entropy

MDPI

Article

Conditional Rényi Divergence Saddlepoint and the Maximization of α-Mutual Information

Changxiao Cai [1] and Sergio Verdú [2,*]

[1] Department of Electrical Engineering, Princeton University, C307 Engineering Quadrangle, NJ 08540, USA; ccai@princeton.edu
[2] Independent Researcher, Princeton, NJ 08540, USA
* Correspondence: verdu@informationtheory.org

Received: 8 August 2019; Accepted: 25 September 2019; Published: 4 October 2019

Abstract: Rényi-type generalizations of entropy, relative entropy and mutual information have found numerous applications throughout information theory and beyond. While there is consensus that the ways A. Rényi generalized entropy and relative entropy in 1961 are the "right" ones, several candidates have been put forth as possible mutual informations of order α. In this paper we lend further evidence to the notion that a Bayesian measure of statistical distinctness introduced by R. Sibson in 1969 (closely related to Gallager's E_0 function) is the most natural generalization, lending itself to explicit computation and maximization, as well as closed-form formulas. This paper considers general (not necessarily discrete) alphabets and extends the major analytical results on the saddle-point and saddle-level of the conditional relative entropy to the conditional Rényi divergence. Several examples illustrate the main application of these results, namely, the maximization of α-mutual information with and without constraints.

Keywords: information measures; relative entropy; conditional relative entropy; mutual information; Rényi divergence; α-mutual information; channel capacity; minimax redundancy

1. Introduction

The Rényi divergence of order α between two probability measures defined on the same measurable space,

$$D_\alpha(P\,\|\,Q) = \frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right)^\alpha \mathrm{d}Q(x), \tag{1}$$

is a useful generalization of the relative entropy $D(P\|Q)$ introduced by Rényi [1] in the discrete case ($\lim_{\alpha\uparrow 1} D_\alpha(P\,\|\,Q) = D(P\|Q)$). Many of the properties satisfied by relative entropy hold for Rényi divergence, such as nonnegativity, convexity, lower semicontinuity, data processing inequality, and additivity for product measures. $D_\alpha(P\|Q)$ can be defined in more generality without requiring $P \ll Q$. A comprehensive survey of the properties satisfied by Rényi divergence can be found in [2]. Just as $D(P\|Q)$, $D_\alpha(P\,\|\,Q)$ provides a useful gauge of the distinctness of P and Q, which has found applications in large deviations problems (such as the asymptotic analysis of hypothesis testing [3–5]), lossless data compression [4,6,7], data transmission through noisy channels [8–10], and statistical physics [11].

If $P_1 \ll P_0$, then Rényi divergence of order $\alpha \in (0,1) \cup (1,\infty)$ can be expressed in terms of relative entropy through [5]

$$(1-\alpha) D_\alpha(P_1\|P_0) = \min_{P \ll P_1} \{\alpha\, D(P\|P_1) + (1-\alpha)\, D(P\|P_0)\}. \tag{2}$$

Although not an f-divergence, there is a one-to-one correspondence between Rényi divergence and Hellinger divergence $\mathscr{H}_\alpha(P\|Q)$ (e.g., [12])

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log\left(1 + (\alpha-1)\mathscr{H}_\alpha(P\|Q)\right). \tag{3}$$

One of the major applications of relative entropy is to quantify statistical dependence in a joint probability measure by means of the mutual information

$$I(X;Y) = D(P_{XY} \,\|\, P_X \times P_Y). \tag{4}$$

The corresponding straight generalization replacing relative entropy by Rényi divergence is also a measure of dependence but has found scant utility so far (see [6,13]). To explore the generalization that we study in this paper, namely α-mutual information, we need to consider the conditional versions of relative entropy and Rényi divergence. These are defined in general for two random transformations $P_{Y|X}$ and $Q_{Y|X}$ and an unconditional probability measure P_X simply as

$$D(P_{Y|X} \,\|\, Q_{Y|X}|P_X) = D(P_{Y|X}P_X\|Q_{Y|X}P_X), \tag{5}$$

$$D_\alpha(P_{Y|X} \,\|\, Q_{Y|X}|P_X) = D_\alpha(P_{Y|X}P_X\|Q_{Y|X}P_X). \tag{6}$$

A major difference between those conditional measures is that while $D(P_{Y|X} \,\|\, Q_{Y|X}|P_X)$ is plainly the expectation $\int D(P_{Y|X=x}\|\, Q_{Y|X=x})\, \mathrm{d}P_X(x)$, the conditional Rényi divergence depends on the function $D_\alpha(P_{Y|X=x}\|\, Q_{Y|X=x})$ in a more involved way. In this paper, the use of the conditional information measures will be circumscribed to the special case in which $Q_{Y|X}$ is actually an unconditional measure. In fact, a more productive way to express mutual information than (4) is the asymmetric expression

$$I(X;Y) = D(P_{Y|X} \,\|\, P_Y|P_X) \tag{7}$$

$$= \min_Q D(P_{Y|X} \,\|\, Q \,|\, P_X). \tag{8}$$

Equation (8) follows from the key additive decomposition formula

$$D(P_{Y|X}\|P_Y|P_X) = D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y), \tag{9}$$

where Q_Y is an arbitrary measure dominating P_Y. We see that (8) is a Bayesian measure of the distinctness of the constellation of probability measures $\{P_{Y|X=x}, x \in \mathcal{A}\}$, sometimes referred to as *information radius*, where the center of gravity of the constellation is none other than P_Y. Equation (8) has proven to be very fertile, particularly when it comes to supremize $I(X;Y)$ with respect to P_X since the ensuing sup min optimization has a saddle-point if and only if there is an input distribution that attains the maximal mutual information. The convexity of $D(P_{Y|X} \,\|\, Q \,|\, P_X)$ in Q and concavity (linearity) in P_X, along with the minimax theorem ensures the existence of the saddle-point whenever the set of allowed input distributions is compact. The Arimoto-Blahut algorithm [14,15] for finding $\max I(X;Y)$ in finite alphabet settings is also inspired by (8).

Mutual information $I(X;Y) = I(P_X, P_{Y|X})$ also possesses a saddle point (assuming convexity and compactness of the corresponding feasible sets) since it is concave in P_X (to see that, nothing better than (9)) and is convex in $P_{Y|X}$. This property has found rich applications in information theory (e.g., [16–18]) but neither it nor its generalization to α-mutual information will not concern us in this paper.

Even if a saddle-point for the conditional relative entropy does not exist, Kemperman [19] showed that sup min can be swapped, thereby establishing the existence of a *saddle value*.

Another well-known application of (8) and the conditional relative entropy saddle-point is the so-called *channel capacity-minimax redundancy* theorem due to Gallager [20] and Ryabko [21] (see also [22,23]), which shows that the maximal mutual information obtained with a finite constellation $\{P_{Y|X=x}, x \in \mathcal{A}\}$ is equal to the minimax redundancy in universal lossless data compression of an unknown source selected from $\{P_{Y|X=x}, x \in \mathcal{A}\}$. Notable generalizations of this result to infinite alphabets without requiring that a distribution maximizing mutual information exists are due to Kemperman [19] and Haussler [24]. Recently, the Rényi counterpart of the channel capacity-minimax redundancy result has also been considered, under various restrictions, in [2,25,26].

The main purpose of this paper is to generalize the saddle-point property of conditional relative entropy and its applications to the maximization of mutual information when relative entropy is replaced by Rényi divergence. Towards that end, we recall the various directions in which mutual information has been generalized using Rényi divergence (see also [27]) :

1. As aforementioned, the straight generalization $D_\alpha(P_{XY} \,\|\, P_X \times P_Y)$ has not yet found wide applicability.
2. In the discrete case and $\alpha \in (0,1) \cup (1,\infty)$, Arimoto [28] proposed the definition of the nonnegative quantity

$$I^{\mathsf{a}}(X;Y) = H_\alpha(X) - H^{\mathsf{a}}_\alpha(X|Y), \tag{10}$$

where the Rényi entropy [1] and Arimoto-Rényi conditional entropy [28] are

$$H_\alpha(X) = \frac{\alpha}{1-\alpha} \log \|P_X\|_\alpha, \tag{11}$$

$$H^{\mathsf{a}}_\alpha(X|Y) = \frac{\alpha}{1-\alpha} \log \mathbb{E}[\|P_{X|Y}(\cdot|Y)\|_\alpha] \tag{12}$$

with the α-norm of a probability mass function denoted as $\|P\|_\alpha = \left(\sum_{x\in\mathcal{A}} P^\alpha(x)\right)^{\frac{1}{\alpha}}$. Arimoto extended his algorithm in [14] to compute what he called the *capacity of order* α,

$$C^{\mathsf{a}}_\alpha = \max_X I^{\mathsf{a}}(X;Y), \tag{13}$$

for finite-alphabet random transformations and showed that there exist codes of rate R and blocklength n whose error probability is upper bounded by

$$\inf_{\alpha\in(\frac{1}{2},1)} \exp\left(\frac{\alpha-1}{\alpha}(C^{\mathsf{a}}_\alpha - R)\right).$$

3. Augustin [29] and, later, Csiszár [4] defined

$$I_\alpha^{\mathsf{c}}(X;Y) = \min_{Q_Y} \mathbb{E}\left[D_\alpha(P_{Y|X}(\cdot|X) \,\|\, Q_Y)\right]. \tag{14}$$

$C_\alpha^{\mathsf{c}} = \max_X I_\alpha^{\mathsf{c}}(X;Y)$ is dubbed the *Augustin capacity of order* α in [30]. Csiszár [4] showed that for $\alpha \in (\frac{1}{2}, 1)$, $I_\alpha^{\mathsf{c}}(X;Y)$ is the intercept on the R-axis of a supporting line of slope $1 - \frac{1}{\alpha}$ of the error exponent function for codes of rate R with constant-composition P_X. Unfortunately, the minimization in (14) is not amenable to explicit solution.

4. For the purpose of coming up with a measure of the similarity among a finite collection of probability measures $\{P_{Y|X=x}, x \in \mathcal{A}\}$, weighted by P_X on $\mathcal{A}$, Sibson [31] proposed the *information radius of order* α as

$$I_\alpha(X;Y) = \min_{Q} D_\alpha(P_{Y|X} \,\|\, Q|P_X). \tag{15}$$

As we will see, the minimization in (15) can be solved explicitly. This is the generalization of mutual information we adopt in this paper and which, as in [27], we refer to as α-mutual information. A word of caution is that in [4], the symbols $I_\alpha(X;Y)$ and $K_\alpha(X;Y)$ are used in lieu of what we denote $I_\alpha^{\mathsf{c}}(X;Y)$ and $I_\alpha(X;Y)$, respectively. $C_\alpha = \max_X I_\alpha(X;Y)$ is dubbed the *Rényi capacity of order* α in [26].

