Robust Generalization via $\alpha-$Mutual Information

Amedeo Roberto Esposito, Michael Gastpar
School of Computer and Communication Sciences
EPFL
{amedeo.esposito, michael.gastpar}@epfl.ch

Ibrahim Issa
Electrical and Computer Engineering Department
American University of Beirut
ii19@aub.edu.lb

***Abstract*—The aim of this work is to provide bounds connecting two probability measures of the same event using Rényi α-Divergences and Sibson's α-Mutual Information, a generalization of respectively the Kullback-Leibler Divergence and Shannon's Mutual Information. A particular case of interest can be found when the two probability measures considered are a joint distribution and the corresponding product of marginals (representing the statistically independent scenario). In this case a bound using Sibson's $\alpha-$Mutual Information is retrieved, extending a result involving Maximal Leakage to general alphabets. These results have broad applications, from bounding the generalization error of learning algorithms to the more general framework of adaptive data analysis, provided that the divergences and/or information measures used are amenable to such an analysis (*i.e.,* are robust to post-processing and compose adaptively). The generalization error bounds are derived with respect to high-probability events but a corresponding bound on expected generalization error is also retrieved.**

***Index Terms*—Rényi-Divergence, Sibson's Mutual Information, Maximal Leakage, Adaptive Data Analysis**

I. Introduction

Let us consider two probability spaces $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ and let $E \in \mathcal{F}$ be a measurable event. Given some divergence between the two distributions $\hat{D}(\mathcal{P}, \mathcal{Q})$ (e.g., KL, Rényi's $\alpha-$Divergence, ...) our aim is to provide bounds of the following shape:

$$\mathcal{P}(E) \leq f(\mathcal{Q}(E)) \cdot g(\hat{D}(\mathcal{P}, \mathcal{Q})), \tag{1}$$

for some functions f, g. E represents some "undesirable" event (e.g., large generalization error), whose measure under $\mathcal{Q}$ is known and whose measure under $\mathcal{P}$ we wish to bound. To that end, we use some notion of "distance" between $\mathcal{P}$ and $\mathcal{Q}$. Of particular interest is the case where $\Omega = \mathcal{X} \times \mathcal{Y}$, $\mathcal{P} = \mathcal{P}_{XY}$ (the joint distribution), and $\mathcal{Q} = \mathcal{P}_X\mathcal{P}_Y$ (product of the marginals). This allows us to bound the likelihood of $E \subseteq \mathcal{X} \times \mathcal{Y}$ when two random variables X and Y are dependent as a function of the likelihood of E when X and Y are independent (typically easier to analyze). Indeed, an immediate application can be found in bounding the generalization error of a learning algorithm and, when the proper measure is chosen, in adaptive data analysis. In order to be used in adaptive data analysis, such measure needs to be robust to post-processing and to compose adaptively (meaning that we can bound the measure between input and output of the composition of a sequence of algorithms if each of them has bounded measure). Results of this form involving mutual information can be found in [1]–[3]. More recently, a different measure satisfying these properties, maximal leakage [4], has been used in [5], [6]. More specifically, it was shown that Equation (1) holds for the following choice of $f(\mathcal{P}_X\mathcal{P}_Y(E)) = \max_y(\mathcal{P}_X(E_y))$ and $g(\hat{D}(\mathcal{P}_{XY}||\mathcal{P}_X\mathcal{P}_Y)) = \exp(\mathcal{L}(X \to Y)) = \mathbb{E}_Y\left(D_\infty(\mathcal{P}_{X|Y}||\mathcal{P}_X)\right) = I_\infty(X;Y)$, where $I_\infty(X;Y)$ is the Sibson mutual information of order infinity. In this work, we derive a general bound in the form of (1) and focus on two interesting special cases. In particular, one specialization of the bound leads to a family of bounds in terms of α-divergences. The other specialization leads to a family of bounds in terms of Sibson's α-mutual information, thus generalizing the previous maximal leakage bound (which corresponds to $\alpha \to \infty$).

