Convexity/Concavity of Renyi Entropy and α-Mutual Information

Siu-Wai Ho
Institute for Telecommunications Research
University of South Australia
Adelaide, SA 5095, Australia
Email: siuwai.ho@unisa.edu.au

Sergio Verdú
Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544, U.S.A.
Email: verdu@princeton.edu

***Abstract*—Entropy is well known to be Schur concave on finite alphabets. Recently, the authors have strengthened the result by showing that for any pair of probability distributions P and Q with Q majorized by P, the entropy of Q is larger than the entropy of P by the amount of relative entropy $D(P||Q)$. This result applies to P and Q defined on countable alphabets. This paper shows the counterpart of this result for the Rényi entropy and the Tsallis entropy. Lower bounds on the difference in the Rényi (or Tsallis) entropy are given in terms of a new divergence which is related to the Rényi (or Tsallis) divergence. This paper also considers a notion of generalized mutual information, namely α-mutual information, which is defined through the Rényi divergence. The convexity/concavity for different ranges of α is shown. A sufficient condition for the Schur concavity is discussed and upper bounds on α-mutual information are given in terms of the Rényi entropy.**

I. Introduction

The Rényi entropy $H_\alpha(P)$ of a probability distribution P of order α was introduced in [1]. The concavity of $H_\alpha(P)$ for P defined on a finite support was shown in [2]. If $\alpha \in (0, 1]$, $H_\alpha(P)$ is strictly concave in P and it is also Schur concave [3]. When $\alpha > 1$, Rényi entropy is neither convex nor concave in general but it is pseudoconcave. Since all concave functions and pseudoconcave functions are quasiconcave, Rényi entropy is Schur concave for $\alpha > 0$ [4].

It is well known that the (Shannon) entropy of finitely-valued random variables is a Schur-concave function [4]. This result was strengthened and generalized for countably infinite random variables in [5], which shows that if a probability distribution Q is majorized by a probability distribution P, their difference in entropy was shown to be lower bounded by the relative entropy $D(P||Q)$. This motivates the search for strengthened results for the Schur-concavity of the Rényi entropy.

In order to strengthen the Schur-concavity of the Rényi entropy, this paper investigates the Schur-concavity of the Tsallis entropy. Tsallis entropy was introduced by Tsallis [6] to study multifractals. The same paper showed that the Tsallis entropy of order β is concave for $\beta > 0$ and convex for $\beta < 0$. The Schur-concavity of the Tsallis entropy was discussed in [7]. Tsallis divergence of order β was introduced in [8] and it is also known as the Hellinger divergence of order β [9]. Indeed, the Tsallis divergence is an f-divergence with $f(x) = \frac{x - x^\beta}{1-\beta}$ and the Tsallis entropy is the corresponding entropy with the form $-\sum_i f(P(i))$ for any probability distribution P.

This paper also considers the convexity of a notion of generalized mutual information, namely α-mutual information. This notion of mutual information was introduced by Sibson [10] in the discrete case, as the *information radius of order* α [11]. Alternative notions of what could be called Rényi mutual information were proposed by Csiszár [12] and Arimoto [13]. α-mutual information is closely related to Gallager's exponent function and plays an important role in a number of applications in channel capacity problems (e.g., [14]).

In Section II, the Schur concavity of the Rényi entropy and the Tsallis entropy are revisited. A new divergence is defined and new bounds on entropy in terms of the new divergence are derived. The α-mutual information is defined and discussed in Section III. Its concavity, convexity, Schur concavity and upper bounds under different settings are illustrated. Due to space limits, some of the proofs are omitted. For simplicity, only countable alphabets are considered in this paper. All $\log$ and $\exp$ have the same arbitrary base. We use $P \ll Q$ to denote that the support of P is a subset of the support of Q. Furthermore, if $P \not\ll Q$, it means that $P \ll Q$ does not hold.

