Robust Generalization via $f-$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*—Given two probability measures P and Q and an event E, we provide bounds on $P(E)$ in terms of $Q(E)$ and $f-$divergences. In particular, the bounds are instantiated when the measures considered are a joint distribution and the corresponding product of marginals. This allows us to control the measure of an event under the joint, using the product of the marginals (typically easier to compute) and a measure of how much the two distributions differ, *i.e.,* an $f-$divergence between the joint and the product of the marginals, also known in the literature as $f-$Mutual Information. The result is general enough to induce, as special cases, bounds involving χ^2-divergence, Hellinger distance, Total Variation, etc. Moreover, it also recovers a result involving Rényi's $\alpha-$divergence. As an application, we provide bounds on the generalization error of learning algorithms via $f-$divergences.**

***Index Terms*—f-Divergences, f-Mutual Information, χ^2-divergence, Maximal Leakage, Generalization Error**

I. Introduction

The generalization error of learning algorithms can be bounded via information measures. Several variants of this insight have been explored in the literature. [1]–[8]. Here, we further develop the perspective taken in our previous work [9]. In that perspective, one interprets a learning algorithm as a randomized mapping that takes as input a data-set, and provides as output a hypothesis (e.g., a classifier). The aim is to retrieve a hypothesis that has good performance both on the training set and on an *independent* test set. Intuitively, if the outcome of the learning algorithm depends too much on its input, then the performance on a new data-point swill be poor. In this case the algorithm is said to *overfit* to the training set. In order to address this issue, a recent line of work tries to quantify this dependence via information measures. Differently from [9], where we considered Rényi's α-Divergences and Sibson's $\alpha-$Mutual Information, in this work[1] we provide an approach that involves $f-$divergences and $f-$Mutual Information. $f-$divergences represent a broad class of dependence measures that generalize Shannon measures (Kullback-Leibler Divergence and Mutual Information), α-divergences and also includes quantities commonly used metrics in statistics, such as the χ^2-divergence, LeCam's Divergence, etc. Given two probability measures $\mathcal{P}$ and $\mathcal{Q}$ and an event E, our aim is to provide bounds of the following shape:

$$\mathcal{P}(E) \leq g(\mathcal{Q}(E)) \cdot h(D_f(\mathcal{P}, \mathcal{Q})), \tag{1}$$

for some $f-$divergence D_f, and functions g, h. 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. In [9], we derive bounds of a similar form where $D_f(\mathcal{P}, \mathcal{Q})$ is replaced by $\alpha-$Mutual Information or $\alpha-$Rényi divergence. Of particular interest is when we consider two random variables X, Y with $\mathcal{P} = \mathcal{P}_{XY}$ (a joint distribution), and $\mathcal{Q} = \mathcal{P}_X\mathcal{P}_Y$ (the product of the marginals). This allows us to bound the likelihood of $E \subseteq \mathcal{X} \times \mathcal{Y}$ when X and Y are dependent as a function of the likelihood of E when X and Y are independent (a scenario that is typically easier to analyze). We can immediately apply such bounds in Statistical Learning and Adaptive Data Analysis. As for statistical learning, such inequalities allow us to control the probability of having a large generalization error with respect to a learning algorithm. The quantity $D_f(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)$ measures how much the outcome of the algorithm depends on its input and represents a measure of *over-fitting*. Intuitively, the smaller the measure, the closer the input and the output are from being independent, and the less the algorithm will overfit. One interesting property that all these bounds share is that such generalization bounds are robust to post-processing. This is a simple consequence of the Data-Processing Inequality that most divergence measures satisfy and simply means that if you process (further) the outcome of your learning algorithm, the dependence on the training set can only decrease. Thus, having a bounded information measure implies having *robust* generalization guarantees.