II. Background And Definitions

A. Sibson's $\alpha-$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 problems [7]. Indeed, it has several useful operational interpretations (e.g., the number of bits by which a mixture of two codes can be compressed, the cut-off rate in block coding and hypothesis testing [8], [9] [10, p. 649]). 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 $\alpha-$Divergence of $\mathcal{P}$ from $\mathcal{Q}$ is defined as follows:

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

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 [8].

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 $\lim_{\alpha \downarrow 1} D_\alpha(\mathcal{P}\|\|Q) = D(\mathcal{P}\|\mathcal{Q})$ [8, Theorem 5]. For an extensive treatment of α-divergences

and their properties we refer the reader to [8]. Starting from the concept of $\alpha-$divergence, Sibson built a generalization of mutual information that retains many interesting properties. The definition is the following [7]:

Definition 2. Let X and Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$, and with marginal distributions $\mathcal{P}_X$ and $\mathcal{P}_Y$, respectively. For $\alpha > 0$, the Sibson's mutual information of order α between X and Y is defined as:

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

Moreover, $\lim_{\alpha\to 1} I_\alpha(X;Y) = I(X;Y)$. On the other hand when $\alpha \to \infty$, we get:

$$I_\infty(X;Y) = \log \mathbb{E}_{\mathcal{P}_Y}\left[\sup_{x:\mathcal{P}_X(x)>0} \frac{\mathcal{P}_{XY}(\{x,Y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{Y\})}\right].$$

For more details on Sibson's α-MI we refer the reader to [7].

B. Learning Theory

In this section, we provide some basic background knowledge on learning algorithms and concepts like generalization error. We are mainly interested in supervised learning, where the algorithm learns a *classifier* by looking at points in a proper space and the corresponding labels.

More formally, suppose we have an instance space $\mathcal{Z}$ and a hypothesis space $\mathcal{H}$. The hypothesis space is a set of functions that, given a data-point $s \in \mathcal{Z}$ give as an output the corresponding label $\mathcal{Y}$. Suppose we are given a training data set $\mathcal{Z}^n \ni S = \{z_1, \dots, z_n\}$ made of n points sampled in an i.i.d fashion from some distribution $\mathcal{P}$. Given some $n \in \mathbb{N}$, a learning algorithm is a (possibly stochastic) mapping $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ that given as an input a finite sequence of points $S \in \mathcal{Z}^n$ outputs some classifier $h = \mathcal{A}(S) \in \mathcal{H}$. In the simplest setting we can think of $\mathcal{Z}$ as a product between the space of data-points and the space of labels, *i.e.,* $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ and suppose that $\mathcal{A}$ is fed with n data-label pairs $(x, y) \in \mathcal{Z}$. In this work we will view $\mathcal{A}$ as a family of conditional distributions $\mathcal{P}_{H|S}$ and provide a stochastic analysis of its generalization capabilities using the information measures introduced above. The goal is to generate a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that has good performance on both the training set and newly sampled points from $\mathcal{X}$. In order to ensure such property the concept of generalization error is introduced.

Definition 3. Let $\mathcal{P}$ be some distribution over $\mathcal{Z}$. Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function. The error (or risk) of a prediction rule h with respect to $\mathcal{P}$ is defined as

$$L_{\mathcal{P}}(h) = \mathbb{E}_{Z\sim\mathcal{P}}[\ell(h, Z)], \tag{4}$$

while, given a sample $S = (z_1, \dots, z_n)$, the empirical error of h with respect to S is defined as

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i). \tag{5}$$

Moreover, given a learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$, its generalization error with respect to S is defined as:

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) = |L_{\mathcal{P}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))|. \tag{6}$$

The definition above considers general loss functions. An important instance for the case of supervised learning is the $0-1$ loss. Suppose again that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and that $\mathcal{H} = \{h | h : \mathcal{X} \to \mathcal{Y}\}$; given a pair $(x, y) \in \mathcal{Z}$ and a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ the loss is defined as follows:

$$\ell(h, (x, y)) = \mathbb{1}_{h(x)\neq y}, \tag{7}$$

where $\mathbb{1}$ is the indicator function. The corresponding errors become:

$$L_{\mathcal{P}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\mathbb{1}_{h(x)\neq y}] = \mathcal{P}(\{(x, y) : h(x) \neq y\}) \tag{8}$$

and

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{h(x_i)\neq y_i}. \tag{9}$$

Another fundamental concept we will need is the sample complexity of a learning algorithm.