II. Schur-Concavity of Rényi Entropy and Tsallis Entropy

Definition 1: For any probability distribution P, the Rényi entropy of order $\alpha \in (0, 1) \cup (1, \infty)$ is defined as

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{i \in \mathcal{A}} P(i)^\alpha \right), \tag{1}$$

where $\mathcal{A}$ is the support of P. By convention, $H_1(P)$ is the Shannon entropy of P.

Definition 2: For any probability distributions P and Q, let $\mathcal{A}$ be the union of the supports of P and Q. The Rényi divergence [1] of order $\alpha \in (0, 1) \cup (1, \infty)$ is defined as

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left(\sum_{i \in \mathcal{A}} P(i)^\alpha Q(i)^{1-\alpha} \right), \tag{2}$$

which becomes infinite if $\alpha > 1$ and $P \not\ll Q$.

Definition 3: Given $(P_X, P_{Y|X}, Q_{Y|X})$, the conditional Rényi divergence of order $\alpha \in (0,1) \cup (1,\infty)$ is

$$\begin{aligned} &D_\alpha(P_{Y|X}||Q_{Y|X}|P_X) && (3)\\ &= D_\alpha(P_{Y|X}P_X||Q_{Y|X}P_X) && (4)\\ &= \frac{1}{\alpha-1}\log\left(\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P^\alpha_{Y|X}(y|x)Q^{1-\alpha}_{Y|X}(y|x)P_X(x)\right), && (5)\end{aligned}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the supports of P_X and P_Y, respectively.

The Tsallis entropy is related to the Rényi entropy through the following one-to-one function. For $\alpha \in (0,1)\cup(1,\infty)$ and $z > 0$, let

$$\varphi_\alpha(z) = \exp(z - \alpha z) \quad (6)$$

and

$$\varphi^{-1}_\alpha(v) = \frac{1}{1-\alpha}\log v. \quad (7)$$

Definition 4: For any probability distribution P, the Tsallis entropy of order $\beta \in (0,1) \cup (1,\infty)$ is defined as

$$S_\beta(P) = \frac{\varphi_\beta(H_\beta(P)) - 1}{1-\beta} = \frac{1}{1-\beta}\left(\sum_{i\in\mathcal{A}} P(i)^\beta - 1\right), \quad (8)$$

where $\mathcal{A}$ is the support of P.

Definition 5: For any probability distributions P and Q, let $\mathcal{A}$ be the support of P. The Tsallis divergence of order $\beta \in (0,1)\cup(1,\infty)$ is defined as

$$\begin{aligned} T_\beta(P||Q) &= \frac{1-\varphi_\beta(-D_\beta(P||Q))}{1-\beta} && (9)\\ &= \frac{1}{1-\beta}\left(1 - \sum_{i\in\mathcal{A}} P(i)^\beta Q(i)^{1-\beta}\right), && (10)\end{aligned}$$

which becomes infinite if $\beta > 1$ and $P \not\ll Q$.

Definition 6: For any probability distributions P and Q and strictly concave function $\phi : [0,1] \to \mathbb{R}$, let

$$\begin{aligned} &\Delta_\phi(P(i), Q(i))\\ =&(P(i) - Q(i))\phi'(Q(i)) + \phi(Q(i)) - \phi(P(i)), && (11)\end{aligned}$$

where $\phi'(x)$ is the derivative of $\phi(x)$ and $i \in \mathcal{A}$ which is the union of the supports of P and Q. Define the divergence

$$W_\phi(P||Q) = \sum_{i\in\mathcal{A}} \Delta_\phi(P(i), Q(i)). \quad (12)$$

Lemma 1: For any probability distributions P and Q and strictly concave function $\phi : [0,1] \to \mathbb{R}$,

$$\Delta_\phi(P(i), Q(i)) \geq 0 \quad (13)$$

and

$$W_\phi(P||Q) \geq 0, \quad (14)$$

with equality if and only if $P = Q$.