A. Related Works

Similar approaches have been considered in [1], [2], [7], [8], [10], [11], where the chosen measure was Mutual Information. A bound involving Maximal Leakage has been extensively studied in [4], [12]. A generalization of the Maximal Leakage result, covering both Rényi's $\alpha-$Divergence and Sibson's $\alpha-$Mutual Information has been presented in [9]. Bounds involving several other measures have been presented in [5], [6]. A different approach consists in bounding the expected generalization error instead, and has been used in [1], [7], [10].

[1]A full version of this work including, in a comprehensive manner, bounds involving Rényi's $\alpha-$divergence, Sibson's $\alpha-$Mutual Information, Maximal Leakage and $f-$Mutual Information can be found on arxiv under the id arXiv:1912.01439.

B. Contributions

Differently from [4], [9], [12], the family of bounds we propose here involves $f-$divergences. Such bounds generalize even further the $\alpha-$Divergence result (although, there is no known direct link between $f-$Divergences and Sibson's Mutual Information instead). This generalization can possibly allow us to circumvent an issue raised in [7], [11] that, in many interesting scenarios where the random variables have infinite support, the Mutual Information between input and output $I(X;Y)$ is infinite. Since $\alpha-$MI is non-decreasing in α [13], every $I_\alpha(X;Y)$ with $\alpha \geq 1$ is such that $I_\alpha(X;Y) \geq I(X;Y)$, and consequently infinite. In this scenario, the bounds presented in [2], [4], [9], [12] become trivial. Using $f-$divergences, and related $f-$Mutual Informations, outside of the $\alpha-$Divergence family it is possible to obtain non-trivial bounds. Indeed [5]:

Example 1. *Let S be a real-valued random vector, given the Markov Chain $S-H-Y$, where the Euclidean norm $\|H\|^2 \leq K$ a.s. and $Y = H + N$ with N Gaussian noise. Because of strong data-processing inequalities [14, Sec. 1.2], the Total Variation distance between the joint and the product of the marginals of S, Y, denoted by $TV(S,Y)$, is strictly less than 1. $I(S;Y)$ may still be infinite.*

Moreover, exploiting a bound involving $I_f(X;Y)$ for a broad enough set of functions f allows to differently measure the dependence between X and Y and can provide different convergence rates [15], [16]. Thus, even though several $f-$divergences may go to 0 with the number of steps (or samples, in the case of a generalization error bound), the rate of convergence can vary along with the sample complexity of learning algorithms. With this drive, we derive a result bounding the measure of an event E under the joint $\mathcal{P}_{XY}$ using the product of the marginals $\mathcal{P}_X\mathcal{P}_Y$ and an $f-$divergence. We also provide some concrete examples using specific f-divergences. As an application, we bound the generalization error of learning algorithms also providing a bound where the divergence measure is guaranteed to always be finite (Hellinger distance). All these large deviations bounds can be extended to expected generalization error results using the same technique as in [17, Lemma 2, Theorem 4].

II. Problem Statement

In this section we will formally define the quantities we will use throughout the paper.

A. Learning Theory

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. 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 $s \in \mathcal{Z}$ outputs the corresponding label $\mathcal{Y}$. Suppose we are given a training data set $\mathcal{Z}^n \ni S = \{z_1, \ldots, z_n\}$ 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. 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 $\mathcal{X}$. To guarantee this, one has to keep the generalization error bounded.

Definition 1. 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

$$L_{\mathcal{P}}(h) = \mathbb{E}_{Z\sim\mathcal{P}}[\ell(h, Z)], \tag{2}$$

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

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i). \tag{3}$$

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

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) = |L_{\mathcal{P}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))|. \tag{4}$$

This definition 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 couple $(x, y) \in \mathcal{Z}$ and a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ the loss is defined as follows:

$$\ell(h, (x, y)) = \mathbb{1}_{h(x)\neq y}, \tag{5}$$

and the corresponding errors become:

$$L_{\mathcal{P}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\mathbb{1}_{h(x)\neq y}] = \mathbb{P}(h(x) \neq y), \tag{6}$$

and

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{h(x_i)\neq y_i}. \tag{7}$$

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

Definition 2. 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 $\mathcal{A} : \mathcal{Z}^m \to \mathcal{H}$ such that, for every distribution $\mathcal{P}$ over the domain $\mathcal{X}$,

$$\mathbb{P}(\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) > \epsilon) \leq \delta. \tag{8}$$

If there is no such m then $m_{\mathcal{H}}(\epsilon, \delta) = \infty$.

