Information Theory: Midterm Exam, 26 April 2019

1. Define the joint typical set as:

$$\begin{aligned}\mathcal{F}_n(\delta) \;:=\; & \Big\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \;: \\ & \left|-\frac{1}{n}\log_2 P_{X^n}(x^n) - H(X)\right| < \delta, \quad \left|-\frac{1}{n}\log_2 P_{Y^n}(y^n) - H(Y)\right| < \delta, \\ & \text{and } \left|-\frac{1}{n}\log_2 P_{X^n,Y^n}(x^n, y^n) - H(X,Y)\right| < \delta\Big\}.\end{aligned}$$

(a) (5%) Show that $|\mathcal{F}_n(\delta)| \le 2^{n(H(X,Y)+\delta)}$.

Hint: $|\mathcal{F}_n(\delta)| \le |\mathcal{G}_n(\delta)|$, where

$$\mathcal{G}_n(\delta) := \left\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \;: \left|-\frac{1}{n}\log_2 P_{X^n,Y^n}(x^n, y^n) - H(X,Y)\right| < \delta\right\}.$$

(b) (5%) Define

$$\mathcal{F}_n(\delta|x^n) := \{y^n \in \mathcal{Y}^n \;:\; (x^n, y^n) \in \mathcal{F}_n(\delta)\}.$$

Let

$$\Lambda_m(\boldsymbol{c}_{m'}, \boldsymbol{c}_m) := P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})\middle|\boldsymbol{c}_m\right)$$

be the probability of Y^n being jointly typical with $\boldsymbol{c}_{m'}$, given that $\boldsymbol{c}_m$ is transmitted. Show that for $m' \ne m$,

$$E[\Lambda_m(\boldsymbol{c}_{m'}, \boldsymbol{c}_m)] \le 2^{-n(I(X;Y)-3\delta)}$$

Hint:

$$\begin{aligned}& E\left[P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})\middle|\boldsymbol{c}_m\right)\right] \\ &= \sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n} P_{X^n}(\boldsymbol{c}_m)P_{X^n}(\boldsymbol{c}_{m'})P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})\middle|\boldsymbol{c}_m\right) \\ &= \sum_{\boldsymbol{c}_m\in\mathcal{X}^n}\sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n} P_{X^n}(\boldsymbol{c}_m)P_{X^n}(\boldsymbol{c}_{m'}) \sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{Y^n|X^n}\left(y^n\middle|\boldsymbol{c}_m\right) \\ &= \sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{X^n}(\boldsymbol{c}_{m'}) \sum_{\boldsymbol{c}_m\in\mathcal{X}^n} P_{X^n}(\boldsymbol{c}_m)P_{Y^n|X^n}\left(y^n\middle|\boldsymbol{c}_m\right) \\ &= \sum_{\boldsymbol{c}_{m'}\in\mathcal{X}^n}\sum_{y^n\in\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{X^n}(\boldsymbol{c}_{m'})P_{Y^n}(y^n) \\ &= \sum_{(\boldsymbol{c}_{m'},y^n)\in\mathcal{F}_n(\delta)} P_{X^n}(\boldsymbol{c}_{m'})P_{Y^n}(y^n)\end{aligned}$$

(c) (5%) Continue from (b). An important step of the proof of Shannon's channel coding theorem is:

$$\lambda_m \le P_{Y^n|X^n}\left(\mathcal{F}_n^c(\delta|\boldsymbol{c}_m)\middle|\boldsymbol{c}_m\right) + \sum_{m'=1,m'\neq m}^{M_n} P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})\middle|\boldsymbol{c}_m\right),$$

where λ_m is the conditional probability of typical set decoding error given that codeword $\boldsymbol{c}_m$ is transmitted. A student hoped to simplify the derivation and extended this step to obtain another upper bound:

$$\lambda_m \le P_{Y^n|X^n}\left(\mathcal{F}_n^c(\delta|\boldsymbol{c}_m)\middle|\boldsymbol{c}_m\right) + \sum_{m'=1}^{M_n} P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})\middle|\boldsymbol{c}_m\right).$$

The student then tried to show that the expected value of the upper bound can be made small by increasing n, but failed. Why did he fail?

Solution.

