On Rényi Measures and Hypothesis Testing

Ofer Shayevitz
Information Theory & Applications Center
University of California, San Diego
La Jolla, CA 92093, USA
ofersha@ucsd.edu

***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 class of measures satisfying some slightly weaker postulates, yet still intuitively appealing as measures of information [2]. Remarkably, this "reversed" line of thought has proved fruitful; the Rényi information measures have been shown to admit 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 $\mathscr{P}(\mathcal{X})$ the set of all probability distributions over $\mathcal{X}$. The support of a distribution $P \in \mathscr{P}(\mathcal{X})$ is the set $S(P) \stackrel{\text{def}}{=} \{x \in \mathcal{X} : P(x) > 0\}$. The (Shannon) *entropy* of $P \in \mathscr{P}(\mathcal{X})$ is[1]

$$H(P) \stackrel{\text{def}}{=} -\sum_{x\in\mathcal{X}} P(x)\log P(x)\,.$$

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

$$D(P_1\|P_2) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} P_1(x)\log\left(\frac{P_1(x)}{P_2(x)}\right)\,.$$

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

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

$$\begin{aligned}(P\circ W)(x,y) &\stackrel{\text{def}}{=} P(x)W(y|x)\\ PW(y) &\stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} P(x)W(y|x)\,.\end{aligned}$$

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

$$D(V\|W|P) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} P(x)D(V(\cdot|x)\|W(\cdot|x))$$

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

$$\begin{aligned}I(P,W) &\stackrel{\text{def}}{=} H(PW) - \sum_{x\in\mathcal{X}} P(x)H(W(\cdot|x))\\ &= \min_Q \sum_{x\in\mathcal{X}} P(x)D(W(\cdot|x)\|Q) \qquad (1)\\ &= \min_Q D(P\circ W\|P\times Q) \qquad (2)\end{aligned}$$

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

$$C(W) \stackrel{\text{def}}{=} \max_P I(P,W)$$

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

[1] We use the conventions $0\log 0 = 0$, and $a\log\frac{a}{0} = 0$ or $+\infty$ according to whether $a = 0$ or $a > 0$ respectively.

The following facts are well known [18].

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

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

(ii) $|\mathscr{P}^n(\mathcal{X})|^{-1} 2^{nH(Q)} \leq |T_Q| \leq 2^{nH(Q)}$.

(iii) $|\mathscr{P}^n(\mathcal{X})| = \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} \leq (n+1)^{|\mathcal{X}|}$.

(iv) For any $\delta > 0$

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

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

$$H_\alpha(P) \stackrel{\text{def}}{=} \frac{1}{1-\alpha} \log \sum_{x\in\mathcal{X}} P(x)^\alpha.$$

We denote by $H_0(P)$, $H_1(P)$ and $H_\infty(P)$ the limits of $H_\alpha(P)$ as α tends to $0, 1$ and ∞, respectively[2]. The *Rényi divergence of order* α between two distributions $P_1, P_2 \in \mathscr{P}(\mathcal{X})$ is[3]

$$D_\alpha(P_1\|P_2) \stackrel{\text{def}}{=} \frac{1}{\alpha-1} \log \sum_{x\in\mathcal{X}} P_1(x)^\alpha P_2(x)^{1-\alpha}.$$

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 α tends to $0, 1$ and ∞, respectively[2]. Note that for $\alpha < 1$, $D_\alpha(P_1\|P_2) < \infty$ if and only if $S(P_1) \cap S(P_2) \neq \emptyset$, and for $\alpha > 1$, $D_\alpha(P_1\|P_2) < \infty$ 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:

$$I_\alpha(P, W) \stackrel{\text{def}}{=} \min_Q \sum_{x\in\mathcal{X}} P(x) D_\alpha(W(\cdot|x)\|Q) \tag{3}$$

corresponding to (1), and

$$K_\alpha(P, W) \stackrel{\text{def}}{=} \min_Q D_\alpha(P \circ W \| P \times Q) \tag{4}$$

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

$$C_\alpha(W) \stackrel{\text{def}}{=} \max_P I_\alpha(P, W)$$

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 $\alpha > 1$,

$$H_\alpha(P) = \min_Q \left\{ \frac{\alpha}{\alpha-1} D(Q\|P) + H(Q) \right\} \tag{5}$$

$$D_\alpha(P_1\|P_2) = \max_{Q\ll P_1} \left\{ \frac{\alpha}{1-\alpha} D(Q\|P_1) + D(Q\|P_2) \right\} \tag{6}$$

$$I_\alpha(P, W) = \max_V \left\{ I(P, V) + \frac{\alpha}{1-\alpha} D(V\|W|P) \right\} \tag{7}$$

$$K_\alpha(P, W) = \max_Q \left\{ I_\alpha(Q, W) + \frac{1}{1-\alpha} D(Q\|P) \right\} \tag{8}$$

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

[2]These limits are known to exist, a fact we reestablish in the sequel.

