On Sibson's α-Mutual Information

Amedeo Roberto Esposito, Adrien Vandenbroucque, Michael Gastpar
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{amedeo.esposito, michael.gastpar}@epfl.ch, adrien.vandenbroucque@alumni.epfl.ch

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834428

***Abstract*—We explore a family of information measures that stems from Rényi's α-Divergences with $\alpha < 0$. In particular, we extend the definition of Sibson's α-Mutual Information to negative values of α and show several properties of these objects. Moreover, we highlight how this family of information measures is related to functional inequalities that can be employed in a variety of fields, including lower-bounds on the Risk in Bayesian Estimation Procedures.**

***Index Terms*—Rényi-Divergence, Sibson's Mutual Information, Information Measures, Bayesian Risk, Estimation**

I. Introduction

Sibson's α-Mutual Information is a generalisation 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. In this work we will extend the definition of I_α in order to include negative values of α. The reason for this extension is tied to a family of functional-analytic inequalities that one can provide for Sibson's α MI for every $\alpha \in \mathbb{R}$ (cf. Section V, Table I).

II. Background and definitions

Throughout the paper we will often use the notion of L^p-norms: let $1 \leq p \leq \infty$ and consider the measurable space $(\Omega, \mathcal{F}, \mu)$, let f be a measurable function with respect to the space, then

$$\|f\|_{L^p(\mu)} = \left(\int |f|^p d\mu \right)^{\frac{1}{p}}. \tag{1}$$

The definition can be extended to values of $p < 1$, however in those cases one does not have norms anymore e.g., if $0 < p < 1$, one recovers a quasi-norm but not a norm.

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 [5]–[7]). It can be defined as follows [6]:

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:

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

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 [6].

It can be shown that if $\alpha > 1$ and $\mathcal{P} \not\ll \mathcal{Q}$ then $D_\alpha(\mathcal{P}\|\mathcal{Q}) = \infty$. The behaviour of the measure for $\alpha \in \{0, 1, \infty\}$ can be defined by continuity. In general, one has that $D_1(\mathcal{P}\|\mathcal{Q}) = D(\mathcal{P}\|\mathcal{Q})$ but if $D(\mathcal{P}\|\mathcal{Q}) = \infty$ or there exists β such that $D_\beta(\mathcal{P}\|\mathcal{Q}) < \infty$ then $\lim_{\alpha \downarrow 1} D_\alpha(\mathcal{P}\|Q) = D(\mathcal{P}\|\mathcal{Q})$ [6, Theorem 5]. For an extensive treatment of α-divergences and their properties we refer the reader to [6]. Starting from Rényi's Divergence and the geometric averaging that it involves, Sibson built the notion of Information Radius [8] which can be seen as a special case of the following quantity [5]:

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

$I_\alpha(X, Y)$ represents a generalisation of Shannon's Mutual Information and possesses many interesting properties [5]. Indeed, $\lim_{\alpha \to 1} I_\alpha(X, Y) = I(X; Y)$. On the other hand when $\alpha \to \infty$, we get: $I_\infty(X, Y) = \log \mathbb{E}_{\mathcal{P}_Y} \left[\sup_{x: \mathcal{P}_X(x) > 0} \frac{\mathcal{P}_{XY}(\{x, Y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{Y\})} \right] = \mathcal{L}(X \to Y)$, where $\mathcal{L}(X \to Y)$ 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 [5], as for Maximal Leakage the reader is referred to [4].

B. Rényi's α-Divergence - Negative Orders

Not much is known or has been explored on Rényi's α-Divergence with $\alpha < 0$. According to Rényi himself [9], only positive orders can be regarded as information measures. One of the reasons behind this statement is probably the fact that, taking an axiomatic approach to measures of information like

the one undertaken in [9], D_α with $\alpha < 0$ satisfies many of the required properties with the opposite sign of inequality. In fact [6]:

Proposition 1. *Let $\alpha < 0$ and μ, ν be two probability measures such that $\nu \ll \mu$ then:*

1) $D_\alpha(\nu\|\mu) \leq 0$ *(as opposed to ≥ 0);*
2) $D_\alpha(\nu\|\mu)$ *is* ***concave*** *in ν (as opposed to* ***convex*** *in μ);*
3) D_α *is* ***upper*** *semi-continuous in the pair (ν, μ) in the topology of set-wise convergence (as opposed to* ***lower*** *semi-continuous) ;*
4) *Let K be a Markov Kernel, then $D_\alpha(K\nu\|K\mu) \geq D_\alpha(\nu\|\mu)$ (as opposed to $D_\alpha(K\nu\|K\mu) \leq D_\alpha(\nu\|\mu)$);*