Definition 4. Fix $\epsilon, \delta \in (0, 1)$. Let $\mathcal{H}$ be a hypothesis class. The sample complexity of $\mathcal{H}$ with respect to (ϵ, δ), denoted by $m_{\mathcal{H}}(\epsilon, \delta)$, is defined as the smallest $m \in \mathbb{N}$ for which there exists a learning algorithm $\mathcal{A}$ such that, for every distribution $\mathcal{P}$ over the domain $\mathcal{X}$ we have that $\mathbb{P}(\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) > \epsilon) \leq \delta$. If there is no such m then $m_{\mathcal{H}}(\epsilon, \delta) = \infty$.

For more details we refer the reader to [11].

III. Main Results

Our main theorem is a general bound on $\mathcal{P}_{XY}(E)$ in terms of $\mathcal{P}_X\mathcal{P}_Y(E)$, parameterized by two real numbers α and α'. For particular choices of α and α', we demonstrate bounds in terms of α-divergence, as well as α-mutual information. The latter is a generalization of the maximal leakage bound in [6].

Theorem 1. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Given $E \in \mathcal{F}$, let $E_y := \{x : (x, y) \in E\}$, i.e., the "fibers" of E with respect to y. Then for any $E \in \mathcal{F}$,*

$$\mathcal{P}_{XY}(E) \leq \left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathcal{P}_X(E_Y)^{\gamma'/\gamma}\right]\right)^{1/\gamma'} \cdot \left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}^{\alpha'/\alpha}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right]\right)^{1/\alpha'}, \tag{10}$$

where $\gamma, \alpha, \gamma', \alpha'$ are such that $1 = \frac{1}{\alpha} + \frac{1}{\gamma} = \frac{1}{\alpha'} + \frac{1}{\gamma'}$, and $\alpha, \gamma, \alpha', \gamma' \geq 1$.

Proof. We have that:

$$\mathcal{P}_{XY}(E) = \mathbb{E}_{\mathcal{P}_{XY}}[\mathbb{1}_E] \tag{11}$$

$$= \mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[\mathbb{1}_E \frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right] \tag{12}$$

$$= \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}\left[\mathbb{1}_{\{X\in E_Y\}}\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right]\right] \tag{13}$$

$$\leq \mathbb{E}_{\mathcal{P}_Y}\Bigg[\left(\mathbb{E}_{\mathcal{P}_X}\left[\mathbb{1}^{\gamma}_{\{X\in E_Y\}}\right]\right)^{1/\gamma}\cdot \left(\mathbb{E}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right)^{1/\alpha}\Bigg] \tag{14}$$

$$= \mathbb{E}_{\mathcal{P}_Y}\left[\mathcal{P}_X(E_Y)^{1/\gamma}\left(\mathbb{E}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right)^{1/\alpha}\right] \tag{15}$$

$$\leq \left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathcal{P}_X(E_Y)^{\gamma'/\gamma}\right]\right)^{1/\gamma'}\cdot \left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{\alpha'/\alpha}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right]\right)^{1/\alpha'}, \tag{16}$$

where (14) and (16) follow from Holder's inequality, given that $\gamma, \alpha, \gamma', \alpha' \geq 1$ and $\frac{1}{\gamma}+\frac{1}{\alpha}=\frac{1}{\gamma'}+\frac{1}{\alpha'}=1$. □

Remark 2. It is clear from the proof that one can similarly bound $\mathbb{E}[g(X,Y)]$ for any positive function $g(X,Y)$ such that $g(X,Y)$ is $\mathcal{P}_X\mathcal{P}_Y$-integrable. But the shape of the bound becomes more complex as one in general does not have that $g(X,Y)^{\gamma} = g(X,Y)$ for every $\gamma \geq 1$.

