ETH Zurich

Department of Mathematics

Generalization Error Bounds via Information-Theoretic Methods and Convex Analysis

Semester Paper

Submitted by:
Gnazzo Angelo

Supervisors:
Dr. Zhivotovskiy Nikita
Postdoctoral Fellow at ETH Zurich
Prof. Dr. Bandeira Afonso
Professor of Mathematics at ETH Zurich

Zurich, June 2022

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Abstract

Generalization error bounds are critical to understanding the performance of machine learning models. The last few years have witnessed important new advancements in the study of these bounds. The purpose of this work is to offer an overview of some of these developments and of the related theoretical tools. Russo and Zou, Xu and Raginsky were the first to obtain bounds that provided an information-theoretic understanding of generalization in learning problems. In this survey we study the generalization error of supervised learning algorithms, first in terms of the mutual information between their input and the output and then, we study the recent generalization of the results of Neu and Lugosi beyond the choice of the mutual information to measure the dependence between the input and the output. They showed that it is possible to replace the mutual information by any strongly convex function of the joint input-output distribution. This new approach provides us generalization bounds that improve the previously known ones or are entirely new. Moreover, we show some bounds constructed in terms of the mutual information between each individual training sample and the output of the learning algorithm, which can be tighter and easier to evaluate in several applications.

Contents

1 Introduction

In the standard model of supervised learning, we are given a set $S_n = \{Z_1, \ldots, Z_n\}$ of n i.i.d data points drawn from a distribution μ (the input) and we consider a learning algorithm that maps this dataset to an output $W_n = \mathcal{A}\left(S_n\right)$. In this framework the learning algorithm can be viewed as a randomized mapping. From now on, we will denote with:

- Z, the instance space (i.e. where the data points can take values)
- W, the hypothesis space (i.e. the space the output belongs to)

In this discussion we will assume these spaces to be measurable.

The learning algorithm is characterized by a Markov kernel $P_{W|S} = k(\cdot, s) = \mathbb{P}(\mathcal{A} \in \cdot)$.

A reasonable choice to study the performance of learning algorithms is to consider a loss function $\ell = W \times Z \to \mathbb{R}_+$. Two key object of interest are:

- training error (or empirical loss) $L_{S_n}\left(W_n\right) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(w, z_i\right)$
- test error (or population risk) $L_\mu(w) := \mathbb{E}[\ell(w, Z)] = \int_Z \ell(w, z)\mu(dz)$

Definition 1.1. *for a learning algorithm characterized by $P_{W|S} = k(\cdot, s)$, the generalization error of the algorithm on μ is defined as:*

$$\operatorname{gen}(W_n, S_n) := L_{S_n}\left(W_n\right) - \mathbb{E}[\ell(W_n, Z)|W_n]$$

Bounding the generalization error is one of the fundamental problems of statistical learning theory because it measures how much the learnt hypothesis suffers from the problem of overfitting.

As a matter of fact, the goal of learning is to ensure that the test error of the output hypothesis of the algorithm is small. From its definition, this is computed under the data generating distribution μ that is often not available and for this reason one considers the training error. Therefore, it is of natural interest to study the discrepancy between training and test error, that is exactly the generalization error and trying to obtain bound on it as accurate as possible.

Typically, there are two main approaches: trying to bound the expected generalization error or finding results in probability using concentration of measure tools. In this article we will try to show some techniques and some important results that makes use of these two approaches.

Defining:

- the centered loss as $\bar{\ell}(w, z) = \ell(w, z) - \mathbb{E}[\ell(w, Z)]$
- the i-th sample loss as $\bar{\ell}_i(w, z) = \bar{\ell}(w, z_i)$
- the j-th partial average loss as $\bar{L}_j = \frac{1}{n}\sum_{i=1}^{j} \bar{\ell}_i\left(w, z\right)$

we can rewrite the expected generalization error as:

$$\mathbb{E}[\operatorname{gen}(W_n, S_n)] = \mathbb{E}[\bar{L}_n(W_n, S_n)] \tag{1.1}$$

where the expectation is taken with respect to the joint probability distribution of (W_n, S_n).

