Strengthened Information-theoretic Bounds on the Generalization Error

Ibrahim Issa, Amedeo Roberto Esposito, Michael Gastpar
EPFL
{ibrahim.issa,amedeo.esposito,michael.gastpar}@epfl.ch,

***Abstract*—The following problem is considered: given a joint distribution P_{XY} and an event E, bound $P_{XY}(E)$ in terms of $P_X P_Y(E)$ (where $P_X P_Y$ is the product of the marginals of P_{XY}) and a measure of dependence of X and Y. Such bounds have direct applications in the analysis of the generalization error of learning algorithms, where E represents a large error event and the measure of dependence controls the degree of overfitting. Herein, bounds are demonstrated using several information-theoretic metrics, in particular: mutual information, lautum information, maximal leakage, and J_∞. The mutual information bound can outperform comparable bounds in the literature by an arbitrarily large factor.**

I. Introduction

One of the main challenges in designing learning algorithms is guaranteeing that they generalize well [1]–[4]. The analysis is made especially hard by the fact that, in order to handle large data sets, learning algorithms are typically adaptive. A recent line of work initiated by Dwork *et al.* [5]–[7] shows that differentially private algorithms provide generalization guarantees. More recently, Russo and Zou [8], and Xu and Raginsky [9], provided an information-theoretic framework for this problem, and showed that the mutual information (between the input and output of the learning algorithm) can be used to bound the generalization error, under a certain assumption. Jiao *et al.* [10] and Issa and Gastpar [11] relaxed this assumption and provided new bounds using new information-theoretic measures.

The aforementioned papers mainly study the expected generalization error. In this paper, we focus instead the *probability* of an undesirable event (e.g., large generalization error in the learning setting). In particular, given an event E and a joint distribution P_{XY}, we bound $P_{XY}(E)$ in terms of $P_X P_Y(E)$ (where $P_X P_Y$ is the product of the marginals of P_{XY}) and a measure of dependence between X and Y.

A bound of this form has been previously derived [12]–[14] where the measure of dependence is mutual information, $I(X;Y)$. We present a new bound in terms of mutual information, which can outperform the existing one by an arbitrarily large factor. Moreover, we prove a new bound using lautum information (a measure introduced by Palomar and Verdú [15]). We demonstrate two further bounds using maximal leakage [16,17] and $J_\infty(X;Y)$ (which was recently introduced by Issa and Gastpar [11]). One advantage of the latter two bounds is that they have a closed-form expression and depend on P_{XY} only through $P_{Y|X}$, hence they are more amenable to analysis.

II. KL Divergence bounds

Let P_{XY} be a joint probability distribution on alphabets $\mathcal{X}\times\mathcal{Y}$, and let $E \subseteq \mathcal{X}\times\mathcal{Y}$ be some ("undesirable") event. We want to bound $P_{XY}(E)$ in terms of $P_X P_Y(E)$ (where $P_X P_Y$ is the product of the marginals induced by the joint P_{XY}) and a measure of dependence between X and Y.

A. Mutual Information Bounds

To this end, consider the following intermediate problem: let P and Q be two probability distributions on an alphabet $\mathcal{Z}$, and let $E \subseteq \mathcal{Z}$ be some event. We will bound $P(E)$ in terms of $Q(E)$ and $D(P||Q)$. Then by replacing P by P_{XY} and Q by $P_X P_Y$, we get a bound for our desired setup in terms of the mutual information $I(X;Y) = D(P_{XY}||P_X P_Y)$.

