

On conditional Sibson's α -Mutual Information

Amedeo Roberto Esposito, Diyuan Wu, Michael Gastpar
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{amedeo.esposito, diyuan.wu, michael.gastpar}@epfl.ch

Abstract—In this work, we analyse how to define a conditional version of Sibson's α -Mutual Information. Several such definitions can be advanced and they all lead to different information measures with different (but similar) operational meanings. We will analyse in detail one such definition, compute a closed-form expression for it and endorse it with an operational meaning while also considering some applications. The alternative definitions will also be mentioned and compared.

Index Terms—Rényi-Divergence, Sibson's Mutual Information, Conditional Mutual Information, Information Measures

I. INTRODUCTION

Sibson's α -Mutual Information is a generalization of Shannon's Mutual Information with several applications in probability, information and learning theory [1]. In particular, it has been used to provide concentration inequalities in settings where the random variables are **not** independent, with applications to learning theory [1]. The measure is also connected to Gallager's exponent function, a central object in the channel coding problem both for rates below and above capacity [2], [3]. Moreover, a new operational meaning has been given to the measure with $\alpha = +\infty$ when a novel measure of information leakage has been proposed in [4], under the name of Maximal leakage. Similarly to I_α , Maximal Leakage has recently found applications in learning and probability theory [1]. However, while Maximal Leakage has a corresponding conditional form [4], Sibson's α -Mutual Information lacks an agreed upon conditional version. In this work we analyse a path that could be taken in defining such a measure and will focus on one specific choice, given in Definition 4 below. We discuss key properties of this choice and endow it with an operational meaning as the error-exponent in a properly defined hypothesis testing problem. Moreover, we hint at some application of this measure to other settings as well. The choice we make is not unique and we will explain how making different choices leads to different information measures, all of them equally meaningful. A conditional version of Sibson's I_α has been presented in [5]. We briefly present their measure in Sec. III-B along with a new result that we believe to be of interest. We then present in Sec. III-C a different choice for conditional I_α . We show some properties of this measure, compare the two objects in Sec. III-E and then discuss a general approach to associate an operational meaning to these measures in Sec. IV. Alternative routes have been considered in [6] where Arimoto's generalisation of the Mutual Information has been considered and a conditional version has been given.

II. BACKGROUND AND DEFINITIONS

Given a function $f : \mathbb{R} \rightarrow [-\infty, +\infty]$ we can define its convex conjugate $f^* : \mathbb{R} \rightarrow [-\infty, +\infty]$ as follows:

$$f^*(\lambda) = \sup_x (\lambda x - f(x)). \quad (1)$$

Given a function f , f^* is guaranteed to be lower semi-continuous and convex. We can re-apply the conjugation operator to f^* and obtain f^{**} . If f is convex and lower semicontinuous then $f = f^{**}$, otherwise all we can say is that $\forall x \in \mathbb{R} \ f^{**}(x) \leq f(x)$. \log denotes the natural logarithm.

A. Sibson's α -Mutual Information

Introduced by Rényi as a generalization of entropy and KL-divergence, α -divergence has found many applications ranging from hypothesis testing to guessing and several other statistical inference and coding problems [7]. Indeed, it has several useful operational interpretations (e.g., hypothesis testing, and the cut-off rate in block coding [8], [9]). It can be defined as follows [8].

Definition 1. Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\alpha > 0$ be a positive real number different from 1. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$ (such a measure always exists, e.g. $\mu = (\mathcal{P} + \mathcal{Q})/2$) and denote with p, q the densities of \mathcal{P}, \mathcal{Q} with respect to μ . The α -Divergence of \mathcal{P} from \mathcal{Q} is defined as follows:

$$D_\alpha(\mathcal{P} \parallel \mathcal{Q}) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu. \quad (2)$$

Remark 1. The definition is independent of the chosen measure μ . It is indeed possible to show that $\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{q}{p}\right)^{1-\alpha} d\mathcal{P}$, and that whenever $\mathcal{P} \ll \mathcal{Q}$ or $0 < \alpha < 1$, we have $\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{p}{q}\right)^\alpha d\mathcal{Q}$, see [8].