For more details we refer the reader to [18, Sections 2-3].

B. $f-$Mutual Information

A generalization of the KL-Divergence can be obtained by considering a generic convex function $f : \mathbb{R}^+ \to \mathbb{R}$, usually with the simple constraint that $f(1) = 0$. The constraint can be ignored as long as $f(1) < +\infty$ by simply considering a new mapping $g(x) = f(x) - f(1)$.

Definition 3. Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a convex function such that $f(1) = 0$. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$. Denoting with p, q the densities of the measures with respect to μ, the $f-$Divergence of $\mathcal{P}$ from $\mathcal{Q}$ is defined as follows:

$$D_f(\mathcal{P}\|\mathcal{Q}) = \int q f\left(\frac{p}{q}\right) d\mu. \tag{9}$$

Note that $f-$divergences are independent from the choice of the dominating measure μ [19]. When absolute continuity between $\mathcal{P}, \mathcal{Q}$ holds, i.e. $\mathcal{P} \ll \mathcal{Q}$, an assumption we will often use, we retrieve the following [19]:

$$D_f(\mathcal{P}\|\mathcal{Q}) = \int f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q}. \tag{10}$$

This generalization includes the KL divergence (by simply setting $f(t) = t\log(t)$). But it also includes:

- Total Variation distance, with $f(t) = \frac{1}{2}|t-1|$;
- Hellinger distance, with $f(t) = (\sqrt{t}-1)^2$;
- Pearson χ^2-divergence, with $f(t) = (t-1)^2$.

f-Mutual Information is defined using $f-$divergence as follows:

Definition 4. Let X and Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$ over a measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{XY})$. Let $(\mathcal{X}, \mathcal{F}_X, \mathcal{P}_X), (\mathcal{Y}, \mathcal{F}_Y, \mathcal{P}_Y)$ be the corresponding probability spaces induced by the marginals. Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a convex function such that $f(1) = 0$. The $f-$Mutual Information between X and Y is defined as:

$$I_f(X;Y) = D_f(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y). \tag{11}$$

If f is strictly convex at 1 and satisfies $f(1) = 0$, then $I_f(X;Y) = 0$ if and only if X and Y are independent [19]. Choosing $f(t) = t\log t$, one recovers the Mutual Information.

III. Main Results

Given two measures $\mathcal{P}, \mathcal{Q}$ and an event E, our main theorem is a bound on $\mathcal{P}(E)$ in terms of $\mathcal{Q}(E)$ and an $f-$Divergence. f has to be an invertible, convex function on $[0, +\infty)$. For ease of exposition, we will always assume that $\mathcal{P} \ll \mathcal{Q}$, *i.e.*, absolute continuity of $\mathcal{P}$ with respect to $\mathcal{Q}$: for every measurable set E if $\mathcal{Q}(E) = 0$ then also $\mathcal{P}(E) = 0$. However, the result can be extended to the case in which absolute continuity fails: in the proof one has to consider densities of $\mathcal{P}, \mathcal{Q}$ with respect to a common dominating mesaure.