Theorem 1: Given $q \in (0,1)$, define $f_q : [q,1] \to \mathbb{R}_+$ as $f_q(p) = D(p||q)$. Then, $f_q(p)$ is a strictly increasing function of p. Given any event E and pair of distributions P and Q with $D(P||Q) \le \log \frac{1}{Q(E)}$,

$$P(E) \le f^{-1}_{Q(E)}\left(D(P||Q)\right). \tag{1}$$

In particular, given an event $E \subseteq \mathcal{X}\times\mathcal{Y}$ and a joint distribution P_{XY} satisfying $I(X;Y) \le \log\frac{1}{P_X P_Y(E)}$,

$$P_{XY}(E) \le f^{-1}_{P_X P_Y(E)}\left(I(X;Y)\right). \tag{2}$$

Proof: Note that $\frac{df_q(p)}{dp} = \log\left(\frac{p}{q}\frac{1-q}{1-p}\right) > 0$ for $p > q$, hence $f_q(p)$ is strictly increasing. Moreover, the range of $f_q(p)$ is $[0, \log(1/q)]$, so (1) is well defined.

If $P(E) \le Q(E)$, then (1) holds trivially since $f^{-1}_{Q(E)}(D(P||Q)) \ge Q(E)$ by the definition of f.

Otherwise, if $P(E) > Q(E)$, then $f_{Q(E)}(P(E)) = D(P(E)||Q(E)) \leq D(P||Q)$, where the second inequality follows from the data processing inequality. Since f_q is strictly increasing, then so is f_q^{-1}. Hence $P(E) \leq f_{Q(E)}^{-1}(D(P||Q))$. ∎

Remark 1: The bound above is tight in the following sense. Let $g : [0,1] \times \mathbb{R}_+ \to [0,1]$ be such that, given any alphabet $\mathcal{Z}$ and event $E \subseteq \mathcal{Z}$, and any two distributions P and Q on $\mathcal{Z}$, $P(E) \leq g(Q(E), D(P||Q))$. Then $g(Q(E), D(P||Q)) \geq f_{Q(E)}^{-1}(D(P||Q))$ if $D(P||Q) \leq \log \frac{1}{Q(E)}$. This is true since given any tuple $(\mathcal{Z}, P, Q, E)$ such that $D(P||Q) \leq \log \frac{1}{Q(E)}$, there exists $(\mathcal{Z}', P', Q', E')$ such that $Q(E') = Q(E)$, $D(P||Q) = D(P'||Q')$, and (1) holds with equality. In particular, choose $\mathcal{Z}' = \{0,1\}$, $E' = \{1\}$, $Q' \sim \text{Ber}(Q(E))$, and $P' \sim \text{Ber}\left(f_{Q(E)}^{-1}(D(P||Q))\right)$.

However, there is no closed form for the bound in (1). The following corollary provides an upper bound in closed form:

Corollary 1: Given $q \in (0, 1/2]$, define $g_q(y) := \log^2(2) + (\log(1-q)+y)(-\log(q)-y)$ and $\hat{f}_q : [0, -\log(q)) \to \mathbb{R}_+$ as follows:

$$\hat{f}_q(y) = \frac{2\log^2(2) + (\log(1-q)+y)\log\frac{(1-q)}{q} + (\log 4)\sqrt{g_q(y)}}{\log^2((1-q)/q) + \log^2(2)}.$$

Then, $\hat{f}_q(y)$ is concave and non-decreasing in y. Moreover, given any event E and pair of distributions P and Q with $D(P||Q) \leq \log \frac{1}{Q(E)}$,

$$P(E) \leq \hat{f}_{Q(E)}(D(P||Q)). \tag{3}$$

In particular, given an event $E \subseteq \mathcal{X} \times \mathcal{Y}$ and a joint P_{XY} satisfying $I(X;Y) \leq \log \frac{1}{P_X P_Y(E)}$,

$$P_{XY}(E) \leq \hat{f}_{P_X P_Y(E)}(I(X;Y)). \tag{4}$$

Proof: Since $g_q(y)$ is concave in y and the square root is concave and non-decreasing, $\sqrt{g_q(y)}$ is concave in y; hence $\hat{f}_q(y)$ is concave in y. To show that it is non-decreasing, consider the derivative (ignoring the positive denominator):

$$\frac{d\hat{f}_q(y)}{dy} = \log\frac{1-q}{q} + \log(4)\frac{-2y - \log(q(1-q))}{2\sqrt{g_q(y)}}.$$

For $y \in [0, -\frac{1}{2}\log(q(1-q))]$, both terms are non-negative (the first is non-negative since $q \leq 1/2$). For $y \in [-\frac{1}{2}\log(q(1-q)), -\log(q)]$, the numerator of the second term is negative and decreasing, and the denominator is positive and decreasing. Hence, it achieves its minimum for $y = -\log(q)$. Since the minimum $\left.\frac{d\hat{f}_q(y)}{dy}\right|_{y=-\log(q)} = 0$, we get that $\frac{d\hat{f}_q(y)}{dy} \geq 0$ for $y \in [0, -\log(q)]$.