In the following, we fix $\beta \in (0,1)\cup(1,\infty)$ and consider the strictly concave function

$$\phi_\beta(x) = \frac{x^\beta - x}{1-\beta}. \quad (15)$$

Its derivative is denoted by $\phi'_\beta(x) = \frac{\beta x^{\beta-1}-1}{1-\beta}$. For $0 \leq x \leq 1$, $\phi'_\beta(x)$ is a decreasing function. Note that the Tsallis divergence can be seen as an f-divergence with $f(x) = -\phi(x)$. The Tsallis entropy is equal to $-\sum_{i\in\mathcal{A}} f(P(i))$.

Without loss of generality, we assume that P and Q are defined on the positive integers and labeled in decreasing probabilities, i.e.,

$$\begin{aligned} P(i) &\geq P(i+1) && (16)\\ Q(i) &\geq Q(i+1). && (17)\end{aligned}$$

We say that Q is majorized by P if for all $k = 1, 2, \ldots$

$$\sum_{i=1}^k Q(i) \leq \sum_{i=1}^k P(i). \quad (18)$$

If Q is majorized by P, then $P \ll Q$. If a real-valued functional $f(\cdot)$ is such that $f(P) \leq f(Q)$ whenever Q is majorized by P, we say that $f(\cdot)$ is Schur-concave. The following result generalizes [5, Theorem 3].

Theorem 1: For $\beta \in (0,1)\cup(1,\infty)$, if Q is majorized by P, then

$$S_\beta(Q) \geq S_\beta(P) + W_{\phi_\beta}(P||Q). \quad (19)$$

Proof: We first consider the case where Q (and therefore P) takes values on a finite integer set $\{1, 2, \ldots, M\}$.

$$\begin{aligned} &S_\beta(Q) - S_\beta(P) - W_{\phi_\beta}(P||Q) && (20)\\ =&\sum_i \phi_\beta(Q(i)) - \sum_i \phi_\beta(P(i)) - W_{\phi_\beta}(P||Q) && (21)\\ =&\sum_i \phi'_\beta(Q(i))(Q(i) - P(i)) && (22)\\ =&\sum_i (Q(i) - P(i)) \sum_{k=i}^{M-1} (\phi'_\beta(Q(k)) - \phi'_\beta(Q(k+1))) +\\ &\sum_{i=1}^M (Q(i) - P(i))\phi'_\beta(Q(M)) && (23)\\ =&\sum_{k=1}^{M-1} (\phi'_\beta(Q(k)) - \phi'_\beta(Q(k+1)))\left(\sum_{i=1}^k Q(i) - \sum_{i=1}^k P(i)\right) && (24)\\ \geq&0, && (25)\end{aligned}$$

where (25) follows from the fact that $\phi'_\beta(x)$ is a decreasing function and Q is majorized by P.

Consider the case that Q is non-zero for all integers. We assume that $S_\beta(Q)$ is finite, as otherwise, there is nothing to prove. We need to take $M \to \infty$ in the foregoing expressions.

Although the last term in the right side of (23) can now be negative, it vanishes due to the finiteness of $S_\beta(Q)$:

$$\sum_{i=1}^{M}(Q(i)-P(i))\phi'_\beta(Q(M)) \geq -\sum_{i=M+1}^{\infty} Q(i)\phi'_\beta(Q(M)) \tag{26}$$

$$\geq -\sum_{i=M+1}^{\infty} Q(i)\phi'_\beta(Q(i)) \tag{27}$$

$$= -\frac{\beta}{1-\beta}\sum_{i=M+1}^{\infty} Q(i)^\beta, \tag{28}$$

where (27) follows from that $\phi'_\beta(x)$ is a decreasing function and (17). Therefore, the Tsallis entropy is a Schur-concave function. ∎

Due to the one-to-one correspondence between the Rényi entropy and the Tsallis entropy through (8), Theorem 1 leads to the following bounds. If $0<\alpha<1$, then

$$\varphi_\alpha(H_\alpha(Q)) - \varphi_\alpha(H_\alpha(P)) \geq (1-\alpha)W_{\phi_\alpha}(P||Q). \tag{29}$$

If $\alpha>1$, then

$$\varphi_\alpha(H_\alpha(P)) - \varphi_\alpha(H_\alpha(Q)) \geq (\alpha-1)W_{\phi_\alpha}(P||Q). \tag{30}$$

These bounds are sufficient to show that the Rényi entropy is Schur-concave. They can further be used to obtain bounds on $H_\alpha(Q)-H_\alpha(P)$ in terms of $W_{\phi_\alpha}(P||Q)$ as follows.