2 Information theoretic methods

The goal of this chapter is to give an overview of the types of bounds on the expected generalization error that one can get using information theoretic methods, following closely the pioneering work of Russo and Zou [1], Xu and Raginsky [2]. The generalization error of a learning algorithm is related and can be determined by its stability properties.

What does it mean for a learning algorithm to be stable?

Definition 2.1. *A learning algorithm is said to be stable if a small change of the input of the algorithm does not change the output of the output much.*

Intuitively, the notion of generalization capability of a stable algorithm captures the idea that its output cannot depend "too much" on any particular training example. The less dependence there is between the output hypothesis W_n and the input dataset S_n, the better the learning algorithm generalizes.

From an information theoretic point of view, the dependence between the input and the output can be naturally measured by the mutual information between them.

Definition 2.2. *The KL divergence between two probability measures μ and ν defined on a common measurable space $(\Omega, \mathcal{F})$ is given by:*

$$\mathcal{D}_{KL}(\mu\|\nu) := \mathbb{E}_\mu\left[\log\left(\frac{\mathrm{d}\,\mu}{\mathrm{d}\,\nu}\right)\right] = \int_\Omega \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu$$

Definition 2.3. *The mutual information between two random variables (X, Y) following laws, respectively, P_X, P_Y and with joint probability law $P_{X,Y}$ is defined as:*

$$I(X;Y) = \mathcal{D}_{KL}(P_{X,Y}||P_X \otimes P_Y)$$

With this definition we are able to proper define the notion of stability:

Definition 2.4. *A learning algorithm is said to be $(\varepsilon, \mu) -$ stable (under the data generating distribution μ) if $I(S_n; W_n) \leq \varepsilon$.*

2.1 Upper-bounding the expected generalization error via $I(S_n; W_n)$

Xu and Raginsky [2] worked in the framework described above, under sub-gaussianity assumption for the loss function. We recall, here the definition of sub-gaussian random variable.

Definition 2.5. *A random variable X with mean μ is called σ sub-gaussian if there is a a positive number σ such that:*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}} \ \forall \lambda \in \mathbb{R}$$

They proved the following result:

Theorem 2.6. *Suppose the loss function $\ell(w,z)$ is σ-subgaussian under μ, $\forall w \in W$, then:*

$$|\mathbb{E}[\text{gen}(W_n, S_n)]| \leq \sqrt{\tfrac{2\sigma^2}{n} I(S_n, W_n)}$$

To prove this main result we need to state and prove the following "decoupling estimate".

Lemma 2.7. *Let X, Y be random variables with joint distribution $P_{X,Y}$. Let $\bar{X}, \bar{Y}$ be independent copies of, respectively, X and Y, such that $P_{\bar{X},\bar{Y}} = P_X \otimes P_Y$. Let f: $X \times Y \to \mathbb{R}$. If f($\overline{X}, \bar{Y}$) is σ-subgaussian under $P_{\bar{X},\bar{Y}}$ then:*

$$|\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y}]| \leq \sqrt{2\sigma^2 I(X,Y)}$$

Proof. (of Lemma 2.7) To prove the Lemma, we use the following representation of Kullback-Leibler divergence, which is attributed to Donsker and Varadhan.

Proposition 2.8. *(Donsker-Varadhan Variational Formula) For any two probabilities measures μ and ν defined on a common measurable space $(\Omega, \mathcal{F})$, we can represent the KL divergence as:*

$$\mathcal{D}_{KL}(\mu \| \nu) = \sup_{\psi \in \mathcal{C}} \left(\int_\Omega \psi \, d\mu - \log \int_\Omega e^\psi \, d\nu \right)$$

where $\mathcal{C}$ is the space of all measurable functions $\psi : \Omega \to \mathbb{R}$ such that $e^\psi \in L^1(\nu)$.

