^{m} p_i \left\{ -\log p_i \right\} \\ &= H_1(P). \end{aligned}$$
We show the last inequality next.
$$\begin{aligned} H_2(P) &= -\log \left\{ \sum_{i=1}^{m} p_i \cdot p_i \right\} \\ &\geq -\log \left\{ \max_i p_i \left(\sum_{i=1}^{m} p_i \right) \right\} \\ &= H_\infty(P). \end{aligned}$$
□

Exercise 5. (List decoding Fano's inequality) Consider discrete random variables X and Y with X taking values in the set $\{0,1\}^k$. Upon observing Y we produce a list $L(Y)$ of size 2^ℓ such that
$$\mathbb{P}(X \in L(Y)) \geq 1 - \varepsilon.$$
Show that
$$H(X|Y) \leq \varepsilon k + (1-\varepsilon)\ell + 1.$$
Hint – Define the random variable $T = \mathbf{1}_{\{X \in L(Y)\}}$. Expand $H(X,Y,T)$ using chain rule.

Solution. Let $T = \mathbf{1}_{\{X \in L(Y)\}}$. Since T is a function of X and Y, we have
$$\begin{aligned} H(X,Y) &= H(X,Y,T) \\ &= H(T) + H(Y|T) + H(X|Y,T). \end{aligned}$$
Since $H(X,Y) = H(Y) + H(X|Y)$ and $H(Y|T) \leq H(Y)$, we have
$$\begin{aligned} H(X|Y) &\leq H(T) + H(X|Y,T) \\ &= H(T) + H(X|Y,L(Y),T) \\ &\leq H(T) + H(X|L(Y),T), \end{aligned}$$

since $L(Y)$ is a function of Y. Now,

$$H(X|L(Y), T = 1) \leq \ell,$$

and

$$H(X|L(Y), T = 0) \leq k - \ell,$$

and hence

$$H(X|L(Y), T) \leq \mathbb{P}[T = 1]\ell + \mathbb{P}[T = 0](k - \ell).$$

The result follows by observing that $\mathbb{P}[T = 1] \leq 1$, $\mathbb{P}[T = 0] \leq \varepsilon$, and $H(T) \leq 1$. □

Exercise 6. (Shearer's lemma) Shearer's lemma is a generalization of the basic inequality

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i).$$

For $S \subseteq [n] = \{1, 2, \ldots\}$, we write $X_S = (X_i : i \in S)$.

a. Prove the lemma: Let $X_1, \ldots, X_n$ be random variables. Let $S_1, \ldots, S_m \subseteq [n]$ be subsets such that each $i \in [n]$ belongs to at least k sets. Then,

$$kH(X_1, \ldots, X_n) \leq \sum_{j=1}^{m} H(X_{S_j}).$$

Solution. Let $S_j = \{i_1, \ldots, i_{s_j}\}$ with $i_1 < \ldots < i_{s_j}$. Then,

$$\begin{aligned} H(X_{S_j}) &= H(X_{i_1}) + H(X_{i_2}|X_{i_1}) + \ldots + H(X_{i_{s_j}}|X_{i_1}, \ldots, X_{i_{s_j-1}}) \\ &\geq H(X_{i_1}|X_1, \ldots, X_{i_1-1}) + H(X_{i_2}|X_1, \ldots, X_{i_2-1}) + \ldots + H(X_{i_{s_j}}|X_1, \ldots, X_{i_{s_j}-1}). \end{aligned}$$

Sum the left side over $j = 1$ to m to obtain $\sum_{i=1}^{m} H(X_{S_j})$. Since each $i \in [n]$ belongs to at least k sets from $S_j, j = 1, \ldots, m$, the sum of the right side over $j = 1$ to m is equal to at least k times the sum $\sum_{i=1}^{n} H(X_i|X^{i-1})$, whereby the result follows. □

b. Suppose n distinct points in $\mathbb{R}^3$ have n_1 distinct projections on the XY-plane, n_2 distinct projections on the XZ-plane, and n_3 distinct projections on the YZ-plane. For two different points, since all three projections cannot be the same, we have $n \leq n_1 n_2 n_3$. Using Shearer's lemma, show that

$$n \leq \sqrt{n_1 n_2 n_3}.$$

Hint – Let $P = (X_1, X_2, X_3)$ be one of the n points picked uniformly at random. Then, $P_1 = (X_1, X_2)$, $P_2 = (X_1, X_3)$, and $P_3 = (X_2, X_3)$ are its three projections.

Solution. By Shearer's lemma, we have

$$2H(P) \leq H(P_1) + H(P_2) + H(P_3).$$

The results follows since $H(P) = \log n$ and $H(P_i) \leq \log n_i, i = 1, 2, 3$. □

Exercise 7. (Shotgun DNA sequencing)[1] DNA sequencing is the basic workhorse of modern day biology and medicine. Shotgun sequencing is the dominant technique used: many randomly located short fragments called reads are extracted from the DNA sequence, and these reads are assembled to reconstruct the original sequence. A basic question is: given a sequencing technology and the statistics of the DNA sequence, what is the minimum number of reads required for reliable reconstruction?

The DNA sequence $s = s_1 s_2 \cdots s_G$ is modeled as an i.i.d. random process of length G with each symbol taking values according to a probability distribution $p = (p_1, p_2, p_3, p_4)$ on the nucleotide alphabet $\{A, C, G, T\}$. A read is a substring of length L from the DNA sequence. The objective of DNA sequencing is to reconstruct the whole sequence s based on N reads from the sequence. The starting location of each read is uniformly distributed on the DNA sequence and are independent from one read to another. We seek to understand the fundamental limits on the two quantities N and L.

a. *Covering:* Argue that for the perfect reconstruction of s, for a fixed L, the collection of reads should *cover* the entire sequence and hence a necessary condition is that $N \geq G/L$.