(a) For $(x^n, y^n) \in \mathcal{G}_n(\delta)$, we have

$$2^{-n(H(X,Y)+\delta)} \le P_{X^n,Y^n}(x^n, y^n) \le 2^{-n(H(X,Y)-\delta)}.$$

Consequently,

$$1 \ge \sum_{(x^n,y^n)\in\mathcal{G}_n(\delta)} P_{X^n,Y^n}(x^n,y^n) = \sum_{(x^n,y^n)\in\mathcal{G}_n(\delta)} 2^{-n(H(X,Y)-\delta)} = 2^{-n(H(X,Y)-\delta)}|\mathcal{G}_n(\delta)|,$$

which immediately gives the desired result.

(b)

$$\begin{aligned}
E\left[P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_{m'})\middle|\boldsymbol{c}_m\right)\right] &= \sum_{(\boldsymbol{c}_{m'},y^n)\in\mathcal{F}_n(\delta)} P_{X^n}(\boldsymbol{c}_{m'})P_{Y^n}(y^n) \\
&\le \sum_{(\boldsymbol{c}_{m'},y^n)\in\mathcal{F}_n(\delta)} 2^{-n(H(X)-\delta)}2^{-n(H(Y)-\delta)} \\
&= |\mathcal{F}_n(\delta)|2^{-n(H(X)-\delta)}2^{-n(H(Y)-\delta)} \\
&\le 2^{n(H(X,Y)+\delta)}2^{-n(H(X)-\delta)}2^{-n(H(Y)-\delta)} \\
&= 2^{-n(H(X)+H(Y)-H(X,Y)-3\delta)} \\
&= 2^{-n(I(X;Y)-3\delta)}.
\end{aligned}$$

(c) Adding the term of $P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_m)\middle|\boldsymbol{c}_m\right)$ actually makes the new upper bound large (in fact, approaching 1) as n goes to infinity as:

$$\begin{aligned}
E\left[P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_m)\middle|\boldsymbol{c}_m\right)\right] &= \sum_{\boldsymbol{c}_m\in\mathcal{X}^n} P_{X^n}(\boldsymbol{c}_m)P_{Y^n|X^n}\left(\mathcal{F}_n(\delta|\boldsymbol{c}_m)\middle|\boldsymbol{c}_m\right) \\
&= P_{X^n,Y^n}\left(\mathcal{F}_n(\delta)\right) > 1-\delta \quad \text{as } n \text{ sufficiently large.}
\end{aligned}$$

2. Now we recall Theorem 3.22 and its proof as follows.

Theorem 3.22 The average rate of every uniquely decodable (UD) D-ary n-th order VLC for a discrete memoryless source $\{X_n\}_{n=1}^{\infty}$ is lower-bounded by the source entropy $H_D(X)$ (measured in D-ary code symbols/source symbol).

Proof: Consider a uniquely decodable D-ary n-th order VLC code for the source $\{X_n\}_{n=1}^{\infty}$

$$f : \mathcal{X}^n \to \{0, 1, \cdots, D-1\}^*$$

and let $\ell(\boldsymbol{c}_{x^n})$ denote the length of the codeword $\boldsymbol{c}_{x^n} = f(x^n)$ for sourceword x^n. Hence,

$$\begin{aligned}
\overline{R}_n - H_D(X) &= \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)\ell(\boldsymbol{c}_{x^n}) - \frac{1}{n} H_D(X^n) \\
&= \frac{1}{n} \left[\sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)\ell(\boldsymbol{c}_{x^n}) - \sum_{x^n \in \mathcal{X}^n} (-P_{X^n}(x^n) \log_D P_{X^n}(x^n)) \right] \\
&= \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \log_D \frac{P_{X^n}(x^n)}{D^{-\ell(\boldsymbol{c}_{x^n})}} \\
&\geq \frac{1}{n} \left[\sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \right] \log_D \frac{\left[\sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)\right]}{\left[\sum_{x^n \in \mathcal{X}^n} D^{-\ell(\boldsymbol{c}_{x^n})}\right]} \\
&\quad \text{(log-sum inequality)} \\
&= -\frac{1}{n} \log \left[\sum_{x^n \in \mathcal{X}^n} D^{-\ell(\boldsymbol{c}_{x^n})} \right] \\
&\geq 0
\end{aligned}$$

where the last inequality follows from the Kraft inequality for uniquely decodable codes and the fact that the logarithm is a strictly increasing function. □

Answer the following questions.

(a) (5%) Reprove Theorem 3.22 by fundamental inequality.

(b) (5%) Based on the proof of Theorem 3.22, argue that if the average codeword length of a UD code equals the source entropy, then $P_{X^n}(x^n) = D^{-\ell(\boldsymbol{c}_{x^n})}$.