Remark 2. We decided to leave the definition of D_α for $\alpha < 0$ unaltered with respect to the positive orders. We thus followed the approach undertaken in [6, Section V] but adapted the definition of I_α for $\alpha < 0$ (cf. Definition 2). An alternative approach would be to change the definition of D_α for negative orders, for instance multiplying D_α by $\operatorname{sign}(\alpha)$, but leave the definition of I_α unaltered. The minus in the definition (for negative α's) would then ensure that all the properties of the information measure are satisfied with the "right" sign of inequality (inverted with respect to Proposition 1). A similar approach can also be undertaken in order to define Rényi's entropy H_α with negative α by simply changing the multiplicative constant from $\frac{1}{1-\alpha}$ to $\frac{1}{\alpha-1}$ for negative orders.

Some of the properties are maintained, like the following:

Lemma 1 ([6, Thm 39])**.** *Let $\alpha \in [-\infty, +\infty]$ and let ν, μ be two probability measures such that $\nu \ll \mu$, then $D_\alpha(\nu\|\mu)$ is non-decreasing in α.*

Moreover, negative orders can be connected to positive orders through the following result:

Lemma 2 ([6, Lemma 10] - Skew Symmetry)**.** *For every $\alpha \in (-\infty, +\infty) \setminus \{0, 1\}$, let let ν, μ be two probability measures such that $\nu \equiv \mu$*

$$D_\alpha(\nu\|\mu) = \frac{\alpha}{1-\alpha} D_{(1-\alpha)}(\mu\|\nu) \tag{4}$$

This relationship, although useful, is restricted to cases where the two measures are equivalent with respect to each other (*i.e.*, absolute continuity holds in both directions: $\nu \ll \mu$ **and** $\mu \ll \nu$). Otherwise, one might incur in settings where one divergence is finite and the other is infinite. An important difference lies in the fact that D_α with $\alpha < 0$ is not concave in the second argument.

Counterexample 1. Consider a discrete setting and point mass functions. Let $\mu_1 = (0.32, 0.68), \mu_2 = (0.5, 0.5), \nu = (0.13, 0.87)$ and $\lambda = 0.4$. One has that with $\alpha = -2$, $D_\alpha(\nu\|\mu_1) = -0.2855$, $D_\alpha(\nu\|\mu_2) = -0.6744$ and, moreover, $-0.5287 = D_\alpha(\nu\|\lambda\mu_1 + (1-\lambda)\mu_2) < \lambda D_\alpha(\nu\|\mu_1) + (1-\lambda) D_\alpha(\nu\|\mu_2) = -0.5188$.

III. Definition

The starting point will be Equation (3). From now on we will consider α to be always strictly negative unless otherwise specified and we will restrict ourselves to discrete probability measures for simplicity. We thus bring forth the following definition:

Definition 2. Let X, Y be two random variables whose joint measure is $\mathcal{P}_{XY}$ and the corresponding marginals are given by $\mathcal{P}_X$ and $\mathcal{P}_Y$. Let $\alpha \in (-\infty, 0)$

$$I_\alpha(X, Y) = -\max_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y). \tag{5}$$

This quantity has, much like Sibson's α-MI with $\alpha > 0$, a closed-form expression given by the following result.

Theorem 1. *Let X, Y be two random variables whose joint measure is $\mathcal{P}_{XY}$ and the corresponding marginals are given by $\mathcal{P}_X$ and $\mathcal{P}_Y$. Let $\alpha < 0$*

$$I_\alpha(X, Y) = -\frac{\alpha}{\alpha - 1} \log \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{\frac{1}{\alpha}}_{\mathcal{P}_X}\left[\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^\alpha\right]\right]. \tag{6}$$