It can be shown that if $\alpha > 1$ and $\mathcal{P} \not\ll \mathcal{Q}$ then $D_\alpha(\mathcal{P} \parallel \mathcal{Q}) = \infty$. The behaviour of the measure for $\alpha \in \{0, 1, \infty\}$ can be defined by continuity. In general, one has that $D_1(\mathcal{P} \parallel \mathcal{Q}) = D(\mathcal{P} \parallel \mathcal{Q})$ but if $D(\mathcal{P} \parallel \mathcal{Q}) = \infty$ or there exists β such that $D_\beta(\mathcal{P} \parallel \mathcal{Q}) < \infty$ then $\lim_{\alpha \downarrow 1} D_\alpha(\mathcal{P} \parallel \mathcal{Q}) = D(\mathcal{P} \parallel \mathcal{Q})$ [8, Theorem 5]. For an extensive treatment of α -divergences and their properties we refer the reader to [8]. Starting from Rényi's Divergence and the geometric averaging that it involves, Sibson built the notion of Information Radius [10]:

Definition 2. Let (μ_1, \dots, μ_n) be a family of probability measures and (w_1, \dots, w_n) be a set of weights s.t. $w_i \geq 0$

for $i = 1, \dots, n$ and such that $\sum_{i=1}^n w_i > 0$. Let $\alpha \geq 1$, the information radius of order α is defined as:

$$\frac{1}{\alpha - 1} \min_{\nu \ll \sum_i w_i \mu_i} \log \left(\sum_i w_i \exp((\alpha - 1)D_\alpha(\mu_i \| \nu)) \right).$$

Suppose now we have two random variables X, Y jointly distributed according to \mathcal{P}_{XY} . It is possible to generalise Def. 2 and see that the information radius is a special case of the following quantity [7]:

$$I_\alpha(X, Y) = \min_{Q_Y} D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X Q_Y). \quad (3)$$

$I_\alpha(X, Y)$ represents a generalisation of Shannon's Mutual Information and possesses many interesting properties [7]. Indeed, $\lim_{\alpha \rightarrow 1} I_\alpha(X, Y) = I(X; Y)$. On the other hand when $\alpha \rightarrow \infty$, we get: $I_\infty(X, Y) = \log \mathbb{E}_{\mathcal{P}_Y} \left[\sup_{x: \mathcal{P}_X(x) > 0} \frac{\mathcal{P}_{XY}(\{x, Y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{Y\})} \right] = \mathcal{L}(X \rightarrow Y)$, where $\mathcal{L}(X \rightarrow Y)$ denotes the Maximal Leakage from X to Y , a recently defined information measure with an operational meaning in the context of privacy and security [4]. For more details on Sibson's α -MI, as well as a closed-form expression, we refer the reader to [7], as for Maximal Leakage the reader is referred to [4].

III. DEFINITION

A. Introduction

The characterisation expressed in (3) represents the foundation of this work. Indeed, using (3) as the definition of Sibson's α -MI allows us to draw parallels with Shannon's Mutual Information. This, in turn, allows us to define, drawing inspiration from Shannon's measures, an analogous conditional version of Sibson's I_α . It is very well known that $I(X; Y) = D(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y)$ as well as $I(X; Y|Z) = D(\mathcal{P}_{XYZ} \| \mathcal{P}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})$. We can thus follow a similar approach in defining a conditional α -Mutual Information: we will estimate the (Rényi's) divergence of the joint \mathcal{P}_{XYZ} from a distribution characterised by the Markov chain $X - Z - Y$ via α -Divergences. Mimicking (3) we will also minimise such divergence with respect to a family of measures. Having now three random variables, we can think of three natural factorisations for \mathcal{P}_{XYZ} (assuming that $X - Z - Y$ holds): $\mathcal{P}_X \mathcal{P}_{Z|X} \mathcal{P}_{Y|Z}$, $\mathcal{P}_Y \mathcal{P}_{Z|Y} \mathcal{P}_{X|Z}$, $\mathcal{P}_Z \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}$. The question then is: which measure should we minimise with respect to, in order to define $I_\alpha(X, Y|Z)$? Natural candidates seem to be the minimisations with respect to $\mathcal{Q}_Z, \mathcal{Q}_{Y|Z}$ and \mathcal{Q}_Y . The matter is strongly connected to the operational meaning that the information measure acquires, alongside with the applications it can provide. Each of the definitions can be useful in specific settings. Keeping this in mind, the purpose of this work is not to compare different definitions in order to find the best one but rather to highlight properties of the different definitions with an operationally driven approach. Each of these measures can be associated to a hypothesis testing problem and a bound relating different measures of the same event (typically a joint and a Markov chain-like distribution). Different applications require different conditional I_α 's. With this drive, let us make

a specific choice for the minimisation and draw a parallel with the others along the way. The random variable whose measure¹ we choose to minimise will be denoted as a superscript.

$$B. I_\alpha^{Y|Z}(X, Y|Z)$$

In [5], conditional α -mutual information was defined as follows:

Definition 3. Let X, Y, Z be three random variables jointly distributed according to \mathcal{P}_{XYZ} . For $\alpha > 0$, a conditional Sibson's mutual information of order α between X and Y given Z is defined as:

$$I_\alpha^{Y|Z}(X, Y|Z) = \min_{Q_{Y|Z}} D_\alpha(\mathcal{P}_{XYZ} \| \mathcal{P}_{X|Z} Q_{Y|Z} \mathcal{P}_Z). \quad (4)$$

It is possible to find a closed-form expression for Def. 3 [5, Section IV.C.2]. This definition is interesting as setting Z equal to a constant allows us to retrieve $I_\alpha(X, Y)$. Moreover, starting from Definition 3 and its closed-form expression one can retrieve the following result.

