Learning and Adaptive Data Analysis via Maximal Leakage

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

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

***Abstract*—There has been growing interest in studying connections between generalization error of learning algorithms and information measures. In this work, we generalize a result that employs the maximal leakage, a measure of leakage of information, and explore how this bound can be applied in different scenarios. The main application can be found in bounding the generalization error. Rather than analyzing the expected error, we provide a concentration inequality. In this work, we do not require the assumption of σ-sub gaussianity and show how our results can be used to retrieve a generalization of the classical bounds in adaptive scenarios (e.g., McDiarmid's inequality for $c-$sensitive functions, false discovery error control via significance level, etc.).**

***Index Terms*—Maximal Leakage, Generalization Error, Adaptive Data Analysis, Differential Privacy, Max-Information, Mutual Information**

I. Introduction

A learning algorithm can be seen as a (possibly randomized) mapping that takes a dataset as input and produces a hypothesis as output (e.g., a classifier). In addition to "explaining" the input data well (according to some loss function), it must perform comparably on a new independent test set. This latter property is called generalization. Intuitively speaking, an algorithm would not generalize well if its output "depends too much" on the input, i.e., it overfits. Thus, it is of interest to quantify this dependence, which also prevents us from directly using well-known results from statistical inference or probability theory (e.g., McDiarmid's inequality).

Our contribution consists in extending and exploring a result [1], [2] that uses *maximal leakage*, a dependence measure originally introduced to quantify the leakage of information from a random variable X to another random variable Y (and denoted by $\mathcal{L}(X\to Y)$) [3]. The basic insight is as follows: if a learning algorithm leaks little information about the training data, then it will generalize well. Moreover, we show that maximal leakage behaves well under composition: we can bound the leakage of a sequence of algorithms if each of them has bounded leakage. This property is particularly useful in the context of adaptive data analysis in which a sequence of algorithms (or tests) are run on the input dataset. Crucially, the choice of the subsequent algorithm depends on the output of previous runs. Furthermore, maximal leakage is robust under post-processing.

Related work: Bounds on the exploration bias and/or the generalization error, using information-theoretic measures, are given in [4]–[7]. A different perspective has been considered in the line of work initiated by Dwork *et al.* [8]–[10], which mainly relies on the notion of *differential privacy* that, besides inducing a suitable notion of stability, provides guarantees under composition. However, differential privacy can be quite restrictive, which led to the introduction of more relaxed notions (that still behave well under composition) such as β-max information, $I^{\beta}_{\infty}(X;Y)$ [9], [11]. Some proofs and supplementary results have been omitted from this extended abstract due to space limitations; these can be found in the full version of the paper [12].

II. Problem Statement

Following the classical framework of statistical learning :

Definition 1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be, respectively, the domain set and the label set. Let $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, and let $\mathcal{H}$ be the hypothesis class, i.e., a set of prediction rules $h:\mathcal{X}\to\mathcal{Y}$. Given $n\in\mathbb{N}$, a learning algorithm $\mathcal{A}$ is a map $\mathcal{A}:\mathcal{Z}^n\to\mathcal{H}$ that, given a sequence of domain points-label pairs $S=((x_1,y_1),\dots,(x_n,y_n))$, outputs some classifier $h=\mathcal{A}(S)\in\mathcal{H}$.*

In order to estimate the capability of a learning algorithm of correctly classifying unseen instances of the domain set, i.e., its *generalization* capability, the concept of generalization error is introduced.

Definition 2. *Let $\mathcal{P}$ be some distribution over $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$. The error (or risk) of a prediction rule $h:\mathcal{X}\to\mathcal{Y}$ with respect to $\mathcal{P}$ is defined as*

$$L_{\mathcal{P}}(h)=\mathbb{E}_{\mathcal{P}}(\mathbb{1}(h(X)\neq Y))=\mathbb{P}_{\mathcal{P}}(h(X)\neq Y), \quad (1)$$

while, given a sample $S=((x_1,y_1),\dots,(x_n,y_n))$, the empirical error of h with respect to S is defined as

$$L_S(h)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(h(x_i)\neq y_i). \quad (2)$$

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

$$\textit{gen-err}_{\mathcal{P}}(\mathcal{A},S)=|L_{\mathcal{P}}(\mathcal{A}(S))-L_S(\mathcal{A}(S))|. \quad (3)$$

978-1-5386-6900-6/19/$31.00 ©2019 IEEE

We are interested in providing a tail bound for the generalization error. In particular, we are interested in bounding the error in terms of maximal leakage. To that end, maximal leakage is defined as follows [3], [13].