Now, let $p := P(E)$ and $q := Q(E)$. Then we can rewrite the inequality $D(p||q) \leq D(P||Q)$ as

$$-\log(1-q) + p\log\left(\frac{1-q}{q}\right) - h(p) \leq D(P||Q), \tag{6}$$

where $h(.)$ is the binary entropy function (in nats). Upper-bounding $h(p) \leq (\log 4)\sqrt{p(1-p)}$, we get

$$-\log(1-q) + p\log\frac{1-q}{q} - (\log 4)\sqrt{p(1-p)} \leq D(P||Q).$$

For ease of notation, let $y := D(P||Q)$ and $\tilde{g}(p)$ be the left-hand side. Then,

$$\frac{d\tilde{g}}{dp} = \log\left(\frac{1-q}{q}\right) - (\log 4)\frac{1-2p}{\sqrt{p(1-p)}}. \tag{7}$$

Hence, there exists p_0 such that $\tilde{g}$ is decreasing on $[0, p_0]$ and increasing on $[p_0, 1]$. Therefore, $\tilde{g}(p) = y$ admits at most two solutions, say $p_1 < p_2$, and $\tilde{g}(p) \leq y \Rightarrow p \leq p_2$. It remains to solve

$$p\log\frac{1-q}{q} - \log(1-q) - (\log 4)\sqrt{p(1-p)} = y. \tag{8}$$

Let $q_1 = \log\frac{1-q}{q}$, and $q_2 = \log(1-q)$. We get

$$\begin{aligned}(pq_1 - q_2 - y)^2 = p(1-p)\log^2(4), &\iff \\ p^2\left(q_1^2 + \log^2(4)\right) - 2p\left(2\log^2(2) + q_1(q_2+y)\right)& \\ + (q_2+y)^2 = 0.&\end{aligned} \tag{9}$$

The discriminant of (9) is given by

$$\begin{aligned}\frac{\Delta}{4} &= \left(2\log^2(2) + q_1(q_2+y)\right)^2 - (q_1^2 + \log^2(4))(q_2+y)^2 \\ &= (q_2+y)(4q_1\log^2(2) - (\log^2(4))(q_2+y)) + 4\log^4(2) \\ &= \left(4\log^2(2)\right)\left(\log^2(2) + (q_2+y)(q_1-q_2-y)\right) \geq 0,\end{aligned}$$

where the inequality follows from the fact that $q_1 - q_2 - y = -\log(q) - y \geq 0$. Hence, the larger root of (9) is given by $\hat{f}_q(p)$, as desired. ∎

1) Comparison with existing bounds: It has been shown [12] [14, Lemma 3.11] [13, Lemma 9] that

$$P(E) \leq \frac{D(P||Q) + \log(2)}{\log(1/Q(E))}. \tag{10}$$

The bound in Corollary 1 can be arbitrarily smaller than (10). That is, let $\tilde{f}_{Q(E)}(D(P||Q))$ be the right-hand side of (10) and consider the calculation shown at the top of the next page.