Proof. We have that for every $\mathcal{Q}_Y$,

$$D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\mathcal{Q}_Y}\left[\mathbb{E}_{\mathcal{P}_X}\left[\mathcal{P}^\alpha_{Y|X}\mathcal{Q}^{-\alpha}_Y\right]\right] \tag{7}$$

$$= \frac{\alpha}{\alpha - 1} \log \mathbb{E}^{\frac{1}{\alpha}}_{\mathcal{Q}_Y}\left[\frac{\mathbb{E}_{\mathcal{P}_X}\left[\mathcal{P}^\alpha_{Y|X}\right]}{Q^\alpha_Y}\right] \tag{8}$$

$$\leq \frac{\alpha}{\alpha - 1} \log \mathbb{E}_{\mathcal{Q}_Y}\left[\frac{\mathbb{E}^{\frac{1}{\alpha}}_{\mathcal{P}_X}\left[\mathcal{P}^\alpha_{Y|X}\right]}{Q_Y}\right] \tag{9}$$

$$= \frac{\alpha}{\alpha - 1} \log \sum_y \mathbb{E}^{\frac{1}{\alpha}}_{\mathcal{P}_X}\left[\mathcal{P}^\alpha_{Y|X}\right] \tag{10}$$

$$= \frac{\alpha}{\alpha - 1} \log \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{\frac{1}{\alpha}}_{\mathcal{P}_X}\left[\left(\frac{\mathcal{P}_{X|Y}}{\mathcal{P}_X}\right)^\alpha\right]\right] \tag{11}$$

$$= -I_\alpha(X, Y). \tag{12}$$

Where (9) follows from the convexity of $x^{\frac{1}{k}}, k < 0$ and Jensen's inequality. This implies that for every $\mathcal{Q}_Y$:

$$-I_\alpha(X, Y) \geq D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y). \tag{13}$$

Moreover one has that, given

$$\mathcal{Q}^\star_Y(y) = \frac{\mathcal{P}_Y(y)\left(\sum_x \mathcal{P}_{X|Y=y}(x)^\alpha \mathcal{P}_X(x)^{1-\alpha}\right)^{\frac{1}{\alpha}}}{\mathbb{E}_{\mathcal{P}_Y}\left[\left(\sum_x \mathcal{P}_{X|Y=y}(x)^\alpha \mathcal{P}_X(x)^{1-\alpha}\right)^{\frac{1}{\alpha}}\right]} \tag{14}$$

then,

$$D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}^\star_Y) = -I_\alpha(X, Y). \tag{15}$$

□

Remark 3. In a setting where both X and Y are discrete random variables then denoting with p_{XY}, p_Y and p_X the joint point mass functions and the corresponding marginals and

denoting with $p_{X|Y=y}$ the conditional point mass functions one has that:

$$I_\alpha(X,Y) = \frac{\alpha}{1-\alpha}\log\sum_y p_Y(y)\left(\sum_x\left(\frac{p_{X|Y=y}(x)}{p_X(x)}\right)^\alpha p_X(x)\right)^{\frac{1}{\alpha}}$$
$$= \frac{\alpha}{1-\alpha}\log\sum_y\left(\sum_x p_{Y|X=x}(y)^\alpha p_X(x)\right)^{\frac{1}{\alpha}}. \quad (16)$$

Theorem 2. *Let X,Y be two discrete random variables whose joint point mass function is p_{XY} and the corresponding marginals are given by p_X and p_Y then*

$$I_{-\infty}(X,Y) = -\log\sum_y\left(\min_{x:p_X(x)>0}\mathcal{P}_{Y|X=x}(y)\right). \quad (17)$$

Proof. The result follows from noticing that one can re-write I_α as follows

$$-I_\alpha(X,Y) = \frac{\alpha}{\alpha-1}\log\mathbb{E}_{\mathcal{P}_Y}\left[\left\|\frac{\mathcal{P}_{X|Y}}{\mathcal{P}_X}\right\|_{L^\alpha(\mathcal{P}_X)}\right], \quad (18)$$

where $\left\|\frac{\mathcal{P}_{X|Y}}{\mathcal{P}_X}\right\|_{L^\alpha(\mathcal{P}_X)}$ denotes the "$\alpha-$norm" with respect to $\mathcal{P}_X$. Taking the limit of $\alpha\to-\infty$ one has that $\frac{\alpha}{\alpha-1}\to 1$ and $\left\|\frac{\mathcal{P}_{X|Y}}{\mathcal{P}_X}\right\|_{L^\alpha(\mathcal{P}_X)} \to \operatorname{ess\,inf}_{\mathcal{P}_X}\frac{\mathcal{P}_{X|Y}}{\mathcal{P}_X}$. The conclusion then follows from simple algebraic manipulations. □