5. Independently, Lapidoth-Pfister [32] and Tomamichel-Hayashi [33] proposed

$$I_\alpha^{\mathsf{l}}(X;Y) = \min_{Q_X} \min_{Q_Y} D_\alpha(P_{XY} \,\|\, Q_X \times Q_Y). \tag{16}$$

and showed that it determines the performance of composite hypothesis tests for independence where the hypothesized joint distribution is known but under the independence hypothesis the marginals are unknown. It was shown in [34] that

$$I_\alpha^{\mathsf{c}}(X;Y) \le I_\alpha^{\mathsf{l}}(X;Y) \le I_\alpha(X;Y). \tag{17}$$

Despite the difference in the definitions of the various versions, it was shown in the discrete setting that [4,28].

$$C_\alpha^{\mathsf{a}} = C_\alpha^{\mathsf{c}} = C_\alpha \tag{18}$$

Therefore, solving for $\max_X I_\alpha(X;Y)$ carries added significance, whenever one of the other definitions is adopted. Note that (17) and (18) imply that $C_\alpha = \max_{P_X} I_\alpha^{\mathsf{l}}(P_X, P_{Y|X})$. A major application for the maximization of $I_\alpha(X;Y)$ is in the large deviation analysis of optimal data transmission codes since the sphere-packing error exponent function and the random-coding error exponent function

$$E_{\mathtt{sp}}(R) = \sup_{\rho \ge 0} \left\{\rho\, C_{\frac{1}{1+\rho}} - \rho\, R\right\}, \tag{19}$$

$$E_{\mathtt{r}}(R) = \sup_{\rho \in [0,1]} \left\{\rho\, C_{\frac{1}{1+\rho}} - \rho\, R\right\}, \tag{20}$$

popularized in [35] and [36], respectively, are upper and lower bounds to the channel reliability function, respectively. A function similar to (20) has recently been shown [37] to yield the large deviations behavior of random coding in the setting of channel resolvability.

The organization of the paper is as follows. Section 2 states the definitions and properties of the various information measures that are used throughout the paper. In particular, we introduce the key notion of α-response to an input probability measure through a given random transformation. In Section 3 we present the main results (with proofs relegated to Section 5) related to the saddle-point and saddle-value of the conditional Rényi divergence, allowing the optimization to be circumscribed to any convex set of input probability measures. The equivalence of the existence of a probability measure that maximizes α-mutual information and the existence of a saddle point is shown and several illustrative examples of the use of this result in the computation of C_α are also given. The fact that a saddle-level exists (i.e., sup min commute) even if there is no input probability measure that achieves the supremum α-mutual information is established, thereby generalizing Kemperman's [19] saddle-level result to Rényi divergence through a different route than that followed in [26].

2. Notation, Definitions and Properties

1. If $(\mathcal{A}, \mathscr{F}, P)$ is a probability space, $X \sim P$ indicates $\mathbb{P}[X \in \mathcal{F}] = P(\mathcal{F})$ for all $\mathcal{F} \in \mathscr{F}$.

2. Let $(\mathcal{A}, \mathscr{F})$ and $(\mathcal{B}, \mathscr{G})$ be measurable spaces, which we refer to as the input and output spaces, respectively, with $\mathcal{A}$ and $\mathcal{B}$ referred to as the input and output *alphabets* respectively. $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$ denotes a *random transformation* from $\mathcal{A}$ to $\mathcal{B}$, i.e. for any $x \in \mathcal{A}$, $P_{Y|X=x}(\cdot)$ is a probability measure on $(\mathcal{B}, \mathscr{G})$, and for any $B \in \mathscr{G}$, $P_{Y|X=\cdot}(B)$ is an $\mathscr{F}$-measurable function. For brevity, we will usually drop mention of the underlying σ-fields. If P is a probability measure on $\mathcal{A}$ and $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$ is a random transformation, the corresponding joint probability measure on $\mathcal{A} \times \mathcal{B}$ is denoted by $P\, P_{Y|X}$ (or, interchangeably, $P_{Y|X} P$). The notation $P \to P_{Y|X} \to Q$ indicates that the output marginal of the joint probability measure $P\, P_{Y|X}$ is denoted by Q.

3. The *relative information* $\imath_{P\|Q}(x)$ between two probability measures P and Q on the same measurable space such that $P \ll Q$ is defined as

$$\imath_{P\|Q}(x) = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(x), \tag{21}$$

where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative of P with respect to Q. The *relative entropy* is

$$D(P\|Q) = \mathbb{E}[\imath_{P_X\|Q_X}(X)], \quad X \sim P. \tag{22}$$

4. Given $P_X \to P_{Y|X} \to P_Y$, the *information density* is defined as

$$\imath_{X;Y}(a;b) = \imath_{P_{Y|X=a}\|P_Y}(b), \quad (a,b) \in \mathcal{A} \times \mathcal{B}. \tag{23}$$

5. Fix $\alpha > 0$, $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$, and a probability measure P_X on $\mathcal{A}$. Then, the output probability measure $P_{Y_{[\alpha]}}$ is called the α-*response* to P_X if

$$\imath_{Y_{[\alpha]}\|Y}(y) = \frac{1}{\alpha} \log \mathbb{E}[\exp(\alpha\, \imath_{X;Y}(X;y) - \kappa_\alpha)], \quad X \sim P_X, \tag{24}$$

where $P_X \to P_{Y|X} \to P_Y$, and κ_α is a scalar that guarantees that $P_{Y_{[\alpha]}}$ is a probability measure. For notational convenience, we omit the dependence of κ_α on P_X and $P_{Y|X}$. Equivalently, if $p_{Y_{[\alpha]}}$ and $p_{Y|X}$ denote the densities with respect to some dominating measure, then (24) becomes

$$p_{Y_{[\alpha]}}(y) = \exp\left(-\frac{\kappa_\alpha}{\alpha}\right) \mathbb{E}^{\frac{1}{\alpha}}\left[p^\alpha_{Y|X}(y|X)\right], \quad X \sim P_X. \tag{25}$$

In particular, the 1-*response* to P_X is P_Y. In [26], the α-response to P_X is dubbed the order α *Rényi mean for prior* P_X.

6. Given two probability measures P and Q on the same measurable space and a scalar $\alpha \in (0,1) \cup (1,\infty)$, the *Rényi divergence* of order α between P and Q is defined as [1]

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int_{\mathcal{A}} p^\alpha q^{1-\alpha} \, \mathrm{d}\mu, \tag{26}$$

where p and q are the Radon-Nikodym derivatives of P and Q, respectively, with respect to a common dominating σ-finite measure μ. We define $D_1(P\|Q) = D(P\|Q)$ as this coincides with the limit from the left at $\alpha = 1$. It is also the limit from the right whenever $D_\alpha(P\|Q) < \infty$ for some $\alpha > 1$. The cases $\alpha = 0$ and $\alpha = \infty$ can be defined by taking the corresponding limits. In this work, we only focus on the simple orders of α, i.e. $\alpha \in (0,1) \cup (1,\infty)$. As we saw in (1), if $P \ll Q$, then (26) becomes

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log\left(\mathbb{E}\left[\exp\left(\alpha\, \imath_{P\|Q}(W)\right)\right]\right), \quad W \sim Q \tag{27}$$

$$= \frac{1}{\alpha - 1} \log\left(\mathbb{E}\left[\exp\left((\alpha - 1)\, \imath_{P\|Q}(V)\right)\right]\right), \quad V \sim P \tag{28}$$

7. If $\alpha \in (0,1) \cup (1,\infty)$, then the binary Rényi divergence of order α is given by

$$d_\alpha(p\|q) = D_\alpha([p\ 1-p]\|[q\ 1-q]) \tag{29}$$

$$= \frac{1}{\alpha - 1} \log\left(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\right). \tag{30}$$

Note that

$$\lim_{\alpha \to 1} d_\alpha\left(p \| \frac{1}{2}\right) = \log 2 - h(p), \tag{31}$$

where the usual binary entropy function is denoted by $h(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$. Given $(p_0, p_1) \in (0,1)^2$, $p_0 \neq p_1$, the solution to $d_\alpha(p_0\|q) = d_\alpha(p_1\|q)$ is

$$q = \left(1 + \left(\frac{p_0^\alpha - p_1^\alpha}{(1-p_1)^\alpha - (1-p_0)^\alpha}\right)^{\frac{1}{1-\alpha}}\right)^{-1}. \tag{32}$$

8. $D_\alpha(P \,\|\, Q) \geq 0$, with equality only if $P = Q$.
9. $D_\alpha(P \,\|\, Q)$ is monotonically increasing with α.
10. While we may have $D(P\|Q) = \infty$ and $P \ll Q$ simultaneously, $D_\alpha(P\|Q) = \infty$ for any $\alpha \in (0,1)$ is equivalent to P and Q being orthogonal. Conversely, if for some $\alpha > 1$, $D_\alpha(P\|Q) < \infty$, then $P \ll Q$.

11. The Rényi divergence satisfies the *data-processing inequality*. If $P_X \to P_{Y|X} \to P_Y$ and $Q_X \to P_{Y|X} \to Q_Y$, then

$$D_\alpha(P_X\|Q_X) \geq D_\alpha(P_Y\|Q_Y). \tag{33}$$

12. Gilardoni [38] gave a strengthened Pinsker's inequality upper bounding the square of the total variation distance by

$$|P-Q|^2 \leq \inf_{\alpha\in(0,1]} \frac{1}{2\alpha} D_\alpha(P\|Q) \tag{34}$$

$$\leq \inf_{\alpha>0} \frac{1}{2\min\{\alpha,1\}} D_\alpha(P\|Q), \tag{35}$$

where we have used the monotonicity in α of the Rényi divergence.

13. The Rényi divergence is lower semicontinuous in the topology of setwise convergence, i.e., if for every event $A \in \mathscr{F}$, $P_n(A) \to P(A)$, and $Q_n(A) \to Q(A)$, then

$$\liminf_{n\to\infty} D_\alpha(P_n\|Q_n) \geq D_\alpha(P\|Q), \quad \alpha \in (0,\infty]. \tag{36}$$

In particular, note that (36) holds if $|P_n - Q_n| \to 0$.

14. In the theory of robust lossless source coding [22,25] the following scalar, called the *α-minimax redundancy* of $P_{Y|X}$, is an important measure of the worst-case redundancy penalty that ensues when the encoder only knows that the data is generated according to one of the probability measures in the collection $\{P_{Y|X=x}, x \in \mathcal{A}\}$:

$$R_\alpha = \inf_{Q_Y} \sup_{x\in\mathcal{A}} D_\alpha(P_{Y|X=x}\|\, Q_Y), \tag{37}$$

where the infimum is over all the probability measures on $\mathcal{B}$.

15. Given input distribution P_X and random transformations $P_{Y|X}, Q_{Y|X}$, the *conditional Rényi divergence* of order $\alpha \in (0,1)\cup(1,\infty)$ is

$$D_\alpha(P_{Y|X}\,\|\,Q_{Y|X}|P_X) = D_\alpha(P_X P_{Y|X}\,\|\,P_X Q_{Y|X}). \tag{38}$$

Although (38) also holds for the familiar $\alpha = 1$ case, in general the conditional Rényi divergence is *not* the arithmetic average of $D(P_{Y|X=x}\|Q_Y)$ with respect to P_X if $\alpha \neq 1$. Instead it's a generalized mean, or a scaled cumulant generating function evaluated at $\alpha - 1$. Specifically, if $X \sim P_X$, then

$$D_\alpha(P_{Y|X}\|Q_Y|P_X) = \frac{1}{\alpha-1}\log \mathbb{E}\left[\exp\left((\alpha-1)D_\alpha(P_{Y|X}(\cdot|X)\|Q_Y)\right)\right]. \tag{39}$$

Regardless of whether $\alpha \in (0,1)$ or $\alpha \in (1,\infty)$, (39) implies that

$$D_\alpha(P_{Y|X}\|Q_Y|P_X) \le \sup_{x\in\mathcal{A}} D_\alpha(P_{Y|X=x}\|Q_Y) \tag{40}$$

$$= \sup_{P_X} D_\alpha(P_{Y|X}\|Q_Y|P_X) \tag{41}$$

with the supremum in (41) over all input probability measures.