Based on the choices of α, α', one can derive different bounds. Two are of particular interests to us and rely on different choices of α'. Choosing $\alpha' = \alpha$ and thus $\gamma' = \gamma$ in Theorem 1, we retrieve:

Corollary 1. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Let $E \in \mathcal{F}$ we have that:*

$$P_{XY}(E) \leq (\mathcal{P}_X\mathcal{P}_Y(E))^{\frac{\alpha-1}{\alpha}}\cdot \exp\left(\frac{\alpha-1}{\alpha}D_{\alpha}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)\right). \tag{17}$$

Choosing $\alpha' \to 1$, which implies $\gamma' \to +\infty$, we retrieve:

Corollary 2. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Given $E \in \mathcal{F}$, we have that:*

$$P_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{1/\gamma}\cdot \tag{18}$$

$$\mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{1/\alpha}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_Y\mathcal{P}_X}\right)^{\alpha}\right]\right] \tag{19}$$

$$= \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{\frac{\alpha-1}{\alpha}}\exp\left(\frac{\alpha-1}{\alpha}I_{\alpha}(X;Y)\right), \tag{20}$$

where $I_{\alpha}(X;Y)$ is the Sibson's mutual information of order α [7].

Remark 3. An in-depth study of $\alpha-$mutual information appears in [7], where a slightly different notation is used. For reference, we can restate Eq. (19) in the notation of [7] to obtain:

$$P_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{1/\gamma}\cdot \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{1/\alpha}_{\mathcal{P}_X}\left[\left(\frac{dP_{Y|X}}{d\mathcal{P}_Y}\right)^{\alpha}\middle|Y\right]\right]. \tag{21}$$

Moreover, for a fixed α due to the property that Holder's conjugates need to satisfy, we have that $\frac{1}{\gamma} = \frac{\alpha-1}{\alpha}$ and the bound in (20) can also be rewritten as:

$$\mathcal{P}_{XY}(E) \leq \exp\left(\frac{\alpha-1}{\alpha}\left(I_{\alpha}(X;Y) + \log\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)\right). \tag{22}$$

Considering the right hand side of (22), because of the non-decreasability of Sibson's $\alpha-$Mutual Information with respect to α [7] we have that, for $1 \leq \alpha_1 \leq \alpha_2$:

$$\frac{\alpha_1-1}{\alpha_1}I_{\alpha_1}(X;Y) \leq \frac{\alpha_2-1}{\alpha_2}I_{\alpha_2}(X;Y). \tag{23}$$

Thus, choosing a smaller α yields a better dependence on $I_{\alpha}(X;Y)$ in the bound, but given that $\frac{1}{\gamma} = \frac{\alpha-1}{\alpha}$ we also have that $\frac{1}{\gamma_1} \leq \frac{1}{\gamma_2}$ and being $\operatorname{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y) \leq 1$ it implies that

$$\left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{\frac{1}{\gamma_1}} \geq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{\frac{1}{\gamma_2}}, \tag{24}$$

with a worse dependence on $\left(\operatorname{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}(E_Y)\right)^{\frac{1}{\gamma}}$ on the bound. This leads to a trade-off between the two quantities. If we focus on Corollary 2, letting $\alpha \to \infty$ we recover a result involving maximal leakage [5], [6], but extending it to general alphabets:

Corollary 3. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Let $E \in \mathcal{F}$ we have that:*

$$P_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)\exp\left(\mathcal{L}\left(X\to Y\right)\right), \tag{25}$$

where $\mathcal{L}\left(X\to Y\right)$ is the maximal leakage [4].

The bound follows from the fact that $\mathcal{L}\left(X\to Y\right) = I_{\infty}(X;Y)$ [12]. A comparison between the bound for maximal leakage and some analogous result obtained for mutual information (through a different approach [1], [2]) can be found in [6].

IV. Applications

In this section, we consider some applications of the above bounds in the context of the generalization error. In the bounds of interest $\mathcal{P}_X(E_y)$ is typically exponentially decaying with the number of samples and the trade-off between α and γ can be explicitly seen in the sample complexity of a learning algorithm:

Corollary 4. *Let $\mathcal{X}\times\mathcal{Y}$ be the sample space and $\mathcal{H}$ be the set of hypotheses. Let $\mathcal{A}: \mathcal{X}^n\times\mathcal{Y}^n \to \mathcal{H}$ be a learning algorithm*

that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, i.e., $S \sim \mathcal{P}^n$. Let ℓ be the $0-1$ loss function as defined in (7)*. Given $\eta \in (0,1)$, let $E = \{(S,h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$. Then,*

