Generalization Error Bounds via Rényi-, f-Divergences and Maximal Leakage

Amedeo Roberto Esposito, *Student Member, IEEE*, Michael Gastpar, *Fellow, IEEE*, and Ibrahim Issa, *Member, IEEE*

***Abstract*— In this work, the probability of an event under some joint distribution is bounded by measuring it with the product of the marginals instead (which is typically easier to analyze) together with a measure of the dependence between the two random variables. These results find applications in adaptive data analysis, where multiple dependencies are introduced and in learning theory, where they can be employed to bound the generalization error of a learning algorithm. Bounds are given in terms of Sibson's Mutual Information, α-Divergences, Hellinger Divergences, and f-Divergences. A case of particular interest is the Maximal Leakage (or Sibson's Mutual Information of order infinity), since this measure is robust to post-processing and composes adaptively. The corresponding bound can be seen as a generalization of classical bounds, such as Hoeffding's and McDiarmid's inequalities, to the case of dependent random variables.**

***Index Terms*— Sibson's Mutual Information, Rényi-Divergence, f-Divergence, maximal leakage, generalization error, 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. Our aim is to provide bounds of the following form:

$$\mathcal{P}(E) \leq \vartheta(\mathcal{Q}(E)) \cdot \varpi(d\mathcal{P}/d\mathcal{Q}), \tag{1}$$

for some functions ϑ, ϖ. 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. $d\mathcal{P}/d\mathcal{Q}$ denotes the Radon-Nikodym derivative of $\mathcal{P}$ with respect to $\mathcal{Q}$ (assuming it exists). $\varpi(d\mathcal{P}/d\mathcal{Q})$ is often going to be a function of some divergence between $\mathcal{P}$ and $\mathcal{Q}$ (e.g., Kullback-Leibler, Rényi's $\alpha-$Divergence, etc.). 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 (a scenario typically much easier to analyze). Such a result can be applied in the analysis of the generalization error of learning algorithms, as well as in adaptive data analysis (with a proper choice of the dependence measure). Adaptive data analysis is a recent field that is gaining attention due to its connection with the "Reproducibility Crisis" [1], [2]. The idea is that, whenever you apply a sequence of analyses to some data (*e.g.*, data-exploration procedures) and each analysis informs the subsequent ones, even though each of these algorithms is guaranteed to generalize well in *isolation*, this may no longer be true when they are *composed* together. The problem that arises with the composition is believed to be connected with the *leakage* of information from the data. The leakage happens because the output of each algorithm becomes an input to the subsequent ones. In order to be used in adaptive data analysis, a measure that provides such bounds 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 the sequence of algorithms if each of them has bounded measure). Results of this form involving mutual information can be found in [3]–[5]. Via inequalities like in (1) we can provide bounds for adaptive mechanisms by treating them as non-adaptive and paying a "penalty" (*e.g.*, a measure of statistical dependency) that estimates how far is the mechanism from being non-adaptive. With this aim, we first provide general bounds in terms of Luxemburg norms, Amemiya norms, and f-mutual information. As corollaries, we derive several families of interesting bounds in the form of (1) with $\mathcal{P} = \mathcal{P}_{XY}$ and $\mathcal{Q} = \mathcal{P}_X \mathcal{P}_Y$:

- a family of bounds involving Sibson's Mutual Information of order α;
- a bound involving Maximal Leakage [6];
- a family of bounds involving the Rényi's and Hellinger divergences of order α;
- a bound involving Hellinger squared distance.

A representation of our results and their connections is given in Figure 1 below. We focus in particular on the bounds involving Maximal Leakage, which is a secrecy metric that has appeared both in the computer security literature [7], and

Manuscript received December 1, 2019; revised October 15, 2020; accepted April 29, 2021. Date of publication May 31, 2021; date of current version July 14, 2021. This work was supported in part by the Swiss National Science Foundation under Grant 169294 and Grant 200364 and in part by the École Polytechnique Fédérale de Lausanne (EPFL). This article was presented in part at the 2019 IEEE International Symposium on Information Theory, in part at the 2019 IEEE Information Theory Workshop, in part at the 2020 International Zürich Seminar, and in part at the 2020 IEEE International Symposium on Information Theory. *(Corresponding author: Amedeo Roberto Esposito.)*

Amedeo Roberto Esposito and Michael Gastpar are with the School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: amedeo.esposito@epfl.ch; michael.gastpar@epfl.ch).

Ibrahim Issa was with the School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. He is now with the Electrical and Computer Engineering Department, American University of Beirut, Beirut 1107 2020, Lebanon (e-mail: ii19@aub.edu.lb).

Communicated by M. Raginsky, Associate Editor for Probability and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3085190.

Digital Object Identifier 10.1109/TIT.2021.3085190

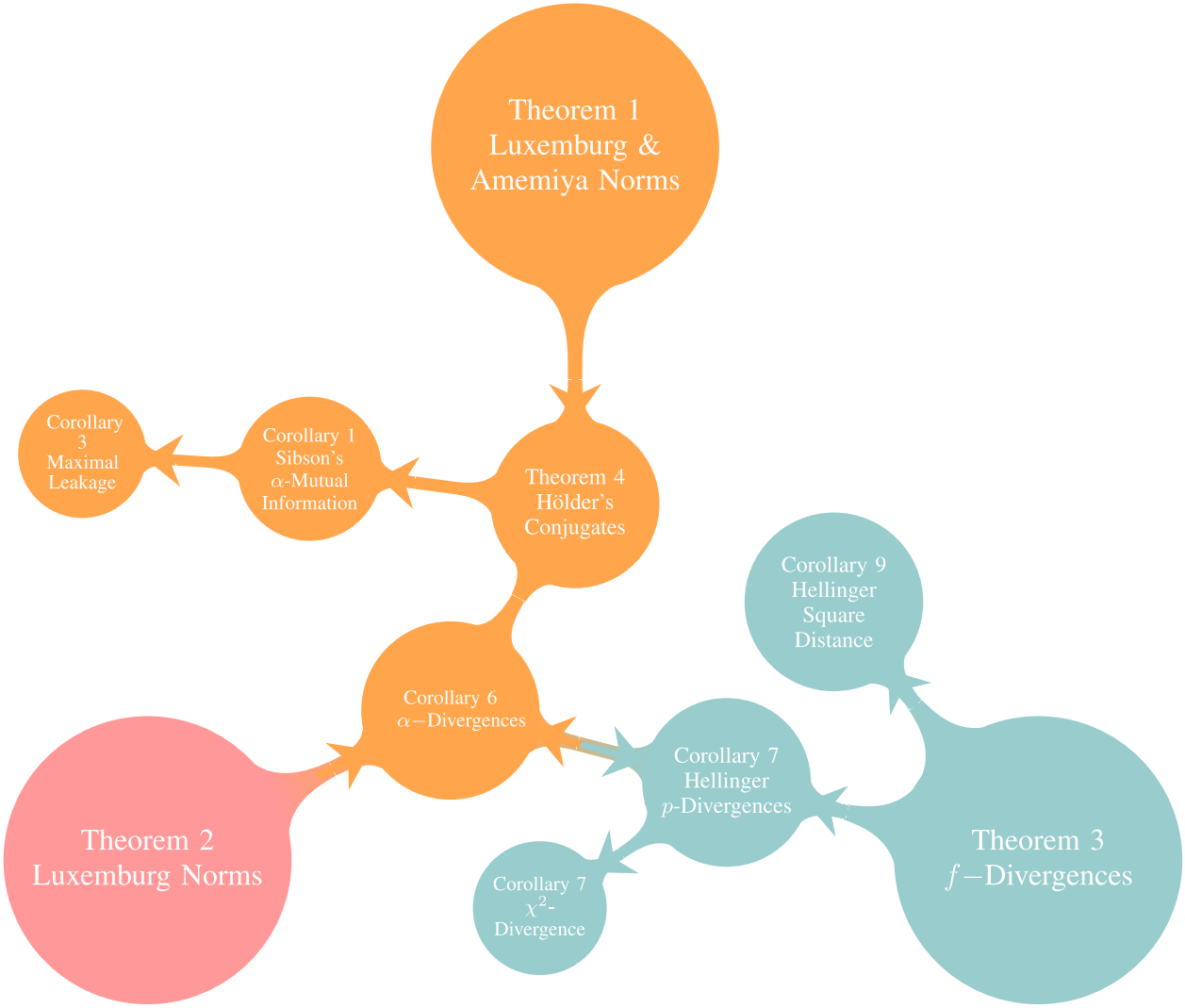

Fig. 1. A graphical representations of our main results and how they connect to each other.

the information theory literature [6]. It quantifies the leakage of information from a random variable X to another random variable Y, and is denoted by $\mathcal{L}(X\to Y)$. The basic insight is as follows: if a learning algorithm leaks little information about the training data, then it will generalize well. Moreover, similarly to differential privacy, maximal leakage behaves well under composition: we can bound the leakage of a sequence of algorithms if each of them has bounded leakage. It is also robust under post-processing. In addition, the expression to compute it is simply given by the following formula (for finite X and Y):

$$\mathcal{L}(X\to Y) = \log \sum_{y} \max_{x:P(x)>0} P_{Y|X}(y|x), \tag{2}$$

making it more amenable to analysis and relatively easy to compute, especially for algorithms whose randomness consists in adding independent noise to the outcomes. Despite the main focus being on a joint distribution and the corresponding product of the marginals, the proof techniques are more general and can be applied to any pair of joint distributions (under a mild condition of absolute continuity). Moreover, the Maximal Leakage result, as well as the bound using infinite-Rényi divergence, reduce to the classical concentration inequalities when independence holds (i.e., $\mathcal{P}_{XY} = \mathcal{P}_X\mathcal{P}_Y$).

A. Further Related Work

In addition to differentially private algorithms, Dwork *et al.* [1] show that algorithms whose output can be described concisely generalize well. They further introduce β-max information to unify the analysis of both classes of algorithms. Consequently, one can provide generalization guarantees for a sequence of algorithms that alternate between differential privacy and short description. In [2], the authors connect β-max information with the notion of approximate differential privacy, but show that there are no generalization guarantees for an arbitrary composition of algorithms that are approximate-DP and algorithms with short description length. With a more information-theoretic approach, bounds on the exploration bias and/or the generalization error are given in [4], [5], [8]–[12], using mutual information and other dependence-measures. Some results have also been found using Wasserstein distance [13], [14].

B. Notation

We will denote by calligraphic letters $\mathcal{P}, \mathcal{Q}$ probability measures and with capital letters X, Y, Z random variables. Given two measures $\mathcal{P}, \mathcal{Q}$, $\mathcal{P} \ll \mathcal{Q}$ denotes the concept of absolute continuity, *i.e.*, for any measurable set E, $\mathcal{Q}(E) = 0 \implies \mathcal{P}(E) = 0$. Given two random variables X, Y over the spaces $\mathcal{X}, \mathcal{Y}$ we will denote by $\mathcal{P}_{XY}$ a joint measure over the product space $\mathcal{X} \times \mathcal{Y}$, while with $\mathcal{P}_X\mathcal{P}_Y$ we will denote the product of the marginals, *i.e.*, for any measurable set $E \subseteq \mathcal{X} \times \mathcal{Y}, \mathcal{P}_X\mathcal{P}_Y(E) = \int_{(x,y)\in E} d\mathcal{P}_X(x)d\mathcal{P}_Y(y)$.

Given a probability measure $\mathcal{P}$ and a random variable X defined over the same space, we will denote with

$$\mathbb{E}_{\mathcal{P}}[X] = \int x d\mathcal{P}(x). \tag{3}$$

Furthermore, given a zero-mean random variable X we say that it is σ^2-sub-Gaussian if the following holds true for every $\lambda \in \mathbb{R}$:

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}. \tag{4}$$

For the remainder of this paper $\log$ is always taken to the base e.

C. Overview

In Section II we define the fundamental objects that will be used in this work:

- In Subsection II-A we consider Rényi's-α Divergences, Sibson's Mutual Information, Maximal Leakage and f-divergences;
- In Subsection II-C we provide an overview of the basic concepts in Learning Theory;

In Section III we prove our most general results:

- Theorem 1 and 2 are the most general one and involves Luxemburg and Amemiya norms;
- Theorem 3 bounds $\mathcal{P}_{XY}(E)$ with a function of $I_f(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)$ and $\mathcal{P}_X\mathcal{P}_Y(E)$;
- Theorem 4 bounds $\mathcal{P}_{XY}(E)$ using norms of $\mathcal{P}_X(E_Y)$ and the Radon-Nikodym derivative $d\mathcal{P}_{XY}/d\mathcal{P}_X\mathcal{P}_Y$.

Then we specialize these results and obtain bounds involving:

- Sibson's $\alpha-$Mutual Information (Section III);
- Maximal Leakage (Section IV);
- Hellinger and α-Divergences (Section V);

In each of these sections we also show how to apply these results to bound the generalization error. In Section VIII we consider the basic definitions of Adaptive Data Analysis and show how some of our results can be employed in the area. To conclude, in Section IX we compare our results with recent results in the literature. Some extension of our bounds to expected generalization error is also considered in Appendix D.

II. Background and Definitions

A. Information Measures

We will now briefly introduce the information measures that we will use to provide bounds. The idea is to try and capture the dependency between two random variables X, Y through some information measure and employ it in order to provide bounds. We will consider X to be the input of a learning algorithm $\mathcal{A}$ and $Y = \mathcal{A}(X)$ the corresponding (random) output. By controlling some measure of dependency, we will control how much the learning algorithm $\mathcal{A}$ is overfitting to the data.

1) Rényi's α-Divergence: Introduced by Rényi in an attempt to generalize the concept of Entropy and KL-Divergence, the α-Divergence has then found many applications over the years in hypothesis testing, guessing and several other statistical inference problems [15]. Indeed, it has several useful operation 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 [16], [17] [18, p. 649]). It can be defined as follows [16]:

Definition 1: Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\alpha > 0$ be a positive real 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:

$$D_\alpha(\mathcal{P}\|\mathcal{Q}) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu. \tag{5}$$

Remark 1: The definition is independent of the chosen measure μ whenever $\infty > \alpha > 0$ and $\alpha \neq 1$. 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 $0 < \alpha < 1$ one has that $\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{p}{q}\right)^{\alpha} d\mathcal{Q}$, see [16].

It can be shown that if $\alpha > 1$ and $\mathcal{P} \not\ll \mathcal{Q}$ then $D_\alpha(\mathcal{P}\|\mathcal{Q}) = \infty$. The behavior of the measure for $\alpha \in \{0, 1, \infty\}$ can be defined by continuity. In particular, we have that $\lim_{\alpha\to 1} D_\alpha(\mathcal{P}\|\mathcal{Q}) = D(\mathcal{P}\|\mathcal{Q})$, *i.e.*, the classical Kullback-Leibler divergence. For an extensive treatment of α-Divergences and their properties we refer the reader to [16].