Remark 4. The quantity $I_{-\infty}$ has already appeared in the literature as "Maximal-Cost Leakage" [4, Thm. 15, Eq. (95)]. Indeed, assuming equivalence between $\mathcal{P}_{XY}$ and $\mathcal{P}_X\mathcal{Q}_Y$ and using the Skew Symmetry of D_α one has that $-D_{-\infty}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y) = D_\infty(\mathcal{P}_X\mathcal{Q}_Y\|\mathcal{P}_{XY})$ and consequently

$$I_{-\infty}(X,Y) = -\max_{\mathcal{Q}_Y} D_{-\infty}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y) \quad (19)$$
$$= \min_{\mathcal{Q}_Y} D_\infty(\mathcal{P}_X\mathcal{Q}_Y\|\mathcal{P}_{XY}) \quad (20)$$
$$= \mathcal{L}^c(X\to Y). \quad (21)$$

IV. Properties

For I_α with $\alpha<0$ we can show the following properties (similar to I_α with $\alpha>0$ [5], [10]):

Theorem 3. *Let X,Y be two random variables such that $\mathcal{P}_{XY}\ll\mathcal{P}_X\mathcal{P}_Y$ and assume that $\alpha<0$, then:*

1) *$I_\alpha(X,Y)\geq 0$ with equality iff X and Y are independent;*
2) *$I_\alpha(X,Y)\neq I_\alpha(Y,X)$;*
3) *Let $0>\alpha_1>\alpha_2$ then $I_{\alpha_1}(X,Y)\leq I_{\alpha_2}(X,Y)$;*
4) *$I_\alpha(X,Y)\leq\mathcal{L}^c(X\to Y)$ for every $-\infty<\alpha<0$;*
5) *Let $X-Y-Z$ be a Markov Chain, $I_\alpha(X,Z)\leq\min\{I_\alpha(X,Y),I_\alpha(Y,Z)\}$;*
6) *$\exp\left(-\frac{\alpha-1}{\alpha}I_\alpha(X,Y)\right)$ is concave in $\mathcal{P}_{Y|X}$ and $I_\alpha(X,Y)$ is convex in $\mathcal{P}_{Y|X}$.*

Proof. 1): We have that $I_\alpha(X,Y) = -D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y^\star)$ and given that $D_\alpha\leq 0$ if $\alpha\leq 0$ we have the non-negativity of I_α. Moreover, if X and Y are independent then $\mathcal{Q}_Y^\star(y)=\mathcal{P}_Y(y)$ and $D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)=0$. On the other hand, if $I_\alpha(X,Y)=0$ then $D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y^\star)=0$ which means that $\mathcal{P}_{Y|X=x}=\mathcal{Q}_Y^\star$, hence Y does not depend on X.
2): It's clear from the definition and shown more concretely in Example 2.
3): We have that, denoting with $\mathcal{Q}_Y^{\alpha_2} = \arg\max_{\mathcal{Q}_Y} D_{\alpha_2}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y)$

$$-I_{\alpha_1}(X,Y) = \max_{\mathcal{Q}_Y} D_{\alpha_1}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y) \quad (22)$$
$$\geq D_{\alpha_1}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y^{\alpha_2}) \quad (23)$$
$$\geq D_{\alpha_2}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y^{\alpha_2}) \quad (24)$$
$$= -I_{\alpha_2}(X,Y) \quad (25)$$

where (24) follows from the non-decreasability of D_α for $\alpha\in[-\infty,\infty]$ [6, Thm. 39].
4): Follows from 3) and the fact that $\mathcal{L}^c(X\to Y) = \lim_{\alpha\to-\infty} I_\alpha(X,Y)$.
5): let $X-Y-Z$ be a Markov chain and let $K_{Z|Y}$ denote the corresponding Markov Kernel $K_{Z|Y}:(\mathcal{Y},\mathcal{F})\to(\mathcal{Z},\hat{\mathcal{F}})$ induced by the transition probabilities $\mathcal{P}_{Z|Y}$. Consequently, given any probability measure $\mathcal{Q}_Y$, one can construct a corresponding $\mathcal{Q}_Z=\mathcal{Q}_YK_{Z|Y}$ where $\mathcal{Q}_Z(z)=\mathbb{E}_{\mathcal{Q}_Y}[\mathcal{P}_{Z|Y}(z)]$. By the Data-Processing Inequality for D_α one has that for every $\mathcal{Q}_Y$:

$$-D_\alpha(\mathcal{P}_{XZ}\|\mathcal{P}_X\mathcal{Q}_Z)\leq -D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y) \quad (26)$$