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - 2n\eta^2\right)\right). \quad (26)$$

Proof. Fix $\eta \in (0,1)$ and $\alpha \geq 1$. Let $\frac{1}{\gamma} = \frac{\alpha-1}{\alpha}$. Let us denote with E_h the fiber of E over h for some $h \in \mathcal{H}$, *i.e.,* $E_h = \{S : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Consider $S, \hat{S} \in \{\mathcal{X} \times \mathcal{Y}\}^n$, where $S = ((x_1,y_1),\ldots,(x_n,y_n))$ and $\hat{S} = ((\hat{x}_1,\hat{y}_1),\ldots,(\hat{x}_n,\hat{y}_n))$. If $S, \hat{S}$ differ only in one position j, *i.e.,* $(x_i,y_i) = (\hat{x}_i,\hat{y}_i)\,\forall i \in [n]\setminus\{j\}$ and $(x_j,y_j) \neq (\hat{x}_j,\hat{y}_j)$ we have that for every $h \in \mathcal{H}$,

$$|L_S(h) - L_{\hat{S}}(h)| \leq \frac{1}{n}. \quad (27)$$

By McDiarmid's inequality [13][Sec. 1.1] and Ineq. (27) we have that for every hypothesis $h \in \mathcal{H}$,

$$\mathcal{P}_S(E_h) \leq 2 \cdot \exp(-2n\eta^2). \quad (28)$$

Then it follows from Corollary 2 and Ineq. (28) that:

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(S;\mathcal{A}(S))\right)\left(2\exp(-2n\eta^2)\right)^{\frac{\alpha-1}{\alpha}}. \quad (29)$$

□

Corollary 5. *Let $\mathcal{X} \times \mathcal{Y}$ be the sample space and $\mathcal{H}$ be the set of hypotheses. Let $\mathcal{A} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, i.e., $S \sim \mathcal{P}^n$. Let ℓ be the $0-1$ loss function. Given $\eta \in (0,1)$, let $E = \{(S,h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$ then, in order to ensure a confidence of $\delta \in (0,1)$, i.e., $\mathbb{P}(E) \leq \delta$, we need a number of samples m satisfying:*

$$m \geq \frac{I_\alpha(S;\mathcal{A}(S)) + \log 2 + \gamma \log\left(\frac{1}{\delta}\right)}{2\eta^2}. \quad (30)$$

Proof. From Corollary 4 we have that

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - 2n\eta^2\right)\right).$$

Fix $\delta \in (0,1)$, our aim is to have that:

$$\exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - 2n\eta^2\right)\right) \leq \delta, \quad (31)$$

solving the inequality wrt n gives us Equation (30). □

Smaller α means that $I_\alpha(S;\mathcal{A}(S))$ will be smaller, but it will imply a larger value for $\gamma = \frac{\alpha}{\alpha-1}$ and thus a worse dependency on $\log(1/\delta)$ in the sample complexity. Let $\mathcal{Z}$ be the sample space and $\mathcal{H}$ be the set of hypotheses. An immediate generalization of Corollary 4 follows by considering loss functions such that for every fixed $h \in \mathcal{H}$, the random variable $l(h,Z)$ (induced by Z) is σ^2-sub Gaussian[1] for some $\sigma > 0$.

[1]Given a random variable X we say that it is σ^2-sub-Gaussian if for every $\lambda \in \mathbb{R}$: $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$.

Corollary 6. *Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{Z}$. Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function such that $\ell(h,Z)$ is σ-sub Gaussian random variable for every $h \in \mathcal{H}$. Given $\eta \in (0,1)$, let $E = \{(S,h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$. Then,*

$$\mathbb{P}(E) \leq \exp\left(\frac{1}{\gamma}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - n\frac{\eta^2}{2\sigma^2}\right)\right). \quad (32)$$

Proof. Fix $\eta \in (0,1)$. Let us denote with E_h the fiber of E over h for some $h \in \mathcal{H}$, *i.e.,* $E_h = \{S : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. By assumption we have that $l(h,Z)$ is $\sigma-$sub Gaussian for every h. We can thus use Hoeffding's inequality for every hypothesis $h \in \mathcal{H}$, and retrieve that for every $h \in \mathcal{H}$:

$$\mathcal{P}_S(E_h) \leq 2 \cdot \exp\left(-n\frac{\eta^2}{2\sigma^2}\right). \quad (33)$$

