Abstract—We provide a variational characterization for the various Rényi information measures via their Shannon counterparts, and demonstrate how properties of the former can be recovered from first principle via the associated properties of the latter. Motivated by this characterization, we give a new operational interpretation for the Rényi divergence in a two-sensor composite hypothesis testing framework.

I. INTRODUCTION

The Shannon Entropy and the Kullback-Leibler divergence play a pivotal role in the study of information theory, large deviations and statistics, arising as the answer to many of the fundamental questions in these fields. Besides their operational importance, these quantities also possess some very natural properties one would expect an information measure to satisfy, a fact that has spurred several different axiomatic characterizations, see [1] and references therein. Motivated by the axiomatic approach, Rényi suggested a more general definition of a Rényi mutual information, has been derived in the context of guessing moments [7], [17], and for one Rényi entropy (of order $\alpha$), this has been observed by the axiomatic approach, Rényi suggested a more general definition of a Rényi mutual information.

Remarkably, this “reversed” line of thought has proved fruitful; yet still intuitively appealing as measures of information [2]. Several operational interpretations, thereby “justifying” their definition. An incomplete list includes [3], [4], [5], [6], [7], [8], [9], [10] for the Rényi entropy, [11], [12], [6], [13], [14] for the Rényi divergence, and [15], [6], [16] for different definitions of a Rényi mutual information.

Interestingly, even though the Shannon measures are a special case of the Rényi measures, the latter can admit a variational characterization in terms of the former. For the Rényi entropy (of order $\alpha < 1$) this has been observed in the context of guessing moments [7], [17], and for one definition of a Rényi mutual information, has been derived in the context of generalized cutoff rates in channel coding [6, Appendix]. Here we further examine relations of that type, and their ramifications. In Section II we give a brief mathematical background. In Section III, we provide a variational characterization for the various Rényi measures via the Shannon measures. In Section IV, we demonstrate how properties of the Rényi measures can be recovered directly from the characterization in a very instructive fashion, via the associated properties of their Shannon counterparts. Finally, motivated by the characterization, we study a two-sensor composite hypothesis testing problem in which the Rényi divergence is shown to play a fundamental role, yielding a new operational interpretation to that quantity. This observation is discussed in Section V.

II. PRELIMINARIES

Let $\mathcal{X}$ be a finite alphabet, and denote by $\mathcal{P}(\mathcal{X})$ the set of all probability distributions over $\mathcal{X}$. The support of a distribution $P \in \mathcal{P}(\mathcal{X})$ is the set $S(P) \eqdef \{ x \in \mathcal{X} : P(x) > 0 \}$. The (Shannon) entropy of $P \in \mathcal{P}(\mathcal{X})$ is

$$H(P) \eqdef -\sum_{x \in \mathcal{X}} P(x) \log P(x).$$

The (Kullback-Leibler) divergence between two distributions $P_1, P_2 \in \mathcal{P}(\mathcal{X})$ is

$$D(P_1 \parallel P_2) \eqdef \sum_{x \in \mathcal{X}} P_1(x) \log \left( \frac{P_1(x)}{P_2(x)} \right).$$

We write $P_1 \ll P_2$ to indicate that $S(P_1) \subseteq S(P_2)$. Note that $D(P_1 \parallel P_2) < \infty$ if and only if $P_1 \ll P_2$.

Let $\mathcal{X}, \mathcal{Y}$ be two finite alphabets. A channel $W : \mathcal{X} \mapsto \mathcal{Y}$ is a set of probability distributions $\{ W(\cdot|x) \in \mathcal{P}(\mathcal{Y}) \}_{x \in \mathcal{X}}$ that maps a distribution $P \in \mathcal{P}(\mathcal{X})$ to the distributions $P \circ W \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and $PW \in \mathcal{P}(\mathcal{Y})$, according to

$$(P \circ W)(x, y) \eqdef P(x)W(y|x),$$

$$PW(y) \eqdef \sum_{x \in \mathcal{X}} P(x)W(y|x).$$

For any two channels $V : \mathcal{X} \mapsto \mathcal{Y}, W : \mathcal{X} \mapsto \mathcal{Y}$, we write