Theorem 1. *Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probabiltiy spaces. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function such that $f(1) = 0$, and assume f is non-decreasing on $[0, +\infty)$. Suppose also that f is such that for every $y \in \mathbb{R}^+$ the set $\{t \geq 0 : f(t) > y\}$ is non-empty, i.e. the generalized inverse, defined as $f^{-1}(y) = \inf\{t \geq 0 : f(t) > y\}$, exists. Let $f^*(t) = \sup_{\lambda \geq 0} \lambda t - f(\lambda)$ be the Fenchel-Legendre dual of $f(t)$ [20, Section 2.2]. Given an event $E \in \mathcal{F}$, we have that:*

$$\mathcal{P}(E) \leq \mathcal{Q}(E) \cdot f^{-1}\left(\frac{D_f(\mathcal{P}\|\mathcal{Q}) + (\mathcal{Q}(E^c))f^*(0)}{\mathcal{Q}(E)}\right). \tag{12}$$

Proof. $\forall \lambda > 0$:

$$\begin{aligned}
\mathcal{P}(E) =& \mathbb{E}_{\mathcal{P}}[\mathbb{1}_E] && (13)\\
=& \mathbb{E}_{\mathcal{Q}}\left[\mathbb{1}_E \frac{d\mathcal{P}}{d\mathcal{Q}}\right] && (14)\\
\overset{(a)}{\leq}& \frac{1}{\lambda}\mathbb{E}_{\mathcal{Q}}\left[f^*(\lambda \mathbb{1}_E) + f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right] && (15)\\
=& \frac{D_f(\mathcal{P}\|\mathcal{Q}) + \mathbb{E}_{\mathcal{Q}}[f^*(\lambda \mathbb{1}_E)]}{\lambda} && (16)\\
\overset{(b)}{\leq}& \frac{D_f(\mathcal{P}\|\mathcal{Q}) + f^*(\lambda)\mathcal{Q}(E) + f^*(0)(\mathcal{Q}(E^c))}{\lambda}, && (17)
\end{aligned}$$

where (a) follows from Young's inequality and where f^* is the Legendre-Fenchel dual of $f$and (b) follows as, being $\mathbb{1}_E \in [0,1]$ and we can write:

$$\begin{aligned}
f^*(\lambda \mathbb{1}_E) &= f^*(\lambda(\mathbb{1}_E + (1 - \mathbb{1}_E)0))) && (18)\\
&\leq \mathbb{1}_E f^*(\lambda) + (1 - \mathbb{1}_E) f^*(0). && (19)
\end{aligned}$$

We can now minimize (17) over all $\lambda > 0$:

$$\begin{aligned}
\mathcal{P}(E) \leq& \inf_{\lambda > 0}\left(\frac{D_f(\mathcal{P}\|\mathcal{Q}) + f^*(\lambda)\mathcal{Q}(E) + (\mathcal{Q}(E^c))f^*(0)}{\lambda}\right)\\
=& \mathcal{Q}(E) \cdot \inf_{\lambda > 0} \frac{\frac{D_f(\mathcal{P}\|\mathcal{Q}) + (\mathcal{Q}(E^c))f^*(0)}{\mathcal{Q}(E)} + f^*(\lambda)}{\lambda} && (20)\\
\overset{(c)}{=}& \mathcal{Q}(E) \cdot f^{-1}\left(\frac{D_f(\mathcal{P}\|\mathcal{Q}) + (\mathcal{Q}(E^c))f^*(0)}{\mathcal{Q}(E)}\right), && (21)
\end{aligned}$$

with (c) following from [20, Lemma 2.4]. In order to use [20, Lemma 2.4] the convex function needs to respect a set of properties. Using the notation of [20], the result is obtained by making the following substitution $\psi = f^*, \psi^* = f$. The properties that the function has to respect in the premise of the Lemma ($f^*(0) = f^{*\prime}(0) = 0$) have the purpose, analyzing the proof, to ensure that f is non-negative, convex and non-decreasing. Since f is convex by assumption, we have that $(f^*)^* = f$ and thus $(f^*)^*$ is convex and non-decreasing by assumption. As for the non-negativity, it is required in order to make sure that for a given $\lambda > 0$, we have that $f(t) \geq \lambda t - f^*(\lambda)$ is unbounded and the set $\{t \geq 0 : f(t) > y\}$ is non-empty for every $y \geq 0$. Thus, the non-negativity of f is a stronger assumption enforced in order to have a well defined generalized inverse $f^{-1}(y) = \inf\{t \geq 0 : f(t) > y\}$, and can be omitted when this is always non-empty. □