Theorem 1. Let $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{F}, \mathcal{P}_{XYZ})$ be a probability space. Let \mathcal{P}_Z and $\mathcal{P}_{X|Z}$ be the induced conditional and marginal distributions. Assume that $\mathcal{P}_{XYZ} \ll \mathcal{P}_Z \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}$. Given $E \in \mathcal{F}$ and $z \in \mathcal{Z}, y \in \mathcal{Y}$, let $E_{z,y} = \{x : (x, y, z) \in E\}$. Then, fixed $\alpha \geq 1$:

$$\mathcal{P}_{XYZ}(E) \leq \mathbb{E}_Z \left[\text{ess sup}_{\mathcal{P}_{Y|Z}} \mathcal{P}_{X|Z}(E_{Z,Y}) \right]^{\frac{\alpha-1}{\alpha}} \cdot \exp \left(\frac{\alpha-1}{\alpha} I_\alpha^{Y|Z}(X, Y|Z) \right). \quad (5)$$

Proof.

$$\mathcal{P}_{XYZ}(E) = \mathbb{E}_{\mathcal{P}_Z \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}} \left[\frac{d\mathcal{P}_{XYZ}}{d\mathcal{P}_Z \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}} \mathbb{1}_E \right] \quad (6)$$

$$\leq \mathbb{E}_{\mathcal{P}_Z}^{\frac{1}{\alpha''}} \left[\mathbb{E}_{\mathcal{P}_{Y|Z}}^{\frac{\alpha''}{\alpha'}} \left[\mathbb{E}_{\mathcal{P}_{X|Z}}^{\frac{\alpha'}{\alpha}} \left[\left(\frac{d\mathcal{P}_{XYZ}}{d\mathcal{P}_Z \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}} \right)^\alpha \right] \right] \right] \cdot \mathbb{E}_{\mathcal{P}_Z}^{\frac{1}{\gamma''}} \left[\mathbb{E}_{\mathcal{P}_{Y|Z}}^{\frac{\gamma''}{\gamma'}} \left[\mathbb{E}_{\mathcal{P}_{X|Z}}^{\frac{\gamma'}{\alpha}} [\mathbb{1}_E] \right] \right] \quad (7)$$

$$\leq \mathbb{E}_Z \left[\text{ess sup}_{\mathcal{P}_{Y|Z}} \mathcal{P}_{X|Z}(E_{Z,Y}) \right]^{\frac{\alpha-1}{\alpha}} \cdot \exp \left(\frac{\alpha-1}{\alpha} I_\alpha^{Y|Z}(X, Y|Z) \right). \quad (8)$$

The first inequality follows from applying Hölder's inequality three times and the six parameters are such that $\frac{1}{\alpha''} + \frac{1}{\gamma''} = \frac{1}{\alpha'} + \frac{1}{\gamma'} = \frac{1}{\alpha} + \frac{1}{\gamma} = 1$. (8) follows from setting $\alpha'' = \alpha$ and $\alpha' = 1$ which imply $\gamma'' = \gamma$ and $\gamma' \rightarrow \infty$. \square

Another property of I_α^Z is that, similarly to unconditional I_α [4], taking the limit of $\alpha \rightarrow \infty$, we have that $I_\alpha^{Y|Z}(X, Y|Z) \xrightarrow{\alpha \rightarrow \infty} \mathcal{L}(X \rightarrow Y|Z)$, leading us to the following:

¹It is clearly possible to minimise over more than one random variable at once, like it has been done in [5], [11] in the context of both regular $I_\alpha(X, Y)$ and conditional $I_\alpha(X, Y|Z)$.

Corollary 1. *Under the same assumptions of Theorem 1:*

$$\mathcal{P}_{XYZ}(E) \leq \mathbb{E}_Z[\text{ess sup}_{\mathcal{P}_{Y|Z}} \mathcal{P}_{X|Z}(E_{Z,Y})] \exp(\mathcal{L}(X \rightarrow Y|Z)). \quad (9)$$

C. $I_\alpha^Z(X, Y|Z)$

As discussed in Section III-A, another natural candidate definition of conditional α -mutual information is the following:

Definition 4. Under the same assumptions of Definition 3:

$$I_\alpha^Z(X, Y|Z) = \min_{\mathcal{Q}_Z} D_\alpha(\mathcal{P}_{XYZ} \| \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z} \mathcal{Q}_Z). \quad (10)$$

To the best of our knowledge Definition 4 has not been considered elsewhere. As for $I_\alpha^{Y|Z}(X, Y|Z)$, it is possible to compute a closed-form expression for $I_\alpha^Z(X, Y|Z)$. We will limit ourselves to discrete random variables for simplicity.