$$D(V||W|P) \eqdef \sum_{x \in \mathcal{X}} P(x)D(V(\cdot|x)||W(\cdot|x)).$$

The (Shannon) mutual information associated with $P$ and $W$ is

$$I(P, W) \eqdef H(PW) - \sum_{x \in \mathcal{X}} P(x)H(W(\cdot|x))$$

$$= \min_{Q} \sum_{x \in \mathcal{X}} P(x)D(W(\cdot|x)||Q) \quad (1)$$

$$= \min_{Q} D(P \circ W || P \times Q) \quad (2)$$

where the identities are well known. The (Shannon) capacity of a channel $W$ is

$$C(W) \eqdef \max_{P} I(P, W).$$

A distribution $P \in \mathcal{P}(\mathcal{X})$ induces a product distribution $P^n \in \mathcal{P}(\mathcal{X}^n)$, where $P^n(x^n) \eqdef \prod_{k=1}^{n} P(x_k)$. The type of a sequence $x^n \in \mathcal{X}^n$ is the distribution $\pi_{x^n} \in \mathcal{P}(\mathcal{X})$ corresponding to the relative frequency of symbols in $x^n$. The set of all possible types of sequences $x^n$ is denoted $\mathcal{P}^n(\mathcal{X})$. The type class of any type $Q \in \mathcal{P}^n(\mathcal{X})$ is the set

$$T_Q \eqdef \{ x^n \in \mathcal{X}^n : \pi_{x^n} = Q \}.$$

1We use the conventions $0 \log 0 = 0$, and $a \log \frac{a}{2} = 0$ or $+\infty$ according to whether $a = 0$ or $a > 0$ respectively.
The following facts are well known [18],

**Lemma 1:** For any type $Q \in \mathcal{P}^n(X)$ and any $x^n \in T_Q$:

(i) $P^n(x^n) = 2^{-n(D(Q||P)+H(Q))}$

(ii) $|\mathcal{P}^n(X)| \leq |T_Q| \leq 2^{nH(Q)}$

(iii) $|\mathcal{P}^n(X)| = \left(\frac{n}{|X|}\right) \leq (n+1)^{|X|}$

(iv) For any $\delta > 0$

$$P^n \left( \{ x^n \in X^n : D(x^n||P) \geq \delta \} \right) \leq |\mathcal{P}^n(X)|2^{-n\delta}$$

Let $\alpha > 0$, $\alpha \neq 1$ throughout. The Rényi entropy of order $\alpha$ of a distribution $P \in \mathcal{P}(X)$ is

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \sum_{x \in X} P(x)^\alpha.$$ We denote by $H_0(P)$, $H_1(P)$ and $H_\infty(P)$ the limits of $H_\alpha(P)$ as $\alpha$ tends to $0$, $1$ and $\infty$, respectively.

The Rényi divergence of order $\alpha$ between two distributions $P_1$, $P_2 \in \mathcal{P}(X)$ is

$$D_\alpha(P_1||P_2) = \frac{1}{\alpha - 1} \log \sum_{x \in X} P(x)^\alpha P_2(x)^{1-\alpha}.$$ We denote by $D_0(P_1||P_2)$, $D_1(P_1||P_2)$ and $D_\infty(P_1||P_2)$ the limits of $D_\alpha(P_1||P_2)$ as $\alpha$ tends to $0$, $1$ and $\infty$, respectively.

Note that for $\alpha < 1$, $D_0(P_1||P_2) < \infty$ if and only if $S(P_1) \cap S(P_2) \neq \emptyset$, and for $\alpha > 1$, $D_\infty(P_1||P_2) < \infty$ if and only if $P_1 \ll P_2$.