Remark 1. A simpler form of the Equation (12) can be found when $-\infty < f^*(0) \leq 0$. Indeed, it is possible to start from Eq. (18) and further upper-bound Ineq. (19) to obtain the following:

$$f^*(\lambda \mathbb{1}_E) \leq \mathbb{1}_E f^*(\lambda). \tag{22}$$

The final shape of the bound would then be:

$$\mathcal{P}(E) \leq \mathcal{Q}(E) \cdot f^{-1}\left(\frac{D_f(\mathcal{P}\|\mathcal{Q})}{\mathcal{Q}(E)}\right). \tag{23}$$

Corollary 1. *Let X, Y be two random variables. Let $(\Omega, \mathcal{F}, \mathcal{P}_{XY}), (\Omega, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces where $\mathcal{F} = \sigma(X, Y)$ (i.e., the σ-algebra generated by (X, Y)). Let f be a convex function satisfying the assumptions of Theorem 1. Given an event $E \in \mathcal{F}$, we have that:*

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot f^{-1}\left(\frac{I_f(X;Y) + (1 - \mathcal{P}_X\mathcal{P}_Y(E))f^*(0)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right). \tag{24}$$

A. Examples

As a first application of Corollary 1, we will retrieve a result shown (through a different approach) in [9, Corollary 1]. We restate it here for ease of reference.

Corollary 2. *Given $E \in \mathcal{F}$ (defined in Corollary 1), we have:*

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

where $\alpha, \gamma \geq 1$ and $1/\alpha + 1/\gamma = 1$.

Proof. Fix $\alpha > 1$ and consider the following convex function:

$$\phi_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}, \tag{25}$$

i.e., the Hellinger Divergence. The restriction of $\phi_\alpha(t)$ to $[0, +\infty)$ is increasing and thus invertible. Since we will consider only ratios between measures, the restriction to the positive real line is sufficient and Corollary 1 is applicable. It follows that:

$$\phi_\alpha^{-1}(t) = \left((\alpha - 1)\, t + 1\right)^{1/\alpha}, \tag{26}$$

and that:

$$\phi_\alpha^*(t) = t\left(\frac{(\alpha - 1)t}{\alpha}\right)^{1/\alpha - 1} - \frac{\left(\frac{(\alpha-1)t}{\alpha}\right)^{\alpha/\alpha - 1}}{\alpha - 1} + \frac{1}{\alpha - 1}, \tag{27}$$

from which we can deduce that:

$$\phi_\alpha^*(0) = \frac{1}{\alpha - 1}. \tag{28}$$

We also have that for a given $\alpha > 0$ and two measures $\mathcal{P}, \mathcal{Q}$ [19]:

$$D_\alpha(\mathcal{P}\|\mathcal{Q}) = \frac{1}{\alpha - 1}\log(1 + (\alpha - 1)D_{f_\alpha}(\mathcal{P}\|\mathcal{Q})), \tag{29}$$

then, with $f = \phi_\alpha$ and computing the right-hand side of Ineq. (24) we retrieve:

$$f^{-1}\left(\frac{I_f(X;Y) + (1 - \mathcal{P}_X\mathcal{P}_Y(E))/(\alpha - 1)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right) \tag{30}$$

$$= \frac{\exp\left(\frac{\alpha-1}{\alpha} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)\right)}{\mathcal{P}_X\mathcal{P}_Y(E)^{1/\alpha}}, \tag{31, 32}$$

To conclude, substitute (32) in (24):

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E)^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha - 1}{\alpha} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)\right). \tag{33}$$

Since $\frac{\alpha-1}{\alpha} = \frac{1}{\gamma}$ is the Holder's conjugate of $\frac{1}{\alpha}$, we recover Corollary 1 of [9]. □