2) Sibson's α-Mutual Information: Starting from the notion of "information radius", Sibson built a generalization of mutual information that retains many interesting properties [19]. Although defined in a different way Sibson's α-Mutual Information $I_\alpha(X, Y)$[1] can be re-defined in terms of α-Divergences [15]:

Definition 2: Let X, Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$. Let $\mathcal{P}_X$ be the corresponding marginal of X (*i.e.*, given a measurable set A, $\mathcal{P}_X(A) = \mathcal{P}_{XY}(A \times \mathcal{Y})$) and let $\mathcal{Q}_Y$ be any probability measure over $\mathcal{Y}$. Let $\alpha > 0$, the Sibson's Mutual Information of order α between X, Y is defined as:

$$I_\alpha(X, Y) = \min_{Q_Y} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X Q_Y). \tag{6}$$

The following, alternative formulation is also useful [15]:

$$I_\alpha(X, Y) = \frac{\alpha}{\alpha - 1} \log \mathbb{E}\left[\mathbb{E}^{\frac{1}{\alpha}}\left[\frac{\mathcal{P}_{Y|X}}{\mathcal{P}_Y}\middle| Y\right]\right] \tag{7}$$

$$= D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y) - D_\alpha(\mathcal{P}_{Y_\alpha}\|\mathcal{P}_Y), \tag{8}$$

[1]Throughout the work we will denote with a comma **asymmetric** information measures (like $I_\alpha(X, Y)$) and with a semicolon **symmetric** information measures (like $I(X; Y)$).

where $\mathcal{P}_{Y_\alpha}$ is the measure minimizing (6). In analogy with the limiting behavior of α-Divergence we have that $\lim_{\alpha\to 1} I_\alpha(X,Y) = I(X;Y)$ while, when $\alpha \to \infty$ we retrieve the following object:

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

To conclude, let us list some of the properties of the measure:

Proposition 1 [15]:

1) **Data Processing Inequality**: given $\alpha > 0$, $I_\alpha(X,Z) \leq \min\{I_\alpha(X,Y), I_\alpha(Y,Z)\}$ if the Markov Chain $X-Y-Z$ holds;
2) $I_\alpha(X,Y) \geq 0$ with equality iff X and Y are independent;
3) Let $\alpha_1 \leq \alpha_2$ then $I_{\alpha_1}(X,Y) \leq I_{\alpha_2}(X,Y)$;
4) Let $\alpha \in (0,1) \cup (1,\infty)$, for a given $\mathcal{P}_X$, $\frac{1}{\alpha-1}\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X,Y)\right)$ is convex in $\mathcal{P}_{Y|X}$;
5) $I_\alpha(X,Y) \leq \min\{\log|X|, \log|Y|\}$;

For an extensive treatment of Sibson's α-MI we refer the reader to [15].

3) Maximal Leakage: A particularly relevant dependence measure, strongly connected to Sibson's Mutual Information is the maximal leakage, denoted by $\mathcal{L}(X\to Y)$. It was introduced as a way of measuring the leakage of information from X to Y, hence the following definition:

Definition 3 (Def. 1 of [6]): Given a joint distribution $\mathcal{P}_{XY}$ on finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, the maximal leakage from X to Y is defined as:

$$\mathcal{L}(X\to Y) = \sup_{U-X-Y-\hat{U}} \log \frac{\mathbb{P}(\{U=\hat{U}\})}{\max_{u\in\mathcal{U}} \mathbb{P}_U(\{u\})}, \quad (9)$$

where U and $\hat{U}$ take values in the same finite, but arbitrary, alphabet.

It is shown in [6, Theorem 1] that, for finite alphabets:

$$\mathcal{L}(X\to Y) = \log \sum_{y\in\mathcal{Y}} \max_{x\in\mathcal{X}:\mathcal{P}_X(x)>0} P_{Y|X}(y|x). \quad (10)$$

If X and Y have a jointly continuous pdf $f(x,y)$, we get [6, Corollary 4]:

$$\mathcal{L}(X\to Y) = \log \int_{\mathbb{R}} \sup_{x:f_X(x)>0} f_{Y|X}(y|x)dy. \quad (11)$$

One can show that $\mathcal{L}(X\to Y) = I_\infty(X;Y)$ i.e., Maximal Leakage corresponds to the Sibson's Mutual Information of order infinity. This allows the measure to retain the properties listed in Proposition 1, furthermore:

Lemma 1 [6]: For any joint distribution $\mathcal{P}_{XY}$ on finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{L}(X\to Y) \geq I(X;Y)$.

Another relevant notion, important for its application to Adaptive Data Analysis, is Conditional Maximal Leakage:

Definition 4 (Conditional Maximal Leakage [6]): Given a joint distribution P_{XYZ} on alphabets $\mathcal{X},\mathcal{Y}$, and $\mathcal{Z}$, define:

$$\mathcal{L}(X\to Y|Z) = \sup_{U:U-X-Y|Z} \log \frac{\mathbb{P}(\{U=\hat{U}(Y,Z)\})}{\mathbb{P}(\{U=\tilde{U}(Z)\})}, \quad (12)$$

where U takes value in an arbitrary finite alphabet and we consider $\hat{U}, \tilde{U}$ to be the optimal estimators of U given (Y,Z) and Z, respectively.

Again, it is shown in [6] that for discrete random variables X,Y,Z:

$$\mathcal{L}(X\to Y|Z) = \log \max_{z:P_Z(z)>0} \sum_y \max_{x:P_{X|Z}(x|z)>0} P_{Y|XZ}(y|xz),$$

and

$$\mathcal{L}(X\to(Y,Z)) \leq \mathcal{L}(X\to Y) + \mathcal{L}(X\to Z|Y). \quad (13)$$

4) f-Mutual Information: Another 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 5: 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:

$$D_f(\mathcal{P}\|\mathcal{Q}) = \int qf\left(\frac{p}{q}\right) d\mu. \quad (14)$$

Despite the fact that the definition uses μ and the densities with respect to this measure, it is possible to show that f-divergences are actually independent from the dominating measure [20]. Indeed, 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 [20]:

$$D_f(\mathcal{P}\|\mathcal{Q}) = \int f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q}. \quad (15)$$

Denoting with $\mathcal{F}_X$ the Sigma-field generated from the random variable X, (i.e., $\sigma(X)$), f-mutual information is defined as follows:

Definition 6: Let X and Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$ over the 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:

$$I_f(X;Y) = D_f(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y). \quad (16)$$

If $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$ we have that:

$$I_f(X;Y) = \int f\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right) d\mathcal{P}_X\mathcal{P}_Y. \quad (17)$$

It is possible to see that, if f satisfies $f(1) = 0$ and it is strictly convex at 1, then $I_f(X;Y) = 0$ if and only if X and Y are independent [20]. This generalization includes the KL (by simply setting $f(t) = t\log(t)$) and allows to retrieve α-Divergences through a one-to-one mapping. But it also includes many more divergences:

- 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 $f(t) = (t-1)^2$.

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 it may help us circumventing issues that commonly used measures, like Mutual Information, may suffer from. Consider for instance the following example [14]: let S be a random vector, via Strong Data-Processing inequalities it is possible to show that, given the Markov Chain $S-H-Y$, where $\|H\| \leq k$ and $Y = H + N$ with N Gaussian noise, the Total Variation distance between the joint and the product of the marginals of S, Y is strictly less than 1, while $I(S;Y)$ may still be infinite. Furthermore, as presented in [21], different divergences between distributions can provide different convergence rates. It has been proved in [22] that it is possible to construct a random walk that converges in $2n\log n$ steps under KL, $n^2\log n$ steps under the χ^2-distance and $n\log n$ in total variation. This shows that 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 obtainable can be quite different and this can possibly impact the sample complexity in the problems we will analyze in later sections.

B. Orlicz Functions and Luxemburg Norms

Let $(\Omega, \mathcal{F}, \mu)$ be a complete and σ-finite measure space and denote with $L^0(\mu)$ the space of all the $\mathcal{F}$-measurable and real valued functions on Ω. Given an Orlicz function, *i.e.*, a convex function $\psi : [0,+\infty) \to [0,+\infty]$ that vanishes at 0 and is not identically 0 or $+\infty$ over the positive real line we can define a functional $I_\psi : L^0(\mu) \to [0,+\infty]$ as $I_\psi(x) = \int_\Omega \psi(|x(t)|)d\mu(t)$. An Orlicz space can then be defined to be [23]:

$$L_\psi(\mu) = \{x \in L^0(\mu) : I_\psi(\lambda x) < +\infty \text{ for some } \lambda > 0\}. \tag{18}$$

The Orlicz space is a Banach space (a complete normed vector space) that can be endowed with several norms: the Luxemburg, Orlicz and Amemiya norm. It can also be showed that Amemiya norms are equivalent to Orlicz norms in general [23]. For the purposes of this paper, let us restrict ourselves to probability spaces and define the corresponding norms with respect to random variables and the expectation operator. In particular, let U be an $\mathcal{F}$-measurable random variable, we can define the Luxemburg norm of U with respect to μ:

$$\|U\|_\psi^\mu = \inf\left\{\sigma > 0 : \mathbb{E}_\mu\left[\psi\left(\frac{|U|}{\sigma}\right)\right] \leq 1\right\} \tag{19}$$

and the Amemiya norm of U with respect to μ:

$$\|U\|_\psi^{A,\mu} = \inf\left\{\frac{\mathbb{E}_\mu\left[\psi(t|U|)\right] + 1}{t} : t > 0\right\}. \tag{20}$$

When the measure is not explicitly specified it corresponds to the probability measure used to define the space where the random variable lives, although we will often need to be more explicit, as we often apply changes of measure. Given these two quantities and the following definition of convex conjugation, one can show the following generalisation of Hölder's inequality:

Definition 7: Given a convex function $\psi : [0,+\infty) \to \mathbb{R}$, define $\psi^\star : [0,+\infty) \to \mathbb{R}$ as

$$\psi^\star(x) = \sup_{\lambda > 0} \lambda x - \psi(\lambda). \tag{21}$$

Lemma 2 [8]: Let ψ be an Orlicz fucntion and $\psi^\star$ denote its conjugate, then for every couple of random variable U, V:

$$\mathbb{E}[UV] \leq \|U\|_\psi \|V\|_{\psi^\star}^A.$$

With $\psi(t) = t^\alpha/\alpha$ (and, consequently, $\psi^*(t) = t^\gamma/\gamma$, with $\frac{1}{\gamma} + \frac{1}{\alpha} = 1$) one recovers Hölder's inequality. For completeness we included a proof of Lemma 2 in Appendix A.

C. Learning Theory

In this section we will 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 $s \in \mathcal{Z}$ 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.*, $\mathcal{Z} = \mathcal{D} \times \mathcal{C}$ and suppose that $\mathcal{A}$ is fed with n pairs data-label $(d, c) \in \mathcal{Z}$. 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 presented so far. The goal is to generate a hypothesis $h : \mathcal{D} \to \mathcal{C}$ 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 8: 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

$$L_\mathcal{P}(h) = \mathbb{E}_{Z\sim\mathcal{P}}[\ell(h, Z)], \tag{22}$$

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

$$L_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, z_i). \tag{23}$$

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

$$\text{gen-err}_\mathcal{P}(\mathcal{A}, S) = |L_\mathcal{P}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))|. \tag{24}$$

The definition just stated considers general loss functions. An important instance for the case of supervised learning is the 0-1 loss. Suppose again that $\mathcal{Z} = \mathcal{D} \times \mathcal{C}$ and that

$\mathcal{H} = \{h|h : \mathcal{D} \to \mathcal{C}\}$, given a couple $(d, c) \in \mathcal{Z}$ and a hypothesis $h : \mathcal{D} \to \mathcal{C}$ the loss is defined as follows:

$$\ell(h, (d, c)) = \mathbb{1}_{h(d)\neq c}, \tag{25}$$

and the corresponding errors become:

$$L_{\mathcal{P}}(h) = \mathbb{E}_{(d,c)\sim\mathcal{P}}[\mathbb{1}_{h(d)\neq c}] = \mathbb{P}(h(d) \neq c). \tag{26}$$

and

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{h(d_i)\neq c_i}. \tag{27}$$

III. General Results

In this section, we present our main results. First, we prove three general bounds on the probability of an event E under a joint distribution $\mathcal{P}_{XY}$ with respect to its probability under the product of the marginals, using notions of Luxemburg norms, Amemiya norms, and f-mutual information. We subsequently derive several interesting corollaries that employ common information measures such as Sibson's mutual information (Section IV), maximal leakage (Section V), α-divergences and Hellinger divergences (Section VI). A particular focus will be given to the bound using maximal leakage, for reasons discussed in the corresponding subsection.

Our first main bound employs the Luxemburg and Amemiya norms. This result is obtained using the generalised Hölder's inequality.

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 $E \in \mathcal{F}$ and two Orlicz functions ψ, φ:

$$\mathcal{P}_{XY}(E) \leq \left\| \left\| \mathbb{1}_{\{X\in E_Y\}} \right\|_{\varphi}^{\mathcal{P}_X} \right\|_{\psi}^{\mathcal{P}_Y} \left\| \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\varphi^\star}^{A,\mathcal{P}_X} \right\|_{\psi^\star}^{A,\mathcal{P}_Y}, \tag{28}$$

where $\mathbb{1}_{\{\}}$ is the indicator function, and for each $y \in \mathcal{Y}$, $E_y := \{x : (x, y) \in E\}$ (*i.e.*, the "fiber" of E with respect to y), and $\varphi^\star$ and $\psi^\star$ are, respectively, the Legendre-Fenchel duals of φ and ψ.

Proof:

$$\mathcal{P}_{XY}(E) = \mathbb{E}_{\mathcal{P}_{XY}}[\mathbb{1}_E] \tag{29}$$

$$= \mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[\mathbb{1}_E \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right] \tag{30}$$

$$= \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}\left[\mathbb{1}_{\{X\in E_Y\}} \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right]\right] \tag{31}$$

$$\overset{(a)}{\leq} \mathbb{E}_{\mathcal{P}_Y}\left[\left\| \mathbb{1}_{\{X\in E_Y\}} \right\|_{\varphi}^{\mathcal{P}_X} \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\varphi^\star}^{A,\mathcal{P}_X}\right] \tag{32}$$

$$\overset{(b)}{\leq} \left\| \left\| \mathbb{1}_{\{X\in E_Y\}} \right\|_{\varphi}^{\mathcal{P}_X} \right\|_{\psi}^{\mathcal{P}_Y} \left\| \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\varphi^\star}^{A,\mathcal{P}_X} \right\|_{\psi^\star}^{A,\mathcal{P}_Y}, \tag{33}$$

where (a) and (b) follow from Lemma 2, *i.e.*, generalised Hölder's inequality. □

Before investigating special cases of the above theorem (yielding explicit bounds in terms of known information measures), we prove a second result that is in the desired form of Equation (1) and only employs the Luxemburg norm. Differently from the first one, this result is obtained from Young-Fenchel's inequality.