Theorem 2. Let $\alpha > 0$ and X, Y, Z be three discrete random variables.

$$I_\alpha^Z(X, Y|Z) = \frac{\alpha}{\alpha-1} \log \sum_z \mathcal{P}_Z(z) \cdot \left(\sum_{x,y} \mathcal{P}_{XY|Z=z}(x,y)^\alpha (\mathcal{P}_{X|Z=z}(x) \mathcal{P}_{Y|Z=z}(y))^{1-\alpha} \right)^{\frac{1}{\alpha}}.$$

The proof follows from the definition of $I_\alpha^Z(X, Y|Z)$ and Sibson's identity [9, Eq. (12)]. Mirroring Section III-B we can state an analogous of Theorem 1 for I_α^Z :

Theorem 3. Let $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{F}, \mathcal{P}_{XYZ})$ be a probability space. Let $\mathcal{P}_{Y|Z}$ and $\mathcal{P}_{X|Z}$ be the induced conditional distributions. Assume that $\mathcal{P}_{XYZ} \ll \mathcal{P}_Z \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}$. Given $E \in \mathcal{F}$ and $z \in \mathcal{Z}$, let $E_z = \{(x, y) : (x, y, z) \in E\}$. Then, fixed $\alpha \geq 1$:

$$\mathcal{P}_{XYZ}(E) \leq \text{ess sup}_{\mathcal{P}_Z} (\mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}(E_z))^{\frac{\alpha-1}{\alpha}} \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha^Z(X, Y|Z)\right). \quad (11)$$

This type of result is useful as it allows us to approximate the probability of E under a joint, with the probability of E under a different measure encoding some independence (typically easier to analyse) — in this specific case, the measure induced by a Markov chain. Such bounds represent, for us, the main application-oriented employment of these measures [1]. Notice that, other than using I_α^Z instead of $I_\alpha^{Y|Z}$, Theorem 3 involves a different essential supremum as compared to Theorem 1. Moving on with the comparison, we have that differently from Definition 3, the information measure we are defining here is symmetric. Moreover, setting Z to a constant in Definition 4 does not allow us to retrieve $I_\alpha(X, Y)$, but rather $D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y)$.

D. An additive SDPI-like inequality

Definition 4 shares some interesting properties with $I_\alpha(X, Y)$. One such property is a rewriting of $I_\alpha(X, Y)$ in terms of D_α . This allows us to leverage the strong data processing inequality (SDPI) for Hellinger integrals of order

α , which in turn allows us to provide an SDPI-like results for I_α^Z . A definition for SDPIs can be found at [12, Def 3.1]

More precisely, we can write

$$\begin{aligned} I_\alpha^Z(X, Y|Z) &= \frac{\alpha}{\alpha-1} \log \mathbb{E}_Z \left[\exp \left(\frac{\alpha-1}{\alpha} D_\alpha(\mathcal{P}_{XY|Z} \| \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}) \right) \right] \\ &= \frac{\alpha}{\alpha-1} \log \mathbb{E}_Z \left[(D_{f_\alpha}(\mathcal{P}_{XY|Z} \| \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}))^{1/\alpha} \right], \end{aligned} \quad (12)$$

where D_{f_α} denotes the Hellinger integral of order α , i.e., given two measures \mathcal{P}, \mathcal{Q} , $D_{f_\alpha}(\mathcal{P} \| \mathcal{Q}) = \mathbb{E}_{\mathcal{Q}} \left[\left(\frac{d\mathcal{P}}{d\mathcal{Q}} \right)^\alpha \right]$. Leveraging Eq. (12) we can state the following.

Theorem 4. Let $\alpha > 1$ and X, Y, W, Z be four random variables such that $(Z, W) - X - Y$ is a Markov chain:

$$I_\alpha^Z(W, Y|Z) \leq \frac{1}{\alpha-1} \log(\eta_{f_\alpha}(\mathcal{P}_{Y|X})) + I_\alpha^Z(W, X|Z), \quad (13)$$

where we denote with $\eta_{f_\alpha}(\mathcal{P}_{Y|X})$ the contraction parameter of the Hellinger integral of order α , i.e., for a given Markov Kernel K , $\eta_{f_\alpha}(K) = \sup_{\mu, \nu \neq \mu} \frac{D_{f_\alpha}(K\mu \| K\nu)}{D_{f_\alpha}(\mu \| \nu)}$ [12, Def. III.1].