Definition 3 ([13])**.** *Given a joint distribution P_{XY} on alphabets $\mathcal{X}$ and $\mathcal{Y}$, define:*

$$\mathcal{L}\left(X\to Y\right) = \sup_{U-X-Y-\hat{U}} \log \frac{\mathbb{P}(\{U=\hat{U}\})}{\max_{u\in\mathcal{U}} \mathbb{P}_U(\{u\})}, \quad (4)$$

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

It is shown in [3, Theorem 1] that, for finite alphabets:

$$\mathcal{L}\left(X\to Y\right) = \log \sum_{y\in\mathcal{Y}} \max_{x\in\mathcal{X}:P_X(x)>0} P_{Y|X}(y|x). \quad (5)$$

If X and Y have a jointly continuous pdf $f(x,y)$, we get [13, Corollary 4]:

$$\mathcal{L}(X\to Y) = \log \int_{\mathbb{R}} \sup_{x:f_X(x)>0} f_{Y|X}(y|x)dy. \quad (6)$$

Definition 4 (Conditional maximal leakage [13])**.** *Given a joint distribution P_{XYZ} on alphabets $\mathcal{X},\mathcal{Y},$ and $\mathcal{Z}$, define:*

$$\mathcal{L}(X\to Y|Z) = \sup_{U:U-X-Y|Z} \log \frac{\mathbb{P}(\{U=\hat{U}(Y,Z)\})}{\mathbb{P}(\{U=\tilde{U}(Z)\})}, \quad (7)$$

where U takes value in an arbitrary finite alphabet, and $\hat{U}$ and $\tilde{U}$ are the optimal (i.e., MAP) estimators of U given (Y,Z) and Z, respectively.

It is shown in [13] that:

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

and $\mathcal{L}\left(X\to(Y,Z)\right) \le \mathcal{L}\left(X\to Y\right) + \mathcal{L}(X\to Z|Y)$. (8)

III. Main Results

Our main result represents a generalization of [1, Theorem 2].

Theorem 1. *Suppose we have two real continuous random variables X,Y. Let f_{XY} be the joint pdf and assume it is continuous, and let $g:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$ be a positive function. Denote with $\bar{X}$ a random variable identically distributed to X but independent from both X and Y. Then,*

$$\mathbb{E}[g(X,Y)] \le \exp\left\{\mathcal{L}\left(X\to Y\right)\right\} \sup_y \mathbb{E}[g(\bar{X},y)]. \quad (9)$$

An analogous result holds if X and Y are discrete, or if $X\in\mathbb{R}^{m_1}$ and $Y\in\mathbb{R}^{m_2}$ for some $(m_1,m_2)\in\mathbb{N}^2$. This can be seen by appropriately making simple modifications to the following proof:

Proof. Let $A=\{x: f(x)>0\}$ and $B=\{y: f(y)>0\}$.

$$\begin{aligned}\mathbb{E}[g(X,Y)] &= \int_y f(y)\int_x f(x|y)g(x,y)dxdy \\ &= \int_{y\in B} f(y)\int_{x\in A}\frac{f(y|x)}{f(y)}f(x)g(x,y)dxdy \\ &\le \int_{y\in B} f(y)\left(\sup_{x\in A}\frac{f(y|x)}{f(y)}\right)\int_x f(x)g(x,y)dxdy \\ &\le \left(\int_{y\in B}\sup_{x\in A} f(y|x)\right)\sup_{y\in B}\mathbb{E}[g(\bar{X},y)]dy. \quad (10)\end{aligned}$$

□

An immediate application is the following: suppose we have an event $E\subseteq\mathcal{X}\times\mathcal{Y}$ and that we want to assess how much the probability of E changes with respect to the statistically independent scenario, if we introduce some dependence of Y on X. For every $y\in\mathcal{Y}$ we can define the event $E_y=\{x:(x,y)\in E\}$. Choosing $g(X,Y)=\mathbb{1}\{X\in E_Y\}$ and applying Theorem 1 we can bound $P_{XY}(E)$ as follows:

Corollary 1.

$$P_{XY}(E) \le \exp(\mathcal{L}\left(X\to Y\right))\cdot\sup_{y\in\mathcal{Y}} P_X(E_y). \quad (11)$$

This subsumes Theorem 2 of [1]. We will now explore how this result can be applied in providing bounds on the generalization error of learning algorithms.