Theorem 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 $E \in \mathcal{F}$ and an Orlicz function ψ:

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot (\psi^{\star\star})^{-1}\left(\frac{1}{\mathcal{P}_X\mathcal{P}_Y(E)}\right) \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\psi}^{\mathcal{P}_X\mathcal{P}_Y} \tag{34}$$

where for $t \geq 0$, $\psi^{-1}(t) := \inf\{s \geq 0 : \psi(s) > t\}$ (and $\psi^{\star\star}$ is defined by applying the transformation in (21) twice).

Proof: Let $\psi^\star : [0, \infty) \to [0, \infty)$ be the transform of ψ defined as follows

$$\psi^\star(t) = \sup_{\lambda>0} \lambda t - \psi(\lambda). \tag{35}$$

Since $\psi(0) = 0$ and ψ is non-negative, it follows that $\psi^\star(0) = 0$. Now, given any $\sigma > 0$ and $t > 0$:

$$\mathcal{P}_{XY}(E) = \mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[\frac{1}{\sigma}\sigma\mathbb{1}_E \frac{\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}}{t} t\right] \tag{36}$$

$$\overset{(a)}{\leq} \frac{t}{\sigma}\mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[\psi^\star(\sigma\mathbb{1}_E) + \psi\left(\frac{\left|\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right|}{t}\right)\right] \tag{37}$$

$$\overset{(b)}{\leq} \frac{t}{\sigma}\left(\psi^\star(\sigma)\mathcal{P}_X\mathcal{P}_Y(E) + \mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[\psi\left(\frac{\left|\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right|}{t}\right)\right]\right), \tag{38}$$

where (a) follows from Young's inequality, and (b) follows from the fact that $\psi^\star(\sigma\mathbb{1}_E) = \mathbb{1}_E\psi^\star(\sigma)$ since $\psi^\star(0) = 0$. Now, by choosing $t = \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\psi}^{\mathcal{P}_X\mathcal{P}_Y}$, we have:

$$\mathcal{P}_{XY}(E) \leq \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\psi}^{\mathcal{P}_X\mathcal{P}_Y} \frac{\psi^\star(\sigma)\mathcal{P}_X\mathcal{P}_Y(E) + 1}{\sigma}. \tag{39}$$

Inequality (39) holds for every $\sigma > 0$, hence we can say that:

$$\mathcal{P}_{XY}(E) \tag{40}$$

$$\leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot \inf_{\sigma>0} \frac{\psi^\star(\sigma) + \frac{1}{\mathcal{P}_X\mathcal{P}_Y(E)}}{\sigma} \cdot \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\psi}^{\mathcal{P}_X\mathcal{P}_Y} \tag{41}$$

$$= \mathcal{P}_X\mathcal{P}_Y(E) \cdot (\psi^{\star\star})^{-1}\left(\frac{1}{\mathcal{P}_X\mathcal{P}_Y(E)}\right) \cdot \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\psi}^{\mathcal{P}_X\mathcal{P}_Y}, \tag{42}$$

where (42) follows from [24, Lemma 2.4] (Note that our $\psi^\star$ plays the role of ψ in Lemma 2.4 and our $\psi^{\star\star}$ plays the role of $\psi^\star$, and the assumption that $\psi^{\star'}(0) = 0$ can be replaced by assuming $\psi^\star$ is non-decreasing which holds true in our case). □

Remark 2: The assumption that ψ is convex, non-decreasing, and non-constant implies that ψ is unbounded, so that ψ^{-1} is well-defined for any t.

Remark 3: With respect to Equation (1) we have that $\mathcal{P} = \mathcal{P}_{XY}, \mathcal{Q} = \mathcal{P}_X\mathcal{P}_Y$, $\vartheta(t) = t(\psi^{\star\star})^{-1}(1/t)$ and $\varpi(t) = \|t\|_{\psi}^{\mathcal{Q}}$. However, Theorem 2 can be applied to any pair of distributions $\mathcal{P}$ and $\mathcal{Q}$ (which do not necessarily correspond to a joint distribution and the product of its marginals).

Remark 4: Note that we defined $\psi^{\star\star}(\lambda)$ as $\sup_{t>0}\{\lambda t - \psi^{\star}(t)\}$. If the supremum was over all $t \in \mathbb{R}$, then we would recover $\psi^{\star\star} = \psi$, but Equation (42) would not necessarily hold. Nevertheless, it is often the case for functions of interest that $\psi^{\star\star}$ (as defined) is equal to ψ, so that the bound becomes

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot \psi^{-1}\left(\frac{1}{\mathcal{P}_X\mathcal{P}_Y(E)}\right) \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{\psi}^{\mathcal{P}_X\mathcal{P}_Y}. \tag{43}$$

Theorem 3: Let $f : [0, +\infty) \to \mathbb{R}$ be a convex function such that $f(1) = 0$, and assume f is non-decreasing on $[0, +\infty)$. Suppose also that f is unbounded, *i.e.*, the generalized inverse, defined as $f^{-1}(y) = \inf\{t \geq 0 : f(t) > y\}$, exists. Given an event $E \in \mathcal{F}$, we have that:

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot f^{-1}\left(\frac{I_f(X,Y) + (1 - \mathcal{P}_X\mathcal{P}_Y(E))f^{\star}(0)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right), \tag{44}$$

where $f^{\star}$ is the Legendre-Fenchel dual of f. Moreover, if $f^{\star}(0) \leq 0$, the bound simplifies to

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot f^{-1}\left(\frac{I_f(X,Y)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right). \tag{45}$$

Proof: Let us denote with $p = \mathcal{P}_{XY}(E), q = \mathcal{P}_X\mathcal{P}_Y(E), \bar{p} = 1 - \mathcal{P}_{XY}(E), \bar{q} = 1 - \mathcal{P}_X\mathcal{P}_Y(E)$. For every $y \geq 0$ we have

$$\begin{aligned} I_f(X,Y) &= D_f(\mathcal{P}_{XY} \| \mathcal{P}_X\mathcal{P}_Y) && (46)\\ &\overset{(a)}{\geq} D_f(\text{Ber}(p) \| \text{Ber}(q)) && (47)\\ &= qf\left(\frac{p}{q}\right) + \bar{q}f\left(\frac{\bar{p}}{\bar{q}}\right) && (48)\\ &\overset{(b)}{\geq} qf\left(\frac{p}{q}\right) + \bar{q}\left(\frac{\bar{p}}{\bar{q}}y - f^{\star}(y)\right) && (49) \end{aligned}$$

where (a) follows from the Data-Processing Inequality for f-divergences and (b) follows from Young's inequality. Choosing $y = 0$ in (49) and re-arranging the terms we retrieve

$$\begin{aligned} &\frac{I_f(X,Y) + \bar{q}f^{\star}(0)}{q} \geq f\left(\frac{p}{q}\right) \iff && (50)\\ &qf^{-1}\left(\frac{I_f(X,Y) + \bar{q}f^{\star}(0)}{q}\right) \geq p. && (51) \end{aligned}$$

□

Remark 5: Alternative proofs can be constructed using the variational representation of f-divergences for a convex function f or an approach similar to the proof of Theorem 2. The interested reader can find these proofs in the Appendix of the arxiv version of this paper [25].

While Theorems 1 and 2 are quite general, computing the Luxemburg or the Amemiya norm can be difficult for most functions. Moreover, our purpose is to retrieve, on the right-hand side of $\mathcal{P}_{XY}(E)$ some function of $\mathcal{P}_X\mathcal{P}_Y(E)$ and an information measure. With this drive, we will now compute some specific instances of these results for certain choices of φ and ψ or f, that allow us to retrieve well-known objects in information theory. The first result we derive is a specific instance of Theorem 1 but it will still be quite general. In particular, it will depend on four parameters $\alpha, \alpha', \gamma, \gamma'$. Different choices of these parameters give rise to bounds involving different Rényi information measures.

Theorem 4: 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 $E \in \mathcal{F}$ and $y \in \mathcal{Y}$, let $E_y = \{x : (x, y) \in E\}$, *i.e.* the "fibers" of E with respect to y. Then,

$$\mathcal{P}_{XY}(E) \leq \mathbb{E}_{\mathcal{P}_Y}^{1/\gamma'}\left[\mathcal{P}_X(E_Y)^{\gamma'/\gamma}\right] \cdot \mathbb{E}_{\mathcal{P}_Y}^{1/\alpha'}\left[\mathbb{E}_{\mathcal{P}_X}^{\alpha'/\alpha}\left[\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right], \tag{52}$$

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

Remark 6: A proof of this result follows from Theorem 1 choosing $\varphi(t) = \frac{t^{\gamma}}{\gamma}$ and $\psi(t) = \frac{t^{\gamma'}}{\gamma'}$ with $\gamma, \gamma' \geq 1$. A more explicit proof can be written using the classical Hölder's inequality twice (similarly to the proof of Theorem 1): once for $\mathcal{P}_X$ and once for $\mathcal{P}_Y$.

Remark 7: It is clear from the proof that one can similarly bound $\mathbb{E}[\hat{g}(X,Y)]$ (instead of $\mathbb{E}[\mathbb{1}_E]$) for any positive function $\hat{g}(X,Y)$ that 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 $\hat{g}(X,Y)^{\gamma} = \hat{g}(X,Y)$ for every $\gamma \geq 1$.

IV. Sibson's Mutual Information

Starting from Theorem 4 and considering the limit as $\alpha' \to 1$, which implies $\gamma' \to +\infty$, we retrieve a bound in terms of Sibson mutual information:

Corollary 1: Given $E \in \mathcal{F}$, we have that:

$$\begin{aligned} &\mathcal{P}_{XY}(E)\\ &\leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_y} \mathcal{P}_X(E_Y)\right)^{1/\gamma} \cdot \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}^{1/\alpha}\left[\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_Y d\mathcal{P}_X}\right)^{\alpha}\right]\right] && (53)\\ &= \left(\operatorname*{ess\,sup}_{\mathcal{P}_y} \mathcal{P}_X(E_Y)\right)^{1/\gamma} \cdot \exp\left(\frac{\alpha - 1}{\alpha} I_{\alpha}(X,Y)\right), && (54) \end{aligned}$$

where $I_{\alpha}(X,Y)$ is the Sibson mutual information of order α [15], and α and γ satisfy $\frac{1}{\alpha} + \frac{1}{\gamma} = 1$.

Remark 8: An in-depth study of α-Mutual Information appears in [15], where a slightly different notation is used. For reference, we can restate Equation (53) in the notation of [15] to obtain:

$$\mathcal{P}_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_y} \mathcal{P}_X(E_Y)\right)^{1/\gamma} \cdot \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}^{1/\alpha}\left[\left(\frac{dP_{Y|X}}{d\mathcal{P}_Y}\right)^{\alpha} \middle| Y\right]\right]. \tag{55}$$

Given that α and γ are Hölder's conjugates, the bound in (54) can be rewritten as:

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

An interesting property of Sibson α-Mutual Information is that it is non-decreasing with respect to α [15]. Considering the right hand side of (56) we have that, for $\alpha_1 \leq \alpha_2$:

$$\frac{\alpha_1 - 1}{\alpha_1} I_{\alpha_1}(X,Y) \leq \frac{\alpha_2 - 1}{\alpha_2} I_{\alpha_2}(X,Y), \tag{56}$$

thus, choosing a smaller α yields a better dependence on $I_\alpha(X,Y)$ in the bound; but given that $\frac{\alpha_1-1}{\alpha_1} \leq \frac{\alpha_2}{\alpha_2-1}$ and $\log \operatorname{ess\,sup}_{\mathcal{P}_y} \mathcal{P}_X(E_Y) \leq 0$, the second term increases for smaller values of α. This leads to a trade-off between the two quantities. We will now explore an interesting application of Corollary 1 that comes from the field of learning theory: generalization error bounds. In such applications, $\mathcal{P}_X(E_y)$ is typically exponentially decaying with the number of samples for every y. Moreover, a different perspective on generalization error bounds, i.e., sample complexity bounds, allow us to see the trade-off between different values of α more explicitly.

A. Generalization Error Bounds

Consider now the learning setup as defined in Section II-C. The next result can be used to give a concentration bound on the generalization error defined in Equation (24):

Corollary 2: 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^2-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,

$$\mathbb{P}(E) \leq 2\exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S,\mathcal{A}(S)) - n\frac{\eta^2}{2\sigma^2}\right)\right). \tag{57}$$

Consequently, in order to ensure a confidence of $\delta \in (0,1)$, i.e. $\mathbb{P}(E) \leq \delta$, it is sufficient to have n samples where

$$n \geq \frac{2\sigma^2}{\eta^2}\left(I_\alpha(S,\mathcal{A}(S)) + \log 2 + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{\delta}\right)\right). \tag{58}$$

Remark 9: The corollary applies to the special case in which $\mathcal{Z} = \mathcal{D} \times \mathcal{C}$ and ℓ is the 0-1 loss function as defined in (25). Indeed, one can show that ℓ is σ^2-sub-Gaussian for $\sigma = \frac{1}{2}$. Moreover, in this case we only need to assume that the samples S are independent, as the use of Hoeffding's inequality in the proof below can be replaced by McDiarmid's inequality (for functions with bounded differences).

Smaller α means that $I_\alpha(S,\mathcal{A}(S))$ will be smaller, but it will imply a worse dependency on $\log(1/\delta)$ in the sample complexity. It is worth noticing that, for fixed α, the sample complexity dependency on $\log(1/\delta)$ is optimal (up to constants). In particular, consider the setup of PAC learning with finite $\mathcal{H}$ with VC dimension d, e.g, assume that $\mathcal{D} = [d]$ and $\mathcal{H} = \{0,1\}^{\mathcal{D}}$, we have that the VC-dimension of $\mathcal{H}$ is d [3], [26]. By [26, Theorem 6.8], we know that the number of necessary samples for learning, in the realizable case, satisfies $n \geq c\frac{d+\log(1/\delta)}{\eta}$, for some constant c. Assume also that $\mathcal{A}$ is the ERM algorithm, in which case the generalization error and the true error are the same and $I_\alpha(S,\mathcal{A}(S)) \leq \log(|\mathcal{H}|) = d$. From (58) for the 0-1 loss we have that $n \geq c\frac{d+\gamma\log(1/\delta)}{\eta^2}$, where $\gamma = \frac{\alpha}{\alpha-1}$. Hence, for a given α, the dependency on δ is optimal. A similar reasoning could be applied to the agnostic case in order to tackle the optimality with respect to η as well.