The Rényi equivalent of the Shannon mutual information has several different definitions, each generalizing a different expansion of the latter, see [6] and references therein. Here we discuss the following two alternatives:

$$I_\alpha(P,W) = \min_{Q} \sum_{x \in X} \sum_{y \in X} P(x)D_\alpha(W(y|x)||Q)$$

corresponding to (1), and

$$K_\alpha(P,W) = \min_{Q} D_\alpha(P \circ W||P \times Q)$$

corresponding to (2). Following [6], we define the capacity of order $\alpha$ of $W$ via (3), i.e.,

$$C_\alpha(W) = \max_{P} I_\alpha(P,W).$$

As it turns out, using $K_\alpha(P,W)$ in the definition above yields the same capacity function [6], a fact we reaffirm in the sequel.

### III. Characterization

In this section, we derive the basic characterization for the various Rényi measures in terms of the Shannon measures.

**Theorem 1:** For $\alpha > 1$,

$$H_\alpha(P) = \min_{Q} \left\{ \frac{\alpha}{\alpha - 1} D(Q||P) + H(Q) \right\}$$

$$D_\alpha(P_1||P_2) = \max_{Q \in \mathcal{P}(X)} \left\{ \frac{\alpha}{1-\alpha} D(Q||P_1) + D(Q||P_2) \right\}$$

$$I_\alpha(P,W) = \max_{V} \left\{ I(P,V) + \frac{\alpha}{1-\alpha} D(V||W||P) \right\}$$

$$K_\alpha(P,W) = \max_{Q} \left\{ I_\alpha(Q,W) + \frac{1}{1-\alpha} D(Q||P) \right\}$$

For $\alpha < 1$, replace min with max and vice versa.

**Remark 1:** The $\alpha < 1$ counterpart of (5) is mentioned in [7, 17]. Both (5) and (6) are simple generalizations, for which we provide an elementary proof. Relation (7) can be found in [6, Appendix], however here we provide a slightly different proof directly via (6). Relation (8) appears to be new.

**Proof:** Let $X_1 \overset{\text{def}}{=} S(P_1)$ and $X_2 \overset{\text{def}}{=} S(P_2)$ for short. We derive a characterization for the functional

$$J_{\alpha,\beta}(P_1,P_2) = -\log \sum_{x \in X_1} P_1(x)^\alpha P_2(x)^\beta$$

for any $\alpha > 0$ and $\beta$. This will yield (5) and (6) in particular, and will also prove useful in the sequel. It is readily verified that the functional is additive, i.e., $J_{\alpha,\beta}(P_1^n,P_2^n) = nJ_{\alpha,\beta}(P_1,P_2)$. Therefore,

$$J_{\alpha,\beta}(P_1,P_2) = -\frac{1}{n} \log \sum_{x^n \in X^n} P_1(x^n)^\alpha P_2(x^n)^\beta$$

$$\leq \min_{Q \in \mathcal{P}^n(X_1)} \{ \alpha D(Q||P_1) + \beta D(Q||P_2) + (\alpha + \beta - 1)H(Q) \}$$

$$+ \frac{|X_1|}{n} \log (n+1)$$

where properties (i) and (ii) of Lemma 1 were used in the first inequality, and property (iii) was used in the second inequality. Similarly,

$$J_{\alpha,\beta}(P_1,P_2) \geq -\frac{1}{n} \log \sum_{Q \in \mathcal{P}^n(X_1)} 2^{-n(\alpha D(Q||P_1) + (\alpha + \beta - 1)H(Q))}$$

$$\geq \min_{Q \in \mathcal{P}^n(X_1)} \{ \alpha D(Q||P_1) + \beta D(Q||P_2) + (\alpha + \beta - 1)H(Q) \}$$

$$- \frac{|X_1|}{n} \log (n+1).$$

$\bigcup_{\alpha} \mathcal{P}^n(X)$ is dense in $\mathcal{P}(X)$, and the objective function is continuous in $Q$ over the compact set $\mathcal{P}(X_1 \cap X_2)$, and equals $\pm \infty$ over $\mathcal{P}(X_1) \setminus \mathcal{P}(X_1 \cap X_2)$ according to sign($\beta$). Thus, taking the limit as $n \to \infty$, we obtain:

$$J_{\alpha,\beta}(P_1,P_2) = \min_{Q \in \mathcal{P}(X_1)} \{ \alpha D(Q||P_1) + \beta D(Q||P_2) + (\alpha + \beta - 1)H(Q) \}.$$ The statement for $H_\alpha(P)$ (resp. $D_\alpha(P_1||P_2)$) now follows by substituting $\beta = 0$ (resp. $\beta = 1 - \alpha$), normalizing by $\alpha - 1$ (resp. $1 - \alpha$), and noting the possible change in sign that replaces min with max. For $H_\alpha(P)$, taking the min or max over all $Q \in \mathcal{P}(X)$ does not change anything.