The proof follows from Eq. (12) and a reasoning similar to [13, Lemma 3] but applied to the D_{f_α} -divergence instead of the KL-divergence.

Remark 2. Notice that data processing inequalities are simply a consequence of the convexity of f [14, Thm 4.2] and $f_\alpha(x) = x^\alpha$ is indeed convex. Hence, although the Hellinger integral is not normalised to be 0 whenever the measures are the same, it does satisfy a DPI. Moreover, the contraction parameter of a strong data-processing inequality is always less than or equal to 1. Hence, $\log(\eta_{f_\alpha}(K)) \leq 0$.

An analogous result of Theorem 4 for Definition 3 does not seem possible.

Remark 3. One can state a result similar to Theorem 4 for unconditional I_α . Specifically, we can write

$$I_\alpha(X, Y) = \frac{\alpha}{\alpha-1} \log \mathbb{E}_Y \left[D_{f_\alpha}^{1/\alpha}(\mathcal{P}_{X|Y} \| \mathcal{P}_X) \right].$$

Since I_α is an asymmetric quantity, we only get the SDPI-like result in one direction. Namely, given the Markov chain $W - X - Y$ we can relate via SDPI $I_\alpha(W, Y)$ and $I_\alpha(X, Y)$ (but, for instance, not $I_\alpha(W, X)$ and $I_\alpha(X, Y)$), as follows:

$$I_\alpha(W, Y) \leq \frac{1}{\alpha-1} \log(\eta_{f_\alpha}(\mathcal{P}_{W|X})) + I_\alpha(X, Y). \quad (14)$$

Theorem 4 and Eq. (14) represent a different from usual SDPI-like inequality. The reason for this is that the (function of the) η parameter is added to the information measure, rather than multiplied. However, one of the main applications of (conditional and not) I_α in bounds requires the exponentiation of the quantity, which brings us back to a multiplicative form. To make this statement more precise, let us state the following:

Corollary 2. *Under the same assumptions of Theorem 4 we have that:*

$$\mathcal{P}_{WYZ}(E) \leq \operatorname{ess\,sup}_{\mathcal{P}_Z} (\mathcal{P}_{W|Z} \mathcal{P}_{Y|Z}(E_Z))^{\frac{\alpha-1}{\alpha}} \cdot (\eta_{f_\alpha}(\mathcal{P}_{Y|X}))^{1/\alpha} \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha^Z(W, X|Z)\right).$$

Corollary 2 follows directly from Theorem 3 and Theorem 4.

Remark 4. A similar result can be derived for unconditional I_α starting from (14) and [1, Corollary 1].

E. Discussion on I_α^Z and $I_\alpha^{Y|Z}$

Let us now use Theorems 1 and 3 as a means of comparison for the two conditional I_α . These results are useful whenever we want to control the joint measure of some event E but we only know how to control it (e.g., via an upper-bound) under some hypothesis of independence [1]. Consider the factorisation of \mathcal{P}_{XYZ} under $X - Z - Y$ to be fixed. In the context of Theorem 1 and 3, according to the measure we know how to control, different conditional I_α 's will appear on the right-hand side of the bound (c.f., Eq. (5), (9) and (11)). For instance, if we assume to be able to control $\operatorname{ess\,sup}_{\mathcal{Q}_Z} (\mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}(E_Z))$ then, Theorem 3 tells us that $I_\alpha^Z(X, Y|Z)$ is the measure to study. If we assume instead that we are able to control terms of the form $\mathbb{E}_{P_Z}[\operatorname{ess\,sup}_{\mathcal{P}_{Y|Z}} \mathcal{P}_{X|Z}(E_{Z,Y})]$ then $I_\alpha^{Y|Z}(X, Y|Z)$ would be the measure to analyse. (Quantities like $\mathbb{E}_{P_Z}[\operatorname{ess\,sup}_{\mathcal{P}_{Y|Z}} \mathcal{P}_{X|Z}(E_{Z,Y})]$, for specific choices of E , are known in the literature as “small-ball probabilities” and have found applications in distributed estimation problems and distributed function computation [13], [15]). More generally, we can find a duality between the measure over which we supremise (on the right-hand side of the bounds) and the corresponding minimisation in the definition of conditional I_α . The same measures also have a fundamental role in defining the hypothesis testing problem that endows the information measure with its operational meaning, as we will see in the next section.

IV. OPERATIONAL MEANING

Drawing inspiration from [5], [11], [16], let us consider the following composite hypothesis testing problem. Fix a pmf \mathcal{P}_{XYZ} , observing a sequence of triples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ we want to decide whether:

- 0) $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ is sampled in an iid fashion from \mathcal{P}_{XYZ} (null hypothesis);
- 1) $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ is sampled in an iid fashion from $\mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}$, where \mathcal{Q}_Z is an arbitrary pmf over the space \mathcal{Z} (alternative hypothesis).