Theorem 2: For P defined on a countable alphabet, the Rényi entropy $H_\alpha(P)$ is a Schur-concave function for $\alpha \in (0,1)\cup(1,\infty)$. Furthermore, if Q is majorized by P, then

a) If $0<\alpha<1$, then in nats

$$H_\alpha(Q)-H_\alpha(P)$$

$$=\varphi_\alpha^{-1}\left(\frac{|\mathcal{A}|^{1-\alpha}}{|\mathcal{A}|^{1-\alpha}-(1-\alpha)W_{\phi_\alpha}(P||Q)}\right) \tag{31}$$

$$\geq \frac{W_{\phi_\alpha}(P||Q)}{|\mathcal{A}|^{1-\alpha}} \tag{32}$$

$$\geq 0, \tag{33}$$

where $\mathcal{A}$ is the support of Q.

b) If $\alpha>1$, then in nats

$$H_\alpha(Q)-H_\alpha(P) \geq \varphi_\alpha^{-1}\left(1-(\alpha-1)W_{\phi_\alpha}(P||Q)\right) \tag{34}$$

$$\geq W_{\phi_\alpha}(P||Q) \tag{35}$$

$$\geq 0. \tag{36}$$

It would be interesting to obtain an inequality similar to [5, Theorem 3] for the Rényi entropy. However, the following example shows that it is not possible.

Example 1: Consider $P=\{0.4,0.35,0.15,0.1\}$ and $Q=\{0.3,0.3,0.3,0.1\}$ so that Q is majorized by P. If $\alpha=0.6$, then $H_\alpha(Q)-H_\alpha(P)>0$ but $H_\alpha(Q)-H_\alpha(P)<D_\alpha(P||Q)$.

To end this section, we illustrate the meaning of $\Delta_\phi(P(i),Q(i))$ and how $W_\phi(P||Q)$ is related to relative

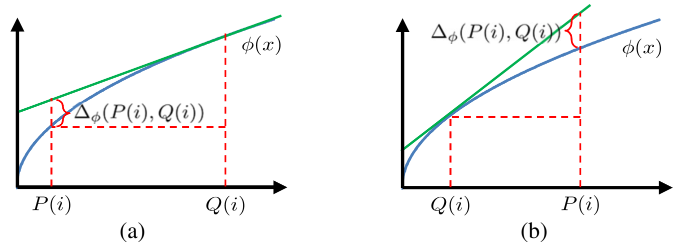

Fig. 1. An illustration of $\phi(x)$ and Δ_ϕ in case a) $P(i)<Q(i)$ and case b) $Q(i)<P(i)$.

entropy and Tsallis divergence. Consider the tangent at $\phi(x)$ when $x=Q(i)$ as shown in Fig. 1. Lemma 1 shows that the tangent is always above $\phi(P(i))$ and the amount is $\Delta_\phi(P(i),Q(i))$.

Theorem 3: For any probability distributions P and Q with $P \not\ll Q$ and $\phi(x)=x\log\frac{1}{x}$,

$$D(P||Q)=W_\phi(P||Q) \tag{37}$$

with the convention $0\log 0=0$. Furthermore, if $\phi_\beta(x)=\frac{x^\beta}{1-\beta}$,

$$\sum_{i\in\mathcal{A}} Q(i)^{1-\beta}\Delta_{\phi_\beta}(P(i),Q(i)) = T_\beta(P||Q), \tag{38}$$

where $\mathcal{A}$ is the support of Q, and hence for $0<\beta<1$,

$$T_\beta(P||Q) \leq W_{\phi_\beta}(P||Q). \tag{39}$$

III. α-Mutual Information

Definition 7: Let $P_X \to P_{Y|X} \to P_Y$, with $X\in\mathcal{X}$ and $Y\in\mathcal{Y}$. The α-mutual information with $\alpha\in(0,1)\cup(1,\infty)$ is