Proof of Corollary 2: 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 $\ell(h,Z)$ is σ^2-sub-Gaussian for every h. We can thus use Hoeffding's inequality and retrieve that for every $h \in \mathcal{H}$:

$$\mathcal{P}_S(E_h) \leq 2 \cdot \exp\left(-n\frac{\eta^2}{2\sigma^2}\right). \tag{59}$$

Then it follows from Corollary 1 and Inequality (59) that:

$$\begin{aligned}\mathbb{P}(E) &\leq \exp\left(\frac{\alpha-1}{\alpha}I_\alpha(S,\mathcal{A}(S))\right) \cdot \left(2\exp\left(-n\frac{\eta^2}{2\sigma^2}\right)\right)^{\frac{1}{\gamma}} \\ &= 2\exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S,\mathcal{A}(S)) - n\frac{\eta^2}{2\sigma^2}\right)\right). \end{aligned} \tag{60}$$

□

V. Maximal Leakage

An interesting special case of Corollary 1 is to let $\alpha \to \infty$. In this scenario, in the right-hand side of Equation (54) we obtain Maximal Leakage [6]. Maximal Leakage has gained growing interest in the last few years and enjoys a series of properties that are of particular interest to us and we will soon analyse. The result will be thus stated independently. Note that considering the other extreme, *i.e.,* $\alpha \to 1$ we retrieve a trivial bound. Indeed, letting $\alpha \to 1$ in any of our results leads to a bound of 1 on $\mathcal{P}_{XY}(E)$. This means that our approach does not provide bounds that exploit either the Kullback-Leibler divergence or the Mutual Information. Nonetheless, we will provide some comparison with an analogous result obtained for Mutual Information (although, derived with a different approach [3]) in Section IX-A.

Corollary 3: Given $E \in \mathcal{F}$, we have that:

$$\mathcal{P}_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_y} \mathcal{P}_X(E_Y)\right) \exp\left(\mathcal{L}\left(X \to Y\right)\right). \tag{61}$$

Proof: The proof follows directly from Corollary 1 and by noting that when $\alpha \to \infty$ then $\gamma \to 1$ and $\mathcal{L}\left(X \to Y\right) = I_\infty(X,Y)$ [6]. □

Remark 10: Corollary 3 can also be proven independently going through the equivalent formulation of D_∞ [16, Theorem 6] and the fact that $\exp(\mathcal{L}\left(X \to Y\right)) = \mathbb{E}_{\mathcal{P}_Y}[\exp(D_\infty(\mathcal{P}_{Y|X=x} \| \mathcal{P}_Y))]$.

This result is particularly useful for the following reasons:

- Maximal Leakage is more amenable to analysis due to its semi-closed form (e.g., it is possible to easily compute the maximal leakage of noise-addition mechanisms);
- The absence of the power $\frac{1}{\gamma}$ in (61) as compared to the right-hand side of (54) allows us to provide a generalization of the classical concentration of measure results in adaptive settings;

- A conditional version of Maximal Leakage allows us to provide adaptive composition results (discussed in Section VIII).

Before discussing examples in which we analyse (simple) schemes with maximal leakage, we first discuss the tightness of the bound.

A. Tightness

We illustrate the bound by first giving three examples where Inequality (61) is met with equality for varying scenarios of dependence: X is independent from Y, X and Y are equal, and X and Y are related but not equal.

Example 1 (Independent Case): Suppose that E is such that $\mathcal{P}_X(E_y) = \zeta$ for all $y \in \mathcal{Y}$. In that case we have that, if X and Y are independent:

$$\zeta = \mathbb{E}_{\mathcal{P}_Y}[\mathcal{P}_X(E_y)] = \mathcal{P}_{XY}(E) \leq \zeta. \tag{62}$$

Example 2 (Strongly Dependent Case): Consider the example presented in [3]: suppose $X = Y \sim \mathcal{U}([n])$ then we have that $\mathcal{L}(X{\to}Y) = \log n$ and if $E = \{(x, y) \in [n] \times [n] | x = y\}$ then,

$$1 = \mathcal{P}_{XY}(E) \leq \frac{1}{n} \cdot n = 1. \tag{63}$$

Example 3: Suppose (X, Y) is a doubly-symmetric binary source with parameter p for some $p < 1/2$. Let $E = \{(x, y) : x = y\}$. Then,

$$1{-}p = \mathcal{P}_{XY}(E) \leq \frac{1}{2}(2(1-p)) = 1 - p. \tag{64}$$

The above examples show that when the worst-case behavior (i.e., $\max_y \mathcal{P}_X(E_y)$) matches with the average-case behavior (i.e., $\mathbb{E}_{\mathcal{P}_Y}[\mathcal{P}_X(E_y)] = \mathcal{P}_X\mathcal{P}_Y(E)$), our bound represents a generalization of the classical concentration of measure inequalities for adaptive settings. This is typically the case in learning scenarios of interest, where we generalise Hoeffding's and McDiarmid's inequalities.

Moreover, the following proposition shows that the bound is tight in the following strong sense: if we want to bound the ratio $\mathcal{P}_{XY}(E)/(\operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_y))$ as a function of $\mathcal{P}_{Y|X}$ only (i.e., independently of $\mathcal{P}_X$ and E), then $\exp\{\mathcal{L}(X{\to}Y)\}$ is the best bound we could get:

Proposition 2: Given finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and a fixed conditional distribution $\mathcal{P}_{Y|X}$, then there exists $\mathcal{P}_X$ and E such that (61) is met with equality. That is,

$$\sup_{E \subseteq \mathcal{X}\times\mathcal{Y}} \sup_{\mathcal{P}_X} \log \frac{\mathcal{P}_{XY}(E)}{\operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_Y)} = \mathcal{L}(X{\to}Y). \tag{65}$$

Proof: Define a function $g : \mathcal{Y} \to \mathcal{X}$ such that $g(y) \in \operatorname{argmax}_{x\in\mathcal{X}} \mathcal{P}_{Y|X}(y|x)$, and let $\mathcal{X}_g \subseteq \mathcal{X}$ be the image of g. Now, let $\mathcal{P}_X$ be the uniform distribution over $\mathcal{X}_g$, and $E = \{(x, y) : x = g(y)\}$. Then, for any $y \in \mathcal{Y}$,

$$E_y = \{g(y)\} \Rightarrow \mathcal{P}_X(E_y) = \frac{1}{|\mathcal{X}_g|}. \tag{66}$$

So we get

$$\mathcal{P}_{XY}(E) = \sum_{(x,y)\in E} \mathcal{P}_{XY}(x, y) \tag{67}$$

$$= \sum_{y\in\mathcal{Y}} \sum_{x\in E_y} \mathcal{P}_X(x)\ \mathcal{P}_{Y|X}(y|x) \tag{68}$$

$$= \sum_{y\in\mathcal{Y}} \mathcal{P}_X(g(y))\mathcal{P}_{Y|X}(y|g(y)) \tag{69}$$

$$= \frac{1}{|\mathcal{X}_g|} \sum_{y\in\mathcal{Y}} \max_x \mathcal{P}_{Y|X}(y|x), \tag{70}$$

where the last equality follows from (66) and the definition of g. □

B. Generalization Error Bounds

We will now explore how this result can be applied in providing bounds on the generalization error of learning algorithms.

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 σ^2-sub-Gaussian random variable for every $h \in \mathcal{H}$. Given $\eta \in (0, 1)$, let $E = \{(S, h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Then,

$$\mathbb{P}(E) \leq 2 \cdot \exp\left(\mathcal{L}(S{\to}\mathcal{A}(S)) - n\frac{\eta^2}{2\sigma^2}\right). \tag{71}$$

Consequently, in order to ensure a confidence of $\delta \in (0, 1)$, i.e. $\mathbb{P}(E) \leq \delta$, it is sufficient to have n samples where

$$n \geq \frac{2\sigma^2}{\eta^2}\left(\mathcal{L}(S{\to}\mathcal{A}(S)) + \log\left(\frac{2}{\delta}\right)\right). \tag{72}$$

The proof follows from Corollary 3 and the same technique used to prove Corollary 2.

Remark 11: Similarly to Corollary 2, this bound applies to the case in which ℓ is the 0-1 loss function with $\sigma = \frac{1}{2}$. Moreover, as discussed following Corollary 2, the dependence on $\log(1/\delta)$ is optimal. Indeed, let us consider the very same example as in the previous section. Let $\mathcal{D} = [d]$ and $\mathcal{H} = \{0, 1\}^{\mathcal{D}}$, we have that the VC-dimension of $\mathcal{H}$ is d. Choosing again $\mathcal{A}$ to be the ERM algorithm we have that $\mathcal{L}(S{\to}\mathcal{A}(S)) = d$. Looking at (72) with $\sigma = 1/2$, we can see how the lack of the $\alpha/(\alpha - 1)$ term (that one can find in (58) instead) allows us to make a clear analogy with the VC-dimension bound stated in [26, Theorem 6.8]. More precisely, from (72) we have that $n \geq \frac{d+\log(2/\delta)}{2\eta^2}$ while [26, Theorem 6.8.3] (realizable case) tells us that $n \geq c\frac{d+\log(1/\delta)}{\eta}$ for some constant c. A similar reasoning could be applied to the agnostic case in order to tackle the optimality with respect to η as well.

Whenever $\mathcal{A}$ is independent from the samples S, we have that $\exp(\mathcal{L}(S{\to}\mathcal{A}(S))) = 1$ and we immediately fall back to the non-adaptive scenario: $\mathbb{P}(E) \leq 2 \cdot \exp\left(-n\frac{\eta^2}{2\sigma^2}\right)$ i.e., Hoeffding's inequality.

C. Analyzing Schemes via Maximal Leakage

A simple way of keeping the Maximal Leakage of an algorithm $\mathcal{A}(X)$ bounded (and thus ensure generalization) is to add noise (e.g., $\hat{Y} = \mathcal{A}(X)+N$ with $\mathcal{A}$ a real-valued function). The proofs for this section can be found in Appendix C.

Lemma 3 (Laplacian Noise): Let $h : \mathcal{X}^n \to \mathbb{R}$ be a function such that $h(x) \in [a, c], a < c \ \forall x \in \mathcal{X}^n$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim Lap(b)$ is such that:

$$\mathcal{L}\left(X \to \mathcal{M}(X)\right) = \log\left(1 + \frac{(c-a)}{b}\right). \tag{73}$$

Similar results can be obtained analyzing different types of noise.

Lemma 4 (Gaussian Noise): Let $h : \mathcal{X}^n \to \mathbb{R}$ be a function such that $\forall x \in \mathcal{X}^n \ h(x) \in [a, c], a < c$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim \mathcal{N}(0, \sigma^2)$ is such that:

$$\mathcal{L}\left(X \to \mathcal{M}(X)\right) = \log\left(1 + \frac{(c-a)}{\sqrt{2\pi\sigma^2}}\right). \tag{74}$$

Lemma 5 (Exponential Noise): Let $h : \mathcal{X}^n \to \mathbb{R}$ be a function such that $\forall x \in \mathcal{X}^n \ h(x) \in [a, c], \ c > 0$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim Exp(\lambda)$ (i.e., $\mathbb{E}[N] = (1/\lambda) = b$) is such that:

$$\mathcal{L}\left(X \to \mathcal{M}(X)\right) = \log\left(1 + \frac{(c-a)}{b}\right). \tag{75}$$

The addition of carefully calibrated noise to control maximal leakage can be used in practice to obtain generalization guarantees of learning algorithms. As an exact analogy to [4, Corollary 4] we can state the following corollary, involving a noisy version of the Empirical Risk Minimization (ERM) algorithm.

Corollary 5: Let us consider the following algorithm:

$$\mathcal{A}(S) = \arg\min_{h \in \mathcal{H}} (L_S(h) + N_h), \tag{76}$$

where N_h is exponential noise drawn independently from the input, added to the empirical risk of each hypothesis on a given data-set S. Suppose $\mathcal{H}$ is countable (i.e., finite or countably infinite), and denote with N_i the noise added to the hypothesis h_i with mean b_i. Then, for every $\eta \in (0, 1)$:

$$\mathbb{P}(\text{gen-err}(\mathcal{A}) \geq \eta) \leq 2\exp\left(\sum_{i=1}^{|\mathcal{H}|} \log\left(1 + \frac{1}{b_i}\right) - 2n\eta^2\right). \tag{77}$$

Choosing $b_i = i^{1.1}/n^{1/3}$, we retrieve:

$$\mathbb{P}(\text{gen-err}(\mathcal{A}) \geq \eta) \leq 2\exp\left(-n(2\eta^2 - 11/n^{2/3})\right). \tag{78}$$

This example shows how simply the maximal leakage bound can be used, in contrast with the mutual information one. Indeed, following the proof of [4, Corollary 4], the mutual information of the same mechanism analysed here is hard to compute directly and the quantity $I(S; H)$ is, in the end, effectively upper-bounded using maximal leakage:

$$I(S; H) \leq \sum_{i=1}^{|\mathcal{H}|} \log\left(1 + \frac{L_\mu(h_i)}{b_i}\right) \tag{79}$$

$$\leq \sum_{i=1}^{|\mathcal{H}|} \log\left(1 + \frac{1}{b_i}\right) = \mathcal{L}\left(S \to H\right). \tag{80}$$

Remark 12: The noisy version of the ERM algorithm does not provide the same guarantees (with respect to the classical ERM) in terms of training error. Having a small generalization error means that the training and testing error are close, but they could both be large. In this case, adding noise we reduce the information measure (by DPI) and as a consequence of our bounds, the generalization error is also reduced. The addition of noise, however, can increase the training error of the new algorithm. In particular, we will no longer choose the empirical loss minimizer ($h^\star = \arg\min_{h \in \mathcal{H}} L_S(h)$) but some hypothesis h that minimizes $L_S(h) + N_h$ and whose error is more or less close to $L_S(h^\star)$, depending on the noise. Hence, while the generalization error may be smaller, both training and testing error could actually be getting larger.

VI. Hellinger and α-Divergences

In this section, we demonstrate new bounds in terms of α-Divergences and f-Divergences, and in particular, Hellinger divergences.

Choosing $\alpha' = \alpha$ and thus $\gamma' = \gamma$ in Theorem 4, we retrieve:

Corollary 6: Given $E \in \mathcal{F}$ and $\alpha > 1$, we have that:

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

This result can, in fact, be proven in several alternative ways:

- via the data processing inequality for D_α;
- via Theorem 2 with $\psi(t) = \frac{t^\alpha}{\alpha}$;
- via Theorem 3 with $f_\alpha(t) = (t^\alpha - 1)/(\alpha - 1)$ (which gives a bound in terms of Hellinger divergences that are in one-to-one mapping with α-divergences [27, Eq. (80)]).