Corollary 2. *Let $\mathcal{X}\times\mathcal{Y}$ be the sample space and $\mathcal{H}$ be the set of hypotheses. Let $\mathcal{A}:\mathcal{X}^n\times\mathcal{Y}^n\to\mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h\in\mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{X}\times\mathcal{Y}$, i.e., $S\sim\mathcal{P}^n$. Given $\eta\in(0,1)$, let $E=\{(S,h): |L_{\mathcal{P}}(h)-L_S(h)|>\eta\}$. Then,*

$$\mathbb{P}(E) \le 2\cdot\exp(\mathcal{L}\left(S\to\mathcal{A}(S)\right) - 2n\eta^2). \quad (12)$$

Proof. Let us denote with E_h the fiber of E over h. By McDiarmid's inequality, with $c=1/n$, we have that for every hypothesis $h\in\mathcal{H}$, $\mathcal{P}_S(E_h)\le 2\cdot\exp(-2n\eta^2)$, as the empirical error defined in equation (2) has sensitivity of $1/n$. Then it follows from Corollary 1 that:

$$\mathbb{P}(E) \le \exp(\mathcal{L}(S\to\mathcal{A}(S)))\cdot 2\exp(-2n\eta^2). \quad (13)$$

□

Whenever $\mathcal{A}$ is independent from the samples S we have that $\exp(\mathcal{L}(S\to\mathcal{A}(S)))=1$, as $\mathbb{P}(\mathcal{A}(S)=y|S=s)=\mathbb{P}(\mathcal{A}(S)=y)$ and the Maximal Leakage is 0. We immediately fall back to the non-adaptive scenario: $\mathbb{P}_{X\times Y}(E)=\mathbb{P}_{XY}(E)\le 1\cdot 2\exp(-2n\eta^2)$ i.e., McDiarmid's inequality with sensitivity $1/n$. A simple way to keep the maximal leakage of an algorithm $\mathcal{A}(X)$ bounded (and thus ensure generalization) is to add noise (e.g., $\hat{Y}=\mathcal{A}(X)+N$). Consider the following mechanism analyzed in [5]:

Example III.1. *Consider a noisy version of the ERM (empirical risk minimization) where $\mathcal{A}(S)=\arg\min_{h\in\mathcal{H}}(L_S(h)+N_h)$, with N_h exponential noise independently added to*

the empirical risk of each hypothesis on S. Suppose that $|\mathcal{H}| = k$ and that $\mathbb{E}[N_i] = i^{1.1}/n^{1/3}$ (with N_i being the noise added to the i-th hypothesis), we have that: $\mathcal{L}(S{\to}H) \leq \sum_{i=1}^{k} \log(1 + n^{1/3}/i^{1.1}) \leq 11 \cdot n^{1/3}$. This implies that $\mathbb{P}(E) \leq 2\exp(-n(2\eta^2 - 11/n^{2/3}))$ and that for every η and with n large enough the bound approaches 0 exponentially fast. The details of the derivation are omitted for space constraints.

Different from [4], [5] our results apply to generic loss functions and do not require the assumption of σ-sub gaussianity. Another interesting application of Corollary 1 may be the following: consider the problem of bounding the probability of making a false discovery, when the statistic to apply is selected with some data dependent algorithm $\mathcal{T}$. Measuring the information leaked from the data through $\mathcal{T}$ with the maximal leakage we retrieve the following:

Corollary 3. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{T}$ be a data dependent algorithm for selecting a test statistic $t \in \mathcal{T}$. Let $\mathbf{X}$ 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 for the test statistic t. 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,*

$$\mathbb{P}(E) \leq \exp(\mathcal{L}(\mathbf{X} \to \mathcal{A}(\mathbf{X}))) \cdot \sigma \tag{14}$$

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 $\delta/\exp(\mathcal{L}(\mathbf{X} \to \mathcal{A}(\mathbf{X}))$. Once again, if $\mathcal{A}$ is independent from $\mathbf{X}$, we recover the bound of σ.