Another interesting application of Corollary 1 is for $f(t) = (t-1)^2$. This function allows us to retrieve the Pearson's χ^2-divergence between two distributions. We will denote, through a slight abuse of notation, with $\chi^2(X, Y) = \chi^2(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)$. The bound we retrieve is the following:

Corollary 3. *Let $f(t) = t^2 - 1$, we have that $I_f(X;Y) = \chi^2(X, Y)$. Let $E \subseteq \mathcal{X} \times \mathcal{Y}$ we have that:*

$$\mathcal{P}_{XY}(E) \leq \sqrt{(\chi^2(X, Y) + 1)\mathcal{P}_X\mathcal{P}_Y(E)} \tag{34}$$

$$\leq \sqrt{\exp\left(\mathcal{L}\left(X \to Y\right)\right)\mathcal{P}_X\mathcal{P}_Y(E)}. \tag{35}$$

Proof. We have that $f^*(t) = t^2/4 + 1$ and thus $f^*(0) = 1$. We also have that $f^{-1}(t) = \sqrt{t+1}$. Applying Corollary 1 we have that:

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E)\sqrt{\frac{\chi^2(X, Y) + (1 - \mathcal{P}_X\mathcal{P}_Y(E))}{\mathcal{P}_X\mathcal{P}_Y(E)} + 1} \tag{36}$$

$$= \sqrt{(\chi^2(X, Y) + 1)\mathcal{P}_X\mathcal{P}_Y(E)}. \tag{37}$$

Equation (35) then follows from [3]:

$$\chi^2(X, Y) \leq \exp\left(\mathcal{L}\left(X \to Y\right)\right) - 1. \tag{38}$$

□

B. Application: Generalization Error

We will now apply Corollary 1 to provide generalization error bounds. Typically, when E is the event of large generalization error, $\mathcal{P}_S\mathcal{P}_{\mathcal{A}(S)}(E)$ is exponentially decaying with the number of samples n used in training. Consequently, as long as $I_f(S; \mathcal{A}(S))$ has a "controlled" growth with respect to n, the probability of having a large generalization error will still decay exponentially fast.

Corollary 4. *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 a σ^2-sub-Gaussian random variable[2], for some σ and 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,*

$$\mathbb{P}(E) \leq \sqrt{2}\exp\left(\frac{1}{2}\left(\log(\chi^2(S, \mathcal{A}(S)) + 1) - n\frac{\eta^2}{2\sigma^2}\right)\right). \tag{39}$$

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. $E_h = \{S : |L_\mathcal{P}(h) - L_S(h)| > \eta\}$.

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

By assumption we have that $\ell(h, Z)$ is a σ^2-sub-Gaussian for every h. We can thus use Hoeffding's inequality for every hypothesis $h \in \mathcal{H}$, and retrieve that for every $h \in \mathcal{H}$:

$$\mathcal{P}_S(E_h) \leq 2 \cdot \exp\left(-n\frac{\eta^2}{2\sigma^2}\right). \tag{40}$$

We also have that

$$\mathcal{P}_S\mathcal{P}_{\mathcal{A}(S)}(E) = \mathbb{E}_{\mathcal{P}_{\mathcal{A}(S)}}\left[\mathcal{P}_S(E_{\mathcal{A}(S)})\right] \leq 2 \cdot \exp\left(-n\frac{\eta^2}{2\sigma^2}\right). \tag{41}$$

Then it follows from Corollary 3 and Inequality (41) that:

$$\mathbb{P}(E) \leq \sqrt{(\chi^2(S, \mathcal{A}(S)) + 1)2\exp\left(-n\frac{\eta^2}{2\sigma^2}\right)} \tag{42}$$

$$= \sqrt{2}\exp\left(\frac{1}{2}\left(\log(\chi^2(S, \mathcal{A}(S)) + 1) - n\frac{\eta^2}{2\sigma^2}\right)\right). \tag{43}$$