[3]For $\alpha > 1$ we adopt the convention where $a^\alpha \cdot 0^{1-\alpha} = 0$ or $+\infty$ according to whether $a = 0$ or $a > 0$ respectively.

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 $\mathcal{X}_1 \stackrel{\text{def}}{=} S(P_1)$ and $\mathcal{X}_2 \stackrel{\text{def}}{=} S(P_2)$ for short. We derive a characterization for the functional

$$J_{\alpha,\beta}(P_1, P_2) \stackrel{\text{def}}{=} -\log \sum_{x\in\mathcal{X}_1} P_1(x)^\alpha P_2(x)^\beta \tag{9}$$

for any $\alpha > 0$ and β. 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,

$$\begin{aligned} J_{\alpha,\beta}(P_1, P_2) &= -\frac{1}{n} \log \sum_{x^n\in\mathcal{X}_1^n} P_1(x^n)^\alpha P_2(x^n)^\beta \\ &\leq -\frac{1}{n} \log \sum_{Q\in\mathscr{P}^n(\mathcal{X}_1)} 2^{-n(\alpha(D(Q\|P_1)+H(Q))+\beta(D(Q\|P_2)+H(Q))} \\ &\qquad \times |\mathscr{P}^n(\mathcal{X}_1)|^{-1} 2^{nH(Q)} \\ &\leq \min_{Q\in\mathscr{P}^n(\mathcal{X}_1)} \{\alpha D(Q\|P_1) + \beta D(Q\|P_2) + (\alpha+\beta-1)H(Q)\} \\ &\qquad + \frac{|\mathcal{X}_1| \log(n+1)}{n} \end{aligned}$$

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

$$\begin{aligned} &J_{\alpha,\beta}(P_1, P_2) \\ &\geq -\frac{1}{n} \log \sum_{Q\in\mathscr{P}^n(\mathcal{X}_1)} 2^{-n(\alpha D(Q\|P_1)+\beta D(Q\|P_2)+(\alpha+\beta-1)H(Q))} \\ &\geq \min_{Q\in\mathscr{P}^n(\mathcal{X}_1)} \{\alpha D(Q\|P_1) + \beta D(Q\|P_2) + (\alpha+\beta-1)H(Q)\} \\ &\qquad - \frac{|\mathcal{X}_1| \log(n+1)}{n}. \end{aligned}$$

$\bigcup_n \mathscr{P}^n(\mathcal{X}_1)$ is dense in $\mathscr{P}(\mathcal{X}_1)$, and the objective function is continuous in Q over the compact set $\mathscr{P}(\mathcal{X}_1 \cap \mathcal{X}_2)$, and equals $\pm\infty$ over $\mathscr{P}(\mathcal{X}_1) \setminus \mathscr{P}(\mathcal{X}_1 \cap \mathcal{X}_2)$ according to $\text{sign}(\beta)$. Thus, taking the limit as $n \to \infty$, we obtain:

$$\begin{aligned} &J_{\alpha,\beta}(P_1, P_2) \\ &\ = \min_{Q\ll P_1} \{\alpha D(Q\|P_1) + \beta D(Q\|P_2) + (\alpha+\beta-1)H(Q)\}. \end{aligned} \tag{10}$$