IV. Adaptive Data Analysis

Other than providing a nice generalization of the classical bounds for adaptive scenarios, maximal leakage can also be employed in adaptive data analysis. The model of adaptive composition we will consider is identical to the setting in [8], [9] and defined as follows:

Definition 5. *Let $\mathcal{X}$ be a set. Let S be a rv over $\mathcal{X}^n$. Let $(\mathcal{A}_1, \ldots, \mathcal{A}_m)$ be a sequence of algorithms such that $\forall i : 1 \leq i \leq m \quad \mathcal{A}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{i-1} \to \mathcal{Y}_i$. Denote with $Y_1 = \mathcal{A}_1(S), Y_2 = \mathcal{A}_2(S, Y_1), \ldots, Y_m = \mathcal{A}_m(S, Y_1, \ldots, Y_{m-1})$. The adaptive composition of $(\mathcal{A}_1, \ldots, \mathcal{A}_m)$ is an algorithm that takes as an input S and sequentially executes the algorithms $(\mathcal{A}_1, \ldots, \mathcal{A}_m)$ as described by the sequence $(Y_i, 1 \leq 1 \leq m)$*

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 1 (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, $\mathcal{L}(X{\to}\mathcal{B}(\mathcal{A}(X))) \leq \mathcal{L}(X{\to}\mathcal{A}(X))$.*

Regarding adaptive composition of two algorithms, we retrieve the following:

Lemma 2 (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 P_X only through its support.

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

$$\forall y \in \mathcal{Y} \quad \mathcal{L}(X \to Z_y) \leq k_2 \iff \tag{15}$$

$$\forall y \in \mathcal{Y} \sum_{z_y \in R_{Z_y}} \max_{x \in R_X} \mathbb{P}(z_y|x) \leq \exp(k_2) \iff \tag{16}$$

$$\forall y \in \mathcal{Y} \sum_{z_y \in R_{Z|Y=y}} \max_{x \in R_X} \mathbb{P}(z|x, y) \leq \exp(k_2). \tag{17}$$

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 equivalent to $\mathbb{P}(z|x, y)$, with $z = \mathcal{B}(x, y)$. Using this observation in the conditional leakage of (8):

$$\mathcal{L}(X \to Z|Y) = \log \max_{y \in R_Y} \sum_{z \in R_{Z|Y=y}} \max_{x \in R_{X|Y=y}} \mathbb{P}(z|x, y) \tag{18}$$

$$\leq \log \max_{y \in R_Y} \sum_{z \in R_{Z|Y=y}} \max_{x \in R_X} \mathbb{P}(z|x, y) \tag{19}$$

$$\leq \log \max_{y \in R_Y} \exp(k_2) = k_2. \tag{20}$$

□

Considering Definition (5), we can show that if the maximal leakage of each algorithm $\mathcal{A}_i$ is bounded then the maximal leakage of the whole composition remains bounded, regardless of the intricate dependencies that the sequence may create:

Lemma 3. *Consider a sequence of algorithms:$(\mathcal{A}_1, \ldots, \mathcal{A}_n)$ such thatfor each $1 \leq i \leq n$, $\mathcal{A}_i : \mathcal{X} \times \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{i-1} \to \mathcal{Y}_i$. Suppose that for all i and for all $(y_1, \ldots, y_{n-1}) \in \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_{i-1}$, $\mathcal{L}(X \to \mathcal{A}_i(X, y_1, \ldots, y_{i-1})) \leq k_i$, then:*

$$\mathcal{L}(X \to (A_1, \ldots, A_n)) = \mathcal{L}(X \to \boldsymbol{A}^n) \leq \sum_{i=1}^{n} k_i. \tag{21}$$

The proof hinges on a generalization of [13, Corollary 2.8], e.g.: $\mathcal{L}(X \to (A_1, \ldots, A_n)) \leq \mathcal{L}(X \to A_1) + \mathcal{L}(X \to A_2|A_1) + \ldots + \mathcal{L}(X \to A_n|(A_1, \ldots, A_{n-1}))$, and similar arguments to the ones in the proof of Lemma 2 [12].