   Solution. Suppose the N reads, each of length L are $L_{i_1}, \dots, L_{i_N}$, where the read L_{i_j} starts at location $i_j \in [G - L + 1]$. For $k = 1, \dots, G$, let $A_k = 1$ if $k \in [i_j, i_j + L]$ for some $i_j, j \in [N]$. Then,

$$\sum_{k=1}^{G} A_k \leq NL.$$

   Assume that $NL < G$. Then, there exists $k^* \in [G]$ for which $A_{k^*} = 0$. Thus, the reads $L_{i_1}, \dots, L_{i_N}$ could have been generated by s or another sequence s' with $s'_i = s_i,\ i \neq k^*$ and $s'_{k^*} \neq s_{k^*}$. Hence, perfect reconstruction is not possible from the given reads. □

b. *An improvement via the coupon collector problem:* The well-known "coupon collector problem" is the following. Suppose we repeatedly and independently sample a random variable that is uniformly distributed over $\{1, 2, \dots, n\}$. How many samples do we need to ensure the sampling of all n numbers? The answer to this question is roughly $n \log n$ (`https://en.wikipedia.org/wiki/Coupon_collector%27s_problem`).

   Now, consider a modified DNA read technique where in each read, you get to observe L independent locations (instead of contiguous locations). Can you use the coupon collector result to get an estimate on the necessary number of reads N for this modified problem? What does it say about the required number of reads for the original problem?

   Solution. For the modified problem, each position $1, \dots, G$ is a coupon and each read provides us with L coupons whereby

$$NL \geq G \ln G.$$

   It is necessary to collect all the coupons for reconstruction for the original problem and hence for the original problem,

$$N \geq \frac{G}{L} \ln G.$$

□

[1] A. Motahari, G. Bresler, and D. Tse, "Information theory of DNA shotgun sequencing." IEEE Transactions on Information Theory 59.10 (2013): 6273-6289.

c. Suppose we have two DNA sequences, the first sequence generated by a uniform distribution on $\{A, C, T, G\}$ and the second by a distribution $(0.5, 0.4, 0.05, 0.05)$. The DNA sequences and the corresponding reads are as follows:

1. Sequence: $ACTGCATAGT$, Reads: TGC, CAT, ACT, TAG, AGT.
2. Sequence: $ACACATACGC$, Reads: ACA, CAC, TAC, ACG, CGC

Impossible to reconstruct?

i. Which among the two sequences can you reconstruct (uniquely) from the reads? Why?

ii. Calculate the Rényi entropy of order 2 (see Ex.4) for both the distributions.

Solution. The first sequence can be reconstructed from the reads while the second cannot be. This is because the second sequence has repeated patterns of length $2 < L$. The Rényi entropy of order 2 for the uniform distribution is

$$H_2(Unif) = \log_2 4 = 2,$$

and for $P = (0.5, 0.4, 0.05, 0.05)$ is

$$H_2(P) = -\log_2[0.5^2 + 0.4^2 + 2 \cdot 0.05^2] \approx 0.725.$$

Recall from Ex.4b that larger the value of H_2, smaller the probability of “collisions” or repeats. □

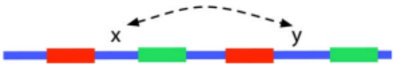

Fig. 4. Two pairs of interleaved repeats of length $L-1$ create ambiguity: from the reads, it is impossible to know whether the sequences **x** and **y** are as shown, or swapped.

d. We observe that even if we have access to all length-L reads of the sequence, *repeats* make reconstruction impossible (see figure). Denoting by S_i^L the length-L subsequence starting at position i, and R_L the number of length-L repeats, we have

$$\mathbb{E}[R_L] = \sum_{1 \le i < j \le G} \mathbb{P}[S_i^L = S_j^L].$$

Justify the following:

$$\mathbb{E}[R_L] > \left(\frac{G^2}{2} - GL\right) e^{-LH_2(P)}.$$

Proof. For a given sequence generated by (p_1, p_2, p_3, p_4), the probability that two specific physically disjoint length-ℓ subsequences are identical is $2^{-\ell H_2(P)}$. In the sum, dropping the terms in which S_i^L and S_j^L overlap, we obtain the desired bound. □

e. *Phase transition:* For $G \gg L$, the above bound may be approximated as

$$\mathbb{E}[R_L] \approx \frac{G^2}{2} e^{-LH_2(P)}.$$

Let $G, L \to \infty$ with $L/\ln G = \overline{L}$, a constant. Conclude that the expected number of repeats approaches zero if

$$\overline{L} > 2/H_2(P)$$

and approaches infinity if

$$\overline{L} < 2/H_2(P).$$

Interpret this result as a prescription for *how large L should be* in order for reconstruction to be successful. Observe that N does not play any role here.

f. Assuming that $\overline{L} > 2/H_2(P)$ and N equals the estimate obtained in part b, conclude that the number of reads (of length L) per nucleotide, given by N/G is roughly $H_2(P)$.

Solution. For reconstruction to be successful, the expected number of repeats should go to 0 and hence we need,

$$L = \overline{L} \ln G > \frac{2 \ln G}{H_2(P)}.$$

If $N = \frac{G}{L} \ln G$, then

$$\frac{N}{G} \leq H_2(P).$$

□