The proof of relations (7) and (8) is relegated to the Appendix.

### IV. Properties Revisited

Many well known properties of the Rényi measures can be derived directly via the characterization in Theorem 1, and the associated properties of the Shannon measures. These alternative derivations seem more instructive, and are sometimes simpler than a direct proof. We shall make no consistent attempt at mathematical rigor in this section, and discuss only a representative sample of properties. For a more exhaustive and rigorous account, see [19].
1. $H_0(P) = \log |S(P)|$: As $\alpha \to 0$ we have $H_n(P) \to \max_{Q} H(Q)$ where the maximization can be restricted to $Q \ll P$ without loss of generality, and is clearly achieved by a uniform distribution over $S(P)$.

2. $H_1(P) = H(P)$. As $\alpha \to 1^+$, the minimization in (5) is attained by $Q \to P$ as otherwise the divergence term blows up. The limit $\alpha \to 1^-$ follows similarly, hence the result.

3. $H_{\infty}(P) = -\log \max_{x \in X} P(x)$: As $\alpha \to \infty$ we have $H_n(P) \to \min_{Q} \{D(Q||P) + H(Q)\}$. The parenthesized sum can be written as $-\sum_{x} P(x) \log P(x)$, and is clearly minimized by $Q(x') = 1$, where $P(x') = \max_{x} P(x)$.

4. $H_n(P)$ is a non-increasing function of $\alpha$: $H_n(P)$ is obtained as a minimization (resp. maximization) of functions that are non-increasing in $\alpha$ over $(0, 1)$ (resp. $(1, \infty)$), hence itself is non-increasing over the two regions. Setting $Q = P$ in Theorem 1 yields $H_n(P) \geq H(P)$ for $\alpha < 1$, and $H_n(P) \leq H(P)$ for $\alpha > 1$, and the statement follows.

5. $D_\alpha(P_1||P_2)$ is convex in $P_2$ for $\alpha > 1$ and any fixed $P_1$, and is convex in the pair $(P_1, P_2)$ for $\alpha < 1$: $D(Q||P_2)$ is convex in $P_2$ for any fixed $Q$, hence so is $\frac{\alpha}{1-\alpha} D(Q||P_1) + D(Q||P_2)$. The statement for $\alpha > 1$ follows since a point-wise maximum of convex functions is convex. For $\alpha < 1$, the convexity of $D(Q||P_1)$ in $(Q, P_1)$ and of $D(Q||P_2)$ in $(Q, P_2)$ implies that $\frac{\alpha}{1-\alpha} D(Q||P_1) + D(Q||P_2)$ is convex in $(P_1, P_2, Q)$.

6. (Data Processing Inequality) For any $P_1, P_2 \in \mathcal{P}(X)$ and channel $W: X \rightarrow Y$, 

$$D_\alpha(P_1 W || P_2 W) \leq D_\alpha(P_1 || P_2).$$

We prove only for $\alpha < 1$. For any $Q \in \mathcal{P}(X)$ we have

$$D_\alpha(P(W)||P_2W) \leq \frac{\alpha}{1-\alpha} D(Q(W)||P_1W) + D(Q(W)||P_2W)$$

$$\leq \frac{\alpha}{1-\alpha} D(Q||P_1) + D(Q||P_2)$$

The first inequality follows from Theorem 1, and the second from the data processing inequality for the Kullback-Leibler divergence [18]. The above holds for any $Q$, and minimizing the right-hand-side over $Q$ yields $D_\alpha(P_1||P_2)$.