□

Let us assume that the loss function is the $0-1$ loss, as defined in Eq. (5). Since $\ell(\cdot) \in [0, 1]$, one can show that for every h, $\ell(h, X)$ is $1/4$-sub-Gaussian. Ineq. (39) then becomes:

$$\mathbb{P}(E) \leq \sqrt{2}\exp\left(\frac{1}{2}\left(\log(\chi^2(S, \mathcal{A}(S)) + 1) - 2n\eta^2\right)\right). \tag{44}$$

An implication of (44) is the following: if $\chi^2(S, \mathcal{A}(S)) < \exp(2n\eta^2) - 1$ then we can guarantee an exponential decay in the probability of having a large generalization error. Doing the same with Ineq. (35), one gets:

$$\mathbb{P}(E) \leq \sqrt{2}\exp\left(\frac{1}{2}\left(\mathcal{L}\left(X{\to}Y\right) - 2n\eta^2\right)\right). \tag{45}$$

From Ineq. (38), one has that every time $\exp(\mathcal{L}\left(X{\to}Y\right)) \leq 2n\eta^2$ then $\chi^2(X, Y) \leq \exp(2n\eta^2) - 1$ and thus, generalization with maximal leakage implies generalization with χ^2. An advantage of Corollary 3 is that $\chi^2(X, Y)$ can be significantly smaller than $\mathcal{L}\left(X{\to}Y\right)$. Indeed:

Example 2. *Let $X \sim Ber(1/2)$ and let $Y = BSC(p)$, with $p < 1/2$. Thus, $P_{Y|X=x}(x) = 1 - p$. In this case $\chi^2(X, Y) = (1-2p)^2$ while $\exp(\mathcal{L}\left(X{\to}Y\right)) - 1 = (1-2p)$. It is easy to see that, since $(1-2p) < 1$ then $(1-2p)^2$ can be much smaller than $(1-2p)$.*

Although χ^2 can be smaller, an advantage in using maximal leakage is that leakage depends on $\mathcal{P}_X$ only through the support. This allows us to provide bounds that depend only loosely on the distribution over the training samples. On the other hand, $\chi^2(X, Y)$ cannot be computed unless one has **full** access to $\mathcal{P}_X$. Such distributions can be very complicated and typically defined on large dimensional spaces (*e.g.*, images, audio-recordings, etc.). In general, one only has access to (and control over) the conditional distributions $P_{Y|X}$ induced by the chosen learning algorithm. This can render the usage of bounds like Ineq. (39) difficult in practice, although the results are tighter in theory. Another important characteristic of maximal leakage is that, as a consequence of the chain rule it satisfies, it composes adaptively [4]. Such property is not known to hold, in general, for either $f-$ or Sibson's $\alpha-$Mutual Information. To complete the discussion about χ^2, let us translate Corollary 4 for the $0-1$ loss function into a sample complexity bound.

Corollary 5. *Under the same assumptions of Corollary 4, but considering the $0-1$ loss, in order to ensure a confidence of $\delta \in (0, 1)$, i.e. $\mathbb{P}(E) \leq \delta$, it is sufficient to have m samples where:*

$$m \geq \frac{\log(\chi^2(X, Y) + 1) + 2\log\left(\frac{\sqrt{2}}{\delta}\right)}{2\eta^2}. \tag{46}$$

Although a fundamental quantity, χ^2 divergence (like mutual information and maximal leakage) is not guaranteed to be bounded. An interesting $f-$divergence that is always guaranteed to be bounded is, instead, the squared Hellinger distance. Using such measure could lead to non-trivial bounds even when other measures become infinite. Assuming that $\mathcal{P}_{XY}(E) \geq \mathcal{P}_X\mathcal{P}_Y(E)$ (typical scenario of interest), we can show the following:

Corollary 6. *Let $E \in \mathcal{F}$ and let $\mathcal{P}_{XY}(E) \geq \mathcal{P}_X\mathcal{P}_Y(E)$, we have that:*

$$\mathcal{P}_{XY}(E) - \mathcal{P}_X\mathcal{P}_Y(E) \leq H^2(X;Y) + 2H(X;Y)\sqrt{\mathcal{P}_X\mathcal{P}_Y(E)} \tag{47}$$

where $H^2(X;Y)$ denotes $H^2(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)$.