From this representation of the KL divergence, it follows that $\forall \lambda \in \mathbb{R}$, evaluating the RHS in our sub-gaussian function f:

$$\mathcal{D}_{KL}(P_{X,Y} \| P_X \otimes P_Y) \geq \mathbb{E}[\lambda f(X,Y)] - \log \mathbb{E}[e^{\lambda f(\bar{X},\bar{Y})}]$$

Then we can apply the definition of sub-gaussian random variable, following from the sub-gaussian assumption on $f(\bar{X},\bar{Y})$. From:

$$\log \mathbb{E}[e^{\lambda(f(\bar{X},\bar{Y}) - \mathbb{E}f(\bar{X},\bar{Y}))}] \leq \tfrac{\lambda^2\sigma^2}{2} \forall \lambda \in \mathbb{R}$$

it follows that:

$$-\log \mathbb{E}[e^{\lambda f(\bar{X},\bar{Y})}] \leq -\lambda \mathbb{E}[f(\bar{X},\bar{Y})] - \tfrac{\lambda^2\sigma^2}{2}$$

Then:

$$\mathcal{D}_{KL}(P_{X,Y} \| P_X \otimes P_Y) \geq \lambda \left(\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y})] \right) - \tfrac{\lambda^2\sigma^2}{2}$$

If we move all the terms to the LHS we get an inequality for a parabola in λ:

$$\tfrac{\lambda^2\sigma^2}{2} - \lambda \left(\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y})] \right) + \mathcal{D}_{KL}(P_{X,Y} \| P_X \otimes P_Y) \geq 0$$

The positivity condition implies that the discriminant must be non-positive:

$$\left(\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y})] \right)^2 \leq 2\sigma^2 \mathcal{D}_{KL}(P_{X,Y} \| P_X \otimes P_Y)$$

which implies the desired result, recalling Definition (2.3):

$$|\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y}]| \leq \sqrt{2\sigma^2 I(X,Y)}$$

QED

Now, we are ready to prove Theorem (2.6):

Proof. (of Theorem 2.6) We start by noting that, if $\ell(w,z)$ is σ-subgaussian $\forall w \in W$ then $f(S,w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$ is $\frac{\sigma}{n}$-subgaussian. In fact, from the definition, $\forall \lambda > 0$:

$$\mathbb{E}\left[e^{\frac{\lambda}{n}\sum_{i=1}^{n}\ell(w,Z_i)}\right] = \mathbb{E}\left[e^{\frac{\lambda}{n}\ell(w,Z_1)} \cdot e^{\frac{\lambda}{n}\ell(w,Z_2)} \cdot \ldots\right]$$

$$= \mathbb{E}\left[e^{\frac{\lambda}{n}\ell(w,Z_1)}\right] \cdot \mathbb{E}\left[e^{\frac{\lambda}{n}\ell(w,Z_2)}\right] \cdot \ldots \leq \left(e^{\frac{\lambda^2\sigma^2}{n^2 2}}\right)^n = e^{\frac{\lambda^2\sigma^2}{2n}}.$$

Combining this result with Lemma (2.7) we recover the result stated in the theorem.

QED

3 Generalization Error Bounds via Convex Analysis

Neu and Lugosi [3] were able to generalize these results. If one considers the information-theoretic generalization bounds that we saw in the previous section, a natural question: what is so special about KL divergence? Can we go beyond it? The KL divergence is of easy interpretability but there could be different choices of "information-theoretic" measures that allows us, in some cases, to control better the generalization error.

Neu and Lugosi [3] are able to find bounds on the generalization error not only in terms of the standard choice of the *mutual information*. As a matter of fact, they show that it's possible to replace the mutual information by any strongly convex function of the joint input-output information and it's possible to replace the subgaussianity condition on the losses by a bound on an appropriately chosen norm capturing the geometry of the dependence measure.