Then it follows from Corollary 2 and Ineq. (33) that:

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(S;\mathcal{A}(S))\right)\left(2\exp\left(-n\frac{\eta^2}{2\sigma^2}\right)\right)^{\frac{\alpha-1}{\alpha}}. \quad (34)$$

□

One important characteristic of these bounds is that they involve information-measures satisfying the data processing inequality [7]. This means that all these results about generalization are **robust** to post-processing, *i.e.,* if the outcome of any learning algorithm with bounded I_α is processed further, the value of the information measure cannot increase. Another desirable property that would render the usage of such measures appealing in Adaptive Data Analysis is the Adaptive Composition property [14]. Alas, the lack of a definition of conditional Sibson's MI does not allows us, for the moment, to fully address the issue and verify whether or not the measure composes adaptively (like Mutual Information and Maximal Leakage [2], [6]). Moreover, a comparison between this and other well-known results in the literature can be found in Table I. One can immediately see that the Sibson's MI bound and, in particular, the Maximal Leakage one, are the ones that most resemble the VC-Dimension bound both in terms of excess probability decay and sample complexity.

V. Bounds on Expected Generalization Error

So far, when analyzing the generalization error, we have only considered high probability bounds, what can these results tell us about the **expected** generalization error? In order to provide a meaningful bound, some assumptions on the quantity $\max_h \mathcal{P}_S(|L_S(h) - \mathbb{E}[L(h)]| > \eta)$ are needed (where S is a random vector of length n, sampled in an iid fashion from some distribution $\mathcal{D}$). More precisely, we will assume this probability to be exponentially decreasing with the number of samples n, as it often happens in the literature [13], [15]. The following result is inspired by [11, p. 419] with a slightly different proof.

TABLE I
COMPARISON BETWEEN BOUNDS

	Robust	Adaptive	Bound	Sample Complexity
$\beta-$Stability [15]	No	No	exp. decay in n	$f(\beta,\eta)\times\log\left(\frac{2}{\delta}\right)$
ϵ-DP [14]	Yes	Yes	$\frac{1}{4}\exp\left(\frac{-n\eta^2}{12}\right)$, $\epsilon\leq\eta/2$	$\frac{12\cdot\log(1/4\delta)}{\eta^2}$
MI [1]	Yes	Yes	$(I(X;Y)+1)/(2n\eta^2-1)$	$I(X;Y)/\eta^2\delta$
Maximal Leakage [6]	Yes	Yes	$2\cdot\exp(\mathcal{L}(X\to Y)-2n\eta^2)$	$(\mathcal{L}\left(X\to Y\right)+\log\left(\frac{2}{\delta}\right))/2\eta^2$
α-Sibson's MI	Yes	Unknown	$\exp(\frac{\alpha-1}{\alpha}(I_\alpha(S;\mathcal{A}(S))+\log 2-2n\eta^2))$	$(I_\alpha(X;Y)+\log 2+\gamma\log\left(\frac{1}{\delta}\right))/2\eta^2$
VC-Dim. K [11]			$2\cdot\exp(\log(K)-2n\eta^2)$	$(\log(K)+\log\left(\frac{2}{\delta}\right))/2\eta^2$

Lemma 1. *Let X be a random variable and let $\hat{x}\in\mathbb{R}$. Suppose that exist $a\geq 0$ and $b\geq e$ such that for every $\eta>0$ $\mathcal{P}_X(|X-\hat{x}|\geq\eta)\leq 2b\exp\left(-\eta^2/a^2\right)$ then $\mathbb{E}\left[|X-\hat{x}|\right]\leq a\left(\sqrt{\log 2b}+\frac{1}{2\sqrt{\log 2b}}\right)$.*

Proof.