$$
\begin{aligned}
&\lim_{q\to 0}\lim_{D(P||Q)\to 0}\frac{\hat{f}_q(D(P||Q))}{\tilde{f}_q(D(P||Q))}\\
&\overset{(a)}{=}\lim_{q\to 0}\frac{\left(2\log^2(2)+q_1q_2+(2\log 2)\sqrt{\log^2(2)-q_2\log(q)}\right)\log(1/q)}{\left(q_1^2+\log^2(2)\right)\log(2)}\\
&=\lim_{q\to 0}\frac{\left(2\log^2(2)+\log^2(1-q)-\log(q)\log(1-q)+(2\log 2)\sqrt{\log^2(2)-\log(1-q)\log(q)}\right)\log(1/q)}{\left(\log^2(1-q)+\log^2(q)-2\log(q)\log(1-q)+\log^2(2)\right)\log(2)}\\
&\overset{(b)}{=}\lim_{q\to 0}\frac{4\log^2(2)\log(1/q)}{\left(\log^2(q)+\log^2(2)\right)\log(2)}\\
&=0,
\end{aligned}
\tag{5}
$$

where in (a) $q_1=\log\frac{1-q}{q}$ and $q_2=\log(1-q)$, and (b) follows from the fact that $\lim_{q\to 0}\log(q)\log(1-q)=0$.

Moreover, one can derive a family of bounds in the form of (10) using the Donsker-Varadhan characterization of the KL divergence. In particular,

$$D(P||Q)=\sup_{f:\mathcal{Z}\to\mathbb{R},\mathbf{E}_Q[e^f]<+\infty}\left\{\mathbf{E}_P[f]-\log\mathbf{E}_Q[e^f]\right\}. \tag{11}$$

Now, let $f=\beta\mathbb{I}\{z\in E\}$ for some $\beta>0$, where $\mathbb{I}\{\}$ is the indicator function. After rearranging terms, we get

$$P(E)\le\frac{D(P||Q)+\log\left(1+(e^\beta-1)Q(E)\right)}{\beta}. \tag{12}$$

Choosing $\beta=\log(1/Q(E))$, we slightly improve (10) by replacing $\log(2)$ with $\log(2-Q(E))$. In fact, we can solve the infimum over $\beta>0$ of the right-hand side of (12). In particular, by [18, Lemma 2.4], the infimum is given by $\ell^{\star-1}(D(P||Q))$, where $\ell^\star$ is the convex conjugate of[1] $\ell(\beta)=\log(1+(e^\beta-1)Q(E))$, and $\ell^{\star-1}(y)=\inf\{t:\ell^\star(t)>y\}$. It turns out that $\ell^\star:\mathbb{R}_+\to\mathbb{R}_+$ is given by

$$\ell^\star(t)=\begin{cases}0, & 0\le t<Q(E),\\ D(t||Q(E)), & Q(E)\le t\le 1,\\ +\infty, & t>1.\end{cases} \tag{13}$$

Now, $P(E)\le\inf\{t:\ell^\star(t)>D(P||Q)\}$. Hence, for $D(P||Q)=0$, $P(E)\le\inf(Q(E),+\infty)=Q(E)$. By noting that $\ell^\star(1)=\log(1/Q(E))$, we get for any $D(P||Q)>\log(1/Q(E))$, $P(E)\le\inf(1,+\infty)=1$. Finally, for $D(P||Q)\in(0,\log(1/Q(E))]$, we get $P(E)\le\{t\in[Q(E),1]:D(t||Q(E))>D(P||Q)\}$, which is equal to $t^\star\in[Q(E),1]$ satisfying $D(t^\star||Q(E))=D(P||Q)$. That is, the bound derived from (13) exactly recovers Proposition 1.

[1]Lemma 2.4 of [18] assumes $\ell''(0)=0$, but the proof goes as is for $\ell''(0)\ge 0$, which is the case here.