The key idea of their work is to analyze the performance of learning algorithms using convex analysis tools that are widely used in the analysis of modern online learning algorithms, as one can ascertain from Orabona [4].

3.1 Upper-bounding the expected generalization error via Convex Dependence Measures

To provide these bounds using convex analysis we need to introduce some notation and some definitions. We denote as:

- $\mathcal{P}(H)$: the set of all probability distributions over a given set H;
- $\mathcal{F}(H)$: the dual set of bounded functions from H to the reals;

- P_{W_n,S_n}: the joint distribution of (W_n, S_n) ;
- P_{W_n}: the marginal distribution of W_n;
- $P_0 = P_{W_n} \otimes \mu^n = P_{W_n} \otimes P_{S_n}$: the product of the marginal distributions;
- $P_{|s} \in \mathcal{P}(W)$: the regular version of the conditional distribution of W_n given $S_n = s$;

We observe, since it will be useful in the following, that the conditional distribution is a linear function of P.

Proposition 3.1. *For any $\lambda \in [0,1]$ and $P, P' \in \mathcal{P}(W \times S)$ the mixture distribution of the two probability measures satisfies:*

$$(\lambda P + (1-\lambda)P')_{|s} = \lambda P_{|s} + (1-\lambda)P'_{|s}.$$

Proof. (of Proposition 3.1) Given $P \in \mathcal{P}(W \times S)$, we indicate by P_S the marginal distribution of S and with $\sigma(S)$ the sigma-algebra generated by S, with $\sigma(W)$ the sigma-algebra generated by W.

The proof of this result follows from an observation: all the distribution in $\mathcal{P}(W \times S)$ have the same marginal with respect to the second variable, i.e. :

$$\forall P, P' \in \mathcal{P}(W \times S) \text{ and } \forall s \in \sigma(S) \text{ we have that } P_S(S = s) = P'_S(S = s)$$

In particular, we can write:

$$P_S(S = s) = P(W \times \{S = s\}) = P'_S(S = s) = P'(W \times \{S = s\})$$

We can define the conditional distribution $P_{|s} \in \mathcal{P}(W)$ as:

$$P_{|s}(w) = \frac{P(\{W = w\} \times \{S = s\})}{P(W \times \{S = s\})} \ \forall w \in \sigma(W)$$

Then, it follows:

$$\begin{aligned}
(\lambda P + (1-\lambda)P')_{|s}(w) &= \frac{(\lambda P + (1-\lambda)P')(\{W = w\} \times \{S = s\})}{(\lambda P + (1-\lambda)P')(W \times \{S = s\})} \\
&= \frac{\lambda P(\{W = w\} \times \{S = s\}) + (1-\lambda)P'(\{W = w\} \times \{S = s\})}{\lambda P(W \times \{S = s\}) + (1-\lambda)P'(W \times \{S = s\})} \\
&= \frac{\lambda P(\{W = w\} \times \{S = s\}) + (1-\lambda)P'(\{W = w\} \times \{S = s\})}{P(W \times \{S = s\})} . \\
&= \frac{\lambda P(\{W = w\} \times \{S = s\})}{P(W \times \{S = s\})} + \frac{(1-\lambda)P'(\{W = w\} \times \{S = s\})}{P(W \times \{S = s\})} \\
&= \lambda P_{|s} + (1-\lambda)P'_{|s}
\end{aligned}$$

. QED

Since we want to provide bounds on the generalization error in terms of a *dependence measure* capturing the dependence between W_n and S_n, which is described by their joint distribution P_{W_n,S_n}, we need to properly state what we mean by it.

Definition 3.2. *We call dependence measure a mapping H: $\mathcal{P}(W \times S) \to \mathbb{R}_+$.*

Definition 3.3. *We call conditional dependence measure a mapping h: $\mathcal{P}(W) \to \mathbb{R}_+$.*

For this discussion we will assume that:

- h is convex on $\mathcal{P}(W)$
- h$(P_{w_n}) = 0$ as P_{w_n} is defined as the marginal of W_n from the joint distribution P_{W_n,S_n}, so it should be not dependent from S_n.
- the dependence measures considered are of the form $H(P) = \mathbb{E}_S[h(P_{|S})]$.