Taking the $\inf$ with respect to $\mathcal{Q}_Y$ leads to

$$-D_\alpha(\mathcal{P}_{XZ}\|\mathcal{P}_X\mathcal{Q}_Z^\star)\leq I_\alpha(X,Y) \quad (27)$$

and given that $I_\alpha(X,Z)\leq -D_\alpha(\mathcal{P}_{XZ}\|\mathcal{P}_X\mathcal{Q}_Z)$ for every $\mathcal{Q}_Z$ the statement follows. The other inequality follows by observing that considering the Kernel determined by $\mathcal{P}_{XZ|XY}$ and denoted by $K_{XZ|YZ}$ one has that $\mathcal{P}_{XZ}=\mathcal{P}_{YZ}K_{XZ|YZ}$ while $\mathcal{P}_X\mathcal{Q}_Z=(\mathcal{P}_X\mathcal{Q}_Y)K_{XZ|YZ}$, consequently by the Data-Processing Inequality for D_α one has that

$$-D_\alpha(\mathcal{P}_{XZ}\|\mathcal{P}_X\mathcal{Q}_Z)\leq -D_\alpha(\mathcal{P}_{YZ}\|\mathcal{P}_Y\mathcal{Q}_Z). \quad (28)$$

The statement follows from a similar argument as above.
6): One can rewrite the expression as follows

$$\exp\left(-\frac{\alpha-1}{\alpha}I_\alpha(X,Y)\right) = \sum_y\left\|\mathcal{P}_{Y|X}\right\|_{L^\alpha(\mathcal{P}_X)}. \quad (29)$$

Similarly to [10, Thm. 11], the concavity follows from the Reverse-Minkowski's inequality and convexity of $I_\alpha(X,Y)$ follows from the concavity just proven and the fact that $-\frac{\alpha}{\alpha-1}\log(x)$ is a non-increasing convex function for a given α (composition of a non-increasing convex function with a concave one gives rise to a convex function [11, Eq (3.10), pag. 84]). □

Following [5] we will now look at some specific choices of X and Y and compute the corresponding values of I_α.

Example 1. Let $X,Y\sim\text{Ber}(1/2)$ and let $\mathbb{P}(Y\neq X)=\delta$ then if $\alpha<0$

$$I_\alpha(X,Y) = I_\alpha(Y,X) = -d_\alpha(\delta\|1/2), \quad (30)$$

where $d_\alpha(p\|q) = \frac{1}{\alpha-1}\log(p^\alpha q^{1-\alpha} + (1-p)^\alpha(1-q)^{1-\alpha})$ denotes the binary α-divergence.

Example 2. Let $X \sim \text{Ber}(1/2)$ and let $Y \in \{0,1,e\}$. Assume also that

$$P_{Y|X=x}(y) = \begin{cases} 1-\delta, & x = y \\ \delta, & y = e \\ 0, & \text{else.} \end{cases} \tag{31}$$

Then, if $\alpha < 0$

$$I_\alpha(X,Y) = -\frac{\alpha}{\alpha-1}\log_2\left(\delta + (1-\delta)2^{\frac{\alpha-1}{\alpha}}\right) \tag{32}$$

$$I_\alpha(Y,X) = -\frac{1}{\alpha-1}\log_2\left(\delta + (1-\delta)2^{\alpha-1}\right) \tag{33}$$

Example 3. Let $X \sim \mathcal{N}(0,\sigma_X^2)$ and $N \sim \mathcal{N}(0,\sigma_N^2)$ with N independent from X then if $-\frac{\sigma_N^2}{\sigma_X^2} < \alpha < 0$

$$I_\alpha(X, X+N) = -\frac{1}{2}\log\left(1 + \alpha\frac{\sigma_X^2}{\sigma_N^2}\right). \tag{34}$$

One can see from the computations that for $\alpha \leq -\frac{\sigma_N^2}{\sigma_X^2}$ the integral does not converge. Notice that using the skew symmetry (Cf. Lemma 1) does not necessarily allow us to compute I_α for negative α's using positive values of α. Negative values are related to the positive ones, but the order of the measures is inverted *i.e.*, $I_\alpha(X,Y) = -D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y^\star)$ would be related to $D_{(1-\alpha)}(\mathcal{P}_X\mathcal{Q}_Y^\star\|\mathcal{P}_{XY})$ rather than $I_{(1-\alpha)}(X,Y)$.

V. Functional Inequalities

Some interesting results can be provided applying reverse Hölder's inequality which are similar to the ones one can provide for I_α with $\alpha > 0$. These results have found applications in Learning Theory, Estimation Theory in a Bayesian setting and Hypothesis Testing [1], [12], [13].