Furthermore, we could compare with the mutual information bound of Russo and Zou [8], and Xu and Raginsky [9]. In particular, by considering $f=\beta\left(\mathbb{I}\{z\in E\}-Q(E)\right)$ for $\beta\in\mathbb{R}$ in (11), we get

$$
\begin{aligned}
D(P||Q)&\ge\beta(P(E)-Q(E))-\log\mathbf{E}_Q\Big[e^{(\mathbb{I}\{Z\in E\}-Q(E))}\Big]\\
&\ge\beta(P(E)-Q(E))-\beta^2/8,
\end{aligned}
$$

where the second inequality follows from the fact that $(\mathrm{Ber}(q)-q)$ is $\frac{1}{4}$-subgaussian (which is true for any random variable whose support has length 1). Since the above inequality holds for any $\beta\in\mathbb{R}$, we get

$$P(E)\le Q(E)+\sqrt{\frac{D(P||Q)}{2}}. \tag{14}$$

Given the form of the 3 bounds, one might expect that (14) outperforms the other two for large values of $D(P||Q)$. This is in fact not true because the range of interest for the right-hand sides is restricted to $[0,1]$. For instance, for small $Q(E)$ and $D(P||Q)=-\log(Q(E))/2$, the bound in (14) is trivial (>1), and the other two bounds are strictly less than 1.

B. Lautum Information Bounds

By considering the data processing inequality $D(q||p)\le D(Q||P)$, we can bound p in terms of q and $D(Q||P)$.

Theorem 2: Given any event E and a pair of distributions P and Q, if $P(E)\le 1/2$, then

$$P(E)\le 1-e^{-h(Q(E))-D(Q||P)}.$$

In particular, given an event $E \subseteq \mathcal{X} \times \mathcal{Y}$ and a joint distribution P_{XY} with $P_{XY}(E) \leq 1/2$,

$$P_{XY}(E) \leq 1 - e^{-h(P_X P_Y(E)) - L(X;Y)}, \qquad (15)$$

where $L(X;Y) := D(P_X P_Y || P_{XY})$ is *the lautum information* [15].

Proof: Set $p = P(E)$ and $q = Q(E)$. As in (6), we can rewrite $D(q||p) \leq D(Q||P)$ as

$$q \log\left(\frac{1-p}{p}\right) - \log(1-p) - h(q) \leq D(Q||P). \quad (16)$$

Since $p \leq 1/2$ (by assumption), we can drop the first term of the left-hand side. Rearranging the inequality then yields Theorem 2. ■

Moreover, we can derive a family of bounds similar to (12) by considering the Donsker-Varadhan representation of $D(Q||P)$:

$$D(Q||P) = \sup_{f:\mathcal{Z}\to\mathbb{R}, \mathbf{E}_P[e^f]<+\infty} \left\{\mathbf{E}_Q[f] - \log \mathbf{E}_P[e^f]\right\}. \quad (17)$$

Now, let $f = -\beta \mathbb{I}\{z \in E\}$ for some $\beta > 0$. Then after rearranging terms, we get for any $\beta > 0$,

$$P(E) \leq \frac{1 - e^{-D(Q||P) - \beta Q(E)}}{1 - e^{-\beta}}. \qquad (18)$$

III. Maximal Leakage Bound

The bounds presented so far in (4) and (15) do not take into account the specific relation of P_{XY} and $P_X P_Y$ as a joint distribution and its marginal. Indeed, they are applications of a more general bound that can be applied to an arbitrary pair of distributions (Corollary 1 and Theorem 2). The following bound does not fall under this category, i.e., it only applies to pairs of distributions forming a joint and marginal.

Theorem 3: Given $\alpha \in [0,1]$, finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, a joint distribution P_{XY} and an event $E \subseteq \mathcal{X} \times \mathcal{Y}$ such that for all $y \in \mathcal{Y}$, $P_X(E_y) \leq \alpha$ where $E_y := \{x : (x,y) \in E\}$, then

$$P_{XY}(E) \leq \alpha \exp\left\{\mathcal{L}(X \to Y)\right\}, \qquad (19)$$

where $\mathcal{L}(X \to Y) = \log \sum_{y\in\mathcal{Y}} \max_{x:P_X(x)>0} P_{Y|X}(y|x)$ is the maximal leakage.

Remark 2: The bound holds more generally but we restrict our attention to finite alphabets to make the presentation of the proof simple.

Remark 3: A similar inequality appeared (without proof) in [19].