The interested reader can find the alternative proofs in the Appendix of the arxiv version of this paper [25]. With respect to Equation (1) we again have that $\mathcal{P} = \mathcal{P}_{XY}$ and $\mathcal{Q} = \mathcal{P}_X\mathcal{P}_Y$ but in this case $\vartheta(t) = t^{1/\gamma}$ and $\varpi(\mathcal{P}/\mathcal{Q})$ does involve a divergence. In particular, $\varpi(t) = \exp\left(\frac{1}{\alpha}\log \mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[t^\alpha\right]\right)$. As Hellinger divergences are of independent interest, and include important objects, like the χ^2-divergence, we restate the bound explicitly in terms of Hellinger divergences. Recall that Hellinger divergences can be characterized by $f_p(t) = (t^p - 1)/(p - 1)$ with $p \in (0, 1) \cup (1, +\infty)$ [27, Eq. (53)]. Theorem 3 can, however, only be applied to $p \in (1, +\infty)$, as $f_p(\cdot)$ is concave for $p \in (0, 1)$. Let $\mathcal{H}_p(X, Y)$ denote the Hellinger divergence of $\mathcal{P}_{XY}$ from $\mathcal{P}_X\mathcal{P}_Y$ and with a slight abuse of notation let $\chi^2(X, Y)$ denote $\chi^2(\mathcal{P}_{XY} || \mathcal{P}_X\mathcal{P}_Y)$. We can now state the following.

Corollary 7: Let $E \subseteq \mathcal{X} \times \mathcal{Y}$ and let $p \in (1, +\infty)$ then $I_{f_p}(X, Y) = \mathcal{H}_p(X, Y)$ and

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E)^{\frac{p-1}{p}} \cdot \left((p-1)\mathcal{H}_p(X, Y) + 1\right)^{1/p}. \tag{81}$$

In particular, for $p = 2$, we have

$$\mathcal{P}_{XY}(E) \leq \sqrt{(\chi^2(X, Y) + 1)\mathcal{P}_X\mathcal{P}_Y(E)} \tag{82}$$

$$\leq \sqrt{\exp\left(\mathcal{L}\left(X \to Y\right)\right)\mathcal{P}_X\mathcal{P}_Y(E)}. \tag{83}$$

The last inequality follows from the fact that $\chi^2(X, Y) \leq \exp\left(\mathcal{L}\left(X \to Y\right)\right) - 1$ (cf. [9]). Applying this result to a learning setting we get:

Corollary 8: 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, for some σ and for every $h\in\mathcal{H}$. Given $\eta\in(0,1)$, let $E=\{(S,h):|L_{\mathcal{P}}(h)-L_S(h)|>\eta\}$ and $p\in(1,+\infty)$. Then,

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

In particular,

$$\mathbb{P}(E)\leq\sqrt{2}\exp\left(\frac{1}{2}\left(\log(\chi^2(S,\mathcal{A}(S))+1)-\frac{n\eta^2}{2\sigma^2}\right)\right), \tag{84}$$

and in order to ensure a confidence of $\delta\in(0,1)$, *i.e.* $\mathbb{P}(E)\leq\delta$, it is sufficient to have n samples where

$$n\geq\frac{2\sigma^2}{\eta^2}\left(\log(\chi^2(S,\mathcal{A}(S))+1)+2\log\left(\frac{\sqrt{2}}{\delta}\right)\right).$$

Remark 13: As before, this result applies to 0-1 loss functions with $\sigma=\frac{1}{2}$.

An implication of (84), say for the 0-1 loss function, 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 Inequality (83), one gets:

$$\mathbb{P}(E)\leq\sqrt{2}\exp\left(\frac{1}{2}\left(\mathcal{L}\left(S\to\mathcal{A}(S)\right)-2n\eta^2\right)\right). \tag{85}$$

Given the relationship between these two measures, one has that every time $\exp(\mathcal{L}(X\to Y))\leq 2n\eta^2$ 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 using $\chi^2(X,Y)$ is that it can be significantly smaller than $\mathcal{L}(X\to Y)$. Indeed:

Example 4: Let $X\sim\text{Ber}(1/2)$ and let $Y=\text{BSC}(p)$, with $p<1/2$. Thus, $P_{Y|X=x}(x)=1-p$. In this case $\chi^2(X,Y)=(1-2p)^2$ while $\exp(\mathcal{L}(X\to Y))-1=(1-2p)$. It is easy to see that, since $(1-2p)<1$ then $(1-2p)^2$ can be much smaller than $(1-2p)$.

On the other hand, an advantage in using maximal leakage is that it 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. $\chi^2(X,Y)$, instead, 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 $\mathcal{P}_{Y|X}$ induced by the chosen learning algorithm. This can render the usage of bounds like Inequality (84) difficult in practice, although tighter in theory. Another important characteristic of maximal leakage is that, as a consequence of the chain rule it satisfies, it composes adaptively (more details in Section VIII). Such property is not known to hold, in general, for either f- or Sibson's α-Mutual Information.

Assuming that $\mathcal{P}_{XY}(E)\geq\mathcal{P}_X\mathcal{P}_Y(E)$ (typical scenario of interest), we can show the following:

Corollary 9: Let $E\in\mathcal{F}$ and assume that $\mathcal{P}_{XY}(E)\geq\mathcal{P}_X\mathcal{P}_Y(E)$, we have that:

$$\mathcal{P}_{XY}(E)-\mathcal{P}_X\mathcal{P}_Y(E)\leq H^2(X;Y)+2H(X;Y)\sqrt{\mathcal{P}_X\mathcal{P}_Y(E)}, \tag{86}$$

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

Proof: Let $f(t)=(\sqrt{t}-1)^2$. We have that $I_f(X,Y)=H^2(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)=H^2(X;Y)$, *i.e.*, the squared Hellinger Distance of the joint from the product of the marginals. Moreover, $f(t)$ is strictly increasing and invertible when restricted to $[1,+\infty)$ and $f^{-1}(t)=t+1+2\sqrt{t}$. (86) follows from Theorem 3 as stated in (45). In particular: starting from Inequality (50), we have that the inverse of f is applied to an inequality of the form $c\geq f\left(\frac{\mathcal{P}_{XY}(E)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right)$. Given that $\frac{\mathcal{P}_{XY}(E)}{\mathcal{P}_X\mathcal{P}_Y(E)}\geq 1$ by assumption and using the invertibility of f on $[+1,\infty)$ we recover (86) after some algebraic manipulations. □

When X and Y are independent, one has that $H(X;Y)=H^2(X;Y)=0$ and Corollary 9 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 $E=\{(x,y)\in[n]\times[n]|x=y\}$,

$$1=\mathcal{P}_{XY}(E)\leq 1-\frac{1}{n^{3/2}}+2\sqrt{\left(1-\frac{1}{n^{3/2}}\right)\frac{1}{n}}. \tag{87}$$

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 [14, 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. To conclude this section, let us then state the generalization error and sample complexity bounds provided by Hellinger distance.

Corollary 10: 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, for some σ and for every $h\in\mathcal{H}$. Given $\eta\in(0,1)$, let $E=\{(S,h):|L_{\mathcal{P}}(h)-L_S(h)|>\eta\}$.

$$\begin{aligned}\mathbb{P}(E)&\leq 2\exp\left(-n\frac{\eta^2}{2\sigma^2}\right)+H^2(S;\mathcal{A}(S))\\&\quad+2^{3/2}H(S;\mathcal{A}(S))\exp\left(-n\frac{\eta^2}{4\sigma^2}\right)\\&\leq 2\exp\left(-n\frac{\eta^2}{2\sigma^2}\right)+H^2(S;\mathcal{A}(S))+2^{3/2}H(S;\mathcal{A}(S)).\end{aligned} \tag{88}$$

and in order to ensure a confidence of $\delta \in (0,1)$, *i.e.* $\mathbb{P}(E) \leq \delta$, it is sufficient to have n samples where

$$n \geq \frac{\log\left(\frac{1}{\delta - H^2(S;\mathcal{A}(S)) + 2^{3/2}H(S;\mathcal{A}(S))}\right)}{4\eta^2}. \tag{89}$$

VII. Comparison of the Results

While the relationship among I_α for various α's is clear, a detailed and complete understanding of the relationship among all the f-divergences and α-divergences is still lacking and many works are trying to address the issue (e.g., [27]). Restricting ourselves to χ^2-like divergences, a summary of our current understanding is the following:

- $I_\alpha(X,Y) = \min_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y) \leq D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)$;
- $I_{\alpha_1}(X,Y) \leq I_{\alpha_2}(X,Y)$ if $\alpha_1 \leq \alpha_2$ [15];
- $\mathcal{L}(X{\to}Y) \geq \log(\chi^2(X,Y)+1) \geq I_2(X,Y)$.

All of this seems to hint, at least in this family of divergences, that the tightest sample-complexity bound would be given by I_2 (where both $1/\alpha = 1/\gamma = 1/2$). Taking $\alpha \in (1,2)$ will also consistently improve the dependency of the bound with respect to the information measure term while rendering the bound closer and closer to 1 and worsening the dependency with respect to δ in the sample complexity, as γ will tend to $+\infty$. The best trade-off between these two quantities remains an open problem. Moreover, the bounds involving other measures, like the Hellinger distance, can be fundamentally different and are not yet well understood. Considering Corollary 10, while the dependency with respect to δ and η seems to be the right one, the role played by the information measure is not as clear. Possibly, finding the optimal f in Theorem 3, for a given behaviour of $\mathcal{P}_X\mathcal{P}_Y(E)$, could also shed some light on whether or not one should consider functions f outside of the χ^2-like family (polynomials). A Taylor expansion argument shows that most f-divergences are, in the end, χ^2-like but, while those measures blow-up in some deterministic settings, others like Total Variation and Hellinger Distance do not. This different behaviour could be key in obtaining the tightest bound in the learning theory framework as well.

VIII. Adaptive Data Analysis

Other than providing a generalization of the classical bounds to adaptive settings, maximal leakage can also be employed in adaptive data analysis. The model of adaptive composition we will be considering is identical to the setting in [1], [2], [28] and defined as follows:

Definition 9 (Adaptive Composition): Let $\mathcal{X}$ be a set. Let S be a random variable over $\mathcal{X}^n$. Let $(\mathcal{A}_1,\ldots,\mathcal{A}_k)$ be a sequence of algorithms such that $\forall i : 1 \leq i \leq k$ $\mathcal{A}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{i-1} \to \mathcal{Y}_i$. Denote with $Y_1 = \mathcal{A}_1(S), Y_2 = \mathcal{A}_2(S,Y_1),\ldots,Y_k = \mathcal{A}_k(S,Y_1,\ldots,Y_{k-1})$. The adaptive composition of $(\mathcal{A}_1,\ldots,\mathcal{A}_k)$ is an algorithm that takes as an input S and sequentially executes the algorithms $(\mathcal{A}_1,\ldots,\mathcal{A}_k)$ as described by the sequence $(Y_i, 1 \leq 1 \leq k)$.

This level of generality allows us to formalize the behavior of a data analyst who, after viewing the previous outcomes of the analyses performed, decides what to do next. A potential analyst would typically execute a sequence of algorithms that are known to have a certain property (e.g., generalize well) when used without adaptivity. The question we would like to address is the following: is this property also maintained by the adaptive composition of the sequence? The answer is not trivial as, for every i, the outcome of $\mathcal{A}_i$ depends both on S and on the previous outputs, that depend on the data themselves. However, when this property is guaranteed by some measure that composes adaptively itself (like differential privacy or, as we will show soon, maximal leakage) then it can be preserved. Indeed, being robust to post-processing, Maximal Leakage allows us to retain the generalization guarantees it provides, regardless of how one may manipulate the outcome of the algorithm:

Lemma 6 (Robustness to Post-Processing): Let $\mathcal{X}$ be the sample space and let X be distributed over $\mathcal{X}$. Let $\mathcal{Y}$ and $\mathcal{Y}'$ be output spaces, and consider $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{B} : \mathcal{Y} \to \mathcal{Y}'$. Then, $\mathcal{L}(X{\to}\mathcal{B}(\mathcal{A}(X))) \leq \mathcal{L}(X{\to}\mathcal{A}(X))$.

The proof is a direct application of the data processing inequality for maximal leakage. The useful implication of this result is as follows: in terms of maximal leakage, any generalization guarantees provided by $\mathcal{A}$ cannot be invalidated by further processing the output of $\mathcal{A}$. Regarding adaptive composition of two algorithms, we retrieve the following:

Lemma 7 (Adaptive Composition of Maximal Leakage): Let $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ be an algorithm such that $\mathcal{L}(X{\to}\mathcal{A}(X)) \leq k_1$. Let $\mathcal{B} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be an algorithm such that for all $y \in \mathcal{Y}$, $\mathcal{L}(X{\to}\mathcal{B}(X,y)) \leq k_2$. Then $\mathcal{L}(X{\to}(\mathcal{A}(X),\mathcal{B}(X,\mathcal{A}(X)))) \leq k_1 + k_2$.

The proof of this lemma relies crucially on the fact that maximal leakage depends on the marginal $\mathcal{P}_X$ only through its support and can be found in Appendix B, along with the other proofs for this section. In order to generalize the result to the adaptive composition of n algorithms, we need to lift the property stated in Equation (13) to more than two random variables.

Lemma 8: Let $k \geq 1$ and $X, A_1, \ldots, A_k$ be random variables.

$$\mathcal{L}(X{\to}(A_1,\ldots,A_k)) \leq \mathcal{L}(X{\to}A_1) + \mathcal{L}(X{\to}A_2|A_1) + \ldots + \mathcal{L}(X{\to}A_k|(A_1,\ldots,A_{k-1})).$$

The proof can be found in Appendix B. An immediate application of Lemma 8 leads us to the following result.

Lemma 9: Consider a sequence of $k \geq 1$ algorithms: $(\mathcal{A}_1,\ldots,\mathcal{A}_k)$ where for each $1 \leq i \leq k$, $\mathcal{A}_i : \mathcal{X} \times \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{i-1} \to \mathcal{Y}_i$. Suppose that for all $1 \leq i \leq k$ and for all $(y_1,\ldots,y_{k-1}) \in \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{i-1}$, $\mathcal{L}(X{\to}\mathcal{A}_i(X,y_1,\ldots,y_{i-1})) \leq j_i$. Then, denoting with $A_1,\ldots,A_k$ the (random) outputs of the algorithms:

$$\mathcal{L}(X{\to}(A_1,\ldots,A_k)) = \mathcal{L}\left(X{\to}A^k\right) \leq \sum_{i=1}^{k} j_i. \tag{90}$$

The conclusion to be drawn is straightforward: given a collection of algorithms that have bounded leakage (and thus good generalizations capabilities) even if the outcome of one of them is used to inform a subsequent analysis (hence,

TABLE I
COMPARISON BETWEEN BOUNDS

	Robust (cf. Lemma 6)	Adaptive (cf. Lemma 7)	Bound	Sample Complexity
β-Stability [29]	No	No	exp. decay in n	$f(\beta,\eta)\times\log\left(\frac{2}{\delta}\right)$
ϵ-DP [1]	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 [3]	Yes	Yes	$(I(S;Y)+1)/(2n\eta^2-1)$	$I(S;Y)/\eta^2\times 1/\delta$
Maximal Leakage	Yes	Yes	$2\cdot\exp(\mathcal{L}\left(S{\to}Y\right)-2n\eta^2)$	$\left(\mathcal{L}\left(S{\to}Y\right)+\log\left(\frac{2}{\delta}\right)\right)/2\eta^2$
α-Sibson's MI	Yes	Unknown	$\exp(\frac{\alpha-1}{\alpha}(I_\alpha(S,Y))+\log 2-2n\eta^2))$	$\left(I_\alpha(S,Y)+\log 2+\gamma\log\left(\frac{1}{\delta}\right)\right)/2\eta^2$
χ^2	Yes	Unknown	$\sqrt{2}\exp\left(\frac{1}{2}\left(\log(\chi^2(S,Y)+1)-2n\eta^2\right)\right)$	$\left(\log(\chi^2(S,Y)+1)+2\log\left(\frac{\sqrt{2}}{\delta}\right)\right)/\eta^2$
VC-Dim. d [26]			$2\cdot\exp(\log(K)-2n\eta^2)$	$\left(d+\log\left(\frac{2}{\delta}\right)\right)/2\eta^2$

creating multiple dependencies on the data) the generalization guarantees of the composition can still be maintained.