We can relate $I_\alpha^Z(X, Y|Z)$ to the error-exponent of the just defined hypothesis testing problem. This can be seen as a more lenient test for markovity where the measure of Z is allowed to vary. Similarly to before, there is a link between which measure is allowed to vary and the minimisation in the definition of conditional I_α . Choosing,

for instance, to minimise over \mathcal{Q}_X allows this measure to vary in the alternative hypothesis. Using Theorem 3 we can already connect I_α^Z to the problem in question. Given a test $T_n : \{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}\}^n \rightarrow \{0, 1\}$, we will denote with p_n^1 (Type-1 error) the probability of wrongfully choosing the hypothesis 1 given that the sequence is distributed according to $\mathcal{P}_{XYZ}^{\otimes n}$, i.e. $p_n^1 = \mathcal{P}_{XYZ}^{\otimes n}(T_n(\{(X_i, Y_i, Z_i)\}_{i=1}^n) = 1)$ and with p_n^2 (Type-2 error) the maximum probability of wrongfully choosing the hypothesis 0 given that the sequence is distributed according to $(\mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})^{\otimes n}$ for some \mathcal{Q}_Z , i.e. $p_n^2 = \sup_{\mathcal{Q}_Z \in \mathcal{P}(\mathcal{Z})} (\mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})^{\otimes n}(T_n(\{(X_i, Y_i, Z_i)\}_{i=1}^n) = 0)$.

Theorem 5. *Let $n > 0$ and $T_n : \{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}\}^n \rightarrow \{0, 1\}$ be a deterministic test, that upon observing the sequence $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ chooses either the null or the alternative hypothesis. Assume that $\exists R > 0 : \forall \mathcal{Q}_Z \in \mathcal{Q}(\mathcal{Z})$ we have $(\mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})^{\otimes n}(T_n(\{(X_i, Y_i, Z_i)\}_{i=1}^n) = 0) \leq \exp(-nR)$. Let also $\alpha \geq 1$,*

$$1 - p_n^1 \leq \exp\left(-\frac{\alpha-1}{\alpha} n(R - I_\alpha^Z(X, Y|Z))\right). \quad (15)$$

Proof. We have that $1 - p_n^1 = \mathcal{P}_{XYZ}^{\otimes n}(T_n(\{(X_i, Y_i, Z_i)\}_{i=1}^n) = 0)$. Starting from Theorem 3:

$$1 - p_n^1 \leq \operatorname{ess\,sup}_{\mathcal{P}_Z^n} \left(\mathcal{P}_{X|Z}^n \mathcal{P}_{Y|Z}^n(E_Z^n) \right)^{1/\gamma} \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha^Z(X^n, Y^n|Z^n)\right). \quad (16)$$

Since we assumed the exponential decay of $(\mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})^{\otimes n}(T_n(\{(X_i, Y_i, Z_i)\}_{i=1}^n) = 0)$ for every \mathcal{Q}_Z we also have that $\operatorname{ess\,sup}_{\mathcal{P}_Z^n} (\mathcal{P}_{X|Z}^n \mathcal{P}_{Y|Z}^n(E_Z^n)) \leq \exp(-nR)$ (consider a measure $\tilde{\mathcal{Q}}_Z$ that puts all the mass on the sequence achieving the essential supremum in (16)). Given the assumption of independence on the triples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ and following a reasoning similar to the one in Eqn. (49) in [7], we have that $I_\alpha^Z(X^n, Y^n|Z^n) = n I_\alpha^Z(X, Y|Z)$. The conclusions then follow from algebraic manipulations of (16). \square

This result implies that if we assume an exponential decay for the type-2 error p_n^2 and $R > I_\alpha^Z(X, Y|Z)$ we have an exponential decay of the probability of correctly choosing the null hypothesis as well. Moreover, for every $n > 0$:

$$\frac{1}{n} \log(1 - p_n^1) \leq -\frac{\alpha-1}{\alpha} (R - I_\alpha^Z(X, Y|Z)). \quad (17)$$

We can conclude that:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log(1 - p_n^1) \leq - \sup_{\alpha \in (1, +\infty]} \frac{\alpha-1}{\alpha} (R - I_\alpha^Z(X, Y|Z)).$$

A. Error exponents

Following the approach undertaken in [11] we can also define an achievable error-exponent pair for the hypothesis testing problem in question.