Hint: For non-negative numbers, a_1, a_2, …, a_n and b_1, b_2, …, b_n,

$$\sum_{i=1}^{n} \left(a_i \log_D \frac{a_i}{b_i} \right) \geq \left(\sum_{i=1}^{n} a_i \right) \log_D \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i},$$

with equality holding iff for all $i = 1, \cdots, n$,

$$\frac{a_i}{b_i} = \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}.$$

(c) (5%) Is it possible to have a source whose binary Huffman code has average codeword length (ACL) equal to the base-2 source entropy but 4-ary Huffman code has average codeword length larger than the base-4 source entropy? If affirmative, give an example; if negative, disprove it.

Note: For the D-ary Huffman code, some zero-probability "dummy" source letters need to be added so that the alphabet size of the expanded source $|\mathcal{X}'|$ is the smallest positive integer greater than or equal to $|\mathcal{X}|$ with

$$|\mathcal{X}'| = 1 \quad (\text{modulo } D-1) \quad (\text{For } D > 2).$$

Solution.

(a)

$$\begin{aligned}
\overline{R}_n - H_D(X) &= \cdots \\
&= \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \log_D \frac{P_{X^n}(x^n)}{D^{-\ell(\mathbf{c}_{x^n})}} \\
&\geq \frac{1}{n} \log_D(e) \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \left(1 - \frac{D^{-\ell(\mathbf{c}_{x^n})}}{P_{X^n}(x^n)}\right) \\
&\quad \text{(fundamental inequality)} \\
&= \frac{1}{n} \log_D(e) \left(\sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) - \sum_{x^n \in \mathcal{X}^n} D^{-\ell(\mathbf{c}_{x^n})} \right) \\
&\geq \frac{1}{n} \log_D(e) \cdot (1-1) = 0,
\end{aligned}$$

where the last inequality follows from the Kraft inequality for uniquely decodable codes.

(b) In the proof of Theorem 3.22, there are two inequalities. Since $\overline{R}_n - H_D(X) = 0$, the first inequality must hold with equality, and the log-sum inequality holds with equality iff $P_{X^n}(x^n) = D^{-\ell(\mathbf{c}_{x^n})}$.

(c) In order to have the ACL of a D-ary Huffman code equal to the base-D source entropy, we must have $P_{X^n}(a_i) = D^{-\ell_i}$ for the ith symbol a_i for some integer ℓ_i. Thus, the answer to the question is affirmative. A quick example will be a source with distribution $\{\frac{1}{2}, \frac{1}{2}\}$. Its base-2 source entropy is 1 and its corresponding binary Huffman code is $\{0, 1\}$ with ACL $= 1$. However, the base-4 source entropy is $\log_4(2) = \frac{1}{2}$ but the ACL of the 4-ary Huffman code is still 1.

3. Let $X_1, X_2, \ldots, X_n, \ldots$ be a stationary discrete random process, where each $X_i \in \mathcal{X}$.

(a) (5%) Let $\{\mathcal{F}_n\}_{n=1}^{\infty}$ be a sequence of sets with each $\mathcal{F}_n \in \mathcal{X}^n$. If

$$\lim_{n \to \infty} \Pr[X^n \in \mathcal{F}_n] = 1,$$

can we use this sequence of sets as the *typical sets* to prove Shannon's source coding theorem? Justify your answer.

(b) (5%) Suppose this sequence of sets satisfies the following two properties:

i. $P_{X^n}(\mathcal{F}_n) > 1-\delta$

ii. $|\mathcal{F}_n| \le 2^{n(H(\mathcal{X})+\delta)}$

Design a simple typical-set encoding as follows:

$$\begin{cases} x^n \to \text{binary-index } x^n \text{ by } k_n \text{ bits}, & \text{when } x^n \in \mathcal{F}_n \\ x^n \to \text{all-zero binary codeword of length } k_n, & \text{when } x^n \notin \mathcal{F}_n \end{cases}$$

where k_n is the number of bits used to index all $x^n \in \mathcal{X}^n$, and is restricted to be only a function of n. Argue that this typical-set encoding can achieve

$$\limsup_{n\to\infty} \frac{k_n}{n} \le H(\mathcal{X}) + \delta \quad \text{and} \quad P_e < \delta,$$

where P_e is the probability of decoding error.