Another interesting application of Corollary 3 in adaptive settings may be the following (same setting of [2]): consider the problem of bounding the probability of making a false discovery, when the statistic to apply is selected with some data dependent algorithm $\mathcal{T}$. In this context, the classical guarantees that allow to upper-bound this probability by the significance value no longer hold. Measuring the information leaked from the data through $\mathcal{T}$ with the maximal leakage we retrieve the following:

Corollary 11: Let $\mathcal{A}:\mathcal{X}^n\to\mathcal{T}$ be a data dependent algorithm for selecting a test statistic $t\in\mathcal{T}$. Let X be a random dataset over $\mathcal{X}^n$. Suppose that $\sigma\in[0,1]$ is the significance level chosen to control the false discovery probability. Denote with E the event that $\mathcal{A}$ selects a statistic such that the null hypothesis is true but its p-value is at most σ. Then,

$$\mathbb{P}(E)\leq\exp(\mathcal{L}\left(X{\to}\mathcal{A}(X)\right))\cdot\sigma.$$

If the analyst wishes to achieve a bound of δ on the probability of making a false discovery in adaptive settings, the significance level σ to be used should be no higher than $\delta/\exp(\mathcal{L}\left(X{\to}\mathcal{A}(X)\right))$. Once again, if $\mathcal{A}$ is independent from X, we recover the known bound of σ.

IX. COMPARISON WITH OTHER BOUNDS

In this section, we compare the proposed new bounds to the existing ones from the literature. A summary is provided in Table I, where the bound involving Hellinger distance has been omitted as very different in shape (both in terms of the high-probability bound and, as a byproduct, in terms of the sample complexity bound). At a glance, from Table I, it is easy to see that most of the information measures bounds (with the sole exception of the Mutual Information one) can have an exponential decay with the number of samples n. This is indeed the desired behaviour with respect to n. The reason is that the event E we consider is the event that the empirical average of some function (empirical risk) evaluated on a sequence of iid random variables (*i.e.*, S, the training samples) diverges from its actual average (risk) more than some constant η. Even in the case where such function is independent from the samples S the decay one typically finds in the literature is at best exponential with respect to n (e.g., McDiarmid's and Hoeffding's inequality). More detailed comparisons will appear in the following subsections.

A. Maximal Leakage and Mutual Information

One interesting result in the field, that connects the generalization error with Mutual Information, under the same assumptions of Corollary 2, is the following (Theorem 8 of [3]):

$$\mathbb{P}(E)\leq\frac{I(S;\mathcal{A}(S))+\log 2}{2n\eta^2-\log 2}. \tag{91}$$

Let us compare this result with Corollary 4 in terms of sample complexity. From Corollary 4, it follows that using a sample size of

$$n\geq\left(\frac{\mathcal{L}\left(S{\to}\mathcal{A}(S)\right)+\log(2/\delta)}{2\eta^2}\right), \tag{92}$$

yields a learner for $\mathcal{H}$ with accuracy η and confidence δ. Using the same reasoning with inequality (91), we get:

$$n\geq\left(\frac{I(S;\mathcal{A}(S))+\log 2+\delta\log 2}{2\eta^2\delta}\right). \tag{93}$$

Since $\mathcal{L}\left(X{\to}Y\right)\geq I(X;Y)$, in the regime where the two measures behave similarly, the reduction in the sample complexity is exponential in δ. The same reasoning can be applied to the sample complexity of I_α for a given $\alpha\in(1,+\infty]$: the exponential improvement in δ remains, although with a worse constant that multiplies the $\log(1/\delta)$ term. Another source of comparison can be found in Example 1 and 2 in Section V-A.

Example 1* (Independent Case): In the same setting as in Example 1 (X independent from Y), from [3, Lemma 15] we retrieve:

$$\zeta=\mathcal{P}_{XY}(E)\leq\frac{1}{-\log(\max_y\mathcal{P}_X(E_y))}=\frac{1}{\log(1/\zeta)}, \tag{94}$$

which is much weaker than the bound $\mathcal{P}_{XY}(E)\leq\zeta$ that can be obtained from (61).

Example 2* (Strongly Dependent Case): In the same setting as in Example 2 ($X=Y\sim\mathcal{U}([n])$) from [3, Lemma 15] we retrieve:

$$1=\mathcal{P}_{XY}(E)\leq 1+\frac{1}{\log n}. \tag{95}$$

However, from (61) we recover:

$$1=\mathcal{P}_{XY}(E)\leq\frac{1}{n}\cdot n=1. \tag{96}$$

Thus, while the bound obtained via Mutual Information is asymptotically tight, the Maximal Leakage bound is met with equality.

B. Maximal Leakage and Differential Privacy

In this section we will compare our results with the generalization guarantees provided by differential privacy (DP). The definition of ϵ-differentially privacy (ϵ-DP) is the following:

Definition 10: Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be a randomised algorithm. $\mathcal{A}$ is ϵ-DP if for every $\mathcal{S} \subseteq \mathcal{Y}$ and every $x, y \in \mathcal{X}^n$ that differ only in one position:

$$\mathbb{P}(\mathcal{A}(x) \in \mathcal{S}) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(y) \in \mathcal{S}). \tag{97}$$

A relationship with Maximal Leakage can be established:

Lemma 10: Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an ϵ-DP randomised algorithm, then

$$\mathcal{L}\left(X \to \mathcal{A}(X)\right) \leq \epsilon \cdot n. \tag{98}$$

Proof: Let $Y = \mathcal{A}(X)$ and assume, for simplicity, that Y is a discrete random variable (the proof for continuous Y follows from very similar arguments). Fix some $\hat{\mathbf{x}} \in \mathcal{X}^n$, $\forall\, \mathbf{x} \in \mathcal{X}^n$ we have that $\mathbf{x}$ and $\hat{\mathbf{x}}$ differ in at most n positions and, iteratively applying the definition of DP, we have that $\mathbb{P}(Y = y|X = \mathbf{x}) \leq e^{\epsilon \cdot n}\mathbb{P}(Y = y|X = \hat{\mathbf{x}})$. Thus:

$$\begin{aligned}\mathcal{L}\left(X \to Y\right) &= \log \sum_{y \in \mathcal{Y}} \max_{\mathbf{x} \in \mathcal{X}^n} \mathbb{P}(Y = y|X = \mathbf{x}) && (99)\\ &\leq \log \sum_{y \in \mathcal{Y}} e^{\epsilon \cdot n}\mathbb{P}(Y = y|X = \hat{\mathbf{x}}) && (100)\\ &= n \cdot \epsilon && (101)\end{aligned}$$

□

This suggests an immediate application of Corollary 4. Indeed, suppose $\mathcal{A}$ is an ϵ-DP algorithm, then:

$$\begin{aligned}\exp(\mathcal{L}\left(X \to Y\right) - 2n\eta^2) &\leq \exp(\epsilon n - 2n\eta^2) && (102)\\ &= \exp(-n(2\eta^2 - \epsilon)). && (103)\end{aligned}$$

In order for the bound to be decreasing with n, we need $2\eta^2 - \epsilon > 0$ leading us to $\epsilon < 2 \cdot \eta^2$, where η represents the accuracy of the learning algorithm and ϵ the privacy parameter. Thus, for fixed η, as long as the privacy parameter is smaller than $2 \cdot \eta^2$, we have guaranteed generalization capabilities for $\mathcal{A}$ with an exponentially decreasing bound. For $\epsilon \leq \eta/2$, it is shown in [28, Theorem 9] that $\mathbb{P}(E) \leq 1/4 \exp\left(-n\eta^2/12\right)$. It is easy to check that, for large enough n, our bound is tighter if $\epsilon \leq \eta^2 \cdot 23/12$.

It is possible to see that enforcing differential privacy on some algorithm $\mathcal{A}$ induces generalization guarantees similar to those stated in Corollary 3: suppose $\mathcal{A}$ is ϵ-DP, with

$$\epsilon \leq \sqrt{\frac{\log(1/\beta)}{2n}}, \tag{104}$$

and let $\max_y \mathcal{P}_X(E_y) \leq \beta$ then [28, Theorem 11]:

$$\mathbb{P}(E) \leq 3\sqrt{\beta}. \tag{105}$$

The results we are providing are qualitatively different. We do not require the imposition of some (possibly very strong) privacy criteria on the algorithm in order to able to analyse its performances. Instead, we propose a way of estimating how the probabilities we are interested in change, measuring the level of dependence through Maximal Leakage. Moreover, given an ϵ-DP algorithm the bound obtained via Inequality (98) can be tighter for certain regimes of ϵ. Indeed, let:

$$\epsilon < \frac{\log\left(3/\sqrt{\beta}\right)}{n} \leq \sqrt{\frac{\log(1/\beta)}{2n}}, \tag{106}$$

using (105) we get a *fixed* bound of $3\sqrt{\beta}$, while with Corollary 3 and Lemma 10 we obtain that:

$$\exp(\mathcal{L}\left(X \to Y\right)) \cdot \beta < \exp\left(\log\left(3/\sqrt{\beta}\right)\right) \cdot \beta = 3\sqrt{\beta}. \tag{107}$$

Hence, whenever the privacy parameter is lower than $1/n \log\left(3/\sqrt{\beta}\right)$ we are able to provide a better bound. Notice that Lemma 10 can be quite loose: using Lemma 3 it is possible to see that for classical mechanisms that imply ϵ-DP, Maximal Leakage can be much lower that $\epsilon \cdot n$. Indeed, using the result proven in Lemma 3, we can find such an example:

Corollary 12: Let $h : \mathcal{X}^n \to \mathbb{R}$ be a function of sensitivity $1/n$ and let $N \sim Lap(1/n\epsilon)$ then the mechanism $\mathcal{M}(x) = h(x) + N$ is ϵ-DP. Without loss of generality we have that $|h(x)| \leq 1$ (e.g. 0-1 loss) and thus:

$$\mathcal{L}\left(X \to \mathcal{M}(X)\right) = \log(1 + \epsilon \cdot n) < \epsilon \cdot n. \tag{108}$$

More importantly, the family of algorithms with bounded Maximal Leakage is not restricted to the differentially private ones. It is easy to see, for instance, that whenever there is a deterministic mapping and ϵ-DP is enforced on it, $\epsilon \geq +\infty$. Trying to relax it to (ϵ, δ)-Differential Privacy does not help either, as one would need $\delta \geq 1$ rendering it practically useless. On the other hand, if the algorithm has a bounded range the Maximal Leakage from input to output is always bounded, since $\mathcal{L}\left(X \to Y\right) \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$. This simple observations allows us to immediately retrieve another result [1, Theorem 9]: $\mathbb{P}(E) \leq |\mathcal{Y}| \cdot \beta$, where β is such that $\mathbb{P}(E_y) \leq \beta$ for every y. Indeed, given a random variable Y with bounded support, $\mathcal{L}\left(X \to Y\right) \leq \log|\mathcal{Y}|$ and from Corollary 3 we have that:

$$\mathbb{P}(E) \leq \max_y \mathbb{P}(E_y) \exp\left(\mathcal{L}\left(X \to Y\right)\right) \leq \beta \cdot |\mathcal{Y}|. \tag{109}$$

This shows how Corollary 3 is more general than both Theorems 6 and 9 of [1].

To conclude the comparison let us now state Corollary 4 with a general sensitivity c:

$$\mathbb{P}(E) \leq 2 \cdot \exp\left(\mathcal{L}\left(X \to Y\right) - \frac{2\eta^2}{c^2 n}\right). \tag{110}$$

By contrast, [1, Cor. 7] states that whenever an algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ outputs a function h of sensitivity c and is $\eta/(cn)$-DP then, denoting with S a random variable distributed over $\mathcal{X}^n$ and with

$$E = \{(S, h) : h(S) - \mathbb{E}(h) \geq \eta\}, \tag{111}$$

we have that:

$$\mathbb{P}(E) \leq 3\exp(-\eta^2/(c^2 n)). \tag{112}$$

It is easy to see that we have a tighter bound whenever the accuracy $\eta > n \cdot c$.

C. Sibson's Mutual Information, Maximal Leakage and Max Information

Another tool used in the line of work started by Dwork et al. [1], [2] is the concept of max-information. The definition is the following:

Definition 11: [1, Def. 10] Let X, Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$ and with marginals $\mathcal{P}_X, \mathcal{P}_Y$. The max-information between X and Y, is defined as follows:

$$I_\infty^M(X,Y) = \log \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{\mathcal{P}_{XY}(\{(x,y)\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{y\})}, \tag{113}$$

while, the β-approximate max-information is defined as:

$$I_\infty^{M,\beta}(X,Y) = \log \sup_{\mathcal{O}\subseteq\mathcal{X}\times\mathcal{Y}, \mathcal{P}_{XY}(\mathcal{O})>\beta} \frac{\mathcal{P}_{XY}(\mathcal{O})-\beta}{\mathcal{P}_X\mathcal{P}_Y(\mathcal{O})}. \tag{114}$$

Remark 14: Notice that we slightly changed the notation from [1] in order to avoid confusion. $I_\infty^M(X,Y)$ does not correspond to Sibsons's I_∞ but it actually corresponds to Rényi's D_∞, *i.e.*, $I_\infty^M(X,Y) = D_\infty(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)$.

One of the main reasons that led to the definition of approximate max-information is related to the generalization guarantees it provides, here recalled for convenience.

Lemma 11: [1, Thm. 13] Let X be a random dataset in $\mathcal{X}^n$ and let $\mathcal{A}: \mathcal{X}^n \to \mathcal{Y}$ be such that for some $\beta \geq 0$, $I_\infty^{M,\beta}(X, \mathcal{A}(X)) = k$. Let $Y = \mathcal{A}(X)$ then, for any event $E \subseteq X^n \times \mathcal{Y}$:

$$\mathcal{P}_{XY}(E) \leq e^k \mathcal{P}_X\mathcal{P}_Y(E) + \beta. \tag{115}$$

The result looks quite similar to Corollary 4, but the two measures, Max-Information and Maximal Leakage, although related, can be quite different. In this section we will analyze the connections and differences between the two measures underlining the corresponding implications.