Definition 5. A pair of error exponents $(E_P, E_Q) \in \mathbb{R}^2$ is called achievable w.r.t the above hypothesis testing problem if there exists a series of tests $\{T_n\}_{n=1}^\infty$ such that ²:

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log \mathcal{P}_{XYZ}^{\otimes n}(T_n(\{(X_i, Y_i, Z_i)\}_{i=1}^n) = 1) > E_P,$$

$$\liminf_{n \rightarrow \infty} \inf_{Q_Z} -\frac{1}{n} \log(\mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})^{\otimes n}(T_n(\{(X_i, Y_i, Z_i)\}_{i=1}^n) = 0) > E_Q.$$

We can then define the error exponent functions [11] $E_P : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ and $E_Q : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ as follows:

$$E_P(E_Q) = \sup\{E_P \in \mathbb{R} : (E_P, E_Q) \text{ is achievable}\} \quad (18)$$

$$E_Q(E_P) = \sup\{E_Q \in \mathbb{R} : (E_P, E_Q) \text{ is achievable}\} \quad (19)$$

It is now possible to relate $I_\alpha^Z(X, Y|Z)$, where $\alpha \in (0, 1]$, with both the Fenchel conjugate of $E_P(\cdot)$, $E_P^*(\cdot)$ and $E_P^{**}(\cdot)$. First, let us characterise $E_P^*(E_Q)$.

Lemma 1.

$$E_P^*(\lambda) = \begin{cases} +\infty, & \text{if } \lambda > 0 \\ \lambda I_{\frac{1}{1-\lambda}}(X, Y|Z), & \text{otherwise.} \end{cases} \quad (20)$$

Proof. Assume $\lambda \leq 0$,

$$\begin{aligned} E_P^*(\lambda) &= \sup_{E_Q \in \mathbb{R}} [\lambda E_Q - E_P(E_Q)] \\ &= \sup_{E_Q \in \mathbb{R}} \left[\lambda E_Q - \inf_{\substack{\mathcal{R}_{XYZ}: \\ D(\mathcal{R}_{XYZ} \| \mathcal{R}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}) \leq E_Q}} D(\mathcal{R}_{XYZ} \| \mathcal{P}_{XYZ}) \right] \\ &\stackrel{(a)}{=} \sup_{E_Q \in \mathbb{R}} \sup_{\substack{\mathcal{R}_{XYZ}: \\ D(\mathcal{R}_{XYZ} \| \mathcal{R}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}) \leq E_Q}} [\lambda E_Q - D(\mathcal{R}_{XYZ} \| \mathcal{P}_{XYZ})] \\ &= \sup_{\mathcal{R}_{XYZ}} \sup_{\substack{E_Q \in \mathbb{R}: \\ E_Q \geq D(\mathcal{R}_{XYZ} \| \mathcal{R}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})}} [\lambda E_Q - D(\mathcal{R}_{XYZ} \| \mathcal{P}_{XYZ})] \\ &\stackrel{(b)}{=} \sup_{\mathcal{R}_{XYZ}} [\lambda D(\mathcal{R}_{XYZ} \| \mathcal{R}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}) - D(\mathcal{R}_{XYZ} \| \mathcal{P}_{XYZ})] \\ &\stackrel{(c)}{=} (\lambda - 1) \inf_{Q_Z} \inf_{\mathcal{R}_{XYZ}} \left[\frac{-\lambda}{1-\lambda} D(\mathcal{R}_{XYZ} \| \mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}) \right. \\ &\quad \left. + \frac{1}{1-\lambda} D(\mathcal{R}_{XYZ} \| \mathcal{P}_{XYZ}) \right] \\ &\stackrel{(d)}{=} \lambda \inf_{Q_Z} D_{\frac{1}{1-\lambda}}(\mathcal{P}_{XYZ} \| \mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}) \\ &\stackrel{(e)}{=} \lambda I_{\frac{1}{1-\lambda}}(X, Y|Z). \end{aligned}$$

Where step (a) follows from an analogous result of [11, Corollary 2] for our testing problem, step (b) follows because, given that $\lambda \leq 0$ then $D(\mathcal{R}_{XYZ} \| \mathcal{Q}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z})$ achieves the maximum. Step (c) follows from an analogous of [11, Lemma 4], (d) follows from [8, Theorem 3] and to conclude (e) follows from Definition 4. For $\lambda > 0$ the reasoning is identical to [11, Lemma 12]. \square

²As pointed out in [11], despite having bounds like in Theorem 5 decaying with two rates E_P, E_Q , we cannot conclude anything on the achievability of the pair.