16. The key additive decomposition formula for the mutual information (9) has a nice counterpart for the α-mutual information [27]. Let $P_X \to P_{Y|X} \to P_Y$ and Q_Y be an arbitrary probability measure on $\mathcal{B}$ such that $P_Y \ll Q_Y$. Then, it is easy to verify that

$$D_\alpha(P_{Y|X}\|P_{Y_{[\alpha]}}|P_X) = D_\alpha(P_{Y|X}\|Q_Y|P_X) - D_\alpha(P_{Y_{[\alpha]}}\|Q_Y), \tag{42}$$

a relationship noted by Sibson [31] in the discrete case.

17. Given $\alpha > 0$, P_X and $P_{Y|X}$, the *α-mutual information* is [27,31]

$$I_\alpha(X;Y) = \min_Q D_\alpha(P_{Y|X} \,\|\, Q|P_X) \tag{43}$$

$$= D_\alpha(P_{Y|X} \,\|\, P_{Y_{[\alpha]}}|P_X) \tag{44}$$

$$= \frac{1}{\alpha-1}\log \mathbb{E}\left[\exp\left((\alpha-1)D_\alpha(P_{Y|X}(\cdot|X)\|P_{Y_{[\alpha]}})\right)\right] \tag{45}$$

$$= D_\alpha(P_{Y|X} \,\|\, P_Y|P_X) - D_\alpha(P_{Y_{[\alpha]}} \,\|\, P_Y), \tag{46}$$

where $P_X \to P_{Y|X} \to P_Y$. It can be checked that the constant in (24) is equal to

$$\kappa_\alpha = (\alpha-1)I_\alpha(X;Y). \tag{47}$$

Note that $I_1(X;Y) = I(X;Y)$ but, in general, $I_\alpha(X;Y) \neq I_\alpha(Y;X)$.

18. An alternative expression for α-mutual information, which will come in handy in our analysis and which does not involve either P_Y or $P_{Y_{[\alpha]}}$ is obtained by introducing an auxiliary probability measure $P_{\bar{Y}}$ dominating the collection $\{P_{Y|X=u}, u \in \mathcal{A}\}$ [27]:

$$I_\alpha(X;Y) = \frac{\alpha}{\alpha-1}\log \mathbb{E}[\mathbb{E}^{\frac{1}{\alpha}}[\exp(\alpha\, \imath_{X;\bar{Y}}(X;\bar{Y}))|\bar{Y}]], \quad (X,\bar{Y}) \sim P_X \times P_{\bar{Y}}, \tag{48}$$

where

$$\imath_{X;\bar{Y}}(x;y) = \log \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_{\bar{Y}}}(y). \tag{49}$$

As usual, sometimes it is convenient to fix σ-finite measures μ_X and μ_Y on the input and output spaces which dominate P_X and $\{P_{Y|X=x} : x \in \mathcal{A}\}$, respectively, and denote their densities with respect to the reference measures by

$$p_{Y|X}(y|x) = \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}\mu_Y}(y), \tag{50}$$

$$p_X(x) = \frac{\mathrm{d}P_X}{\mathrm{d}\mu_X}(x). \tag{51}$$

Then, we can write α-mutual information as

$$I_\alpha(P_X, P_{Y|X}) = \frac{\alpha}{\alpha - 1} \log \int_{\mathcal{B}} \left(\int_{\mathcal{A}} p^\alpha_{Y|X}(y|x) p_X(x) \, \mathrm{d}\mu_X(x) \right)^{\frac{1}{\alpha}} \mathrm{d}\mu_Y(y). \tag{52}$$

19. In the special case of discrete alphabets,

$$E_0(\rho, P_X, P_{Y|X}) = \rho \, I_{\frac{1}{1+\rho}}(X;Y), \tag{53}$$

where the left side is the familiar Gallager function defined in [36] for $\rho \in (0,1)$ as

$$E_0(\rho, P_X, P_{Y|X}) = -\log \sum_{y \in \mathcal{B}} \left(\sum_{x \in \mathcal{A}} P_X(x) P^{\frac{1}{1+\rho}}_{Y|X}(y|x) \right)^{1+\rho}. \tag{54}$$

20. Fix $\alpha > 0$, $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$, and a collection $\mathcal{P}$ of probability measures on the input space. Then, we denote

$$C_\alpha(\mathcal{P}) = \sup_{P_X \in \mathcal{P}} I_\alpha(P_X, P_{Y|X}). \tag{55}$$

When $\mathcal{P}$ is the set of all input measures we write simply C_α, dubbed the *Rényi capacity* in [26]. C_α is a measure of the similarity of the family $\{P_{Y|X=x}, x \in \mathcal{A}\}$, which plays an important role in the analysis of the fundamental limits of information transmission through noisy channels, particularly in the regime of exponentially small error probability. For a long time (e.g., [39]) the *cutoff rate* $C_{\frac{1}{2}}$ was conjectured to be the maximal rate for which reliable codes with manageable decoding complexity can be found. The zero-error capacity of the discrete memoryless channel with feedback is equal to either zero or [40]

$$C_{0\mathsf{f}} = \max_X I_0(X;Y), \tag{56}$$

depending on whether there is $(a_1, a_2) \in \mathcal{A}^2$ such that $P_{Y|X}(\cdot|a_1) \perp P_{Y|X}(\cdot|a_2)$.

21. The related quantity $\max_{P_X} I_{\frac{1}{\alpha}}(P_X, P^\alpha_{Y|X})$ arises in the study of the fundamental limits of guessing and task completion under mismatch [41,42].

22. While $D(P\|Q)$ is *convex* in the pair (P, Q), the picture for Rényi divergence is somewhat more nuanced:

(a) If $\alpha \in (0,1)$, then $D_\alpha(P \,\|\, Q)$ is convex in (P,Q).

(b) If $\alpha > 0$, then $D_\alpha(P \,\|\, Q)$ is convex in Q for all P, (see [4]).

23. For any fixed pair $(P_{Y|X}, Q_{Y|X})$, $D_\alpha(P_{Y|X}\|Q_{Y|X}|P_X)$ is concave (resp. convex) in P_X if $\alpha \geq 1$ (resp. $\alpha \in (0,1]$) (see [43]).

24. The α-mutual information $I_\alpha(P_X, P_{Y|X})$ is concave in P_X for any fixed $P_{Y|X}$ and $\alpha > 1$ (see [43]). If $\alpha \in (0,1) \cup (1,\infty)$, then the following monotonically increasing function of $I_\alpha(P_X, P_{Y|X})$ is concave in P_X

$$\Gamma_\alpha \left(I_\alpha(P_X, P_{Y|X}) \right) = \frac{1}{\alpha - 1} \varphi_{\frac{1}{\alpha}} \left(I_\alpha(P_X, P_{Y|X}) \right), \tag{57}$$

where $\varphi_\alpha(z) = \exp(z - \alpha z)$ (see [10,43]).

3. Conditional Rényi Divergence Game

As can be expected from (43), when maximizing α-mutual information, for fixed $P_{Y|X}$, with respect to the input probability measure, it is interesting to consider a zero-sum game with payoff function

$$D_\alpha(P_{Y|X} \,\|\, Q|P_X)$$

such that one player tries to maximize it by choosing $P_X \in \mathcal{P}$, where $\mathcal{P}$ is a given collection of input probability measures, and the other player tries to minimize it by choosing the probability measure $Q \in \mathcal{Q}$ on the output space. Balancing simplicity and generality and motivated by applications, while we allow $\mathcal{P}$ to be a proper subset of the set of all input probability measures, we assume that there are no restrictions in the choice of the output probability measure, and therefore $\mathcal{Q}$ stands for the whole collection of probability measures on the output space. This game also arises in the determination of the worst-case redundancy in (37). In Section 3.1 we consider the important special case in which there exists an input distribution that attains the supremum in (55). In the more general scenario in which the supremum may not be achieved, we cannot identify a saddle point but we can indeed swap sup and min as we show in Section 3.2.

3.1. Saddle point

We begin by showing that the maximal α-mutual information input distribution and its α-response form a saddle point.

Theorem 1. *Let $\mathcal{P}$ be a convex set of probability distributions on $\mathcal{A}$ and $\mathcal{Q}$ be the set of all probability distributions on $\mathcal{B}$. Let $\alpha \in (0,1) \cup (1,\infty)$. Suppose that there exists some $P_X^* \in \mathcal{P}$ such that*

$$I_\alpha(P_X^*, P_{Y|X}) = \max_{P_X \in \mathcal{P}} I_\alpha(P_X, P_{Y|X}) < \infty, \tag{58}$$

and denote the α-response to P_X^* by $P_{Y_{[\alpha]}}^*$. Then, for any $(P_X, Q_Y) \in \mathcal{P} \times \mathcal{Q}$,

$$D_\alpha(P_{Y|X}\|P_{Y_{[\alpha]}}^*|P_X) \leq D_\alpha(P_{Y|X}\|P_{Y_{[\alpha]}}^*|P_X^*) \tag{59}$$

$$\leq D_\alpha(P_{Y|X}\|Q_Y|P_X^*). \tag{60}$$

Conversely, if $(P_X^*, P_{Y_{[\alpha]}}^*)$ is a saddle point of $D_\alpha(P_{Y|X}\| \cdot |\cdot)$, namely, (59)–(60) are satisfied, then P_X^* maximizes the α-mutual information.

Remark 1. *Assuming that $\mathcal{P}$ includes δ_x (unit mass at $x \in \mathcal{A}$), (59) implies that for any $x \in \mathcal{A}$,*

$$D_\alpha(P_{Y|X=x} \| P^*_{Y_{[\alpha]}}) \le \max_{P_X \in \mathcal{P}} I_\alpha(P_X, P_{Y|X}). \tag{61}$$

We can easily obtain corollaries to Theorem 1 that elucidate useful properties of the saddle point.

Corollary 1. *Let $\alpha \in (0,1) \cup (1,\infty)$. Under the assumptions in Theorem 1, for any $P_X \in \mathcal{P}$, we have*

$$D_\alpha(P_{Y_{[\alpha]}} \| P^*_{Y_{[\alpha]}}) \le I_\alpha(P^*_X, P_{Y|X}) - I_\alpha(P_X, P_{Y|X}) < \infty, \tag{62}$$

where $P_{Y_{[\alpha]}}$ is the α-response to P_X. Moreover, $P_{Y_{[\alpha]}} = P^*_{Y_{[\alpha]}}$ if, in addition to P^*_X, P_X also attains $C_\alpha(\mathcal{P}) = \max_{P_X \in \mathcal{P}} I_\alpha(P_X, P_{Y|X})$.