Hence, 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 (thus creating multiple dependencies on the data)

the generalization guarantees of the composition can still be maintained.

V. Comparison with other bounds

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 [14], [15]:

$$\mathbb{P}(E) \leq \frac{I(S;\mathcal{A}(S)) + \log 2}{2n\eta^2 - \log 2}. \tag{22}$$

Let us compare this result with Corollary 2 in terms of sample complexity.

Definition 6. *Fix $\eta, \delta \in (0,1)$. Let $\mathcal{H}$ be a hypothesis class. The sample complexity of $\mathcal{H}$ with respect to (η, δ), denoted by $m_{\mathcal{H}}(\epsilon, \delta)$, is defined as the smallest $n \in \mathbb{N}$ for which there exists a learning algorithm $\mathcal{A}$ such that, for every distribution $\mathcal{P}$ over the domain $\mathcal{X}$ yields*

$$\mathbb{P}(\textit{gen-err}_{\mathcal{P}}(\mathcal{A}, S) > \eta) \leq \delta. \tag{23}$$

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

From Corollary 2, it follows that using a sample size of $n \geq \left(\frac{\mathcal{L}(S\to\mathcal{A}(S))+\ln(1/\delta)}{\eta^2}\right)$ yields a learner for $\mathcal{H}$ with accuracy η and confidence δ and this, in turn, implies that $m_{\mathcal{H}}(\eta,\delta) = O\left(\frac{\mathcal{L}(S\to\mathcal{A}(S))+\ln(1/\delta)}{\eta^2}\right)$. Using the same reasoning with inequality (22), we get $m_{\mathcal{H}}(\eta,\delta) = O\left(\frac{I(\mathcal{A}(S);S)}{\eta^2}\cdot\frac{1}{\delta}\right)$. Provided that $\mathcal{L}(X\to Y) \geq I(X;Y)$ [3], in the regime where the two measures behave similarly, the reduction in the sample complexity is exponential in δ. Moreover, as shown in [14], if we consider the case where $\mathcal{X} = [d]$ and $\mathcal{H} = \{0,1\}^{\mathcal{X}}$, we have that the VC-dimension of $\mathcal{H}$ is d and, since $\mathcal{L}(S \to \mathcal{A}(S)) \leq \log(|\mathcal{H}|) \leq d$, our bound recovers exactly the VC-dimension bound [16], which is always sharp.

B. Maximal Leakage and Differential Privacy

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

Definition 7. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be a randomized algorithm. $\mathcal{A}$ is (ϵ,δ)-differentially private if for every $\mathcal{S} \subseteq \mathcal{Y}$ and every $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ that differ only in one position:*

$$\mathbb{P}(\mathcal{A}(\mathbf{x}) \in \mathcal{S}) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(\mathbf{y}) \in \mathcal{S}) + \delta \tag{24}$$

A relationship with maximal leakage can be established:

Lemma 4. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an ϵ-Differentially Private randomized algorithm, if $\mathbf{X}$ is a discrete random vector distributed over $\mathcal{X}^n$ and $Y = \mathcal{A}(\mathbf{X})$ is a discrete r.v. then $\mathcal{L}(\mathbf{X} \to \mathcal{A}(\mathbf{X})) \leq \epsilon \cdot n$.*

The proof can be found in [12]. This suggests an immediate application of Corollary 2. Indeed, suppose $\mathcal{A}$ is an ϵ-DP algorithm, then:

$$\exp(\mathcal{L}(X \to Y) - 2n\eta^2) \leq \exp(-n(2\eta^2 - \epsilon)) \tag{25}$$