(c) (5%) Further suppose that other than the two properties in (c), the typical sets satisfies

$$(\forall\, x^n \in \mathcal{F}_n)\ P_{X^n}(x^n) \le 2^{-n(H(\mathcal{X})-\delta)}.$$

Now for an alternative encoder that wishes to use only k'_n bits to encode the binary source stream x^n for each n, where

$$\frac{k'_n}{n} \le H(\mathcal{X}) - 2\delta,$$

show that its probability of correct decoding P'_c is upper bounded by $\delta + 2^{-n\delta}$.
Hint: Let $\mathcal{S}_n$ be the set of source streams x^n that can be correctly decoded; then, its size must be upper bounded by $2^{k'_n} \le 2^{n(H(\mathcal{X})-2\delta)}$.

Solution.

(a) The answer is "not necessarily." For example, if we let $\mathcal{F}_n = \mathcal{X}^n$, then $\Pr[X^n \in \mathcal{F}_n] = 1$ for every n; apparently, such choice cannot be used to prove Shannon's source coding theorem.

In fact, we also additionally need that $|\mathcal{F}_n|$ is close to $2^{nH(\mathcal{X})}$, for which such sequence of sets should exist according to Shannon's source coding theorem, where $H(\mathcal{X})$ is the entropy rate of the source.

(b) The number of bits required for this encoder must be upper-bounded by

$$k_n \le \left\lceil \log_2\left(2^{n(H(\mathcal{X})+\delta)} + 1\right)\right\rceil,$$

which implies

$$\limsup_{n\to\infty} \frac{k_n}{n} \le \limsup_{n\to\infty} \frac{1}{n}\left\lceil \log_2\left(2^{n(H(\mathcal{X})+\delta)} + 1\right)\right\rceil = H(\mathcal{X}) + \delta.$$

Since only the all-zero codeword cannot be recovered back to the original source symbols, the error rate satisfies

$$P_e \le P_{X^n}(\mathcal{F}_n^c) < \delta.$$

(c) The probability of correct block decoding satisfies

$$\begin{aligned}
1 - P_e' &= \sum_{x^n \in \mathcal{S}_n} P_{X^n}(x^n) \\
&= \sum_{x^n \in \mathcal{S}_n \cap \mathcal{F}_n^c} P_{X^n}(x^n) + \sum_{x^n \in \mathcal{S}_n \cap \mathcal{F}_n} P_{X^n}(x^n) \\
&\leq P_{X^n}(\mathcal{F}_n^c) + |\mathcal{S}_n \cap \mathcal{F}_n| \cdot \max_{x^n \in \mathcal{F}_n} P_{X^n}(x^n) \\
&< \delta + |\mathcal{S}_n| \cdot \max_{x^n \in \mathcal{F}_n} P_{X^n}(x^n) \\
&< \delta + 2^{n(H(X)-2\delta)} \cdot 2^{-n(H(X)-\delta)} \\
&= \delta + 2^{-n\delta}.
\end{aligned}$$

4. Define the *divergence typical set* as

$$\mathcal{A}_n(\delta) := \left\{ x^n \in \mathcal{X}^n : \left| \frac{1}{n} \log_2 \frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} - D(P_X \| P_{\hat{X}}) \right| < \delta \right\}.$$

It can be shown that for any sequence x^n in $\mathcal{A}_n(\delta)$,

$$P_{X^n}(x^n) 2^{-n(D(P_X\|P_{\hat{X}})-\delta)} > P_{\hat{X}^n}(x^n) > P_{X^n}(x^n) 2^{-n(D(P_X\|P_{\hat{X}})+\delta)}.$$

(a) (5%) Prove that

$$P_{\hat{X}^n}(\mathcal{A}_n(\delta)) \leq 2^{-n(D(P_X\|P_{\hat{X}})-\delta)} P_{X^n}(\mathcal{A}_n(\delta)).$$

(b) (5%) Show that for any $\mathcal{B}_n \in \mathcal{X}^n$,

$$P_{\hat{X}^n}(\mathcal{B}_n) \geq 2^{-n(D(P_X\|P_{\hat{X}})+\delta)} P_{X^n}(\mathcal{B}_n \cap \mathcal{A}_n(\delta)).$$

Hint: Use (a).

Solution.