$$I_\alpha(X;Y)=\min_Q D_\alpha(P_{Y|X}\|Q|P_X) \tag{40}$$

$$=\frac{\alpha}{\alpha-1}\log\sum_{y\in\mathcal{Y}}\left(\sum_{x\in\mathcal{X}}P_X(x)P^\alpha_{Y|X=x}(y)\right)^{\frac{1}{\alpha}}. \tag{41}$$

By convention, $I_1(X;Y)$ is the mutual information.

Some properties of $I_\alpha(X;Y)$ related to the parameter α are summarized in the following theorem.

Theorem 4: For any fixed P_X on $\mathcal{X}$ and $P_{Y|X}:\mathcal{X}\to\mathcal{Y}$,

a) $I_\alpha(X;Y)$ is continuous with $\alpha\in(0,\infty)$ and

$$\lim_{\alpha\to 0} I_\alpha(X;Y) = -\sup_{y\in\mathcal{Y}}\log\left(\sum_{x:P_{Y|X}(y|x)>0}P_X(x)\right). \tag{42}$$

$$\lim_{\alpha\to 1} I_\alpha(X;Y)=I_1(X;Y). \tag{43}$$

$$\lim_{\alpha\to\infty} I_\alpha(X;Y)=\log\left(\sum_{y\in\mathcal{Y}}\sup_{x:P_X(x)>0}P_{Y|X}(y|x)\right). \tag{44}$$

b) $I_\alpha(X;Y)$ is increasing with α.

Proof: a). The limit in (42) can be obtained by using $L_\infty-$norm. For any $\epsilon > 0$, there exists a sufficiently small $0 < \alpha$ such that

$$\left(\sum_{x:P_{Y|X}(y|x)>0} P_X(x)\right) - \epsilon < \sum_{x\in\mathcal{X}} P_X(x)P^\alpha_{Y|X=x}(y) \quad (45)$$

$$< \sum_{x:P_{Y|X}(y|x)>0} P_X(x). \quad (46)$$

Since $\epsilon > 0$ is arbitrary, using L_∞-norm gives

$$\lim_{\alpha\to 0}\left(\sum_y\left(\sum_{x\in\mathcal{X}} P_X(x)P^\alpha_{Y|X=x}(y)\right)^{\frac{1}{\alpha}}\right)^\alpha \quad (47)$$

$$= \sup_{y\in\mathcal{Y}} \sum_{x:P_{Y|X}(y|x)>0} P_X(x). \quad (48)$$

So the limit in (42) follows.

The limits in (43) and (44) can be easily verified by using L'Hôpital's rule. Due to (43), $I_\alpha(X;Y)$ is continuous with $\alpha \in (0,\infty)$.

b). For $\alpha < \beta$, denote $Q^*_\beta = \operatorname{argmin}_Q D_\beta(P_{Y|X}||Q|P_X)$ so that

$$I_\alpha(X;Y) \le D_\alpha(P_{Y|X}||Q^*_\beta|P_X) \quad (49)$$

$$\le D_\beta(P_{Y|X}||Q^*_\beta|P_X) \quad (50)$$

$$= I_\beta(X;Y), \quad (51)$$

where (49) and (50) follow from (40) and [15, Theorem 3], respectively. ■

Provided that $C_{0f} > 0$, Theorem 4 intriguingly reveals that the zero-error capacity of a discrete memoryless channel with feedback [16] is equal to

$$C_{0f} = \sup_X I_0(X;Y), \quad (52)$$

where I_0 is defined as the continuous extension of I_α.

Some upper bounds on α-mutual information in terms of Rényi entropy are illustrated. These bounds provide some insights about generalizing conditional entropy.