Lemma 12: Let $\mathcal{A}: \mathcal{X}^n \to \mathcal{Y}$ be a randomised algorithm such that $I_\infty^M(X, \mathcal{A}(X)) \leq k$. Then, $\mathcal{L}(X{\to}\mathcal{A}(X)) \leq k$.

Proof: Denote with $Y = \mathcal{A}(X)$. Having a bound of k on the Max-Information of $\mathcal{A}$ means that for all $x \in \mathcal{X}^n$, and $y \in \mathcal{Y}, \mathbb{P}(Y = y|X = x) \leq e^k \cdot \mathbb{P}(Y = y)$ and this implies that $\mathcal{L}(X{\to}Y) \leq k$. □

More generally, we can say the following.

Lemma 13: $I_\infty^M(X,Y) \geq D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y) \geq I_\alpha(X,Y)$ for every $\alpha \in [1, +\infty]$.

Proof: We have that $I_\infty^M(X,Y) = D_\infty(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y) \geq \mathcal{L}(X{\to}Y) \geq I_\alpha(X,Y)$ for any $\alpha \in [1, +\infty)$. □

With respect to β-approximate max-information instead, we can state the following.

Lemma 14: Let $\mathcal{A}: \mathcal{X}^n \to \mathcal{Y}$ be a randomised algorithm. Let X be a random variable distributed over $\mathcal{X}^n$ and let $Y = \mathcal{A}(X)$. Suppose X, Y are discrete random variables and denote with $\mathcal{P}_{XY}$ the joint distribution and with $\mathcal{P}_X, \mathcal{P}_Y$ the corresponding marginals. For any $\beta \in (0,1)$ and $\alpha \in (1, +\infty]$

$$I_\infty^{M,\beta}(X, \mathcal{A}(X)) \leq \frac{\alpha-1}{\alpha} I_\alpha(X, \mathcal{A}(X)) + \log\left(\frac{1}{\beta}\right). \tag{116}$$

Proof: Fix any $\beta > 0$. Using [1, Lemma 18] we have that if

$$\mathcal{P}_{XY}\left(\left\{(x,y) \in \mathcal{X}\times\mathcal{Y} \middle| \frac{\mathcal{P}_{XY}(\{x,y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{y\})} \geq e^k\right\}\right) \leq \beta,$$

then

$$I_\infty^{M,\beta}(X,Y) \leq k.$$

Denote with $Y = \mathcal{A}(X)$, and with $F = \left\{(x,y) \in \mathcal{X}\times\mathcal{Y} \middle| \frac{\mathcal{P}_{XY}(\{x,y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{y\})} \geq \frac{\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X,Y)\right)}{\beta}\right\}$, then

$$\begin{aligned}
\mathcal{P}_{XY}(F) &\leq \frac{\mathbb{E}_{\mathcal{P}_{XY}}\left[\frac{\mathcal{P}_{XY}(\{X,Y\})}{\mathcal{P}_X(\{X\})\mathcal{P}_Y(\{Y\})}\right]\cdot\beta}{\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X,Y)\right)} \\
&\leq \frac{\mathbb{E}_{\mathcal{P}_Y}\left[\left(\mathbb{E}_{\mathcal{P}_X}\left[\left(\frac{\mathcal{P}_{Y|X}(\{Y\})}{\mathcal{P}_Y(\{Y\})}\right)^\alpha \middle| Y\right]\right)^{1/\alpha}\right]\cdot\beta}{\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X,Y)\right)} \\
&= \beta.
\end{aligned}$$

Hence, $I_\infty^{M,\beta}(X, \mathcal{A}(X)) \leq \log\left(\frac{\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X,Y)\right)}{\beta}\right) = \frac{\alpha-1}{\alpha} I_\alpha(X,Y) + \log\left(\frac{1}{\beta}\right)$.

Taking the limit at $\alpha \to \infty$ one also gets that $I_\infty^{M,\beta}(X, \mathcal{A}(X)) \leq \mathcal{L}(X{\to}Y) + \log\left(\frac{1}{\beta}\right)$. □

The role played by β can lead to undesirable behaviors of β-approximate max-information. The following example shows how β-approximate max-information can be unbounded while. On the other hand, in the discrete case, the Maximal Leakage between two random variables is always guaranteed to be bounded.

Example 5: Let us fix a $\beta \in (0,1)$. Suppose $X \sim \text{Ber}(2\beta)$. We have that $\mathcal{L}(X{\to}X) = \log|\text{supp}(X)| = \log 2$. For the β-approximate max-information we have: $I_\infty^{M,\beta}(X,X) \geq \log((2\beta-\beta)/\beta^2) = \log(1/\beta)$. It can thus be arbitrarily large.

Another interesting characteristic of max-information is that, differently from differential privacy, it can be bounded even if we have deterministic algorithms: this observation is implied by the connection with what in the literature is known as "description length" of an algorithm, and synthesized in the following result [1]: Let $\mathcal{A}: \mathcal{X}^n \to \mathcal{Y}$ be a randomised algorithm, for every $\beta > 0$,

$$I_\infty^{M,\beta}(\mathcal{A}, n) \leq \log\left(\frac{|\mathcal{Y}|}{\beta}\right). \tag{117}$$

In contrast, with Sibson's I_α for every $\alpha \in [1, +\infty)$ we have that

$$I_\alpha(X, \mathcal{A}(X)) \leq \mathcal{L}(X{\to}\mathcal{A}(X)) \leq \log(|\mathcal{Y}|). \tag{118}$$

Clearly, being β typically very small in the key applications, the corresponding multiplicative factors in the bounds are $(|Y|/\beta)$ and $|\mathcal{Y}|$, and the difference between the two quantities can be substantial. It is also worth noticing that Equation (117) can be seen as a consequence of Lemma 14 and Equation (118). The difference between the two measures is not uniquely restricted to deterministic mechanisms. The following is a simple example of a randomised mapping where Maximal Leakage is smaller than β-approximate-max-information, for small β.

Example 6: Consider $X \sim \text{Ber}(1/2)$ and a random variable Y with support $\mathcal{Y} = \{0, 1, e\}$. Consider also the following randomised mapping: $\mathbb{P}(Y = e|X = x) = \xi$ and

$\mathbb{P}(Y = x|X = x) = 1 - \xi$. That is, Y can be interpreted as passing X through a binary erasure channel with erasure probability ξ. In this case, the Maximal Leakage is $\mathcal{L}(X \to Y) = \log(2 - \xi)$ [6]; while, for β-Approximate max-information one finds (after a series of computations) that: $I_\infty^{M,\beta}(X, Y) = \log(2 \cdot \max\{(1-\xi-\beta)/(1-\xi), (1-\beta)/(1+\xi)\})$; It is easy to see how for a fixed ξ and for β going to 0, Approximate Max-Information approaches $\log 2$ while Maximal Leakage is strictly smaller.

X. Conclusion

Our aim was to bound the probability of an event E under the joint distribution $\mathcal{P}_{XY}$ via information measures and the probability of the same event under the product of the marginals $\mathcal{P}_X\mathcal{P}_Y$. We started presenting bounds involving Luxemburg and Amemiya norms. We then particularised one of these results as a family of bounds characterized by four parameters $\alpha, \gamma, \alpha', \gamma' \geq 1$, constrained by the following equality $\frac{1}{\alpha} + \frac{1}{\gamma} = \frac{1}{\alpha'} + \frac{1}{\gamma'} = 1$ (*i.e.*, Hölder's conjugates). We explicit and analyze the following choices of parameters:

- with $\alpha' = \alpha$ and $\gamma' = \gamma$ we retrieve a family of bounds involving the Rényi's divergence of order α. A rewriting of this result allowed us to also recover a bound involving p-Hellinger divergences with $p \in (1, +\infty)$;
- with $\alpha' \to 1$ and consequently, $\gamma' \to \infty$ we retrieve a family of bounds involving Sibson's Mutual Information of order α;
- with $\alpha' \to 1$, $\gamma' \to \infty$, $\alpha \to \infty$ and $\gamma \to 1$, we retrieve a bound involving Maximal Leakage;

We also provided a family of bounds involving f-divergences where f is an invertible convex function. We focused in particular on Maximal Leakage, since its semi-closed form and the dependence on $\mathcal{P}_X$ only through the support make it more amenable to analysis. Moreover, we show that the measure is robust under post-processing and composes adaptively. The robustness to post-processing is true for any information measure satisfying the data-processing inequality. However, since we currently lack a definition of conditional Sibson's MI or f-mutual information it is not possible to verify whether or not these other measures also compose adaptively. Another interesting property of Maximal Leakage, instead, is that the bound it provides represents a generalisation of the classical concentration inequalities in adaptive mechanisms. The comparison with the other approaches showed how this measure is less strict than Differential Privacy and yet still provides strong generalization guarantees. We also showed how, in regimes where Mutual Information and Maximal Leakage behave similarly, the leakage bound provides an exponential improvement in the sample complexity. In general, one can also see that the sample complexity induced by I_α and Maximal Leakage is actually optimal with respect to δ in the realizable case. This shows how information measures can play a role similar to the VC-dimension, but tailored to the specific algorithm rather than the hypothesis class itself. Indeed, while the VC-dimension is a property of $\mathcal{H}$ only, information measures depend also on the samples and on (the distribution induced by) the algorithm. Some bounds on expected generalization error are also provided but, probably as an artifact of the analysis, they are generally worse (for finite number of samples n) than the ones that use Mutual Information [4], [5], [11].

Appendix A
Proof of the Generalised Hölder's Inequality

Let us recall the statement:

Let ψ be an Orlicz fucntion and $\psi^\star$ denote its Legendre-Fenchel dual (*i.e.*, $\psi^\star(x) = \sup_\lambda \lambda x - \psi(\lambda)$), then for every pair of random variable U, V:

$$\mathbb{E}[UV] \leq \|U\|_\psi \|V\|_{\psi^\star}^A. \tag{119}$$

Proof: For every $\sigma, t > 0$ we have that:

$$\mathbb{E}[UV] = \mathbb{E}\left[\sigma \frac{U}{\sigma} \frac{1}{t} V t\right] \tag{120}$$

$$\overset{(c)}{\leq} \frac{\sigma}{t} \mathbb{E}\left[\psi\left(\frac{|U|}{\sigma}\right) + \psi^\star(|V|t)\right] \tag{121}$$

where (c) follows from Young's inequality for convex functions. Choosing $\sigma = \|U\|_\psi$:

$$\mathbb{E}[UV] \leq \frac{\|U\|_\psi}{t} \mathbb{E}\left[\psi\left(\frac{|U|}{\|U\|_\psi}\right) + \psi^\star(|V|t)\right] \tag{122}$$

$$\overset{(d)}{\leq} \|U\|_\psi \frac{1 + \mathbb{E}\left[\psi^\star(|V|t)\right]}{t}. \tag{123}$$

(d) follows from the definition of Luxemburg norm, *i.e.*, $\mathbb{E}\left[\psi\left(|U|/\|U\|_\psi\right)\right] \leq 1$. Taking the infimum with respect to t in (123) gives us (119) by definition of Amemiya norm. □

Appendix B
Properties of Maximal Leakage

In this appendix we will provide proofs for the properties of Maximal Leakage. Let us start with the Adaptive Composition of the measure and let us recall the statement for reference:

Let $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ be an algorithm such that $\mathcal{L}(X \to \mathcal{A}(X)) \leq k_1$. Let $\mathcal{B} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be an algorithm such that for all $y \in \mathcal{Y}$, $\mathcal{L}(X \to \mathcal{B}(X, y)) \leq k_2$.

Then $\mathcal{L}(X \to (\mathcal{A}(X), \mathcal{B}(X, \mathcal{A}(X)))) \leq k_1 + k_2$.

The proof of this lemma relies crucially on the fact that maximal leakage depends on the marginal $\mathcal{P}_X$ only through its support.

Proof: Let us denote with R_X the support of a random variable X. If we consider the second constraint in our assumption and denoting with $Z_y = \mathcal{B}(X, y)$, we get:

$$\forall y \in \mathcal{Y} \quad \mathcal{L}(X \to Z_y) \leq k_2 \iff \tag{124}$$

$$\forall y \in \mathcal{Y} \sum_{z_y \in R_{Z_y}} \max_{x \in R_X} \mathbb{P}(z_y|x) \leq \exp(k_2) \iff \tag{125}$$

$$\forall y \in \mathcal{Y} \sum_{z_y \in R_{Z|Y=y}} \max_{x \in R_X} \mathbb{P}(z|x, y) \leq \exp(k_2). \tag{126}$$

The last step holds, since every y generates a family of conditional distributions $\mathbb{P}(z_y|x)$ through $\mathcal{B}$ and this probability

is just $\mathbb{P}(z|x,y)$, with $z = \mathcal{B}(x,y)$. Using this observation in the conditional leakage of (13):

$$\mathcal{L}(X{\to}Z|Y) = \log \max_{y\in R_Y} \sum_{z\in R_{Z|Y=y}} \max_{x\in R_{X|Y=y}} \mathbb{P}(z|x,y) \tag{127}$$

$$\leq \log \max_{y\in R_Y} \sum_{z\in R_{Z|Y=y}} \max_{x\in R_X} \mathbb{P}(z|x,y) \tag{128}$$

$$\leq \log \max_{y\in R_Y} \exp(k_2) \tag{129}$$

$$= k_2, \tag{130}$$

leading us to the desired bound. □

Let us now show the generalization of this property to k random variables. The statement reads:

Let $k \geq 1$ and $X, A_1, \ldots, A_k$ be random variables.

$$\mathcal{L}(X{\to}(A_1,\ldots,A_k)) \leq \mathcal{L}(X{\to}A_1) + \mathcal{L}(X{\to}A_2|A_1) + \ldots + \mathcal{L}(X{\to}A_k|(A_1,\ldots,A_{k-1})).$$

Proof:

$$\mathcal{L}(X{\to}(A_1,\ldots,A_k)) = \mathcal{L}\left(X{\to}A^k\right) \tag{131}$$

$$= \mathcal{L}\left(X{\to}(A^{k-1},A_k)\right), \tag{132}$$

then the result follows from recursively applying the same argument to $\mathcal{L}\left(X{\to}A^{k-1}\right)$. □

APPENDIX C
EXAMPLES

A. Proof of Lemma 3

We will now compute the value of the Maximal Leakage for an additive noise mechanism, where the noise is a Laplace random variable. Recall the statement of the lemma 3 is:

Let $h : \mathcal{X}^n \to \mathbb{R}$ be a function such that $h(x) \in [a,c]$, $a < c\ \forall x \in \mathcal{X}^n$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim Lap(b)$ is such that:

$$\mathcal{L}(X{\to}\mathcal{M}(X)) = \log\left(1 + \frac{(c-a)}{2b}\right) \tag{133}$$

Let $Y = g(X) + N$, starting from Equation (11),

$$\exp(\mathcal{L}(X{\to}Y)) = \int_{\mathbb{R}} \sup_{x:f_X(x)>0} f_{Y|X}(y|x)dy \tag{134}$$

$$= \int_{\mathbb{R}} \sup_{x:f_X(x)>0} f_N(y-h(x))dy \tag{135}$$

$$= \frac{1}{2b}\left(\int_{-\infty}^{+\infty} \sup_{x:\mathcal{P}_X(x)>0} \exp\left(\frac{-|y-h(x)|}{b}\right)dy\right) \tag{136}$$

$$= \frac{1}{2b}\left(\int_{-\infty}^{a} \exp\left(\frac{-|y-a|}{b}\right)dy + \int_a^c dy\right) \tag{137}$$

$$+ \frac{1}{2b}\left(\int_c^{+\infty} \exp\left(\frac{-|y-c|}{b}\right)dy\right) \tag{138}$$

$$= \frac{1}{2b}\left(\int_{-\infty}^{0} \exp\left(\frac{-|z|}{b}\right)dz + (c-a)\right) \tag{139}$$

$$+ \frac{1}{2b}\left(\int_0^{+\infty} \exp\left(\frac{-|w|}{b}\right)dw\right) \tag{140}$$

$$= \frac{1}{2b}\left((c-a) + 2\int_0^{+\infty} \exp\left(\frac{-w}{b}\right)dw\right) \tag{141}$$

$$= \frac{1}{2b}((c-a)+2b) = \left(1 + \frac{(c-a)}{2b}\right). \tag{142}$$

■

The proofs of the other additive noise mechanisms (Gaussian and Exponential) follow a similar approach to the one used to prove Lemma 3.