(a)

$$\begin{aligned}
P_{\hat{X}^n}(\mathcal{A}_n(\delta)) &= \sum_{x^n \in \mathcal{A}_n(\delta)} P_{\hat{X}^n}(x^n) \\
&\leq \sum_{x^n \in \mathcal{A}_n(\delta)} P_{X^n}(x^n) 2^{-n(D(P_X\|P_{\hat{X}})-\delta)} \\
&= 2^{-n(D(P_X\|P_{\hat{X}})-\delta)} \sum_{x^n \in \mathcal{A}_n(\delta)} P_{X^n}(x^n) \\
&= 2^{-n(D(P_X\|P_{\hat{X}})-\delta)} P_{X^n}(\mathcal{A}_n(\delta))
\end{aligned}$$

(b)

$$\begin{aligned}
P_{\hat{X}^n}(\mathcal{B}_n) &\geq P_{\hat{X}^n}(\mathcal{B}_n \cap \mathcal{A}_n(\delta)) \\
&= \sum_{x^n \in \mathcal{B}_n \cap \mathcal{A}_n(\delta)} P_{\hat{X}^n}(x^n) \\
&\geq \sum_{x^n \in \mathcal{B}_n \cap \mathcal{A}_n(\delta)} P_{X^n}(x^n) 2^{-n(D(P_X \| P_{\hat{X}})+\delta)} \\
&= 2^{-n(D(P_X \| P_{\hat{X}})+\delta)} P_{X^n}(\mathcal{B}_n \cap \mathcal{A}_n(\delta)).
\end{aligned}$$

5. For the minimization of differentiable convex function $f(\boldsymbol{x})$ over the convex set

$$\mathcal{Q} = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1 \text{ and each } x_i \geq 0 \right\},$$

we note that

i) each inequality constraint $g_i(\boldsymbol{x}) = -x_i < 0$ is affine (and hence convex),

ii) the equality constraint $h(\boldsymbol{x}) = \sum_{i=1}^{n} x_i - 1 = 0$ is affine,

iii) $g_i(\boldsymbol{x})$ and $h(\boldsymbol{x})$ are both differentiable;

hence, the strong duality holds if, and only if, the KKT condition (given below) holds.

$$\text{KKT condition:} \begin{cases} g_i(\boldsymbol{x}) \leq 0, \quad \lambda_i \geq 0, \quad \lambda_i g_i(\boldsymbol{x}) = 0 & i = 1, \ldots, n \\ h(\boldsymbol{x}) = 0 \\ \frac{\partial L}{\partial x_k}(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\nu}) = \frac{\partial f}{\partial x_k}(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_k}(\boldsymbol{x}) + \nu \frac{\partial h}{\partial x_k}(\boldsymbol{x}) = 0 & k = 1, \ldots, n \end{cases}$$

(a) (5%) Show that the above KKT condition can be equivalently simplified to

$$\text{KKT condition:} \begin{cases} h(\boldsymbol{x}) = 0 & \text{(C1)} \\ \begin{cases} \dfrac{\partial f}{\partial x_k}(\boldsymbol{x}) + \nu \dfrac{\partial h}{\partial x_k}(\boldsymbol{x}) = 0, & \text{if } x_k > 0; \\ \dfrac{\partial f}{\partial x_k}(\boldsymbol{x}) + \nu \dfrac{\partial h}{\partial x_k}(\boldsymbol{x}) \geq 0, & \text{if } x_k = 0; \end{cases} \quad k = 1, \ldots, n & \text{(C2)} \end{cases}$$

Consequently, one only needs to deal with a single Lagrange multiplier ν in the minimization manipulation.

(b) (5%) Why isn't it theoretically sound to verify that $f(\boldsymbol{x})$ is a convex function over a *non-convex* set $\tilde{\mathcal{Q}}$?

(c) (5%) Let $f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \ln(x_i)$. Determine $\boldsymbol{x}^\diamond = \boldsymbol{x}^\diamond(\nu)$ that satisfies the sub-condition (C2) in (a).

(d) (5%) Based on the answer in (c), determine $\boldsymbol{x}^*$ and ν^* that satisfy the sub-condition (C1) in (a).

(e) (5%) From (d), what are the values of the Lagrange multipliers $\{\lambda_i^*\}$ that fulfill the original KKT condition?

Solution.

(a) First, we note that $g_i(\boldsymbol{x}) = -x_i$, and hence the first sub-condition becomes

$$x_i \geq 0, \quad \lambda_i \geq 0, \quad \text{and} \quad x_i\lambda_i = 0.$$

Then, we note that the 3rd sub-condition dictates

$$\frac{\partial f}{\partial x_k}(\boldsymbol{x}) + \nu\frac{\partial h}{\partial x_k}(\boldsymbol{x}) = -\sum_{i=1}^{m}\lambda_i\frac{\partial g_i}{\partial x_k}(\boldsymbol{x}) = \lambda_k$$

and hence we can combine the 1st sub-condition and the 3rd sub-condition into one condition as:

$$\begin{cases} \dfrac{\partial f}{\partial x_k}(\boldsymbol{x}) + \nu\dfrac{\partial h}{\partial x_k}(\boldsymbol{x}) = 0, & \text{if } x_k > 0; \\ \dfrac{\partial f}{\partial x_k}(\boldsymbol{x}) + \nu\dfrac{\partial h}{\partial x_k}(\boldsymbol{x}) \geq 0, & \text{if } x_k = 0. \end{cases}$$