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 $Q \in \mathscr{P}(\mathcal{X})$ 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 $H_\alpha(P) \to \max_Q H(Q)$ 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. $H_\infty(P) = -\log\max_{x\in\mathcal{X}} P(x)$: As $\alpha \to \infty$ we have $H_\alpha(P) \to \min_Q\{D(Q\|P) + H(Q)\}$. The parenthesized sum can be written as $-\sum_x Q(x)\log P(x)$, and is clearly minimized by $Q(x') = 1$, where $P(x') = \max_x P(x)$.
4. *$H_\alpha(P)$ is a non-increasing function of α.*: $H_\alpha(P)$ is obtained as a minimization (resp. maximization) of functions that are non-increasing in α over $(0,1)$ (resp. $(1,\infty)$), hence itself is non-increasing over the two regions. Setting $Q = P$ in Theorem 1 yields $H_\alpha(P) \geq H(P)$ for $\alpha < 1$, and $H_\alpha(P) \leq H(P)$ 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 pointwise 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). The result now follows since minimizing a convex function over a convex set ($\mathscr{P}(S(P_1))$ in this case) preserves convexity.
6. (*Data Processing Inequality*) *For any $P_1, P_2 \in \mathscr{P}(\mathcal{X})$ and channel $W : \mathcal{X} \mapsto \mathcal{Y}$,*
$$D_\alpha(P_1W\|P_2W) \leq D_\alpha(P_1\|P_2)\,.$$
We prove only for $\alpha < 1$. For any $Q \in \mathscr{P}(\mathcal{X})$ we have
$$\begin{aligned} &D_\alpha(P_1W\|P_2W) \\ &\quad\leq \frac{\alpha}{1-\alpha}D(QW\|P_1W) + D(QW\|P_2W) \\ &\quad\leq \frac{\alpha}{1-\alpha}D(Q\|P_1) + D(Q\|P_2) \end{aligned}$$
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 phenomena. The sensors may generally have different sampling rates with some ratio $\lambda > 0$, i.e., for each sample provided by Sensor 1, λ samples are provided by Sensor 2. When the phenomena is present, it is observed at Sensor 1 as i.i.d. samples from an unknown distribution P_1 in some given family $\mathbf{P_1} \subseteq \mathscr{P}(\mathcal{X})$, and at Sensor 2 as i.i.d. samples from an unknown distribution P_2 in some given family $\mathbf{P_2} \subseteq \mathscr{P}(\mathcal{X})$. When the phenomena is absent, both sensors observe i.i.d. samples from a common unknown "ambient noise" distribution Q in some given family $\mathbf{Q} \subseteq \mathscr{P}(\mathcal{X})$. 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[4] are from Sensor 2. A *decision rule* corresponds to a set $\Omega_n \subseteq \mathcal{X}^n$, which is allowed to be a function of the families $\mathbf{P_1}, \mathbf{P_2}, \mathbf{Q}$, but not of the actual (P_1, P_2, Q). The decision rule declares "phenomena" if the sample vector lies in Ω_n, and "no phenomena" otherwise. The *miss-detection* and *false-alarm* error probabilities associated with Ω_n and a triplet (P_1, P_2, Q) are

$$p_{\text{MD}}(\Omega_n|P_1,P_2) \stackrel{\text{def}}{=} P^{(n)}\left(\mathcal{X}^n \setminus \Omega_n\right)$$
$$p_{FA}(\Omega_n|Q) \stackrel{\text{def}}{=} Q^n(\Omega_n)$$

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

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

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

$$E^*_{MD} \stackrel{\text{def}}{=} \sup_{\Omega\in\mathscr{F}} \inf_{P_1\in\mathbf{P_1}, P_2\in\mathbf{P_2}} E_{MD}(\Omega|P_1,P_2)$$

where

$$\mathscr{F} \stackrel{\text{def}}{=} \left\{\Omega : \lim_{n\to\infty} p_{FA}(\Omega_n|Q) = 0\,,\ \forall Q\in\mathbf{Q}\right\}\,.$$

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

$$D_\alpha(\mathbf{P}\|\mathbf{P}') \stackrel{\text{def}}{=} \inf_{P\in\mathbf{P}, P'\in\mathbf{P}'} D_\alpha(P\|P')\,. \tag{11}$$

Furthermore, write $\mathbf{Q}^*$ for the closure of the family of all distributions of the form

$$Q^*(x) = \frac{P_1(x)^{\frac{1}{1+\lambda}}P_2(x)^{\frac{\lambda}{1+\lambda}}}{\sum_{x\in\mathcal{X}} P_1(x)^{\frac{1}{1+\lambda}}P_2(x)^{\frac{\lambda}{1+\lambda}}}$$

for some $P_1 \in \mathbf{P_1}, P_2 \in \mathbf{P_2}$.