Theorem 5: For any given $(P_X, P_{Y|X})$ and $0 < \alpha < \infty$,

$$H_{\frac{1}{\alpha}}(P_X) = I_\alpha(X;X) \ge I_\alpha(P_X, P_{Y|X}) \quad (53)$$

Proof: It is easy to verify the equality in (53). The inequality can be seen by the following. Let $U = X$ so that $X -\square- U -\square- Y$ forms a Markov chain. The inequality in (53) follows from [14, Theorem 5.2] for $\alpha \ne 1$ and [17] for $\alpha = 1$. ■

The difference between the sides in (53) is a good candidate for the conditional Rényi entropy for which several candidates have been suggested in the past (e.g., [12], [13]).

Since $H_\alpha(P_X)$ is decreasing in α [2], the following corollary holds.

Corollary 6: For any given $(P_X, P_{Y|X})$ and $0 < \alpha \le 1$,

$$H_\alpha(P_X) \ge I_\alpha(P_X, P_{Y|X}). \quad (54)$$

The following result shows the convexity (or concavity) of the conditional Rényi divergence, which proves to be useful to investigate the properties of α-mutual information.

Theorem 7: Given $(P_X, P_{Y|X}, Q_{Y|X})$, the conditional Rényi divergence $D_\alpha(P_{Y|X}||Q_{Y|X}|P_X)$ is

a) concave in P_X for $\alpha \ge 1$ and quasiconcave in P_X for $0 < \alpha < \infty$.

b) convex in $(P_{Y|X}, Q_{Y|X})$ for $0 < \alpha \le 1$ and quasiconvex in $(P_{Y|X}, Q_{Y|X})$ for $0 < \alpha < \infty$.

Proof: a). For $0 < \alpha < 1$, $\frac{1}{\alpha-1}\log a$ is a decreasing function in a. The quasiconcavity of $I_\alpha(X;Y)$ can be seen from (5). For $\alpha > 1$, $\frac{1}{\alpha-1}\log a$ is concave in a so that the concavity of $I_\alpha(X;Y)$ can be seen from (5). Since $I_1(X;Y)$ is concave in P_X, part a) is verified.

Part b) is a consequence of (4) together with [15, Theorem 13] and [15, Theorem 11]. ■

Theorem 8: Let $P_X \to P_{Y|X} \to P_Y$. For fixed $P_{Y|X}$,

a) $I_\alpha(P_X, P_{Y|X}) = I_\alpha(X;Y)$ is concave in P_X for $\alpha \ge 1$.

b) $I_\alpha(P_X, P_{Y|X}) = I_\alpha(X;Y)$ is quasiconcave in P_X for $0 < \alpha < \infty$.

Proof: Consider $0 < \lambda < 1$ and $\bar\lambda = 1 - \lambda$. We first prove part a). $I_1(X;Y)$ is known to be concave in P_X for fixed $P_{Y|X}$ [17]. Consider any probability distributions P_{X_0} and P_{X_1} and let

$$Q^* = \operatorname*{argmin}_Q D_\alpha(P_{Y|X}\|Q|\bar\lambda P_{X_0} + \lambda P_{X_1}). \quad (55)$$

Theorem 7 can be applied to show part a) as follows:

$$I_\alpha(\bar\lambda P_{X_0} + \lambda P_{X_1}, P_{Y|X})$$

$$= D_\alpha(P_{Y|X}\|Q^*|\bar\lambda P_{X_0} + \lambda P_{X_1}) \quad (56)$$

$$\ge \bar\lambda D_\alpha(P_{Y|X}\|Q^*|P_{X_0}) + \lambda D_\alpha(P_{Y|X}\|Q^*|P_{X_1}) \quad (57)$$

$$\ge \bar\lambda \min_Q D_\alpha(P_{Y|X}\|Q|P_{X_0}) + \lambda \min_Q D_\alpha(P_{Y|X}\|Q|P_{X_1}). \quad (58)$$

Similarly, part b) can also be shown by using Theorem 7. ■

Example 2: Although $I_\alpha(P_X, P_{Y|X})$ is concave in P_X for $\alpha \ge 1$, it can be shown that in general it is not Schur concave by considering $\lambda = 0.3$, $\alpha = 2$, $1 - P_{X_0}(0) = P_{X_0}(1) = 0.4$, $P_{X_1}(0) = P_{X_1}(1) = 0.5$, and

$$P_{Y|X} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}. \quad (59)$$

However, the Schur concavity can be preserved if $P_{Y|X}$ describes a symmetric channel. In other words, all the rows of the probability transition matrix $P_{Y|X}$ are permutations of other rows and so are the columns. Since all symmetric quasiconcave functions are Schur concave [4], Theorem 8 implies the following corollary.