Proof of Corollary 1. For any $P_X \in \mathcal{P}$,

$$\begin{aligned} I_\alpha(P_X, P_{Y|X}) &= D_\alpha(P_{Y|X} \| P_{Y_{[\alpha]}} | P_X) && (63)\\ &= D_\alpha(P_{Y|X} \| P^*_{Y_{[\alpha]}} | P_X) - D_\alpha(P_{Y_{[\alpha]}} \| P^*_{Y_{[\alpha]}}) && (64)\\ &\le D_\alpha(P_{Y|X} \| P^*_{Y_{[\alpha]}} | P^*_X) - D_\alpha(P_{Y_{[\alpha]}} \| P^*_{Y_{[\alpha]}}) && (65)\\ &= I_\alpha(P^*_X, P_{Y|X}) - D_\alpha(P_{Y_{[\alpha]}} \| P^*_{Y_{[\alpha]}}), && (66) \end{aligned}$$

where (64) and (65) follow from (42) and (59), respectively. Since Rényi divergence is nonnegative, $D_\alpha(P_{Y_{[\alpha]}} \| P^*_{Y_{[\alpha]}}) = 0$ if P_X also attains $C_\alpha(\mathcal{P})$. □

Therefore, Corollary 1 implies that the α-responses to all the maximal α-mutual information input distributions must be identical. Moreover, if $\alpha > 1$, then every α-response to any input distribution satisfies $P_{Y_{[\alpha]}} \ll P^*_{Y_{[\alpha]}}$.

If $\mathcal{P}$ is the space of all probability distributions on $\mathcal{A}$, then we can get the following corollary.

Corollary 2. *Unconstrained maximization of α-mutual information. Suppose that $\alpha \in (0,1) \cup (1,\infty)$ and $\mathcal{P}$ contains all probability mass functions on the discrete alphabet $\mathcal{A}$. Fix $P_{Y|X}\colon \mathcal{A} \to \mathcal{B}$. For any input distribution $\bar{P}_X$, denote its support by $\bar{\mathcal{A}}_X \subset \mathcal{A}$ and the corresponding α-response by $\bar{P}_{Y_{[\alpha]}}$.*

A necessary and sufficient condition for $\bar{P}_X$ to achieve $\max_X I(X;Y) < \infty$ is

$$\max_{a \in \bar{\mathcal{A}}_X} D_\alpha(P_{Y|X=a} \| \bar{P}_{Y_{[\alpha]}}) = \min_{a \in \bar{\mathcal{A}}_X} D_\alpha(P_{Y|X=a} \| \bar{P}_{Y_{[\alpha]}}) \ge \max_{a \in \bar{\mathcal{A}}^c_X} D_\alpha(P_{Y|X=a} \| \bar{P}_{Y_{[\alpha]}}). \tag{67}$$

Proof of Corollary 2.

- $\max_X I_\alpha(X;Y) = I_\alpha(\bar{P}_X, P_{Y|X}) \Rightarrow$ (67): Regardless of whether $\alpha > 1$ or $\alpha < 1$, we see from (45) that if there exists some $x_0 \in \bar{\mathcal{A}}_X$ such that

$$D_\alpha\left(P_{Y|X=x_0} \| \bar{P}_{Y_{[\alpha]}}\right) < \max_{P_X \in \mathcal{P}} I_\alpha(P_X, P_{Y|X}), \tag{68}$$

then $I_\alpha(\bar{P}_X, P_{Y|X}) < \max_{P_X \in \mathcal{P}} I_\alpha(P_X, P_{Y|X})$, which contradicts the assumed optimality of $\bar{P}_X$. Moreover, if there exists some $x_0 \in \bar{\mathcal{A}}^c_X$ such that (68) holds with the strict inequality reversed, then (59) would be violated, again contradicting the assumed optimality of $\bar{P}_X$.

- (67) $\Rightarrow \max_X I_\alpha(X;Y) = I_\alpha(\bar{P}_X, P_{Y|X})$: Again, we see from (45) that if (67) is satisfied, then (59) is satisfied. Since $\bar{P}_{Y_{[\alpha]}}$ is the α-response to $\bar{P}_X$, (59) is also satisfied, and the converse part in Theorem 1 results in the optimality of $\bar{P}_X$.

□

Remark 2. *According to Corollary 2, if some input distribution P_X^* achieves C_α, we know the α-response output distribution $P_{Y_{[\alpha]}}^*$ is equidistant in $D_\alpha(\cdot\|P_{Y_{[\alpha]}}^*)$ to any of the output distributions in the collection*

$$\mathcal{S} = \{P_{Y|X=x}, \quad P_X^*(x) > 0\} \subset \mathcal{Q}. \tag{69}$$

Moreover, we know that the optimal α-response output distribution is actually unique even if there exist several optimal input distributions. So the key is to find the unique centroid of $\mathcal{S}$ when the distance is measured by the Rényi divergence. In contrast to the maximization of the mutual information, the optimal α-response output distribution is no longer a mixture of the conditional output distributions.

Remark 3. *Corollary 2 enables us to recover Gallager's finite alphabet result in Theorem 5.6.5 of [44], which characterizes the maximal α-mutual information input distribution if $\alpha \in (0,1)$ when both $\mathcal{A}$ and $\mathcal{B}$ are finite. The optimal input distribution P_X^* must satisfy the following condition:*

$$\sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} P_{Y|X}^\alpha(y|x) P_X^*(x)\right)^{\frac{1-\alpha}{\alpha}} P_{Y|X}^\alpha(y|u) \geq \sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} P_{Y|X}^\alpha(y|x) P_X^*(x)\right)^{\frac{1}{\alpha}}, \tag{70}$$

with equality for all u such that $P_X^*(u) > 0$. To verify this condition, note that Corollary 2 requires that

$$\exp(\kappa_\alpha^*) = \exp\left((\alpha-1)C_\alpha\right) \leq \sum_{y\in\mathcal{B}} P_{Y|X=u}^\alpha(y) \left(P_{Y_{[\alpha]}}^*(y)\right)^{1-\alpha}, \tag{71}$$

with equality if $P_X^*(u) > 0$, and where κ_α^* stands for the normalizing constant in (24) with $P_X \leftarrow P_X^*$. Upon substitution of (25) with $P_X \leftarrow P_X^*$, we obtain (70). The assumption of finite output alphabet can be easily dispensed with to obtain the more general optimality condition

$$\mathbb{E}\left[\Psi_\alpha^{1-\alpha}(Y^*) \exp\left(\alpha\, \imath_{X;Y}^*(u;Y^*)\right)\right] \geq \mathbb{E}\left[\Psi_\alpha(Y^*)\right] \tag{72}$$

with equality for all u such that $P_X^*(u) > 0$. In (72), $\imath_{X;Y}^*$ stands for the information density corresponding to $P_X^* \to P_{Y|X} \to P_Y^*$ and

$$\Psi_\alpha(y) = \mathbb{E}^{\frac{1}{\alpha}}\left[\exp\left(\alpha\, \imath_{X;Y}^*(X^*;y)\right)\right], \quad X^* \sim P_X^* \tag{73}$$

If $\alpha > 1$, condition (72) holds with the inequality reversed.

Remark 4. *When $\mathcal{B}$ is finite, it was shown in [2,4,25] that for any $\alpha \in [0,\infty]$,*

$$C_\alpha = R_\alpha, \tag{74}$$

where R_α is defined in (37). This is now established without imposing finiteness conditions, as long as there is an input that achieves the maximal α-mutual information because

$$\begin{aligned} R_\alpha &\le C_\alpha && (75)\\ &= \max_{P_X \in \mathcal{P}} \min_{Q_Y \in \mathcal{Q}} D_\alpha(P_{Y|X} \| Q_Y | P_X) && (76)\\ &\le \min_{Q_Y \in \mathcal{Q}} \max_{P_X \in \mathcal{P}} D_\alpha(P_{Y|X} \| Q_Y | P_X) && (77)\\ &\le R_\alpha, && (78) \end{aligned}$$

where (75) follows from particularizing (59) to deterministic P_X, and (78) follows from (40).

3.2. Minimax identity

In this section we drop the assumption that there exists an input probability measure that attains the maximal α-mutual information and show that the conditional Rényi divergence still satisfies a minimax identity, even if a saddle point does not exist.

Theorem 2. *Let $\mathcal{P}$ be a convex set of probability distributions on $\mathcal{A}$ and $\mathcal{Q}$ be the set of all probability distributions on $\mathcal{B}$. We have the minimax equality:*

$$\begin{aligned} C_\alpha(\mathcal{P}) &= \sup_{P_X \in \mathcal{P}} \min_{Q_Y \in \mathcal{Q}} D_\alpha(P_{Y|X} \| Q_Y | P_X) && (79)\\ &= \min_{Q_Y \in \mathcal{Q}} \sup_{P_X \in \mathcal{P}} D_\alpha(P_{Y|X} \| Q_Y | P_X). && (80) \end{aligned}$$

Furthermore, if $C_\alpha(\mathcal{P}) < \infty$, then there exists a unique element in $\mathcal{Q}$ attaining the minimum in (80).

The assumption of convexity in Theorem 2 is not superfluous, as the following example illustrates.

Example 1. *Let $\mathcal{A} = \mathcal{B} = \mathbb{N}$ and $Y = X + N$, where N is a geometric random variable on the nonnegative integers with positive mean and independent of X. Let $\mathcal{P}$ be the non-convex set of all the deterministic probability measures on $\mathcal{A}$. In this case, the left side of (80) is zero, while the right side is infinity. To see this, note that for any $Q_Y \in \mathcal{Q}$ and $n \in \mathbb{N}$, it follows from the data processing inequality applied to the binary deterministic transformation $1\{Y \ge n\}$ that*

$$D_\alpha(P_{Y|X=n} \| Q_Y) \ge d_\alpha\left(1 \| Q_Y(\{n+1, \ldots\})\right), \tag{81}$$

whose right side diverges as $n \to \infty$. Therefore, for any $Q_Y \in \mathcal{Q}$, (39) results in

$$\sup_{P_X \in \mathcal{P}} D_\alpha(P_{Y|X} \| Q_Y | P_X) = \infty. \tag{82}$$

Continuing with the theme in Remark 4, Theorem 2 extends the validity of $R_\alpha = C_\alpha$ without requiring the existence of the maximal α-mutual information input distribution. It was conjectured in [2] (and proved in [26]) that for $\alpha \in (0, \infty)$, if $R_\alpha < \infty$ and $\mathcal{B}$ is finite or countable, there exists a unique redundancy-achieving distribution

$$Q_Y^* = \arg\min_{Q_Y} \sup_{x \in \mathcal{A}} D_\alpha(P_{Y|X=x} \| Q_Y) \tag{83}$$

and for all probability measures Q_Y on the output space,

$$\sup_{x\in\mathcal{A}} D_\alpha(P_{Y|X=x}\|Q_Y) \geq C_\alpha + D_\alpha(Q_Y^*\|Q_Y). \tag{84}$$

We can prove the conjecture easily with the help of Theorem 2.