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]

$$E^*_{MD} = D(\mathbf{Q}\|\mathbf{P_1})$$

which can be achieved by the decision rule

$$\Omega_n = \left\{x^n : \inf_{Q\in\mathbf{Q}} D(\pi_{x^n}\|Q) \geq \delta_n\right\}\,. \tag{12}$$

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

Example 3: Suppose $\mathbf{P_1}$ and $\mathbf{P_2}$ have disjoint supports, i.e., $S(P_1) \cap S(P_2) = \emptyset$ for all $P_1 \in \mathbf{P_1}$ and $P_2 \in \mathbf{P_2}$. Then $E^*_{MD} = \infty$ regardless of $\mathbf{Q}$. 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 $Q \in \mathscr{P}(\mathcal{X})$.

[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 $\mathbf{P_1}$ and $\mathbf{P_2}$, 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 $\mathbf{Q}$, this measure is related to a Rényi divergence between the two families.

Theorem 2: For any choice of $\mathbf{P_1}, \mathbf{P_2}, \mathbf{Q}$ and λ,

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

with equality if and only if the closure of $\mathbf{Q}$ has an nonempty intersection with the associated $\mathbf{Q}^*$.

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

$$E^*_{MD} = (1+\lambda)^{-1} \left(D(Q\|\mathbf{P_1}) + \lambda D(Q\|\mathbf{P_2}) \right) .$$

Achievability follows by letting $\Omega^{(1)}_{n_1}$ and $\Omega^{(2)}_{n_2}$ be the optimal per-sensor decision rules as in (12), and setting

$$\Omega_n \stackrel{\text{def}}{=} \left\{ (x^{n_1}, y^{n_2}) : x^{n_1} \in \Omega^{(1)}_{n_1} \;\text{ or }\; y^{n_2} \in \Omega^{(2)}_{n_2} \right\} . \qquad (13)$$

The converse is a simple generalization of the standard single-sensor case [20]. Let $\Omega' = \{\Omega'_n\}$ be any sequence of decision rules achieving a vanishing false-alarm probability. For $i \in \{1,2\}$, let Γ_{n_i} denote the union of all n_i-dimensional type classes T_{Q_i} where $Q_i \in \mathscr{P}^{n_i}(\mathcal{X})$ satisfies $D(Q_i\|Q) \leq \delta_{n_i}$. By Lemma 1 property (iv), we have $Q^n(\Gamma_{n_1} \times \Gamma_{n_2}) \to 1$ as $n \to \infty$. Since by our assumption $Q^n(\mathcal{X}^n \setminus \Omega'_n) \to 1$, then $Q^n((\Gamma_{n_1} \times \Gamma_{n_2}) \setminus \Omega'_n) \geq \frac{1}{2}$ (say) for any n large enough. Thus, there must exist a pair of types $(Q_{1,n}, Q_{2,n}) \in \Gamma_{n_1} \times \Gamma_{n_2}$ such that $Q^n((T_{Q_{1,n}} \times T_{Q_{2,n}}) \setminus \Omega'_n) \geq \frac{1}{2} Q^n(T_{Q_{1,n}} \times T_{Q_{2,n}})$. Since both Q^n and $P^{(n)}$ are constant over $T_{Q_{1,n}} \times T_{Q_{2,n}}$, the same inequality holds for $P^{(n)}$. Therefore,

$$\begin{aligned} -\frac{1}{n} \log P^{(n)}(\mathcal{X}^n \setminus \Omega'_n) & \\ &\leq -\frac{1}{n} \log P^{(n)}((T_{Q_{1,n}} \times T_{Q_{2,n}}) \setminus \Omega'_n) \\ &\leq -\frac{1}{n} \log \frac{1}{2} P^{(n)}(T_{Q_{1,n}} \times T_{Q_{2,n}}) \\ &\leq (1+\lambda)^{-1} \left(D(Q_{1,n}\|P_1) + \lambda D(Q_{2,n}\|P_2) \right) \\ &\qquad + \frac{1 + 2|\mathcal{X}| \log(n+1)}{n} \end{aligned}$$

where properties (i)-(iii) of Lemma 1 were used in the last inequality. Letting $n \to \infty$, and recalling that $D(Q_{i,n}\|Q) \to 0$ which implies $D(Q_{i,n}\|P_i) \to D(Q\|P_i)$, the converse follows.