$$\mathbb{E}\left[|X-\hat{x}|\right]=\int_0^{+\infty}\mathcal{P}_X(|X-\hat{x}|\geq\eta)d\eta \tag{35}$$

$$\leq\int_0^{+\infty}\min\left(1,2b\exp\left(-\eta^2/a^2\right)\right)d\eta \tag{36}$$

$$=\int_0^{\sqrt{a^2\log 2b}}d\eta+\int_{\sqrt{a^2\log 2b}}^{+\infty}2b\exp(-\frac{\eta^2}{a^2})d\eta \tag{37}$$

$$\leq a\left(\sqrt{\log 2b}+\frac{1}{2\sqrt{\log 2b}}\right). \tag{38}$$

□

Theorem 2. *Let $\mathcal{A}:\mathcal{Z}^n\to\mathcal{H}$ be a learning algorithm and let $I_\alpha(S;\mathcal{A}(S))$ be the dependence measure chosen. Suppose that the loss function $l:\mathcal{Z}\times\mathcal{H}\to\mathbb{R}$ is such that $\forall h\mathcal{P}_{S\sim\mathcal{D}^n}(|L_S(h)-\mathbb{E}[L(h)]|>\eta)\leq 2\exp\left(-\frac{\eta^2}{2\sigma^2}n\right)$ for some $\sigma>0$ (e.g. $l(h,Z)$ is σ^2-sub-Gaussian), then:*

$$\mathbb{E}\left[|L_S(H)-\mathbb{E}[L(H)]|\right]\leq \tag{39}$$

$$\sqrt{\frac{2\sigma^2\gamma}{n}}\left(\sqrt{\frac{\log(2)+I_\alpha(S;\mathcal{A}(S))}{\gamma}}+\frac{1}{2\sqrt{\frac{\log 2+I_\alpha(S;\mathcal{A}(S))}{\gamma}}}\right). \tag{40}$$

Proof. The proof is a simple application of Lemma 1 and Corollary 6 with $a=\sqrt{2\gamma\sigma^2}/\sqrt{n}$ and with $b=2^{\frac{1}{\gamma}-1}\exp\left(\frac{I_\alpha(\mathcal{A}(S);S)}{\gamma}\right)$. □

An interesting application of Theorem 2 can be found by considering $\mathcal{L}\left(S\to\mathcal{A}(S)\right)$ and the $0-1$ loss (hence, $1/4$-sub-Gaussian).

Corollary 7. *Let $\mathcal{A}:\mathcal{Z}^n\to\mathcal{H}$. Consider the $0-1$ loss, then $\forall h\mathcal{P}_{S\sim\mathcal{D}^n}(|L_S(h)-\mathbb{E}[L(h)]|>\eta)\leq 2\exp\left(-2\eta^2n\right)$, and:*

$$\mathbb{E}\left[|L_S(H)-\mathbb{E}[L(H)]|\right]\leq \tag{41}$$

$$\frac{1}{\sqrt{2n}}\left(\sqrt{\log 2+\mathcal{L}\left(S\to\mathcal{A}(S)\right)}+\frac{1}{2\sqrt{\log 2+\mathcal{L}\left(S\to\mathcal{A}(S)\right)}}\right). \tag{42}$$

REFERENCES

[1] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," ser. Proceedings of Machine Learning Research, vol. 83. PMLR, 07–09 Apr 2018, pp. 25–55.

[2] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, p. 2521–2530.

[3] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 51. PMLR, 09–11 May 2016, pp. 1232–1240.

[4] I. Issa, S. Kamath, and A. B. Wagner, "An operational measure of information leakage," in *2016 Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 234–239.

[5] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *2019 IEEE International Symposium on Information Theory, ISIT Paris, France, July 7-12*, 2019.

[6] A. R. Esposito, M. Gastpar, and I. Issa, "Learning and adaptive data analysis via maximal leakage," in *IEEE Information Theory Workshop, ITW 2019, Visby, Gotland, Sweden, Aug 25-28*, 2019.

[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. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information 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] P. D. Grünwald, *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[11] S. Shalev-Shwartz and S. Ben-David., *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[12] I. Issa, A. B. Wagner, and S. Kamath, "An Operational Approach to Information Leakage," *ArXiv e-prints*, jul 2018.

[13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[14] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. Cambridge, MA, USA: MIT Pressf, 2015.

[15] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, 3 2002.