B. Proof of Corollary 5

Suppose the hypothesis space is countable and let $k := |\mathcal{H}|$ (could be infinite). Suppose also that $\mathbb{E}[N_i] = b_i$ [4] (with N_i being the noise added to the i-th hypothesis). Since the choice of the hypothesis depends only on the noisy empirical errors, the following is a Markov Chain $S - (L_S(h_i))_{i\in[k]} - (L_S(h_i)+N_i)_{i\in[k]} - H$. Then by the data-processing inequality for Maximal Leakage:

$$\mathcal{L}(S{\to}H) \leq \mathcal{L}\left((L_S(h_i))_{i\in[k]}{\to}(L_S(h_i)+N_i)_{i\in[k]}\right). \tag{143}$$

Also, denoting with $X_i = L_S(h_i)$ and with $Y_i = X_i + N_i$:

$$\exp(\mathcal{L}((X_1,\ldots,X_k){\to}(Y_1,\ldots,Y_k)) \tag{144}$$

$$= \int\cdots\int_{-\infty}^{+\infty} \max_{x^n} f(y^n|x^n)dy^n \tag{145}$$

$$= \int\cdots\int_{-\infty}^{+\infty} \max_{x^n}\left(\prod_{i=1}^k f_{N_i}(y_i-x_i)\right)dy^n \tag{146}$$

$$= \int\cdots\int_{-\infty}^{+\infty} \max_{x^n}\left(\prod_{i=1}^k \frac{1}{b_i}e^{-(y_i-x_i)/b_i}\right)dy^n \tag{147}$$

$$= \prod_{i=1}^k \int_{-\infty}^{+\infty} \max_{x_i}\left(\frac{1}{b_i}e^{-(y_i-x_i)/b_i}\right)dy \tag{148}$$

$$= \prod_{i=1}^k \left(1 + \frac{1}{b_i}\right). \tag{149}$$

Equation (149), along with Corollary 4, implies that:

$$\mathbb{P}(\text{gen-err}(\mathcal{A}) \geq \eta) \leq 2\exp\left(\mathcal{L}(S{\to}H) - 2n\eta^2\right) \tag{150}$$

$$= 2\exp\left(\sum_{i=1}^k \log\left(1+\frac{1}{b_i}\right) - 2n\eta^2\right). \tag{151}$$

Now, suppose that $b_i = i^{1.1}/n^{1/3}$,

$$\mathcal{L}(S{\to}H) \leq \sum_{i=1}^k \log(1 + n^{1/3}/i^{1.1}) \tag{152}$$

$$\leq n^{1/3}\sum_{i=1}^{+\infty}\frac{1}{i^{1.1}} \tag{153}$$

$$\leq (n^{1/3})\cdot 11. \tag{154}$$

We have that

$$\mathbb{P}(\text{gen-err} \geq \eta) \leq 2\exp(-n(2\eta^2 - 11/n^{2/3})). \tag{155}$$

APPENDIX D
EXPECTED GENERALIZATION ERROR

Given the generalization error bounds proposed so far, one may ask how these reflect in results on the expected value of the generalization error. To give a meaningful result one needs to make some assumptions on the probability of our event E. In particular, we will assume this probability to be exponentially decreasing with the number of samples n (as it often happens in the literature [24], [29]). This section will focus only on Sibson's α-Mutual Information. It is possible to extend these results also to f-Mutual Information. However, in order to do so, one needs more information on f. For example, using the same techniques and starting from Corollary 8, one can state a bound on the expected generalization error for p-Hellinger divergences. It is, however, unclear how to derive a result involving all increasing and convex functions f. The following result is inspired by [26, p. 419] with a different (slightly improved, for our purposes) proof.

Lemma 15: Let X be a random variable and let $\hat{x} \in \mathbb{R}$. Suppose that there 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 $\mathbb{E}\left[|X - \hat{x}|\right] \leq a \min\left\{3\sqrt{\log b}, 2\sqrt{\log 2b}\right\}$.

Proof: Since $|X - \hat{x}|$ is a positive random variable we have that

$$\mathbb{E}\left[|X - \hat{x}|\right] = \int_0^{+\infty} \mathcal{P}_X(|X - \hat{x}| \geq \eta) d\eta. \tag{156}$$

Since for small values of η the exponential bound may be exceedingly loose, instead of trivially upper-bounding (156) we do the following:

$$\begin{aligned}
\mathbb{E}\left[|X - \hat{x}|\right] &= \int_0^{+\infty} \mathcal{P}_X(|X - \hat{x}| \geq \eta) d\eta && (157)\\
&\leq \int_0^{+\infty} \min\left(1, 2b \exp\left(-\eta^2/a^2\right)\right) d\eta && (158)\\
&= \int_0^{\sqrt{a^2 \log 2b}} d\eta + \int_{\sqrt{a^2 \log 2b}}^{+\infty} 2b \exp(-\eta^2/a^2) d\eta && (159)\\
&\leq \sqrt{a^2 \log 2b} && (160)\\
&\quad + \frac{a^2}{\sqrt{a^2 \log 2b}} \int_{\sqrt{a^2 \log 2b}}^{+\infty} \frac{2b\eta}{a^2} \exp(-\eta^2/a^2) d\eta && (161)\\
&= a\left(\sqrt{\log 2b} + \frac{1}{\sqrt{\log 2b}}\right) && (162)\\
&\leq a \min\left\{3\sqrt{\log b}, 2\sqrt{\log 2b}\right\}. && (163)
\end{aligned}$$

□

Theorem 5: Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ be a learning algorithm and let $I_\alpha(S, \mathcal{A}(S))$ (*i.e.*, Sibson's Mutual Information of order α) be the dependence measure chosen. Suppose that the loss function $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ 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. $\ell(h, Z) - \mathbb{E}[\ell(h, Z)]$ is σ^2-sub-Gaussian for each h), then:

$$\mathbb{E}\left[|L_S(H) - \mathbb{E}[L(H)]|\right] \leq \sqrt{\frac{8\sigma^2(\log(2) + I_\alpha(S, \mathcal{A}(S)))}{n}}.$$

Proof: The proof is a simple application of Lemma 15 and Corollary 2 with $a = \sqrt{2\gamma\sigma^2}/\sqrt{n}$ and $b = 2^{\frac{1}{\gamma}-1} \exp\left(\frac{1}{\gamma} I_\alpha(S, \mathcal{A}(S))\right)$. □

Remark 15: Notice that, even though we provide a concrete example (Theorem 5) that uses σ^2-sub-Gaussianity the assumption is not strictly necessary. Lemma 15 only requires that the probability of X diverging from $\hat{x}$ decays exponentially fast. This can be true also for other classes of random variables, like sub-Weibull random variables with an opportune choice of parameters [30].

An important result, obtained through a different route, is the bound on the expected generalization error via Mutual Information (Theorem 1 of [4]). We restate it here for ease of reference. Under the assumption that $\ell(h, Z)$ is σ^2-sub Gaussian for each h:

$$\left|\mathbb{E}\left[L_S(H) - \mathbb{E}[L(H)]\right]\right| \leq \sqrt{\frac{2\sigma^2}{n} I(S; \mathcal{A}(S))}. \tag{164}$$

In the spirit of comparison, let us also state a similar bound using Theorem 5 but with $\alpha \to \infty$ and using $3a\sqrt{\log b}$ as a bound on the expected value. Setting $a = \sqrt{\frac{2\sigma^2}{n}}$, $b = \exp(\mathcal{L}\left(S \to \mathcal{A}(S)\right))$ one retrieves the following:

$$\begin{aligned}
\left|\mathbb{E}\left[L_S(H) - \mathbb{E}[L(H)]\right]\right| &\leq \mathbb{E}\left[|L_S(H) - \mathbb{E}[L(H)]|\right] && (165)\\
&\leq 3\sqrt{\frac{2\sigma^2}{n} \mathcal{L}\left(S \to \mathcal{A}(S)\right)}. && (166)
\end{aligned}$$

We have seen before (c.f., Section IX-A) that Sibsons's α-MI brings an exponential improvement in the dependency over δ when considering the bound on the large deviation event. However, the measure does not seem to bring any improvement in controlling the expected generalization error. Looking at (166) one immediately sees that (other than being a constant away from Inequality (164)) Inequality (166) is always going to provide a worse bound, as $I_\alpha(S, \mathcal{A}(S)) \geq I(S; \mathcal{A}(S))$ for every $\alpha > 1$. This represents, perhaps, an artifact of the analysis used to prove Lemma 15.

REFERENCES

[1] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 2. Cambridge, MA, USA: MIT Press, 2015, pp. 1–29.

[2] R. Rogers, A. Roth, A. Smith, and O. Thakkar, "Max-information, differential privacy, and post-selection hypothesis testing," in *Proc. IEEE 57th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2016, pp. 487–494.

[3] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," *Proc. Mach. Learn. Res.*, vol. 83, pp. 25–55, Apr. 2018.

[4] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2521–2530.

[5] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," *Proc. Mach. Learn. Res.*, vol. 51, pp. 1232–1240, May 2016.

[6] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[7] C. Braun, K. Chatzikokolakis, and C. Palamidessi, "Quantitative notions of leakage for one-try attacks," *Electron. Notes Theor. Comput. Sci.*, vol. 249, pp. 75–91, Aug. 2009.

[8] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1475–1479.

[9] I. Issa and M. Gastpar, "Computable bounds on the exploration bias," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 576–580.
[10] A. R. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2018, pp. 7245–7254.
[11] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 121–130, May 2020.
[12] A. Pensia, V. Jog, and P.-L. Loh, "Generalization error bounds for noisy, iterative algorithms," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 546–550.
[13] A. T. Lopez and V. Jog, "Generalization error bounds using Wasserstein distances," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.
[14] H. Wang, M. Diaz, J. C. S. S. Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 577–581.
[15] S. Verdú, "α-mutual information," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2015, pp. 1–6.
[16] T. van Erven and P. Harremoes, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.
[17] I. Csiszar, "Generalized cutoff rates and Renyi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995.
[18] P. D. Grünwald, *The Minimum Description Length Principle* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2007.
[19] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969, doi: 10.1007/BF00537520.
[20] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.
[21] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Stat. Rev.*, vol. 70, no. 3, pp. 419–435, Dec. 2002.
[22] F. E. Su, "Methods for quantifying rates of convergence for random walks on groups," M.S. thesis, Dept. Math., Harvard Univ., Cambridge, MA, USA, 1995.
[23] H. Hudzik and L. Maligranda, "Amemiya norm equals Orlicz norm in general," *Indagationes Math.*, vol. 11, no. 4, pp. 573–585, Dec. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0019357700800269
[24] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.
[25] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and maximal leakage," 2019. *arXiv:1912.01439*. [Online]. Available: http://arxiv.org/abs/1912.01439
[26] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[27] I. Sason and S. Verdu, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[28] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, "Preserving statistical validity in adaptive data analysis," in *Proc. 47th Annu. ACM Symp. Theory Comput.*, New York, NY, USA, Jun. 2015, pp. 117–126.
[29] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, Mar. 2002.
[30] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel, "Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier-tailed distributions," *Stat*, vol. 9, no. 1, p. e318, 2020.

Amedeo Roberto Esposito (Student Member, IEEE) received the B.S. and M.S. degrees in computer science from the Università degli studi di Salerno, Fisciano, Italy, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the School of Computer and Communication Sciences, École Polytechnique Fédérale Lausanne (EPFL), Switzerland. His research interests are in information theory, learning theory, and probability theory.

Michael Gastpar (Fellow, IEEE) received the Dipl.El.-Ing. degree from the Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland, in 1997, the M.S. degree in electrical engineering from the University of Illinois at Urbana–Champaign, Urbana, IL, USA, in 1999, and the Doctorat ès Sciences degree from the École Polytechnique Fédérale Lausanne(EPFL), Switzerland, in 2002.

He was a Student in engineering and philosophy with The University of Edinburgh and the University of Lausanne. From 2003 to 2011, he was an assistant and a tenured Associate Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Since 2011, he has been a Professor with the School of Computer and Communication Sciences, Ecole Polytechnique Fédérale Lausanne (EPFL). He was a Professor with the Delft University of Technology, The Netherlands, and a Researcher with the Mathematics of Communications Department, Bell Labs, Murray Hill, NJ, USA, and Lucent Technologies, Murray Hill. His research interests are in network information theory and related coding, and signal processing techniques, with applications to sensor networks and neuroscience.

Dr. Gastpar received the IEEE Communications Society and Information Theory Society Joint Paper Award in 2013 and the EPFL Best Thesis Award in 2002. He served as the Technical Program Committee Co-Chair for the 2010 and 2021 International Symposia on Information Theory (Austin, TX, USA, and Melbourne, Australia). He was an Associate Editor of Shannon Theory for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2008 to 2011. He was an Information Theory Society Distinguished Lecturer from 2009 to 2011.

Ibrahim Issa (Member, IEEE) received the Ph.D. degree from the School of Electrical Engineering, Cornell University, Ithaca, NY, USA, in 2017. In January 2019, he joined the Electrical and Computer Engineering Department, American University of Beirut, as an Assistant Professor. From August 2017 to December 2018, he was with the Laboratory for Information in Networked Systems, Swiss Federal Institute of Technology Lausanne, as a Post-Doctoral Researcher. His research interests include privacy and security, information theory, machine learning, and quantum information theory. He received the Outstanding ECE Ph.D. Thesis Research Award for his thesis on Information Leakage.