Proof. The proof follows from Theorem 1 as stated in Ineq. (23) applied with $f(t) = (\sqrt{t} - 1)^2$ with $t \geq 1$. Such f gives us $I_f(X;Y) = H^2(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y) = H^2(X;Y)$, *i.e.*, the squared Hellinger Distance. □

When X and Y are independent, one has that $H(X;Y) = H^2(X;Y) = 0$ and Corollary 6 recovers $\mathcal{P}_{XY}(E) = \mathcal{P}_X\mathcal{P}_Y(E)$. On the other hand, if $Y = X \sim \mathcal{U}([n])$ then, if $E = \{(x, y) \in [n] \times [n] | x = y\}$,

$$1 = \mathcal{P}_{XY}(E) \leq 1 - \frac{1}{n^{3/2}} + 2\sqrt{\left(1 - \frac{1}{n^{3/2}}\right)\frac{1}{n}}. \tag{48}$$

Thus, the bound is asymptotically tight even when Y depends strongly on X. Regardless, the same reasoning that compared Maximal Leakage to χ^2 applies: computing $H(X;Y)$ requires access to the marginal distributions $\mathcal{P}_X, \mathcal{P}_Y$ and can be very complicated. Indeed, even for simple additive noise channels, no closed form expression is known for $H(X^n;Y)$ (or even for $TV(X^n;Y)$). In the context of learning instead, even for simple gradient descent mechanisms [5, Example 2], computing such measures can be very hard. It is, in general, possible to bound the divergence measures for every $\mathcal{P}_X$ (*e.g.*, maximizing over all the possible $\mathcal{P}_X$), but this often implies using Maximal Leakage as a distribution-independent upper-bound on the chosen measure [17, Remark 7], [3, Corollary 2].

REFERENCES

[1] 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.

[2] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," in *Proceedings of Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, F. Janoos, M. Mohri, and K. Sridharan, Eds., vol. 83. PMLR, 07–09 Apr 2018, pp. 25–55.

[3] I. Issa and M. Gastpar, "Computable bounds on the exploration bias," in *2018 IEEE International Symposium on Information Theory, ISIT Vail, CO, USA, June 17-22, 2018*, 2018, pp. 576–580.

[4] A. R. Esposito, M. Gastpar, and I. Issa, "Learning and Adaptive Data Analysis via Maximal Leakage," *IEEE Information Theory Workshop, ITW 2019, Visby, Gotland, Sweden, Aug 25-28*, 2019.

[5] H. Wang, M. Diaz, J. C. S. S. Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 577–581, 2019.

[6] A. T. Lopez and V. S. Jog, "Generalization error bounds using Wasserstein distances," *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2018.

[7] A. Pensia, V. Jog, and P.-L. Loh, "Generalization error bounds for noisy, iterative algorithms," *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550, 2018.

[8] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, pp. 1–1, 2020.

[9] A. R. Esposito, M. Gastpar, and I. Issa, "Robust generalization via α-Mutual Information," *2020 International Zurich Seminar on Information and Communication (IZS), Feb 26-28*, 2020.

[10] 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, A. Gretton and C. C. Robert, Eds., vol. 51. Cadiz, Spain: PMLR, 09–11 May 2016, pp. 1232–1240.

[11] A. R. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," *CoRR*, vol. abs/1806.03803, 2018.

[12] 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)*, July 2019, pp. 582–586.

[13] 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.

[14] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 35–55, 2016.

[15] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Internat. Statist. Rev.*, pp. 419–435, 2002.

[16] F. E. Su, *Methods for Quantifying Rates of Convergence for Random Walks on Groups, PhD Thesis*. Harvard University, 1995.

[17] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-Divergences and Maximal Leakage," 2019.

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

[19] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theor.*, vol. 52, no. 10, pp. 4394–4412, 2006.

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