Corollary 9: For any fixed symmetric $P_{Y|X}$, $I_\alpha(P_X, P_{Y|X}) = I_\alpha(X;Y)$ is Schur concave in P_X for $0 < \alpha < \infty$.

We now fix P_X and consider the behavior of $I_\alpha(P_X, \cdot)$.

Theorem 10: Let $P_X \to P_{Y|X} \to P_Y$. For fixed P_X,

a) $I_\alpha(P_X, P_{Y|X}) = I_\alpha(X;Y)$ is convex in $P_{Y|X}$ for $0 < \alpha \le 1$.

b) $I_\alpha(P_X, P_{Y|X}) = I_\alpha(X;Y)$ is quasiconvex in $P_{Y|X}$ for $0 < \alpha < \infty$.

Proof: Consider $0 < \lambda < 1$ and $\bar{\lambda} = 1 - \lambda$. We first prove part a). $I_1(X;Y)$ is known to be convex in $P_{Y|X}$ [17]. Consider any conditional probability distributions $P_{Y_0|X}$ and $P_{Y_1|X}$ and let

$$Q_i^* = \underset{Q}{\operatorname{argmin}}\, D_\alpha(P_{Y_i|X} \| Q | P_X) \tag{60}$$

for $i = 0$ or 1. Due to $0 < \alpha < 1$ and Theorem 7,

$$\begin{aligned}
& I_\alpha(P_X, \lambda P_{Y_1|X} + \bar{\lambda} P_{Y_0|X}) && (61)\\
=& \min_Q D_\alpha(\lambda P_{Y_1|X} + \bar{\lambda} P_{Y_0|X} || Q | P_X) && (62)\\
\le& D_\alpha(\lambda P_{Y_1|X} + \bar{\lambda} P_{Y_0|X} || \lambda Q_1^* + \bar{\lambda} Q_0^* | P_X) && (63)\\
\le& \lambda D_\alpha(P_{Y_1|X} || Q_1^* | P_X) + \bar{\lambda} D_\alpha(P_{Y_0|X} || Q_0^* | P_X) && (64)\\
=& \lambda I_\alpha(P_X, P_{Y_1|X}) + \bar{\lambda} I_\alpha(P_X, P_{Y_0|X}). && (65)
\end{aligned}$$

Similarly, Part b) can be verifed by using Theorem 7. ■

In Theorems 8 and 10, we have made different assumptions on the ranges of α. The following examples justify that those assumptions are not superfluous.

Example 3: Let $\lambda = 0.3$, $1 - P_{X_0}(0) = P_{X_0}(1) = 1$, $P_{X_1}(0) = P_{X_1}(1) = 0.5$, and

$$P_{Y|X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{66}$$

$I_{0.1}(P_X, P_{Y|X})$ is not concave in P_X and $I_{0.3}(P_X, P_{Y|X})$ is not convex in P_X.

Example 4: Let $\lambda = 0.3$, $P_X(0) = P_X(1) = 0.5$,

$$P_{Y_0|X_0} = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix} \quad \text{and} \quad P_{Y_1|X_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{67}$$

$I_2(P_X, P_{Y|X})$ is not concave in $P_{Y|X}$ and $I_3(P_X, P_{Y|X})$ is not convex in $P_{Y|X}$.