7. $K_\alpha(P, W) \leq I_\alpha(P, W)$ for $\alpha < 1$, and $K_\alpha(P, W) \geq I_\alpha(P, W)$ for $\alpha > 1$: Immediate by substituting $Q = P$ in the variational characterization of $K_\alpha(P, W)$.

8. $C_\alpha(W) = \max_{P} K_\alpha(P, W)$: We prove only for $\alpha > 1$. Maximize the right-hand-side of (8) over all $P$; note that $P = Q$ is the maximizer for any fixed $Q$.

V. A COMPOSITE HYPOTHESIS TESTING PROBLEM

Suppose two sensors monitor the occurrence of some phenomenon. The sensors may generally have different sampling rates with some ratio $\lambda > 0$, i.e., for each sample provided by Sensor 1, $\lambda$ samples are provided by Sensor 2. When the phenomenon is present, it is observed at Sensor 1 as i.i.d. samples from an unknown distribution $P_1$ in some given family $P_1 \subseteq \mathcal{P}(X)$, and at Sensor 2 as i.i.d. samples from an unknown distribution $P_2$ in some given family $P_2 \subseteq \mathcal{P}(X)$. When the phenomenon is absent, both sensors observe i.i.d. samples from a common unknown “ambient noise” distribution $Q$ in some given family $Q \subseteq \mathcal{P}(X)$. The samples obtained form the sensors are assumed to be mutually independent under each hypothesis.

Suppose we are given $n$ samples from the two Sensors together, where the first $n_1$ samples are from Sensor 1, and the last $n_2 = \lambda n_1$ samples are from Sensor 2. A decision rule corresponds to a set $\Omega_n \subseteq \Omega^n$, which is allowed to be a function of the families $P_1, P_2, Q$, but not of the actual $(P_1, P_2, Q)$. The decision rule declares “phenomena” if the sample vector lies in $\Omega_n$, and “no phenomena” otherwise. The miss-detection and false-alarm error probabilities associated with $\Omega_n$ and a triplet $(P_1, P_2, Q)$ are

$$p_{MD}(\Omega_n|P_1, P_2) \overset{\text{def}}{=} p^{(n)}(X^n \setminus \Omega_n)$$

$$p_{FA}(\Omega_n|Q) \overset{\text{def}}{=} Q^n(\Omega_n)$$

where $p^{(n)} \overset{\text{def}}{=} P_1^{n_1} \times P_2^{n_2}$. The miss-detection exponent associated with a sequence $\Omega = \{\Omega_n\}_{n=1}^\infty$ of decision rules is

$$E_{MD}(\Omega|P_1, P_2) \overset{\text{def}}{=} \liminf_{n \rightarrow \infty} -\frac{1}{n} \log p_{MD}(\Omega_n|P_1, P_2).$$

We will be interested here in maximizing the worst-case mistedetection exponent while guaranteeing a vanishing false-alarm probability, over all feasible $(P_1, P_2, Q)$. Namely, we will consider

$$E_{MD}^* \overset{\text{def}}{=} \sup_{\Omega \in \mathcal{F}} \inf_{P_1 \in P_1, P_2 \in P_2} E_{MD}(\Omega|P_1, P_2)$$

where

$$\mathcal{F} \overset{\text{def}}{=} \left\{ \Omega : \lim_{n \rightarrow \infty} p_{FA}(\Omega_n|Q) = 0, \forall Q \in \mathcal{Q} \right\}.$$ 

In what follows, let $\delta_n \overset{\text{def}}{=} \frac{|X||\log n|}{n}$, and for any two families $P, P' \subseteq \mathcal{P}(X)$, define

$$D_\alpha(P||P') \overset{\text{def}}{=} \sup_{P \in P} \inf_{P' \in P'} D_\alpha(P||P') \overset{\text{def}}{=} \sup_{P \in P} \inf_{P' \in P'} D_\alpha(P||P').$$

Furthermore, write $Q'$ for the closure of the family of all distributions of the form

$$Q'(x) = \sum_{x' \in X} P_1(x)\frac{x}{x'} P_2(x)\frac{x'}{x}$$

for some $P_1 \in P_1, P_2 \in P_2$.