Now, we can prove the connection to $E_P^{**}(\cdot)$.³

Theorem 6. Given $E_Q, E_P \in \mathbb{R}$

$$E_P^{**}(E_Q) = \sup_{\alpha \in (0,1]} \frac{1-\alpha}{\alpha} (I_\alpha(X, Y|Z) - E_Q), \quad (21)$$

$$E_Q^{**}(E_P) = \sup_{\alpha \in (0,1]} \left(I_\alpha(X, Y|Z) - \frac{\alpha}{1-\alpha} E_P \right). \quad (22)$$

Proof.

$$E_P^{**}(E_Q) = \sup_{\lambda \in \mathbb{R}} (\lambda E_Q - E_P^*(\lambda)) \quad (23)$$

$$\stackrel{(f)}{=} \sup_{\lambda \leq 0} (\lambda E_Q - E_P^*(\lambda)) \quad (24)$$

$$\stackrel{(g)}{=} \sup_{\lambda \leq 0} (\lambda E_Q - I_{\frac{1}{1-\lambda}}(X, Y|Z)) \quad (25)$$

$$\stackrel{(h)}{=} \sup_{\alpha \in (0,1]} \frac{1-\alpha}{\alpha} (I_\alpha(X, Y|Z) - E_Q). \quad (26)$$

Where (f) follows from $E_P^*(\lambda) = +\infty$ for $\lambda > 0$, (g) follows from Lemma 1 and (h) by setting $\alpha = \frac{1}{1-\lambda}$. The proof of (22) follows from similar arguments. \square

V. CONCLUSIONS

We have considered the problem of defining a conditional version of Sibson's α -Mutual Information. Drawing inspiration from an equivalent formulation of $I_\alpha(X, Y)$ as $\min_{Q_Y} D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X Q_Y)$ we saw how several of these propositions can be made for a $I_\alpha(X, Y|Z)$. Two have already been analysed in [5]. We proposed here a general approach that allows to connect to each such measure:

- 1) a bound, allowing to approximate the probability $\mathcal{P}_{XYZ}(E)$ with the probability of E under a product distribution induced by the Markov chain $X - Z - Y$;
- 2) an operational meaning as the error exponent of a hypothesis testing problem where the alternative hypothesis is a markov-like distribution and some measures are allowed to vary.

A simple relationship between the hypothesis testing problem and the information measure can already be found using the bound described in 1), without requiring any extra machinery. To conclude, the usefulness of a measure clearly comes from its applications and ease of computability. While the latter remains the same for all the possible conditional I_α the former can vary according to the definition. With this in mind, the various definitions are equally meaningful and it seems reasonable to use the conditional I_α that best suits the specific application at hand.

ACKNOWLEDGMENT

The work in this paper was supported in part by the Swiss National Science Foundation under Grants 169294 and 200364.

³Notice that $E_P^{**}(\cdot)$ is not guaranteed to be equal to $E_P(\cdot)$. Indeed, it is possible to find examples where the function is not convex and thus, all we retrieve is a lower bound on E_P [11, Example 14].

REFERENCES

- [1] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via rényi-, f-divergences and maximal leakage," *Accepted for Publication in IEEE Transactions on Information Theory*, 2021. [Online]. Available: <http://arxiv.org/abs/1912.01439>
- [2] R. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Transactions on Information Theory*, vol. 11, no. 1, pp. 3–18, 1965.
- [3] R. G. Gallager, *Information Theory and Reliable Communication*. USA: John Wiley & Sons, Inc., 1968.
- [4] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, 2020.
- [5] M. Tomamichel and M. Hayashi, "Operational interpretation of Rényi information measures via composite hypothesis testing against product and markov distributions," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1064–1082, 2018.
- [6] J. Liao, L. Sankar, O. Kosut, and F. P. Calmon, "Robustness of maximal α -leakage to side information," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 642–646.
- [7] S. Verdú, " α -mutual information," in *2015 Information Theory and Applications Workshop, ITA 2015, San Diego, CA, USA, February 1-6, 2015*, 2015, pp. 1–6.
- [8] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.
- [9] I. Csiszar, "Generalized cutoff rates and Rényi's information measures," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 26–34, Jan 1995.
- [10] R. Sibson, "Information radius," *Z. Wahrscheinlichkeitstheorie verw Gebiete* 14, pp. 149–160, 1969.
- [11] A. Lapidath and C. Pfister, "Testing against independence and a Rényi information measure," in *2018 IEEE Information Theory Workshop (ITW)*, 2018, pp. 1–5.
- [12] M. Raginsky, "Strong data processing inequalities and ϕ -sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.
- [13] A. Xu and M. Raginsky, "Information-theoretic lower bounds for distributed function computation," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2314–2337, 2017.
- [14] Y. Wu, "Lecture notes on: Information-theoretic methods for high-dimensional statistics," 2020.
- [15] A. Xu and M. Raginsky, "Information-theoretic lower bounds on bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, 2017.
- [16] A. Lapidath and C. Pfister, "Two measures of dependence," *Entropy*, vol. 21, no. 8, 2019. [Online]. Available: <https://www.mdpi.com/1099-4300/21/8/778>