In order for the bound to be decreasing with n, we need $\epsilon < 2\cdot\eta^2$, where η represents the accuracy of the generalization error 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 [8, Theorem 9] that $\mathbb{P}(E) \leq \frac{1}{4}\exp\left(-n\eta^2/12\right)$. For large enough n, our bound is tighter if $\epsilon \leq \frac{23}{12}\eta^2$. It is also possible to see that enforcing differential privacy on some algorithm $\mathcal{A}$ induces generalization guarantees similar to those stated in Corollary 1: suppose $\mathcal{A}$ is ϵ-DP, with $\epsilon \leq \sqrt{\frac{\ln(1/\beta)}{2n}}$, and let $\max_y P_X(E_y) \leq \beta$ then [8, Theorem 11]:

$$\mathbb{P}(E) \leq 3\sqrt{\beta}. \tag{26}$$

The results we are providing are different in the sense that, measuring the dependence of Y on X through leakage we can use it as a multiplicative factor in estimating how far the measure of E (with respect to the joint) is from the maximum measure of E_y, over all the $y \in \mathcal{Y}$ (with respect to the marginal distribution of X). Our bound is thus tailored to the dependence that $\mathcal{A}$ induces on X, while the one in (26) is fixed. Furthermore we do not need to impose any constraint on the distributions of $\mathcal{A}(X)|X = x$ to provide it. To highlight the difference, suppose that $\epsilon < \frac{\log(3/\sqrt{\beta})}{n}$, using (26) we get a bound of $3\sqrt{\beta}$, while with Corollary 1 and Lemma 4 we obtain that:

$$\exp(\mathcal{L}(\mathbf{X} \to Y)) \cdot \beta < \exp\left(\log\left(3/\sqrt{\beta}\right)\right) \cdot \ \beta = 3\sqrt{\beta}. \tag{27}$$

Hence, whenever the privacy parameter is lower than $\frac{\log(3/\sqrt{\beta})}{n}$ we are able to provide a better bound. Also notice that Lemma 4 can be quite loose: it is possible to see that for classical mechanisms that imply ϵ-DP, maximal leakage can be much lower that $\epsilon \cdot n$.

Example V.1. *Let $f : \mathcal{X}^n \to \mathbb{R}$ be a function of sensitivity $1/n$ and let $N \sim Lap(1/n\epsilon)$ then the mechanism $\mathcal{M}(x) = f(x) + N$ is $\epsilon-$DP. Without loss of generality we have that $|f(x)| \leq 1$ (e.g. 0-1 loss) and this implies $\mathcal{L}(X \to \mathcal{M}(X)) = \log(1+\epsilon\cdot n) < \epsilon\cdot n$. The details of the derivation are omitted for space constraints.*

More importantly, the family of algorithms with bounded maximal leakage is not restricted to the differentially private ones. Indeed, whenever there is a deterministic mapping and ϵ-Differential Privacy is enforced on it, a lower bound on ϵ of $+\infty$ is retrieved. Trying to relax it to $(\epsilon,\delta)-$Differential Privacy does not help either, as one would need $\delta \geq 1$ rendering it practically useless, while if the algorithm has a bounded range the maximal leakage is always bounded, since $\mathcal{L}(X \to Y) \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$. This simple observation allows us to immediately retrieve another result [9, Theorem 9]: $\mathbb{P}(E) \leq |\mathcal{Y}|\cdot\beta$, showing how Corollary 1 is more general than both Theorems 6 and 9 of [9].

To conclude let us now state Corollary 2 with a general sensitivity c (thus, more general loss functions than the $0-1$ loss): $\mathbb{P}(E) \leq 2\cdot\exp\left(\mathcal{L}(\mathbf{X}\to Y) - \frac{2\eta^2}{c^2 n}\right)$. [9, Corollary 7]

states that whenever an algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ outputs a function f of sensitivity c and is $\eta/(cn)-$DP then, denoting with S a random variable distributed over $\mathcal{X}^n$ and with $E = \{(S, f) : f(S) - \mathbb{E}(f) \geq \eta\}$ we have that $\mathbb{P}(E) \leq 3\exp(-\eta^2/(c^2 n))$. Our bound is tighter whenever $\eta > n \cdot c$.