As a result, it is now clear that for a general $\mathbf{Q}$

$$E^*_{MD} \leq (1+\lambda)^{-1} \inf_{Q \in \mathbf{Q}} \left(D(Q\|\mathbf{P_1}) + \lambda D(Q\|\mathbf{P_2}) \right) . \qquad (14)$$

The decision rule (13) above (with $\Omega^{(1)}_{n_1}$ and $\Omega^{(2)}_{n_2}$ now taking the infimum over the family $\mathbf{Q}$) will generally fail to achieve the upper bound in (14), and may even not attain a vanishing miss-detection probability. For instance, if $\mathbf{P_1} = \{P_1\}$, $\mathbf{P_2} = \{P_2\}$ and $\mathbf{Q} = \{P_1, P_2\}$, then $p_{\text{MD}}(\Omega_n | P_1, P_2) \to 1$, whereas the upper bound (14) is positive if $P_1 \neq P_2$. Clearly, the problem is that each sensor makes its own binary decision before those are combined, not taking into account that Q is common. This shortcoming is easily corrected by the following modified decision rule:

$$\widetilde{\Omega}_n = \left\{ (x^{n_1}, y^{n_2}) : \inf_{Q \in \mathbf{Q}} \max\left\{ D(\pi_{x^{n_1}}\|Q), D(\pi_{y^{n_2}}\|Q) \right\} \geq \delta'_n \right\}$$

where $\delta'_n = \max(\delta_{n_1}, \delta_{n_2})$.

Let us show that this rule attains the upper bound in (14). For any $Q \in \mathbf{Q}$, $\widetilde{\Omega}_n$ is contained in the set of all vectors (x^{n_1}, y^{n_2}) for which either $D(\pi_{x^{n_1}}\|Q) \geq \delta'_n$ or $D(\pi_{y^{n_2}}\|Q) \geq \delta'_n$. Thus, using Lemma 1 property (iv) together with the union bound, we obtain

$$\begin{aligned} p_{FA}(\widetilde{\Omega}_n | Q) &\leq |\mathscr{P}^{n_1}(\mathcal{X})| 2^{-n_1 \delta'_n} + |\mathscr{P}^{n_2}(\mathcal{X})| 2^{-n_2 \delta'_n} \\ &\leq \binom{n_1 + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} n_1^{-|\mathcal{X}|} + \binom{n_2 + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} n_2^{-|\mathcal{X}|} \end{aligned}$$

hence $p_{FA}(\widetilde{\Omega}_n | Q) \to 0$ as $n \to \infty$, for any $Q \in \mathbf{Q}$.

Define the set $\Pi_n \subseteq \mathscr{P}^{n_1}(\mathcal{X}) \times \mathscr{P}^{n_2}(\mathcal{X})$ of all the type pairs (Q_1, Q_2) for which there exists some $Q \in \mathbf{Q}$ such that $D(Q_1\|Q) < \delta'_n$ and $D(Q_2\|Q) < \delta'_n$. By definition, $\mathcal{X}^n \setminus \widetilde{\Omega}_n$ is a union of all type classes products pertaining to Π_n. Therefore, using properties (i)-(iv) of Lemma 1 again, we get

$$\begin{aligned} -\frac{1}{n} \log P^{(n)}(\mathcal{X}^n \setminus \widetilde{\Omega}_n) & \\ &= -\frac{1}{n} \log \sum_{(Q_1, Q_2) \in \Pi_n} P_1^{n_1}(T_{Q_1}) \cdot P_2^{n_2}(T_{Q_2}) \\ &\geq (1+\lambda)^{-1} \min_{(Q_1, Q_2) \in \Pi_n} \left(D(Q_1\|P_1) + \lambda D(Q_2\|P_2) \right) \\ &\qquad - \frac{2|\mathcal{X}| \log(n+1)}{n} . \end{aligned}$$

Let $(Q_{1,n}, Q_{2,n})$ achieve the minimum above. Then by definition there exists $Q_n \in \mathbf{Q}$ such that $D(Q_{i,n}\|Q_n) < \delta'_n \to 0$ for $i \in \{1,2\}$, which implies that $D(Q_{n,i}\|P_i) \to D(Q_n\|P_i)$. Hence for any $P_1 \in \mathbf{P_1}, P_2 \in \mathbf{P_2}$,

$$E_{MD}(\widetilde{\Omega} | P_1, P_2) \geq (1+\lambda)^{-1} \inf_{Q \in \mathbf{Q}} \left(D(Q\|P_1) + \lambda D(Q\|P_2) \right)$$