Although Examples 3 and 4 show some unpleasant behaviour of $I_\alpha(P_X, P_{Y|X})$, the Arimoto-Blahut algorithm [18] can still be applied to maximize $I_\alpha(P_X, P_{Y|X})$ for $\alpha > 0$. The following theorem, motivated by [14], illustrates that some optimization problems about $I_\alpha(P_X, P_{Y|X})$ can be converted into convex optimization problems.

Theorem 11: Consider $\alpha \in (0,1) \cup (1, \infty)$. Let

$$f_\alpha(P_X, P_{Y|X}) = \frac{1}{\alpha - 1} \varphi_{\frac{1}{\alpha}}(I_\alpha(P_X, P_{Y|X})) \tag{68}$$

$$= \frac{1}{\alpha - 1} \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} P_X(x) P^\alpha_{Y|X=x}(y) \right)^{\frac{1}{\alpha}}, \tag{69}$$

where $\frac{1}{\alpha-1}$ presents in (68) such that $\frac{1}{\alpha-1}\varphi_{\frac{1}{\alpha}}(z)$ is a monotonic increasing function in z.

a) For a fixed $P_{Y|X}$,

$$\sup_{P_X} I_\alpha(P_X, P_{Y|X}) = \varphi^{-1}_{\frac{1}{\alpha}} \left((\alpha - 1) \sup_{P_X} f_\alpha(P_X, P_{Y|X}) \right), \tag{70}$$

where $f_\alpha(P_X, P_{Y|X})$ is concave in P_X.

b) For a fixed P_X,

$$\inf_{P_{Y|X}} I_\alpha(P_X, P_{Y|X}) = \varphi^{-1}_{\frac{1}{\alpha}} \left((\alpha - 1) \inf_{P_{Y|X}} f_\alpha(P_X, P_{Y|X}) \right), \tag{71}$$

where $f_\alpha(P_X, P_{Y|X})$ is convex in $P_{Y|X}$.

Proof: Since $\frac{1}{\alpha-1}\varphi_{\frac{1}{\alpha}}(z)$ is a monotonic increasing function in z, (70) and (71) follow from the relation in (68). The function $f_\alpha(P_X, P_{Y|X})$ in (69) is concave in P_X because $\frac{1}{\alpha-1} z^{\frac{1}{\alpha}}$ is concave in z. The function $f_\alpha(P_X, P_{Y|X})$ in (69) is convex in $P_{Y|X}$ due to the Minkowski inequality for $\alpha > 1$ and the reverse Minkowski inequality for $\alpha < 1$. ■

Remark: Note that convexity/concavity were mistakenly reversed in [14, Theorem 5].

References

[1] A. Rényi, "On measures of entropy and information," *Fourth Berkeley Symp. on Mathematical Statistics and Probability*, pp. 547–561, 1961.

[2] M. Ben-Bassat and J. Raviv, "Rényi's entropy and the probability of error," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 324–331, May 1978.

[3] Y. He, A. B. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1211–1220, 2003.

[4] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*. Springer, 2010.

[5] S.-W. Ho and S. Verdú, "On the interplay between conditional entropy and error probability," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 5930–5942, 2010.

[6] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of statistical physics*, vol. 52, no. 1-2, pp. 479–487, 1988.

[7] S. Furuichi, K. Yanagi, and K. Kuriyama, "Fundamental properties of Tsallis relative entropy," *Journal of Mathematical Physics*, vol. 45, no. 12, pp. 4868–4877, 2004.

[8] C. Tsallis, "Generalized entropy-based criterion for consistent testing," *Physical Review E*, vol. 58, no. 2, p. 1442, 1998.

[9] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, Oct 2006.

[10] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[11] S. Verdú, "α-mutual information," *2015 Information Theory and Applications Workshop*, 2015.

[12] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. on Information Theory*, vol. 41, no. 1, pp. 26–34, 1995.

[13] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," *Proc. Coll. Math. Soc. János Bolyai.*, 1975.

[14] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1327–1333, 2010.

[15] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[16] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, 1956.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2012.

[18] S. Arimoto, "Computation of random coding exponent functions," *IEEE Transactions on Information Theory*, vol. 22, no. 6, pp. 665–671, 1976.