C. Maximal Leakage and Max Information

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

Lemma 5 ([9])**.** *Let $\mathbf{X}$ be a random dataset in $\mathcal{X}^n$ and let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be such that for some $\beta \geq 0$, $I_\infty^\beta(\mathbf{X}, \mathcal{A}(\mathbf{X})) = k$ then, for any event $\mathcal{O} \subseteq X^n \times \mathcal{Y}$:*

$$\mathbb{P}_{(\mathbf{X}, \mathcal{A}(\mathbf{X}))}(\{(\mathbf{X}, \mathcal{A}(\mathbf{X})) \in \mathcal{O}\}) \leq \tag{28}$$

$$e^k \cdot \mathbb{P}_{\mathbf{X} \times \mathcal{A}(\mathbf{X})}(\{(\mathbf{X}, \mathcal{A}(\mathbf{X})) \in \mathcal{O}\}) + \beta \tag{29}$$

The result looks quite similar to Corollary 2, but the two measures, max-information and maximal leakage, although related, can be quite different. Indeed:

Lemma 6. *Let X, Y be two discrete random variables such that $I_\infty(X;Y) \leq k$, then, $\mathcal{L}(X \to Y) \leq k$.*

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

Lemma 7. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be a randomized algorithm. Let $\mathbf{X}$ be a discrete rv distributed over $\mathcal{X}^n$ and let $Y = \mathcal{A}(\mathbf{X})$. For any $\beta \in (0, 1)$*

$$I_\infty^\beta(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq \mathcal{L}(\mathbf{X} \to \mathcal{A}(\mathbf{X})) + \log\left(\frac{1}{\beta}\right). \tag{30}$$

The role played by β can lead to undesirable behaviours of β-approx MI. The following example shows how β-approx MI can be unbounded while, in the discrete case, the maximal leakage between two random variables is always bounded by the logarithm of the smaller cardinality.

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

Another interesting property of max-information is that, differently from differential privacy, it can be bounded even for deterministic algorithms [9]: let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be a randomized algorithm, for every $\beta > 0$,

$$I_\infty^\beta(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq \log\left(\frac{|\mathcal{Y}|}{\beta}\right). \tag{31}$$

By contrast, with maximal leakage we have

$$\mathcal{L}(\mathbf{X} \to \mathcal{A}(\mathbf{X})) \leq \log(|\mathcal{Y}|) \tag{32}$$

Clearly, being $0 < \beta$ typically very small in the key applications, the corresponding multiplicative factors in the bounds are $(|\mathcal{Y}|/\beta)$ and $|\mathcal{Y}|$, and the difference between the two bounds can be substantial. It is also worth noticing that (31) can be seen as a consequence of Lemma 7 and (32). To conclude, the difference between the two measures is not uniquely restricted to deterministic mappings:

Example V.3. *Consider $X \sim Ber(1/2)$ and $Y = BEC_\alpha(X)$ (the output of a BEC, with erasure probability α, when X is transmitted). More formally, we have that $\mathcal{Y} = \{0, 1, e\}$ and the following randomized mapping: $\mathbb{P}(Y = e|X = x) = \alpha$ and $\mathbb{P}(Y = x|X = x) = 1 - \alpha$. In this case, the maximal leakage is $\mathcal{L}(X \to Y) = \log(2 - \alpha)$ [13]; while, for β-approximate max-information one finds : $I_\infty^\beta(X;Y) = \log(2 \cdot \max\{(1-\alpha-\beta)/(1-\alpha), (1-\beta)/(1+\alpha)\})$; for a fixed α and for β going to 0, approximate max-information approaches $\log 2$ while maximal leakage is strictly smaller.*

References

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

[2] A. D. Smith, "Information, privacy and stability in adaptive data analysis," *CoRR*, vol. abs/1706.00820, 2017.

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

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

[5] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," p. 2521–2530, 2017.

[6] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 1475–1479.

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

[8] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, "Preserving statistical validity in adaptive data analysis," in *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*. New York, NY, USA: ACM, 2015, pp. 117–126.

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

[10] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, ser. STOC '16. New York, NY, USA: ACM, 2016, pp. 1046–1059.

[11] R. M. Rogers, A. Roth, A. D. Smith, and O. D. Thakkar, "Max-information, differential privacy, and post-selection hypothesis testing," *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–494, 2016.

[12] A. R. Esposito, M. Gastpar, and I. Issa, "A new approach to adaptive data analysis and learning via maximal leakage," 2019. [Online]. Available: https://arxiv.org/abs/1903.01777

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

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

[15] E. A. Arutjunjan, "Bounds for the Exponent of the Probability of Error for a Semicontinuous Memoryless Channel," *Probl. Peredachi Inf.*, vol. 4, no. 4, pp. 37–48, 1968.

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