Maximal leakage has recently appeared in the information theory literature [17] as an operational measure of information leakage:

Definition 1: Given a joint distribution P_{XY} on finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, the *maximal leakage* from X to Y is defined as

$$\mathcal{L}(X \to Y) = \sup_{U - X - Y - \hat{U}} \log \frac{\mathbf{Pr}\left(U = \hat{U}\right)}{\max_{u\in\mathcal{U}} P_U(u)},$$

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

That is, given X and Y, $\mathcal{L}(X \to Y)$ is given by (the logarithm of) the multiplicative increase of the probability of guessing any (possibly randomized) function of X by observing Y (as compared with no observations). Hence, as a leakage metric, one can view maximal leakage as controlling the degree of dependence between the input and the output.

Proof of Theorem 3: Fix $y \in \mathcal{Y}$ satisfying $P_Y(y) > 0$, and consider the pair of distributions $P_{X|Y=y}$ and P_X:

$$\begin{aligned}\exp\{D_\infty(P_{X|Y=y}||P_X)\} &= \sup_{A\subseteq\mathcal{X}} \frac{P_{X|Y=y}(A)}{P_X(A)} \\ &= \max_{x:P_{X|Y}(x|y)>0} \frac{P_{X|Y}(x|y)}{P_X(x)}.\end{aligned}$$

where the equalities follow from [20, Theorem 6]. Hence,

$$\begin{aligned}P_{X|Y=y}(E_y) &\leq \alpha \max_{x:P_{X|Y}(x|y)>0} \frac{P_{X|Y}(x|y)}{P_X(x)} \\ &= \alpha \max_{x:P_{X|Y}(x|y)>0} \frac{P_{Y|X}(y|x)}{P_Y(y)}.\end{aligned}$$

Now,

$$\begin{aligned}P_{XY}(E) &= \mathbf{E}_Y\left[P_{X|Y=y}(E_y)\right] \\ &\leq \alpha \sum_{y:P_Y(y)>0} \max_{x:P_{X|Y}(x|y)>0} P_{Y|X}(y|x) \\ &\overset{(a)}{=} \alpha \sum_{y:P_Y(y)>0} \max_{x:P_X(x)>0} P_{Y|X}(y|x) \\ &= \alpha \sum_{y\in\mathcal{Y}} \max_{x:P_X(x)>0} P_{Y|X}(y|x)\end{aligned}$$

where (a) follows from the following (readily verifiable) facts:

$P_Y(y) > 0$ and $P_{X|Y}(x|y) > 0 \Rightarrow P_X(x) > 0$,
$P_Y(y) > 0$ and $P_{X|Y}(x|y) = 0 \Rightarrow P_{Y|X}(y|x) = 0$. ■

The bound of Theorem 3 outperforms the bound in (10) if and only if

$$\frac{e^{\mathcal{L}(X\to Y)}}{I(X;Y) + \log 2} \leq \frac{1}{\alpha \log(1/\alpha)}.$$

In applications of interest, the input consists of n i.i.d samples, and α is exponentially small. The above inequality thus holds in certain cases of interest.

One advantage of the bound of Theorem 3 is that it depends on a partial description of $P_{Y|X}$ only. By contrast, maximizing the mutual information bounds over P_X would not yield a closed-form solution. Hence, the above bound is simpler to analyze than the mutual information bounds. Moreover, for fixed P_X, the bound is convex in $P_{Y|X}$. In the next subsection, we present a bound with similar properties.

IV. J_∞-BOUND

Theorem 4: Given $\alpha \in [0, 1/2]$, finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, a joint distribution P_{XY} and an event $E \subseteq \mathcal{X} \times \mathcal{Y}$ such that for all $y \in \mathcal{Y}$, $P_X(E_y) \leq \alpha$ where $E_y := \{x : (x, y) \in E\}$, then

$$P_{XY}(E) \leq \alpha \left(2(1-\alpha) J_\infty(X;Y) + 1\right), \tag{20}$$

where $J_\infty(X;Y) = \frac{1}{2} \sum_{y \in \mathcal{Y}} \left(\max_x P_{Y|X}(y|x) - \min_x P_{Y|X}(y|x)\right)$ [11].