Proof. Let $\mathcal{P}$ be the convex set of all probability measures on $\mathcal{A}$. Since $C_\alpha = R_\alpha < \infty$, by Theorem 2, we know there exists a unique $P^*_{Y_{[\alpha]}}$ such that $\sup_{P_X\in\mathcal{P}} D_\alpha(P_{Y|X}\|P^*_{Y_{[\alpha]}}|P_X) = C_\alpha$, which implies that $P^*_{Y_{[\alpha]}}$ is precisely the unique redundancy-achieving distribution in (83). Moreover, as shown in the proof of Theorem 2, we can find a sequence $\{P_{X_n}\}_{n\geq 1}$ in $\mathcal{P}$ such that $I_\alpha(P_{X_n}, P_{Y|X}) \to C_\alpha$ as $n\to\infty$ and such that the corresponding α-responses $P_{Y_{n[\alpha]}}$ converge to $P^*_{Y_{[\alpha]}}$ in the total variation metric. Pick an arbitrary $Q_Y \in \mathcal{Q}$. If $\alpha > 1$ and $P_{Y|X=x} \not\ll Q_Y$ for some $x\in\mathcal{A}$, then $\sup_{x\in\mathcal{A}} D_\alpha(P_{Y|X=x}\|Q_Y) = \infty$ and (84) holds. Otherwise, by (42) we always have

$$D_\alpha(P_{Y|X}\|Q_Y|P_X) = D_\alpha(P_{Y|X}\|P^*_{Y_{[\alpha]}}|P_X) + D_\alpha(P^*_{Y_{[\alpha]}}\|Q_Y). \tag{85}$$

For any $n\geq 1$, since $\mathcal{P}$ includes all probability measures on $\mathcal{A}$, we have

$$\begin{aligned}
\sup_{x\in\mathcal{A}} D_\alpha(P_{Y|X=x}\|Q_Y) &= \sup_{P_X\in\mathcal{P}} D_\alpha(P_{Y|X}\|Q_Y|P_X) && (86)\\
&\geq D_\alpha(P_{Y|X}\|Q_Y|P_{X_n}) && (87)\\
&= D_\alpha(P_{Y|X}\|P_{Y_{n[\alpha]}}|P_{X_n}) + D_\alpha(P_{Y_{n[\alpha]}}\|Q_Y) && (88)\\
&= I_\alpha(P_{X_n}, P_{Y|X}) + D_\alpha(P_{Y_{n[\alpha]}}\|Q_Y), && (89)
\end{aligned}$$

where (88) is due to (42). Taking the limit as $n\to\infty$, the lower-semicontinuity of the Rényi divergence ensures that

$$\sup_{x\in\mathcal{A}} D_\alpha(P_{Y|X=x}\|Q_Y) \geq C_\alpha + D_\alpha(P^*_{Y_{[\alpha]}}\|Q_Y), \tag{90}$$

and therefore the sought-after Q_Y^* is none other that $P^*_{Y_{[\alpha]}}$, the unique maximal α-mutual information output distribution. □

4. Finding C_α

In this section, we present a number of examples to illustrate how the results in Section 3 can be used to maximize the α-mutual information with respect to the input distribution. It is instructive to contrast the present approach with the maximization of α-mutual information invoking the KKT conditions, which is feasible in both the case $\alpha > 1$ in which the functional is concave with respect to the input distribution, and the case $\alpha\in(0,1)$ in which a monotonically increasing function of α-mutual information is concave. Simple finite-alphabet examples of such approach can be found in [44] when dealing with the E_0 functional in (54). Thanks to Theorem 1 it is possible to avoid taking derivatives of any functionals.

Example 2 (Binary symmetric channel)**.** *Let the input and output alphabet be* $\mathcal{A} = \mathcal{B} = \{0,1\}$ *and the random transformation be*

$$P_{Y|X} = \begin{bmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{bmatrix}, \quad \delta\in[0,1]. \tag{91}$$

Let's try the input distribution $P_X^*(0) = P_X^*(1) = 0.5$. *Then, according to* (25), *the* α*-response output distribution is also equiprobable* $P^*_{Y_{[\alpha]}}(0) = P^*_{Y_{[\alpha]}}(1) = 0.5$. *Since*

$$D_\alpha(P_{Y|X=0}\|P^*_{Y_{[\alpha]}}) = D_\alpha(P_{Y|X=1}\|P^*_{Y_{[\alpha]}}) = d_\alpha\left(\delta \,\|\, \tfrac{1}{2}\right) \tag{92}$$

the conditions of Corollary 2 *are met,* P_X^* *attains the maximal* α*-mutual information and therefore,*

$$C_\alpha = d_\alpha\left(\delta \,\|\, \tfrac{1}{2}\right) = \log 2 + \frac{1}{\alpha - 1}\log\left(\delta^\alpha + (1-\delta)^\alpha\right), \tag{93}$$

which satisfies, according to (31)

$$\lim_{\alpha\to 1} C_\alpha = \log 2 - h(\alpha), \tag{94}$$

$$\lim_{\alpha\to\infty} C_\alpha = \log\left(2\max\{\delta, 1-\delta\}\right). \tag{95}$$

Example 3 (Binary erasure channel.)**.** *Let the input/output alphabets be* $\mathcal{A} = \{0,1\}$ *and* $\mathcal{B} = \{0, \mathrm{e}, 1\}$, *and the random transformation be*

$$P_{Y|X} = \begin{bmatrix} 1-\delta & 0 \\ \delta & \delta \\ 0 & 1-\delta \end{bmatrix}, \quad \delta \in [0,1]. \tag{96}$$

(Departing from usual practice, columns/rows represent input/output letters respectively, i.e. probability vectors are column vectors, although for typographical convenience we show them as row vectors in the text.)

The α*-response output distribution to* $P_X^*(0) = P_X^*(1) = 0.5$ *is*

$$P^*_{Y_{[\alpha]}}(0) = P^*_{Y_{[\alpha]}}(1) = \frac{1-\delta}{\delta 2^{\frac{1}{\alpha}} + 2(1-\delta)}, \tag{97}$$

$$P^*_{Y_{[\alpha]}}(\mathrm{e}) = \frac{\delta 2^{\frac{1}{\alpha}}}{\delta 2^{\frac{1}{\alpha}} + 2(1-\delta)}. \tag{98}$$

By symmetry,

$$C_\alpha = D_\alpha(P_{Y|X=0}\|P^*_{Y_{[\alpha]}}) = D_\alpha(P_{Y|X=1}\|P^*_{Y_{[\alpha]}}) \tag{99}$$

$$= \frac{1}{\alpha-1}\log\left((1-\delta)^\alpha\left(\frac{(1-\delta)}{\delta 2^{\frac{1}{\alpha}} + 2(1-\delta)}\right)^{1-\alpha} + \delta^\alpha\left(\frac{\delta 2^{\frac{1}{\alpha}}}{\delta 2^{\frac{1}{\alpha}} + 2(1-\delta)}\right)^{1-\alpha}\right) \tag{100}$$

$$= \frac{1}{\alpha-1}\log\left(\frac{1-\delta+2^{\frac{1-\alpha}{\alpha}}\delta}{\left(\delta 2^{\frac{1}{\alpha}} + 2(1-\delta)\right)^{1-\alpha}}\right) \tag{101}$$

$$= \frac{1}{\alpha-1}\log\left(\frac{2^{\frac{\alpha-1}{\alpha}}(1-\delta)+\delta}{\left(\delta + 2^{\frac{\alpha-1}{\alpha}}(1-\delta)\right)^{1-\alpha}}\right) \tag{102}$$

$$= \frac{\alpha}{\alpha-1}\log\left((1-\delta)2^{\left(1-\frac{1}{\alpha}\right)} + \delta\right), \tag{103}$$

which satisfies (in bits)

$$\lim_{\alpha\to 1} C_\alpha = 1-\delta, \tag{104}$$

$$\lim_{\alpha\to\infty} C_\alpha = \log_2(2-\delta). \tag{105}$$

Example 4 (Binary asymmetric channel)**.** *Let the input and output alphabet be* $\mathcal{A}=\mathcal{B}=\{0,1\}$ *and the random transformation be*

$$P_{Y|X} = \begin{bmatrix} 1-\delta_0 & \delta_1 \\ \delta_0 & 1-\delta_1 \end{bmatrix}, \quad (\delta_0,\delta_1)\in[0,1]^2. \tag{106}$$

If $\delta_0+\delta_1=1$*, then* $I_\alpha(X;Y)=0$ *for any input distribution. We will assume* $\delta_0+\delta_1<1$*. Otherwise, we can just relabel the output alphabet* $(0,1)\leftarrow(1,0)$*, or equivalently* $(\delta_0,\delta_1)\leftarrow(1-\delta_0,1-\delta_1)$*. The condition*

$$D_\alpha(P_{Y|X=0}\|P^*_{Y_{[\alpha]}}) = D_\alpha(P_{Y|X=1}\|P^*_{Y_{[\alpha]}}) \tag{107}$$

is now $d_\alpha\left(1-\delta_0\|P^*_{Y_{[\alpha]}}(0)\right) = d_\alpha\left(\delta_1\|P^*_{Y_{[\alpha]}}(0)\right)$*, which, in view of* (32) *yields*

$$P^*_{Y_{[\alpha]}}(0) = \left(1+\left(\frac{(1-\delta_0)^\alpha-\delta_1^\alpha}{(1-\delta_1)^\alpha-\delta_0^\alpha}\right)^{\frac{1}{1-\alpha}}\right)^{-1}. \tag{108}$$

We can verify from (25) *that this corresponds to the* α*-response to* $P^*_X(1)=p^*=1-P^*_X(0)$*, where* $p^*\in[0,1]$ *is the solution to*

$$\frac{\delta_0^\alpha(1-p)+(1-\delta_1)^\alpha p}{(1-\delta_0)^\alpha(1-p)+\delta_1^\alpha p} = \left(\frac{(1-\delta_0)^\alpha-\delta_1^\alpha}{(1-\delta_1)^\alpha-\delta_0^\alpha}\right)^{\frac{\alpha}{1-\alpha}}. \tag{109}$$

Then,

$$C_\alpha = D_\alpha(P_{Y|X=0}\|P^*_{Y_{[\alpha]}}) \tag{110}$$

$$= \frac{1}{\alpha-1}\log\left((1-\delta_0)^\alpha(1-\delta_1)^\alpha-\delta_0^\alpha\delta_1^\alpha\right)+\log\left(\left((1-\delta_0)^\alpha-\delta_1^\alpha\right)^{\frac{1}{1-\alpha}}+\left((1-\delta_1)^\alpha-\delta_0^\alpha\right)^{\frac{1}{1-\alpha}}\right). \tag{111}$$

which satisfies

$$\lim_{\alpha\to 1} C_\alpha = \log\left(\exp\left(\frac{h(\delta_0)}{1-\delta_1-\delta_0}\right)+\exp\left(\frac{h(\delta_1)}{1-\delta_1-\delta_0}\right)\right)-\frac{(1-\delta_1)h(\delta_0)+(1-\delta_0)h(\delta_1)}{1-\delta_1-\delta_0}, \tag{112}$$

and

$$\lim_{\alpha\to\infty} C_\alpha = \log(2-\delta_0-\delta_1). \tag{113}$$

Example 5 (Z channel)**.** *Let the input and output alphabet be* $\mathcal{A}=\mathcal{B}=\{0,1\}$ *and the random transformation be*

$$P_{Y|X} = \begin{bmatrix} 1 & \delta \\ 0 & 1-\delta \end{bmatrix}, \quad \delta\in[0,1]. \tag{114}$$

Since this is the special case $(\delta_0, \delta_1) = (0, \delta)$ *of the binary asymmetric channel we obtain*

$$C_\alpha = \log\left(1 + \left(\frac{1-\delta^\alpha}{(1-\delta)^\alpha}\right)^{\frac{1}{1-\alpha}}\right). \tag{115}$$

The limit

$$\lim_{\alpha\to 1} C_\alpha = \log\left(1 - \delta^{\frac{1}{1-\delta}} + \delta^{\frac{\delta}{1-\delta}}\right) \tag{116}$$

coincides with the capacity of the Z-channel originally derived in [45].