Example 1: The case where $\lambda = 0$ (single sensor) corresponds to a classical setting of composite hypothesis testing. It is well known that in this case [20]

$$E_{MD}^* = D(Q||P_1)$$

which can be achieved by the decision rule

$$\Omega_n = \left\{ x^n : \inf_{Q \in \mathcal{Q}} D(\pi x^n||Q) \geq \delta_n \right\} \overset{\text{def}}{=} \Omega_n$$

(12)

Example 2: If $P_1 \cap P_2 \cap Q \neq \emptyset$, then $E_{MD}^* = 0$ for any $\lambda$.

Example 3: Suppose $P_1$ and $P_2$ have disjoint supports, i.e., $S(P_1) \cap S(P_2) = \emptyset$ for all $P_1 \in P_1$ and $P_2 \in P_2$. Then $E_{MD}^* = \infty$ regardless of $Q$. This is achieved by a simple decision rule that declares “phenomena” when the empirical supports of the samples from the sensors are disjoint, and “no phenomena” otherwise. Clearly, this rule has a zero miss-detection probability for any $n$. It is also easy to see that its false-alarm probability tends to zero exponentially for any $Q \in \mathcal{P}(X)$.

4 For brevity, we disregard integer issues.
Generally, one would expect the optimal miss-detection exponent to be related to some measure of disparity between the families $P_1$ and $P_2$, quantifying the fact that the noise $Q$ cannot mimic both $P_1$ and $P_2$ too well at the same time. As it turns out, at least in the worst case sense over the choice of $Q$, this measure is related to a Rényi divergence between the two families.

**Theorem 2:** For any choice of $P_1$, $P_2$, $Q$, and $\lambda$,

$$E_{MD}^* \geq \lambda(1 + \lambda)^{-1} D_{\frac{1}{1 + \lambda}}(P_1 \| P_2)$$

with equality if and only if the closure of $Q$ has an empty intersection with the associated $Q^*$.

**Proof:** Consider first the case where $Q = \{Q\}$. Let us show that

$$E_{MD}^* = (1 + \lambda)^{-1} (D(Q\|P_1) + \lambda D(Q\|P_2)) .$$

Achievability follows by letting $\Omega_{n1}$ and $\Omega_{n2}$ be the optimal per-sensor decision rules as in (12), and setting

$$\Omega_n = \{ x^{n1}, y^{n2} : x^{n1} \in \Omega_{n1}^{(1)} \text{ or } y^{n2} \in \Omega_{n2}^{(2)} \} .$$

The converse is a simple generalization of the standard single-sensor case [20]. Let $\Omega = \{ \Omega_n \}$ be any sequence of decision sets achieving a vanishing false-alarm probability. For $i \in \{1, 2\}$, let $\Gamma_n$ denote the union of all $n$-dimensional type classes $T_{Q_i}$, where $Q_i \in \mathcal{P}_n(\mathcal{X})$ satisfies $D(Q_i\|Q) \leq \delta_n$. By Lemma 1 property (iv), we have $Q^n(\Gamma_n \times \Gamma_{n2}) \to 1$ as $n \to \infty$. Since by our assumption $Q_n(\mathcal{X}\times \Omega_n) \to 1$, then $Q^n(\Gamma_n \times \Omega_n) \to 1$ as $n \to \infty$. Thus, there must exist a growing number of pairs $(Q_{n1}, Q_{n2}) \in \Gamma_n \times \Gamma_{n2}$ such that $Q^n(T_{Q_{n1}} \times T_{Q_{n2}}) \to 1$ for $i \in \{1, 2\}$, which implies that $D(Q_{ni}\|P_i) \to D(Q_{n}\|P_i)$. Hence for any $P_1 \in P_1$, $P_2 \in P_2$,

$$E_{MD}(\Omega_{n1}, P_1, P_2) \geq (1 + \lambda)^{-1} \inf_{Q \in \mathcal{P}_n(\mathcal{X})} (D(Q\|P_1) + \lambda D(Q\|P_2)) .$$

where $\delta_n = \max(\delta_{n1}, \delta_{n2})$.