Therefore, $\widetilde{\Omega}$ attains the upper bound in (14), and thus

$$\begin{aligned} E^*_{MD} &= (1+\lambda)^{-1} \inf_{Q \in \mathbf{Q}} \left(D(Q\|\mathbf{P_1}) + \lambda D(Q\|\mathbf{P_2}) \right) \qquad (15) \\ &\geq (1+\lambda)^{-1} \min_{Q \in \mathscr{P}(\mathcal{X})} \left(D(Q\|\mathbf{P_1}) + \lambda D(Q\|\mathbf{P_2}) \right) \\ &= \lambda(1+\lambda)^{-1} D_{\frac{1}{1+\lambda}}(\mathbf{P_1}\|\mathbf{P_2}) \end{aligned}$$

where the inequality is on account of Theorem 1. Note that for the $\alpha < 1$ counterpart of (6), minimizing over $Q \in \mathscr{P}(\mathcal{X})$ instead of $Q \ll P_1$ changes nothing. Necessary and sufficient conditions for an equality can be easily verified, see [19]. ∎

The lower bound in Theorem 2 is independent of the noise family $\mathbf{Q}$, hence the Rényi divergence between the families $\mathbf{P_1}$ and $\mathbf{P_2}$ admits an operational interpretation as the optimal worst-case miss-detection exponent (up to a constant) when the noise distribution Q is completely unknown (i.e., $\mathbf{Q} = \mathscr{P}(\mathcal{X})$), or more generally, when Q can take values in the "worst noise" set $\mathbf{Q}^*$. In other cases this serves only as a lower bound, and the strictly larger exponent is given by (15). It is possible to interpret this latter exponent as a (limit of a) generalized form of the Rényi divergence, taking into account also the family $\mathbf{Q}$, see [19]

Appendix

Proof of relations (7) and (8): . As in [6], the minimum in (3) and (4) can be replaced with an infimum over distributions

Q with $S(Q) = \mathcal{Y}$, merely excluding possibly infinite values. This will be implicit below. For $\alpha > 1$, we have

$$\begin{aligned}
&I_\alpha(P,W) \\
&\stackrel{(a)}{=} \inf_Q \sum_{x\in\mathcal{X}} P(x) \max_{R\ll W(\cdot|x)} \left(\frac{\alpha}{1-\alpha} D(R\|W(\cdot|x)) + D(R\|Q) \right) \\
&= \inf_Q \max_V \sum_{x\in\mathcal{X}} P(x) \left(\frac{\alpha}{1-\alpha} D(V(\cdot|x)\|W(\cdot|x)) \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad \left. + D(V(\cdot|x)\|Q) \right) \\
&\stackrel{(b)}{=} \max_V \inf_Q \left(\frac{\alpha}{1-\alpha} D(V\|W|P) \right. \\
&\qquad\qquad\qquad\qquad \left. + \sum_{x\in\mathcal{X}} P(x) D(V(\cdot|x)\|Q) \right) \\
&\stackrel{(c)}{=} \max_V \left\{ I(P,V) + \frac{\alpha}{1-\alpha} D(V\|W|P) \right\}
\end{aligned} \tag{16}$$

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 Q for any fixed V. Hence, $\max$ and $\inf$ can be interchanged [21, Theorem 4.2]. Concavity in V follows by writing each of the summands as $[D(V(\cdot|x)\|Q) - D(V(\cdot|x)\|W(\cdot|x))] + \frac{1}{1-\alpha} D(V(\cdot|x)\|W(\cdot|x))$, 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 $V : \mathcal{X} \mapsto \mathcal{Y}$ changes nothing. The simpler derivation for $\alpha < 1$ is similar. To establish (8), write:

$$\begin{aligned}
&K_\alpha(P,W) \\
&\quad \stackrel{(a)}{=} \inf_Q \max_{P'\circ V} \left\{ \frac{\alpha}{1-\alpha} D(P'\circ V\|P\circ W) \right. \\
&\qquad\qquad\qquad\qquad \left. + D(P'\circ V\|P\times Q) \right\} \\
&\quad \stackrel{(b)}{=} \max_{P'\circ V} \inf_Q \left\{ \frac{\alpha}{1-\alpha} D(P'\circ V\|P\circ W) \right. \\
&\qquad\qquad\qquad\qquad \left. + D(P'\circ V\|P\times Q) \right\} \\
&\quad = \max_{P'\circ V} \inf_Q \left\{ \frac{\alpha}{1-\alpha} D(P'\circ V\|P\circ W) + D(P'\|P) \right. \\
&\qquad\qquad\qquad\qquad \left. + D(P'\circ V\|P'\times Q) \right\} \\
&\quad \stackrel{(c)}{=} \max_{P'\circ V} \left\{ \frac{\alpha}{1-\alpha} D(P'\circ V\|P\circ W) + D(P'\|P) \right. \\
&\qquad\qquad\qquad\qquad \left. + I(P',V) \right\} \\
&\quad = \max_{P'\circ V} \left\{ \frac{\alpha}{1-\alpha} D(V\|W|P') + \frac{1}{1-\alpha} D(P'\|P) \right. \\
&\qquad\qquad\qquad\qquad \left. + I(P',V) \right\} \\
&\quad \stackrel{(d)}{=} \max_{P'} \left\{ I_\alpha(P',W) + \frac{1}{1-\alpha} D(P'\|P) \right\}
\end{aligned}$$

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 $P' \in \mathscr{P}(\mathcal{X})$ changes nothing. The simpler derivation for $\alpha < 1$ is similar. ■

REFERENCES

[1] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, 2008.

[2] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Sympos. Math. Stat. and Prob.*, 1960, vol. 1, pp. 547–561.

[3] L.L. Campbell, "A coding theorem and Rényi's entropy," *Information and Control*, vol. 8, no. 4, pp. 423 – 429, 1965.

[4] A. Rényi, "On the foundations of information theory," *Review of the International Statistical Institute*, vol. 33, no. 1, pp. 1–14, 1965.

[5] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Info. Theory*, vol. IT-14, pp. 490 – 501, May 1968.

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

[7] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. on Info. Theory*, vol. 42, no. 1, pp. 99 –105, Jan. 1996.

[8] C.H. Bennett, G. Brassard, C. Crepeau, and U.M. Maurer, "Generalized privacy amplification," *IEEE Trans. on Info. Theory*, vol. 41, no. 6, pp. 1915 –1923, Nov. 1995.

[9] U. Erez and R. Zamir, "Error exponents of modulo-additive noise channels with side information at the transmitter," *IEEE Trans. on Info. Theory*, vol. 47, no. 1, pp. 210 –218, Jan. 2001.

[10] O. Shayevitz, E. Meron, M. Feder, and R. Zamir, "Delay and redundancy in lossless source coding," *IEEE Trans. on Info. Theory*, submitted (available online arXiv:1012.4225).

[11] R. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. on Info. Theory*, vol. 11, no. 1, pp. 3 – 18, Jan. 1965.

[12] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. of the 48th Allerton Conference on Communication, Control, and Computing*, 2010.

[13] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Multiple source adaptation and the Rényi divergence," in *Proc. of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

[14] T. van Erven and P. Harremoes, "Rényi divergence and majorization," in *Proc. of the International Symposium on Information Theory*, 2010, pp. 1335 –1339.

[15] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," in *Topics in information theory (Second Colloq., Keszthely, 1975)*, pp. 41–52. Colloq. Math. Soc. János Bolyai, Vol. 16. North-Holland, Amsterdam, 1977.

[16] A. Ingber, I. Leibowitz, R. Zamir, and M. Feder, "Distortion lower bounds for finite dimensional joint source-channel coding," in *Proc. of the International Symposium on Information Theory*, July 2008, pp. 1183–1187.

[17] N. Merhav and E. Arikan, "The Shannon cipher system with a guessing wiretapper," *IEEE Trans. on Info. Theory*, vol. 45, no. 6, pp. 1860 –1866, Sept. 1999.

[18] I. Csiszár and J. Körner, *Information theory : Coding theorems for discrete memoryless systems*, 1986.

[19] O. Shayevitz, "A note on a characterization of Rényi measures and its relation to composite hypothesis testing," *IEEE Trans. on Info. Theory*, submitted (available online arXiv:1012.4401).

[20] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 2004.

[21] M. Sion, "On general minimax theorems," *Pac. J. Math.*, vol. 8, pp. 171–176, 1958.