The next example illustrates a case in which there are multiple optimal input distributions.

Example 6. *Let* $\mathcal{A} = \{0,1,2\}$, $\mathcal{B} = \{0,1,2,3\}$ *and the random transformation be*

$$P_{Y|X} = \begin{bmatrix} \frac{1}{2}-\delta & \delta & \frac{1}{2}-\delta \\ \delta & \frac{1}{2}-\delta & \delta \\ \delta & \frac{1}{2}-\delta & \delta \\ \frac{1}{2}-\delta & \delta & \frac{1}{2}-\delta \end{bmatrix}, \quad \delta \in \left[0, \tfrac{1}{2}\right]. \tag{117}$$

Let $P_X^0 = [\frac{1}{2}\ \ \frac{1}{2}\ \ 0]$ *and* $P_X^1 = [0\ \ \frac{1}{2}\ \ \frac{1}{2}]$. *Its easy to verify that the corresponding* α*-responses are the equiprobable distribution on* $\mathcal{B}$. *To verify that* P_X^0 *and* P_X^1 *attain the maximal* α*-mutual information, denote* $P^*_{Y_{[\alpha]}} = \left[\frac{1}{4}\ \ \frac{1}{4}\ \ \frac{1}{4}\ \ \frac{1}{4}\right]$. *For all* $x \in \mathcal{A}$, *we have*

$$D_\alpha\left(P_{Y|X=x} \| P^*_{Y_{[\alpha]}}\right) = D_\alpha\left(\left[\delta\ \ \frac{1}{2}-\delta\ \ \frac{1}{2}-\delta\ \ \delta\right] \Big\| \left[\frac{1}{4}\ \ \frac{1}{4}\ \ \frac{1}{4}\ \ \frac{1}{4}\right]\right) \tag{118}$$

$$= \frac{2\alpha-1}{\alpha-1}\log 2 + \log\left(\delta^\alpha + \left(\frac{1}{2}-\delta\right)^\alpha\right) \tag{119}$$

$$= C_\alpha, \tag{120}$$

where (120) *follows from Corollary 2.*

In the next example C_α is constant in α.

Example 7 (Additive phase noise)**.** *Let the input and output alphabet be* $\mathcal{A} = \mathcal{B} = [0, 2\pi)$ *and the random transformation be* $Y = (X + N) \mod 2\pi$, *where* N *is independent of* X *and is uniform on the interval* $[-\theta_0, \theta_0]$ *with* $\theta_0 \in (0, \pi]$. *Suppose* P_X^* *is uniform on* $[0, 2\pi)$, *it is easy to verify that* $P^*_{Y_{[\alpha]}}$ *is also uniform on* $[0, 2\pi)$. *Invoking* (26), *we obtain*

$$D_\alpha\left(P_{Y|X=x} \| P^*_{Y_{[\alpha]}}\right) = \log\frac{\pi}{\theta_0}, \quad x \in \mathcal{A}, \tag{121}$$

which according to Theorem 1 must be equal to C_α *attained by* P_X^*.

Example 8 (Additive Gaussian noise). *Let $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $Y = X + N$, where $N \sim \mathcal{N}(0, \sigma^2)$ is independent of X. Fix $\alpha \in (0,1)$ and $P > 0$. Suppose that the set, $\mathcal{P}$, of allowable input probability measures on $\mathcal{A}$ consists of those that satisfy*

$$\mathbb{E}\left[\exp_{\mathrm{e}}\left(-\frac{\alpha(1-\alpha)X^2}{2\left(\alpha^2 P+\sigma^2\right)}\right)\right] \geq \sqrt{\frac{\alpha^2 P+\sigma^2}{\alpha P+\sigma^2}}. \tag{122}$$

By completing the square, it is easy to verify that $P_X^* \sim \mathcal{N}(0, P)$ satisfies (122) with equality. Furthermore, its α-response is $P_{Y_{[\alpha]}^*} \sim \mathcal{N}(0, \alpha P + \sigma^2)$. To show that P_X^* does indeed attain $C_\alpha(\mathcal{P})$, first we compute

$$D_\alpha\left(P_{Y|X=x} \| P_{Y_{[\alpha]}^*}^*\right) = \frac{1}{2}\log\left(1+\frac{\alpha P}{\sigma^2}\right) - \frac{1}{2(1-\alpha)}\log\left(\frac{\alpha P+\sigma^2}{\alpha^2 P+\sigma^2}\right) + \frac{1}{2}\frac{\alpha x^2}{\alpha^2 P+\sigma^2}\log \mathrm{e}. \tag{123}$$

With (122) and (123), it is straightforward to see that if $P_X \in \mathcal{P}$ then

$$D_\alpha\left(P_{Y|X} \| P_{Y_{[\alpha]}^*}^* | P_X\right) \leq D_\alpha\left(P_{Y|X} \| P_{Y_{[\alpha]}^*}^* | P_X^*\right). \tag{124}$$

Consequently, Theorem 1 establishes that P_X^* achieves the maximal α-mutual information, which using (39) is given by

$$C_\alpha(\mathcal{P}) = \frac{1}{2}\log\left(1+\frac{\alpha P}{\sigma^2}\right). \tag{125}$$

5. Proofs

5.1. Proof of Theorem 1

We deal with the converse statement first. If (59)–(60) are satisfied then $(P_X^*, P_{Y_{[\alpha]}^*}^*)$ is a saddle point, and therefore,

$$I_\alpha(P_X^*, P_{Y|X}) = \max_{P_X \in \mathcal{P}} \min_{Q_Y \in \mathcal{Q}} D_\alpha(P_{Y|X} \| Q_Y | P_X) \tag{126}$$

$$= \min_{Q_Y \in \mathcal{Q}} \max_{P_X \in \mathcal{P}} D_\alpha(P_{Y|X} \| Q_Y | P_X). \tag{127}$$

which, by definition of α-mutual information, implies that P_X^* attains $\max_{P_X \in \mathcal{P}} I_\alpha(P_X, P_{Y|X})$. To show that the optimal input and its α-response must form a saddle point, first we deal with the case $\alpha > 1$, in which we can use the concavity of the conditional Rényi divergence. Choose arbitrary $\nu \in (0,1)$ and $P_X \in \mathcal{P}$. Let $P_\nu = \nu P_X + (1-\nu)P_X^*$ and denote its α-response by

$$Q_{Y_{[\alpha]}}^{(\nu)} = \arg\min_{Q_Y \in \mathcal{Q}} D_\alpha(P_{Y|X} \| Q_Y | P_\nu). \tag{128}$$

Therefore, $Q^{(0)}_{Y_{[\alpha]}} = P^*_{Y_{[\alpha]}}$. We have

$$\begin{aligned} I_\alpha(P^*_X, P_{Y|X}) &\geq I_\alpha(P_\nu, P_{Y|X}) && (129)\\ &= D_\alpha(P_{Y|X}\|Q^{(\nu)}_{Y_{[\alpha]}}|P_\nu) && (130)\\ &\geq \nu D_\alpha(P_{Y|X}\|Q^{(\nu)}_{Y_{[\alpha]}}|P_X) + (1-\nu)D_\alpha(P_{Y|X}\|Q^{(\nu)}_{Y_{[\alpha]}}|P^*_X) && (131)\\ &\geq \nu D_\alpha(P_{Y|X}\|Q^{(\nu)}_{Y_{[\alpha]}}|P_X) + (1-\nu)\min_{Q_Y\in\mathcal{Q}} D_\alpha(P_{Y|X}\|Q_Y|P^*_X) && (132)\\ &= \nu D_\alpha(P_{Y|X}\|Q^{(\nu)}_{Y_{[\alpha]}}|P_X) + (1-\nu)I_\alpha(P^*_X, P_{Y|X}), && (133)\end{aligned}$$

where (129) is due to the assumption of the optimality of P^*_X, (131) holds because $D_\alpha(P_{Y|X}\|Q_Y|P_X)$ is concave in P_X for $\alpha > 1$ and (130) and (133) are due to the definition of α-mutual information.

It follows that

$$D_\alpha(P_{Y|X}\|Q^{(\nu)}_{Y_{[\alpha]}}|P_X) \leq I_\alpha(P^*_X, P_{Y|X}) = D_\alpha(P_{Y|X}\|P^*_{Y_{[\alpha]}}|P^*_X). \tag{134}$$

Since $|Q^{(\nu)}_{Y_{[\alpha]}} - Q^{(0)}_{Y_{[\alpha]}}| \to 0$ as $\nu \to 0$, the lower semicontinuity of Rényi divergence and (134) imply (59).

We now show the desired result for $\alpha \in (0,1)$. In this case, the method of proof is easy to adapt to the $\alpha > 1$ case but it is more cumbersome than the foregoing proof, which is able to capitalize on the concavity of the conditional Rényi divergence. The starting point is the expression in (52) which we write as

$$\begin{aligned} I_\alpha(P_X, P_{Y|X}) &= \frac{\alpha}{\alpha-1}\log f(p_X), && (135)\\ f(r) &= \int_{\mathcal{B}}\left(\int_{\mathcal{A}} p^\alpha_{Y|X}(y|x)r(x)\,\mathrm{d}\mu_X(x)\right)^{\frac{1}{\alpha}}\mathrm{d}\mu_Y(y), && (136)\end{aligned}$$

where we have defined the functional f on the convex cone $\bar{\mathcal{P}}$ of nonnegative functions r on the input space such that $r(x) = \beta\frac{\mathrm{d}P}{\mathrm{d}\mu_X}(x)$ for some $\beta \geq 0$ and $P \in \mathcal{P}$. Recall from (48) that when the argument is a density, then we have

$$f(p_X) = \mathbb{E}[\mathbb{E}^{\frac{1}{\alpha}}[\exp(\alpha\, \imath_{X;\bar{Y}}(X;\bar{Y}))|\bar{Y}]], \quad (X,\bar{Y}) \sim P_X \times P_{\bar{Y}}. \tag{137}$$

By virtue of the convexity of $(\cdot)^{\frac{1}{\alpha}}$, f is a convex functional. Its directional (Gateaux) derivative is given by (note that the assumed finiteness of $I_\alpha(X;Y)$ allows swapping of differentiation and integration by means of the dominated convergence theorem)

$$\begin{aligned} f'(r;q) &= \frac{\mathrm{d}}{\mathrm{d}\delta}f(r+\delta q)|_{\delta=0} && (138)\\ &= \frac{1}{\alpha}\int_{\mathcal{B}}\left(\int_{\mathcal{A}} p^\alpha_{Y|X}(y|x)r(x)\,\mathrm{d}\mu_X(x)\right)^{\frac{1-\alpha}{\alpha}}\left(\int_{\mathcal{A}} p^\alpha_{Y|X}(y|x)q(x)\,\mathrm{d}\mu_X(x)\right)\mathrm{d}\mu_Y(y). && (139)\end{aligned}$$

Define the Lagrangian

$$L(r,\lambda) = f(r) - \lambda\left(\int_{\mathcal{A}} r(x)\,\mathrm{d}\mu_X(x) - 1\right). \tag{140}$$