Theorem 4. *Let X, Y be two random variables whose joint measure is $\mathcal{P}_{XY}$ and the corresponding marginals are given by $\mathcal{P}_X$ and $\mathcal{P}_Y$. For every $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+$ one has that*

$$\mathbb{E}_{\mathcal{P}_{XY}}[f(X,Y)] \geq \mathbb{E}_{\mathcal{P}_Y}^{\frac{1}{\beta'}}\left[\mathbb{E}_{\mathcal{P}_X}^{\frac{\beta'}{\beta}}\left[f(X,Y)^\beta\right]\right] \tag{35}$$

$$\cdot\, \mathbb{E}_{\mathcal{P}_Y}^{\frac{1}{\alpha'}}\left[\mathbb{E}_{\mathcal{P}_X}^{\frac{\alpha'}{\alpha}}\left[\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^\alpha\right]\right], \tag{36}$$

where $\frac{1}{\alpha}+\frac{1}{\beta} = 1 = \frac{1}{\alpha'}+\frac{1}{\beta'}$ and $\alpha, \alpha' < 1$.

Proof.

$$\mathbb{E}_{\mathcal{P}_{XY}}[f(X,Y)] = \mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[f(X,Y)\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)\right] \tag{37}$$

$$\geq \mathbb{E}_{\mathcal{P}_Y}\Bigg[\mathbb{E}_{\mathcal{P}_X}^{\frac{1}{\beta}}\left[f(X,Y)^\beta\right] \tag{38}$$

$$\cdot\, \mathbb{E}_{\mathcal{P}_X}^{\frac{1}{\alpha}}\left[\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^\alpha\right]\Bigg] \tag{39}$$

$$\geq \mathbb{E}_{\mathcal{P}_Y}^{\frac{1}{\beta'}}\left[\mathbb{E}_{\mathcal{P}_X}^{\frac{\beta'}{\beta}}\left[f(X,Y)^\beta\right]\right] \tag{40}$$

$$\cdot\, \mathbb{E}_{\mathcal{P}_Y}^{\frac{1}{\alpha'}}\left[\mathbb{E}_{\mathcal{P}_X}^{\frac{\alpha'}{\alpha}}\left[\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^\alpha\right]\right], \tag{41}$$

where each inequality follows from applying the reverse Hölder's inequality (which, in turn, requires the positivity of f). □

Taking the limit of $\alpha' \to 1$ (which implies $\beta' \to -\infty$) leads to the following bound involving $I_\alpha(X,Y)$ with $\alpha < 1$:

Corollary 1. *Consider the same setting as in Theorem 4. For every $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+$ and for every $\alpha < 1$ and $\beta = \frac{\alpha-1}{\alpha}$ one has that*

$$\mathbb{E}_{\mathcal{P}_{XY}}[f(X,Y)] \geq \operatorname*{ess\,inf}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}^{\frac{1}{\beta}}\left[f(X,Y)^\beta\right]\right] \tag{42}$$

$$\cdot \exp\left(\text{sign}(\alpha)\cdot\frac{\alpha-1}{\alpha}I_\alpha(X,Y)\right). \tag{43}$$

Corollary 1 holds for every non-negative function f and it is thus suited to provide lower-bounds in settings like the one described in Section VI. Common applications of information-measures in these settings typically require the employment of Markov's inequality in order to relate the expected value of a loss function with the information-measure (cf. [14, Thm. 1]). Using I_α with $\alpha < 0$ this is not necessary, as one can employ Corollary 1 directly. However, given $0 < \alpha < 1$ (which in turn implies $\beta < 0$), if there exists an x with positive measure with respect to $\mathcal{P}_X$ and such that for every y such that $\mathcal{P}_Y(\{y\}) > 0$, $f(x,y) = 0$, one recovers a trivial lower-bound on $\mathbb{E}_{\mathcal{P}_{XY}}[f(X,Y)]$. This prevents us from setting $f = \mathbb{1}_E$ and from recovering a bound that involves probabilities. Bounding the probability of an event under the joint using the product of the marginals can be useful when bounding the probability of having a large generalisation error [1] (or, again, in Bayesian Risk settings [12], [14]). It is thus important to understand if and when these results can be retrieved. One can see that whenever $\alpha < 0$ (which implies $0 < \beta < 1$) one can plug-in indicator functions in Corollary 1 and provide the following lower-bound connecting the probability of any event E with respect to the joint $\mathcal{P}_{XY}$ and the product of the marginals $\mathcal{P}_X\mathcal{P}_Y$:

Corollary 2. *Let X, Y be two random variables and consider the probability spaces $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY})$ and $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$. Let $E \in \mathcal{F}$ and, given $y \in \mathcal{Y}$, denote with $E_y = \{x : (x,y) \in E\}$, then, for every $\alpha < 0$*

$$\mathcal{P}_{XY}(E) \geq \min_y \mathcal{P}_X(E_y)^{\frac{1}{\beta}}\cdot\exp\left(-\frac{\alpha-1}{\alpha}I_\alpha(X,Y)\right) \tag{44}$$

$$= \exp\left(\frac{1}{\beta}\left(\log(\min_y \mathcal{P}_X(E_y)) - I_\alpha(X,Y)\right)\right). \tag{45}$$

Taking the limit of $\alpha \to -\infty$ one recovers the following:

$$\mathcal{P}_{XY}(E) \geq \min_y \mathcal{P}_X(E_y)\exp(-I_{-\infty}(X,Y)) \tag{46}$$

$$= \min_y \mathcal{P}_X(E_y)\exp(-\mathcal{L}^c(X\to Y)). \tag{47}$$

Let us compare, through Table I, Corollary 1, Corollary 2 and a straight-forward generalisation of [1, Corollary 1] that we will now state for reference:

TABLE I
BEHAVIOUR OF THE BOUNDS EXPRESSED IN COROLLARY 1, COROLLARY 2, COROLLARY 3 AND [1, COROLLARY 1]

Behaviour of the Bound $\mathbb{E}_{\mathcal{P}_{XY}}[f(X,Y)] \lessgtr h_\beta(f(X,Y)) \cdot g(I_\alpha(X,Y))$			
Admitted values of α	$\alpha<0 \implies 0<\beta<1$	$0<\alpha<1 \implies \beta<0$	$\alpha>1 \implies \beta>1$
Information-Measure $g(I_\alpha)$	$\exp((1-\alpha)/\alpha \cdot I_\alpha(X,Y))$	$\exp((\alpha-1)/\alpha \cdot I_\alpha(X,Y))$	$\exp((\alpha-1)/\alpha \cdot I_\alpha(X,Y))$
Multiplicative Term $h_\beta(f)$	$\min_y \mathbb{E}_{P_X}^{\frac{1}{\beta}}\left[f(X,y)^\beta\right]$	$\min_y \mathbb{E}_{P_X}^{\frac{1}{\beta}}\left[f(X,y)^\beta\right]$	$\max_y \mathbb{E}_{P_X}^{\frac{1}{\beta}}\left[f(X,y)^\beta\right]$
Multiplicat. Term $h_\beta(\mathbb{1}_E)$	$\min_y(P_X(E_y))^{\frac{1}{\beta}}$	cannot be provided	$\max_y(P_X(E_y))^{\frac{1}{\beta}}$
Inequality	$\mathbb{E}_{\mathcal{P}_{XY}}[f] \geq h_\beta(f)\cdot g(I_\alpha(X,Y))$	$\mathbb{E}_{\mathcal{P}_{XY}}[f] \geq h_\beta(f)\cdot g(I_\alpha(X,Y))$	$\mathbb{E}_{\mathcal{P}_{XY}}[f] \leq h_\beta(f)\cdot g(I_\alpha(X,Y))$
References	Corollary 1 and Corollary 2	Corollary 1	Corollary 3 and [1, Corollary 1]

Corollary 3. *Consider the same setting as in Theorem 4. For every $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+$ and for every $\alpha > 1$ and $\beta = \frac{\alpha-1}{\alpha}$ one has that*

$$\mathbb{E}_{\mathcal{P}_{XY}}[f(X,Y)] \leq \operatorname*{ess\,sup}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}^{\frac{1}{\beta}}\left[f(X,Y)^\beta\right]\right] \tag{48}$$

$$\cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X,Y)\right). \tag{49}$$

We will now briefly explore immediate applications of the results presented in Section V in lower-bounding the Bayesian Risk in estimation procedures.