Proof: The theorem follows from Theorem 1 and Corollary 1 of [11]. In particular, following the same proof steps as in [11], one can show that for any function[2] $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$,

$$\begin{aligned} &\left|\mathbf{E}_{P_{XY}}[f(X,Y)] - \mathbf{E}_{P_X P_Y}[f(X,Y)]\right| \leq \\ &\left(\max_y \mathbf{E}_{P_X}[|f(X,y) - \mu_y|]\right) J_\infty(X;Y), \end{aligned} \tag{21}$$

where $\mu_y := \mathbf{E}_{P_X}[f(X,y)]$. Now, set $f(x,y) = \mathbb{I}\{(x,y) \in E\}$. Then, $\mathbf{E}_{P_{XY}}[f(X,Y)] = P_{XY}(E)$, $\mathbf{E}_{P_X P_Y}[f(X,Y)] = P_X P_Y(E) \leq \alpha$, and $\mathbf{E}_{P_X}[f(X,y)] = P_X(E_y)$. Moreover,

$$\begin{aligned} \mathbf{E}_{P_X}[|f(X,y) - P_X(E_y)|] &= 2P_X(E_y)\left(1 - P_X(E_y)\right) \\ &\leq \alpha(1-\alpha), \end{aligned}$$

where the last inequality follows from the assumption that $P_X(E_y) \leq \alpha \leq \frac{1}{2}$. Then, it follows from (21) that

$$P_{XY}(E) - P_X P_Y(E) \leq 2\alpha(1-\alpha) J_\infty(X;Y). \tag{22}$$

The theorem follows by noting that $P_X P_Y(E) \leq \alpha$. ∎

ACKNOWLEDGMENT

This work was supported in part by the Swiss National Science Foundation under Grant 169294.

[2] In [11], the authors consider $X = (X_1, \cdots, \mathcal{X}_n)$, $\mathcal{Y} = \{1, 2, \cdots, n\}$, and $f(X,Y) = X_Y$. Nevertheless, the proof of (21) remains the same.

REFERENCES

[1] J. P. Ioannidis, "Why most published research findings are false," *PLoS medicine*, vol. 2, no. 8, p. e124, 2005.

[2] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological science*, vol. 22, no. 11, pp. 1359–1366, 2011.

[3] M. Hardt and J. Ullman, "Preventing false discovery in interactive data analysis is hard," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 454–463.

[4] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2635–2670, 2010.

[5] 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*. ACM, 2015, pp. 117–126.

[6] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," *CoRR*, vol. abs/1506.02629, 2015. [Online]. Available: http://arxiv.org/abs/1506.02629

[7] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic Stability for Adaptive Data Analysis," *ArXiv e-prints*, Nov. 2015.

[8] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Artificial Intelligence and Statistics*, 2016, pp. 1232–1240.

[9] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *arXiv preprint arXiv:1705.07809*, 2017.

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

[11] I. Issa and M. Gastpar, "Computable bounds on the exploration bias," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2018, pp. 576–580.

[12] 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.

[13] 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. [Online]. Available: http://proceedings.mlr.press/v83/bassily18a.html

[14] V. Feldman and T. Steinke, "Calibrating noise to variance in adaptive data analysis," *arXiv preprint arXiv:1712.07196*, 2017.

[15] D. P. Palomar and S. Verdú, "Lautum information," *IEEE transactions on information theory*, vol. 54, no. 3, pp. 964–975, 2008.

[16] C. Braun, K. Chatzikokolakis, and C. Palamidessi, "Quantitative notions of leakage for one-try attacks," *Electronic Notes in Theoretical Computer Science*, vol. 249, pp. 75–91, 2009.

[17] I. Issa, S. Kamath, and A. B. Wagner, "An operational measure of information leakage," in *Proc. of 50th Ann. Conf. on Information Sciences and Systems (CISS)*, Mar. 2016.

[18] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

[20] T. van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.