Since f is convex and (135) is maximized by P_X^* among all probability measures on the convex set $\mathcal{P}$, there exists some $\lambda_0 \geq 0$ such that

$$\max_{\lambda \geq 0} \min_{r} L(r, \lambda) = L(p_X^*, \lambda_0), \tag{141}$$

where the minimization is over the convex cone $\bar{\mathcal{P}}$. It follows from standard convex optimization (e.g., see p. 227 in Reference [46]) that the Gateaux derivative of $L(\cdot, \lambda_0)$ at p_X^* in the direction of any $q \in \bar{\mathcal{P}}$ satisfies

$$L'(p_X^*; q, \lambda_0) \geq 0, \tag{142}$$

with equality if $q = p_X^*$. Invoking (139), we obtain

$$\begin{aligned} L'(p_X^*; q, \lambda_0) =& \frac{1}{\alpha} \int_{\mathcal{B}} \left(\int_{\mathcal{A}} p_{Y|X}^{\alpha}(y|x) p_X^*(x) \, \mathrm{d}\mu_X(x) \right)^{\frac{1-\alpha}{\alpha}} \left(\int_{\mathcal{A}} p_{Y|X}^{\alpha}(y|x) q(x) \, \mathrm{d}\mu_X(x) \right) \mathrm{d}\mu_Y(y) \\ & - \lambda_0 \left(\int_{\mathcal{A}} p_X^*(x) \, \mathrm{d}\mu_X(x) - 1 \right). \end{aligned} \tag{143}$$

Specializing (142) and its condition for equality to $q \leftarrow p_X$ we obtain

$$L'(p_X^*; p_X, \lambda_0) \geq L'(p_X^*; p_X^*, \lambda_0) \tag{144}$$

which, upon substitution of (143), becomes

$$f(p_X^*) \leq \int_{\mathcal{A}} \int_{\mathcal{B}} p_{Y|X}^{\alpha}(y|x) \left(\int_{\mathcal{A}} p_{Y|X}^{\alpha}(y|x) p_X^*(x) \, \mathrm{d}\mu_X(x) \right)^{\frac{1-\alpha}{\alpha}} p_X(x) \, \mathrm{d}\mu_Y(y) \, \mathrm{d}\mu_X(x). \tag{145}$$

Taking $\frac{1}{\alpha - 1} \log(\cdot)$ of both sides of (145), invoking (25) and (47), the inequality is reversed and we obtain

$$D_\alpha(P_{Y|X} \| P_{Y_{[\alpha]}}^* | P_X) + \frac{1-\alpha}{\alpha} I_\alpha(X^*; Y^*) \leq \frac{1}{\alpha} I_\alpha(X^*; Y^*), \tag{146}$$

which upon rearranging is the sought-after inequality (59).

5.2. *Proof of Theorem 2*

In order to show

$$\inf_{Q_Y \in \mathcal{Q}} \sup_{P_X \in \mathcal{P}} D_\alpha(P_{Y|X} \| Q_Y | P_X) \leq \sup_{P_X \in \mathcal{P}} \min_{Q_Y \in \mathcal{Q}} D_\alpha(P_{Y|X} \| Q_Y | P_X) = C_\alpha(\mathcal{P}), \tag{147}$$

we construct $Q_Y^* \in \mathcal{Q}$ such that

$$D_\alpha(P_{Y|X} \| Q_Y^* | P_X) \leq C_\alpha(\mathcal{P}), \quad \forall P_X \in \mathcal{P}. \tag{148}$$

Moreover, Q_Y^* is indeed the minimizer in the leftmost side of (147) and we may replace the inf with min therein.

The construction of Q_Y^* follows a Cauchy-sequence approach in the proof of Kemperman's result in [47]. Let $\{P_{X_n}\}_{n \geq 1}$ be a sequence of probability distributions in $\mathcal{P}$ such that

$$\lim_{n \to \infty} I_\alpha(P_{X_n}, P_{Y|X}) = C_\alpha(\mathcal{P}). \tag{149}$$

Fix an arbitrary $P_X \in \mathcal{P}$ and let $\mathcal{P}_n \subset \mathcal{P}$ denote the convex hull of $\{P_X, P_{X_1}, \ldots, P_{X_n}\}$, which is a compact set. Although $I_\alpha(\cdot, P_{Y|X})$ may not be concave for $\alpha \in (0,1)$, recall from (57) that the monotonically increasing function Γ_α is such that $\Gamma_\alpha(I_\alpha(\cdot, P_{Y|X}))$ is concave. So there exists some $P^*_{X_n} \in \mathcal{P}_n$ that attains $C_\alpha(\mathcal{P}_n)$. Thus for any $n \geq 1$,

$$I_\alpha(P_{X_n}, P_{Y|X}) \leq I_\alpha(P^*_{X_n}, P_{Y|X}) \leq C_\alpha(\mathcal{P}) \tag{150}$$

by the definition of $C_\alpha(\mathcal{P})$. The asymptotic optimality of the sequence X_n implies that $I_\alpha(P^*_{X_n}, P_{Y|X}) = C_\alpha(\mathcal{P}_n)$ also converges to $C_\alpha(\mathcal{P})$.

Denote by $P^*_{Y_{n[\alpha]}}$ the α-response to $P^*_{X_n}$. Then for any $m \geq n \geq 1$, we have

$$\begin{aligned} I_\alpha(P^*_{X_n}, P_{Y|X}) &= D_\alpha(P_{Y|X} \| P^*_{Y_{n[\alpha]}} | P^*_{X_n}) && (151)\\ &= D_\alpha(P_{Y|X} \| P^*_{Y_{m[\alpha]}} | P^*_{X_n}) - D_\alpha(P^*_{Y_{n[\alpha]}} \| P^*_{Y_{m[\alpha]}}) && (152)\\ &\leq D_\alpha(P_{Y|X} \| P^*_{Y_{m[\alpha]}} | P^*_{X_m}) - D_\alpha(P^*_{Y_{n[\alpha]}} \| P^*_{Y_{m[\alpha]}}) && (153)\\ &= I_\alpha(P^*_{X_m}, P_{Y|X}) - D_\alpha(P^*_{Y_{n[\alpha]}} \| P^*_{Y_{m[\alpha]}}), && (154) \end{aligned}$$

where (151) and (154) are due to the definition of α-mutual information, (152) follows from (42), and (153) holds because of Theorem 1 applied to $\mathcal{P}_m$ since $P^*_{X_n} \in \mathcal{P}_m$ as $\mathcal{P}_n \subset \mathcal{P}_m$ for $m \geq n$. Rearranging the end-to-end inequality in (151)–(154) results in

$$D_\alpha(P^*_{Y_{n[\alpha]}} \| P^*_{Y_{m[\alpha]}}) \leq I_\alpha(P^*_{X_m}, P_{Y|X}) - I_\alpha(P^*_{X_n}, P_{Y|X}). \tag{155}$$

But since $I_\alpha(P^*_{X_n}, P_{Y|X})$ converges to $C_\alpha(\mathcal{P})$, it is a Cauchy sequence, i.e.

$$\left| I_\alpha(P^*_{X_n}, P_{Y|X}) - I_\alpha(P^*_{X_m}, P_{Y|X}) \right| \to 0, \quad n, m \to \infty. \tag{156}$$

Hence, (155) ensures that $P^*_{Y_{n[\alpha]}}$ is also a Cauchy sequence in the sense that $D_\alpha(P^*_{Y_{n[\alpha]}} \| P^*_{Y_{m[\alpha]}}) \to 0$ as $n, m \to \infty$. By the generalized Pinkser's inequality (35), $P^*_{Y_{n[\alpha]}}$ is also a Cauchy sequence in total variation distance, i.e., $|P^*_{Y_{n[\alpha]}} - P^*_{Y_{m[\alpha]}}| \to 0$ as $n, m \to \infty$. Since the space of probability measures is complete in the total variation distance, $\{P^*_{Y_{n[\alpha]}}\}_n$ must possess a limit point, which we denote by $P^*_{Y_{[\alpha]}}$.

Now, by Theorem 1 applied to $\mathcal{P}_n$, we have

$$D_\alpha(P_{Y|X} \| P^*_{Y_{n[\alpha]}} | P_X) \leq D_\alpha(P_{Y|X} \| P^*_{Y_{n[\alpha]}} | P^*_{X_n}) \leq C_\alpha(\mathcal{P}), \tag{157}$$

and since $D_\alpha(P\|Q)$ is lower-semicontinuous in (P, Q) for $\alpha > 0$, taking limits as $n \to \infty$ of (157), we obtain

$$D_\alpha(P_{Y|X} \| P^*_{Y_{[\alpha]}} | P_X) \leq C_\alpha(\mathcal{P}). \tag{158}$$

Next we show that (158) holds for all $P_X \in \mathcal{P}$, in other words, the limit point $P^*_{Y_{[\alpha]}}$ does not depend on the initial choice of $P_X \in \mathcal{P}$. Choose an arbitrary distribution $Q_X \in \mathcal{P}, Q_X \neq P_X$, and introduce the following

notation: $\mathcal{P}'_n$ is the convex hull of $\{Q_X, P_X, P_{X_1}, \ldots, P_{X_n}\}$; $Q^*_{X_n}$ is a maximizer of $I_\alpha(P_X, P_{Y|X})$ in $\mathcal{P}'_n$; its α-response is $Q^*_{Y_{n[\alpha]}}$; and $Q^*_{Y_{[\alpha]}}$ is the limit of the sequence $\{Q^*_{Y_{n[\alpha]}}\}_n$. Then we have

$$D_\alpha(P^*_{Y_{n[\alpha]}}\|Q^*_{Y_{n[\alpha]}}) = D_\alpha(P_{Y|X}\|Q^*_{Y_{n[\alpha]}}|P^*_{X_n}) - D_\alpha(P_{Y|X}\|P^*_{Y_{n[\alpha]}}|P^*_{X_n}) \tag{159}$$

$$\leq D_\alpha(P_{Y|X}\|Q^*_{Y_{n[\alpha]}}|Q^*_{X_n}) - D_\alpha(P_{Y|X}\|P^*_{Y_{n[\alpha]}}|P^*_{X_n}) \tag{160}$$

$$= I_\alpha(Q^*_{X_n}, P_{Y|X}) - I_\alpha(P^*_{X_n}, P_{Y|X}), \tag{161}$$

where (159) holds because of (42); (160) is due to Theorem 1 applied to $\mathcal{P}'_n$ and the fact that $P^*_{X_n} \in \mathcal{P}_n \subset \mathcal{P}'_n$ for any $n \geq 1$; and (161) is because of the definition of the α-mutual information.

The same argument that led to the conclusion that $I_\alpha(P^*_{X_n}, P_{Y|X}) \to C_\alpha(\mathcal{P})$ establishes that $I_\alpha(Q^*_{X_n}, P_{Y|X}) \to C_\alpha(\mathcal{P})$. Therefore, taking limits as $n \to \infty$ in (159)–(161) and applying the lower-semicontinuity of $D_\alpha(P\|Q)$ again, we obtain

$$D_\alpha(P^*_{Y_{[\alpha]}}\|Q^*_{Y_{[\alpha]}}) = 0, \tag{162}$$

and therefore $P^*_{Y_{[\alpha]}} = Q^*_{Y_{[\alpha]}}$. Since the limiting output distribution is the same whether we use $\mathcal{P}_n$ or $\mathcal{P}'_n$ and according to the latter the roles of P_X and Q_X are identical, we conclude that had we defined $\mathcal{P}_n$ with Q_X instead of P_X, we would have reached the same limiting output distribution and (158) holds for all $P_X \in \mathcal{P}$. So we have constructed Q^*_Y satisfying (148).