VI. BAYESIAN RISK

Let $\mathcal{W}$ denote the parameter space and assume that we have access to a prior distribution over this space $\mathcal{P}_W$. Suppose then that we observe W through the family of distributions $\mathcal{P} = \{\mathcal{P}_{X|W=w} : w \in \mathcal{W}\}$. Given a function $\phi : \mathcal{X} \to \hat{\mathcal{W}}$, one can then estimate W from the observations $X \sim \mathcal{P}_{X|W}$ via $\phi(X) = \hat{W}$. Let us denote with $\ell : \mathcal{W}\times\hat{\mathcal{W}} \to \mathbb{R}^+$ a loss function, the Bayesian risk is defined as:

$$R = \inf_\phi \mathbb{E}[\ell(W,\phi(X)] = \inf_\phi \mathbb{E}[\ell(W,\hat{W})]. \tag{50}$$

Corollary 4. *Consider the Bayesian framework just described, the following must hold for every $\alpha < 0$ and every $\hat{W} = \phi(X^n)$:*

$$R \geq \exp\left(\frac{1}{\beta}\left(-I_\alpha(W,X^n) + \log\left(\operatorname*{ess\,inf}_{P_{\hat{W}}} \mathbb{E}_{P_W}\left[\ell(W,\hat{W})^\beta\right]\right)\right)\right). \tag{51}$$

Moreover, taking the limit of $\alpha \to -\infty$ one recovers the following:

$$R \geq \rho \exp\left(-\mathcal{L}^c(W\to X^n)\right) \operatorname*{ess\,inf}_{\mathcal{P}_{\hat{W}}} \mathbb{E}_{P_W}\left[\ell(W,\hat{W})\right]. \tag{52}$$

Proof. The result follows from Corollary 1, Property 5 and the fact that $W - X^n - \hat{W}$ is a Markov Chain. □

One can also derive a result using the technique explored in [12], [14]:

Corollary 5. *Consider the Bayesian framework just described, the following must hold for every $\rho > 0$, $\alpha < 0$ and every $\hat{W} = \phi(X^n)$:*

$$R \geq \rho \exp\left(-\frac{1}{\beta} I_\alpha(W,X^n)\right)\left(1 - L_W(\rho)\right)^{\frac{1}{\beta}}, \tag{53}$$

where $L_W(\rho) = \max_{\hat{w}} \mathcal{P}_W(\ell(W,\hat{w}) \leq \rho)$, also known in the literature as "small-ball probability". Moreover, taking the limit of $\alpha \to -\infty$ one recovers the following:

$$R \geq \rho \exp\left(-\mathcal{L}^c(W\to X^n)\right)\left(1 - L_W(\rho)\right). \tag{54}$$

Proof. We start by observing that for every ρ, by Markov's inequality

$$\mathbb{E}[\ell(W,\hat{W})] \geq \rho\left(\mathbb{P}(\ell(W,\hat{W}) \geq \rho)\right). \tag{55}$$

The statement then follows from further lower-bounding $\mathbb{P}(\ell(W,\hat{W}) \geq \rho)$ with Corollary 2 and noticing that $\min_{\hat{w}} \mathcal{P}_W(\ell(W,\hat{w}) \geq \rho) = \min_{\hat{w}}(1 - \mathcal{P}_W(\ell(W,\hat{w}) \leq \rho)) = 1 - L_W(\rho)$. □

VII. CONCLUSION

In this work we extended the definition of Sibson's α-Mutual Information to negative values of α. In order to have a properly defined object and to be consistent with the axiomatic properties that an information measure should satisfy (according to Rényi [9]), we slightly adapted the original definition from (3) for $\alpha > 0$ to (5) for $\alpha < 0$. We presented a sequence of properties that these objects satisfy and we connected the information-measures to a family of functional inequalities (similarly to I_α with $\alpha > 1$ [1, Thm 4, Remark 7, Cor 1]). This family of inequalities exists, although not always with the same sign of inequality, for every $\alpha \in \mathbb{R}$ as shown in Table I. When $\alpha > 1$ and through these inequalities, one can bound the probability of having a large generalisation error and the Bayesian Risk in Estimation procedures [1], [12]. Here, as an example of application, we considered once again the Bayesian Risk setting and showed how this family of information measures can be employed in such a framework even if $\alpha < 1$.

ACKNOWLEDGMENT

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

REFERENCES

[1] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and Maximal Leakage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, 2021.

[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] 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.

[6] 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.

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

[8] R. Sibson, "Information radius," *Z. Wahrscheinlichkeitstheorie verw Gebiete 14*, pp. 149–160, 1969.

[9] A. Rényi, "On measures of entropy and information," *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, vol. 1, pp. 547–561, 1960.

[10] S.-W. Ho and S. Verdú, "Convexity/concavity of renyi entropy and α-mutual information," in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 745–749.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[12] A. R. Esposito and M. Gastpar, "Lower-bounds on the bayesian risk in estimation procedures via Sibson's α-mutual information," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 748–753.

[13] A. R. Esposito, D. Wu, and M. Gastpar, "On conditional Sibson's α-Mutual Information," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 1796–1801.

[14] 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.