(b) Because $\lambda \cdot \boldsymbol{x} + (1-\lambda)\cdot\tilde{\boldsymbol{x}}$ may not lie in the non-convex $\tilde{\mathcal{Q}}$ even if both $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ are.

(c) For the given $f(\boldsymbol{x})$, the sub-condition (C2) becomes:

$$\begin{cases} [1+\ln(x_k)] + \nu = 0, & x_k > 0; \\ [1+\ln(x_k)] + \nu \geq 0, & x_k = 0; \end{cases} \text{for } 1 \leq k \leq n.$$

This implies

$$x_k^\diamond = e^{-\nu-1} \text{ for } 1 \leq k \leq n.$$

Note that if $x_\ell^\diamond$ equals zero for some specific ℓ, then we must have $\nu \geq -1-\ln(x_k) = \infty$, which implies $x_k^\diamond = e^{-\nu-1} = 0$, and a contradiction to the setting of $x_k^\diamond > 0$ is resulted.

(d) $\sum_{i=1}^n e^{-\nu-1} = 1$ implies $\nu^* = -1+\ln(n)$. Hence, $x_k^* = \frac{1}{n}$ for $1 \leq k \leq n$.

(e) Since $x_k^* > 0$ for $1 \leq k \leq n$, $\lambda_k^* = 0$ for $1 \leq k \leq n$.

6. Answer the following questions. Only a direct answer is required and no justification is needed.

(a) (5%) What are the three axioms raised by Shannon for the measurement of information?

(b) (5%) What is the limsup and liminf of the sequence $\{a_n = (-1)^n \cdot (1 + \frac{1}{n})\}$?

(c) (5%) If a sequence of (random) observations $\{x_n\}$ on a phenomenon constitutes a stationary process, does the strong law of large number hold?

Solution.

(a) i) Monotonicity in event probability

ii) Additivity for independent events

iii) Continuity in event probability

(b) $\limsup_{n\to\infty} a_n = 1$ and $\liminf_{n\to\infty} a_n = -1$

(c) No. The sample average does not guarantee to converge to the ensemble average (but to a random variable).

7. (5%) Prove the log-sum inequality in Problem 2(b) in terms of the fundamental inequality. Give the necessary and sufficient condition under which equality holds.

Hint: Subtract one side from the other side and apply the fundamental inequality: For any $x > 0$ and $D > 1$, we have that

$$\log_D(e) \cdot \left(1 - \frac{1}{x}\right) \le \log_D(x) \le \log_D(e) \cdot (x - 1),$$

with equality holding if, and only if, $x = 1$.

Solution.

$$\begin{aligned}
&\sum_{i=1}^{n}\left(a_i \log_D \frac{a_i}{b_i}\right) - \left(\sum_{i=1}^{n} a_i\right) \log_D \frac{\sum_{j=1}^{n} a_j}{\sum_{k=1}^{n} b_k} \\
&= \sum_{i=1}^{n} a_i \log_D \left(\frac{a_i}{b_i}\frac{\sum_{k=1}^{n} b_k}{\sum_{j=1}^{n} a_j}\right) \\
&\ge \log_D(e) \sum_{i=1}^{n} a_i \left(1 - \frac{b_i}{a_i}\frac{\sum_{j=1}^{n} a_j}{\sum_{k=1}^{n} b_k}\right) \\
&= \log_D(e) \left(\sum_{i=1}^{n} a_i - \sum_{i=1}^{n} b_i \cdot \frac{\sum_{j=1}^{n} a_j}{\sum_{k=1}^{n} b_k}\right) \\
&= \log_D(e) \left(\sum_{i=1}^{n} a_i - \sum_{j=1}^{n} a_j\right) \\
&= 0
\end{aligned}$$

with equality holding iff for all $i = 1, \cdots, n$,

$$\frac{b_i}{a_i}\frac{\sum_{j=1}^{n} a_j}{\sum_{k=1}^{n} b_k} = 1, \quad \text{i.e.,} \quad \frac{a_i}{b_i} = \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}.$$