Finally, we show that $P^*_{Y_{[\alpha]}}$ is the only element that achieves

$$\inf_{Q_Y \in \mathcal{Q}} \sup_{P_X \in \mathcal{P}} D_\alpha(P_{Y|X}\|Q_Y|P_X).$$

Arguing by contradiction, suppose that there exists another $\widehat{P_Y}$ such that

$$\sup_{P_X \in \mathcal{P}} D_\alpha(P_{Y|X}\|\widehat{P_Y}|P_X) = C_\alpha(\mathcal{P}). \tag{163}$$

As earlier in the proof, let $\{P_{X_n} \in \mathcal{P}\}_n$ be a sequence satisfying (149), and denote the corresponding α-responses by $P_{Y_{n[\alpha]}}$. Then, invoking (42) again we have

$$D_\alpha(P_{Y|X}\|P_{Y_{n[\alpha]}}|P_{X_n}) + D_\alpha(P_{Y_{n[\alpha]}}\|\widehat{P_Y}) = D_\alpha(P_{Y|X}\|\widehat{P_Y}|P_{X_n}) \tag{164}$$

$$\leq C_\alpha(\mathcal{P}) < \infty, \tag{165}$$

where the inequality follows from (163). Using (149) we obtain

$$D_\alpha(P_{Y_{n[\alpha]}}\|\widehat{P_Y}) \leq C_\alpha(\mathcal{P}) - D_\alpha(P_{Y|X}\|P_{Y_{n[\alpha]}}|P_{X_n}) \tag{166}$$

$$\to 0, \tag{167}$$

and by (35), it follows that

$$|P_{Y_{n[\alpha]}} - \widehat{P_Y}| \to 0. \tag{168}$$

Furthermore, we established above that

$$|P_{Y_{n[\alpha]}} - P^*_{Y_{[\alpha]}}| \to 0. \tag{169}$$

So, by the triangle inequality, we conclude that $\widehat{P_Y} = P^*_{Y_{[\alpha]}}$.

6. Conclusions

The supremization of α-mutual information with respect to the input distribution plays an important role in various information theoretic settings, most notably in the error exponent of optimal codes operating below capacity. We show that the optimal (if it exists) input distribution, together with its α-response, form a saddle-point of the conditional Rényi divergence, and vice versa, the existence of the saddle point ensure the existence of a maximal α-mutual information input distribution. The application of this result to various discrete and non-discrete settings illustrates the power and generality of this tool, which mirrors a similar result enjoyed by conditional relative entropy; However, the proof of the latter result is much easier due to the more convenient structure of the objective function. Regardless of whether there exists an input distribution maximizing α-mutual information, there always exists a unique optimal output distribution, which is the limit of the α-responses of any asymptotically optimal sequence of input distributions. Furthermore, a saddle-value exists and

$$\sup_{P_X} \min_{Q} D_\alpha(P_{Y|X} \,\|\, Q|P_X) = \min_{Q} \sup_{P_X} D_\alpha(P_{Y|X} \,\|\, Q|P_X) \tag{170}$$

even if we restrict the feasible set of input distributions to be an arbitrary convex subset. These results lend further evidence to the notion that, out of all the available Rényi-generalizations of mutual information, the α-mutual information defined as in (43) is the most convenient and insightful, although $I^{\mathsf{c}}_\alpha(X;Y)$ is also of considerable interest particularly in the error exponent analysis of channels with cost constraints.

Author Contributions: Both authors contributed to the conceptualization, investigation, results, original draft preparation, editing, revision and response to reviewers.

Funding: This work was partially supported by the US National Science Foundation under Grant CCF-1016625, and in part by the Center for Science of Information, an NSF Science and Technology Center under Grant CCF-0939370.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rényi, A. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*; Neyman, J., Ed.; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
2. Van Erven, T.; Harremoës, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820.
3. Blahut, R.E. Hypothesis testing and information theory. *IEEE Trans. Inf. Theory* **1974**, *20*, 405–417.
4. Csiszár, I. Generalized cutoff rates and Rényi's information measures. *IEEE Trans. Inf. Theory* **1995**, *41*, 26–34.
5. Shayevitz, O. On Rényi measures and hypothesis testing. In Proceedings of the 2011 IEEE International Symposium on Information Theory, Saint Petersburg, Russia, 31 July–5 August 2011; pp. 894–898.
6. Harremoës, P. Interpretations of Rényi entropies and divergences. *Phys. A Stat. Mech. Its Appl.* **2006**, *365*, 57–62.
7. Sason, I.; Verdú, S. Improved bounds on lossless source coding and guessing moments via Rényi measures. *IEEE Trans. Inf. Theory* **2018**, *64*, 4323–4346.
8. Haroutunian, E.A. Estimates of the exponent of the error probability for a semicontinuous memoryless channel. *Probl. Inf. Transm.* **1968**, *4*, 29–39.
9. Haroutunian, E.A.; Haroutunian, M.E.; Harutyunyan, A.N. Reliability criteria in information theory and in statistical hypothesis testing. *Found. Trends Commun. Inf. Theory* **2007**, *4*, 97–263.

10. Polyanskiy, Y.; Verdú, S. Arimoto channel coding converse and Rényi divergence. In Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing, Allerton, IL, USA, 29 September–1 October 2010; pp. 1327–1333.
11. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
12. Liese, F.; Vajda, I. *Convex Statistical Distances*; Number 95 in Teubner Texte zur Mathematik; Teubner: Leipzig, Germany, 1987.
13. Tridenski, S.; Zamir, R.; Ingber, A. The Ziv-Zakai-Renyi bound for joint source-channel coding. *IEEE Trans. Inf. Theory* **2015**, *61*, 429–4315.
14. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 14–20.
15. Blahut, R.E. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473.
16. Blachman, N.M. Communication as a game. *Proc. IRE Wescon* **1957**, *2*, 61–66.
17. Borden, J.M.; Mason, D.M.; McEliece, R.J. Some information theoretic saddlepoints. *SIAM J. Control Optim.* **1985**, *23*, 129–143.
18. Lapidoth, A.; Narayan, P. Reliable communication under channel uncertainty. *IEEE Trans. Inf. Theory* **1998**, *44*, 2148–2177.
19. Kemperman, J. On the Shannon capacity of an arbitrary channel. In *Koninklijke Nederlandse Akademie van Wetenschappen, Indagationes Mathematicae*; Elsevier: Amsterdam, The Netherlands, 1974; Volume 77, No. 2, pp. 101–115.
20. Gallager, R.G. *Source Coding With Side Information and Universal Coding*; Technical Report LIDS-P-937; Lab. Information Decision Systems, Massachusetts Institute of Technology: Cambridge, MA, USA, 1979.
21. Ryabko, B. Encoding a source with unknown but ordered probabilities. *Probl. Inf. Transm.* **1979**, *15*, 134–138.
22. Davisson, L.; Leon-Garcia, A. A source matching approach to finding minimax codes. *IEEE Trans. Inf. Theory* **1980**, *26*, 166–174.
23. Ryabko, B. Comments on "A Source Matching Approach to Finding Minimax Codes". *IEEE Trans. Inf. Theory* **1981**, *27*, 780–781.
24. Haussler, D. A general minimax result for relative entropy. *IEEE Trans. Inf. Theory* **1997**, *43*, 1276–1280.
25. Yagli, S.; Altuğ, Y.; Verdú, S. Minimax Rényi redundancy. *IEEE Trans. Inf. Theory* **2018**, *64*, 3715–3733.
26. Nakiboğlu, B. The Rényi capacity and center. *IEEE Trans. Inf. Theory* **2019**, *65*, 841–860.
27. Verdú, S. α-mutual information. In Proceedings of the 2015 Information Theory and Applications Workshop, San Diego, CA, USA, 1–6 February 2015; pp. 1–6.
28. Arimoto, S. Information Measures and Capacity of Order α for Discrete Memoryless Channels. In *Topics in Information Theory, Proceedings of the Coll. Math. Soc. János Bolyai*; Bolyai: Keszthely, Hungary, 1975; pp. 41–52.
29. Augustin, U. Noisy Channels. Ph.D. Thesis, Universität Erlangen-Nürnberg, Erlangen, Germany, 1978.
30. Nakiboglu, B. The Augustin capacity and center. *arXiv* **2018**, arXiv:1606.00304.
31. Sibson, R. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **1969**, *14*, 149–160.
32. Lapidoth, A.; Pfister, C. Two measures of dependence. *Entropy* **2019**, *21*, 778.
33. Tomamichel, M.; Hayashi, M. Operational interpretation of Rényi information measures via composite hypothesis testing against product and Markov distributions. *IEEE Trans. Inf. Theory* **2018**, *64*, 1064–1082.
34. Aishwarya, G.; Madiman, M. Remarks on Rényi versions of conditional entropy and mutual information. In Proceedings of the 2019 IEEE International Symposium on Information Theory, Paris, France, 7–12 July 2019; pp. 1117–1121.
35. Shannon, C.E.; Gallager, R.G.; Berlekamp, E. Lower bounds to error probability for coding on discrete memoryless channels, I. *Inf. Control* **1967**, *10*, 65–103.
36. Gallager, R.G. A simple derivation of the coding theorem and some applications. *IEEE Trans. Inf. Theory* **1965**, *11*, 3–18.
37. Yagli, S.; Cuff, P. Exact soft-covering exponent. *IEEE Trans. Inf. Theory* **2019**, *65*, 6234–6262.

38. Gilardoni, G.L. On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. *IEEE Trans. Inf. Theory* **2010**, *56, no. 11*, 5377–5386.
39. Massey, J.L. Coding and modulation in digital communications. In Proceedings of the 1974 International Zurich Seminar on Digital Communications, Zurich, Switzerland, 12–15 March 1974; pp. E2(1)–E2(4).
40. Shannon, C.E. The zero error capacity of a noisy channel. *IRE Trans. Inf. Theory* **1956**, *2*, 8–19.
41. Sundaresan, R. Guessing under source uncertainty. *IEEE Trans. Inf. Theory* **2007**, *53*, 269–287.
42. Bunte, C.; Lapidoth, A. Encoding tasks and Rényi entropy. *IEEE Trans. Inf. Theory* **2014**, *60*, 5065–5076.
43. Ho, S.W.; Verdú, S. Convexity/concavity of Rényi entropy and α-mutual information. In Proceedings of the 2015 IEEE International Symposium on Information Theory, Hong Kong, China, 14–19 June 2015; pp. 745–749.
44. Gallager, R.G. *Information Theory and Reliable Communication*; John Wiley and Sons: New York, USA, 1968; Volume 2.
45. Verdú, S. Channel Capacity. In *The Electrical Engineering Handbook*, 2nd ed.; IEEE Press: Piscataway, NJ, USA; CRC Press: Boca Raton, FL, USA, 1997; Chapter 73.5, pp. 1671–1678.
46. Luenberger, D. *Optimization by vector space methods*; John Wiley and Sons: Hoboken, NJ, USA, 1997.
47. Polyanskiy, Y.; Wu, Y. Lecture Notes on Information Theory. 2017. Available online: http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf (accessed on 30 April 2019).