Let us show that this rule attains the upper bound in (14). For any $Q \in \mathcal{Q}_n$, $\Omega_n$ is contained in the set of all vectors $(x^{n1}, y^{n2})$ for which either $D(\pi_{x^{n1}}(Q) \geq \delta_{n1}$ or $D(\pi_{y^{n2}}(Q) \geq \delta_{n2}$. Thus, using Lemma 1 property (iv) together with the union bound, we obtain

$$p_{FA}(\Omega_n, Q) \leq |\mathcal{P}_{n1}(\mathcal{X})|2^{-n_1\delta_n'0} + |\mathcal{P}_{n2}(\mathcal{X})|2^{-n_2\delta_n'0}.$$
Q with \( S(Q) = \mathcal{Y} \), merely excluding possibly infinite values. This will be implicit below. For \( \alpha > 1 \), we have

\[
I_\alpha(P, W) = \inf_Q \max_{P \in \mathcal{P}^1(X)} \left( \sum_{x \in X} P(x) \max_{R \in \mathcal{W}(|x|)} \left( \frac{1}{1-\alpha} D(R | W(|x|)) + D(R | Q) \right) \right)
\]

\[
= \inf_Q \max_{P \in \mathcal{P}^1(X)} \left( \sum_{x \in X} P(x) \left( \frac{1}{1-\alpha} D(V(|x|) | W(|x|)) + D(V(|x|) | Q) \right) \right)
\]

\[
= \max_V \inf_Q \left( \frac{\alpha}{1-\alpha} D(V | W) + \sum_{x \in X} P(x) D(V(|x|) | Q) \right)
\]

\[
= \max_V \left( I(P, V) + \frac{\alpha}{1-\alpha} D(V | W) \right)
\]

The maximization is taken over all channels \( V \) such that \( P \circ V \ll P \circ W \). The equalities above are justified as follows:

(a) by virtue of Theorem 1.

(b) the objective function is continuous and concave in \( V \) over a compact set for any fixed \( Q \), and convex in \( V \) for any fixed \( V \). Hence, \( \max \) and \( \inf \) can be interchanged [21, Theorem 4.2]. Concavity of \( V \) follows by writing each of the subaddends as \( D(V(|x|) | Q) - D(V(|x|) | W(|x|)) + \frac{1}{1-\alpha} D(V(|x|) | W(|x|)) \), which is the sum of a linear function and a concave function in \( V \) (for \( \alpha > 1 \)).

(c) on account of (1).

This establishes (7) for \( \alpha > 1 \), where we note that taking the last \( \max \) over all channels \( V : X \rightarrow Y \) changes nothing. The simpler derivation for \( \alpha < 1 \) is similar. To establish (8), write:

\[
K_\alpha(P, W)
\]

\[
= \inf_Q \max_{P \circ V} \left\{ \frac{\alpha}{1-\alpha} D(P' | W | P \circ W) + D(P' | P \circ V \times Q) \right\}
\]

\[
= \max_{P \circ V} \inf_Q \left\{ \frac{\alpha}{1-\alpha} D(P' | W | P \circ V) + D(P' | P \circ V \times Q) \right\}
\]

\[
= \max_{P' \circ V} \left\{ \frac{\alpha}{1-\alpha} D(P' | W | P \circ V) + D(P' | P) \right\}
\]

\[
= \max_{P' \circ V} \left\{ I(P', V) \right\}
\]

\[
= \max_{P' \circ V} \left\{ I(P', W) + \frac{1}{1-\alpha} D(P' | P) \right\}
\]

The maximization is over all \( P' \) and \( V \) such that \( P' \circ V \ll P \circ W \). Equalities (a) and (b) are justified similarly to their counterparts in (16), while (c) and (d) follows from (2) and (7) respectively. This establishes (8) for \( \alpha > 1 \), where we note that taking the last \( \max \) over all \( P' \in \mathcal{P}(X) \) changes nothing. The simpler derivation for \( \alpha < 1 \) is similar.

\section*{References}