For the sake of a clear exposition, let's recall the definitions of convexity of a function (see, e.g. Hiriart-Urruty and Lemaréchal [5]).

We denote, for any function $g \in \mathcal{F}(W)$, its expectation under a distribution $Q \in \mathcal{P}(W)$ as the following bi-linear map:

$$\langle Q, g \rangle = \mathbb{E}_{W \sim Q}[g(W)] \tag{3.1}$$

Definition 3.4. *We say that the function h is convex if $\forall Q, Q' \in \mathcal{P}(W)$, $\forall \lambda \in [0, 1]$ we have that: $h(\lambda Q + (1 - \lambda)Q') \leq \lambda h(Q) + (1 - \lambda)h(Q')$. This implies that, $\forall Q, Q' \in \mathcal{P}(W)$ $\exists g \in \mathcal{F}(W)$ such that $h(Q) \geq h(Q') + \langle g, Q - Q' \rangle$.*

The set of all functions g satisfying this property have a special name:

Definition 3.5. *We define the subdifferential of h at Q' as:*

$$\partial h(Q') = g \in \mathcal{F}(W) : h(Q) \geq h(Q') + \langle g, Q - Q' \rangle$$

The elements $g \in \partial h(Q')$ are called subgradients.

Definition 3.6. *We say that a conditional measure h is α-strongly convex with respect to a norm $\| \cdot \|$ on the space of (signed) finite measures over W if it satisfies:*

$$h(Q) \geq h(Q') + \langle g, Q - Q' \rangle + \tfrac{\alpha}{2} \|Q - Q'\|^2$$

Having introduced the definitions of convexity, notice that, from the assumption that the dependence measures considered are of the form $H(P) = \mathbb{E}_S[h(P_{|S})]$, H is convex in its argument $P \in P(W \times S)$ due to $P_{|S}$ being linear in P (from Proposition 3.1)

Since they will be of key importance in the proof of the main result of Neu and Lugosi [3] we introduce some definitions regarding *convex duality*.

Definition 3.7. *Let H a normed space with norm $\|\cdot\|$ and let $\mathcal{F}(H)$ be its dual space. The dual norm of a function $f \in \mathcal{F}(H)$ is defined as:*

$$\|f\|_* = \sup_{F:\|F\|\leq 1} \langle f, F\rangle$$

In our framework, we will mostly consider distributions $Q \in \mathcal{P}(W \times S)$ and the dual norm of functions $f \in \mathcal{F}(W \times S)$ (for instance, the loss function).

Definition 3.8. *The Fenchel conjugate of a function (not necessarily convex) $f : X \to \mathbb{R}$ is the function $f^* : \mathcal{F}(X) \to \mathbb{R}$ defined by*

$$f^*\left(x^*\right) := \sup_{x \in X} \left\{\langle x, x^*\rangle - f(x)\right\}$$

The operation $f \to f^*$ is also called a Fenchel-Legendre transform. The function f^* is convex. Observe that the conjugate operation is order-reversing : for functions $f, g : X \to [-\infty, +\infty]$, the inequality $f \geq g$ implies $f^* \leq g^*$.

In our framework, we call "overfitting potential" the Fenchel conjugate of the dependence measure H on the set $\mathcal{P}(W \times S)$, i.e. :

$$\Phi(f) = H^*(f) = \sup_{P \in \mathcal{P}(W \times S)} \{\langle P, f\rangle - H(P)\} \tag{3.2}$$

The potential Φ has some basic properties that will be useful in the following. Note that, whenever Φ is bounded it has a nonempty subdifferential $\partial\Phi(f)$, consisting of the convex hull of the maximizers of $\langle P, f\rangle - H(P)$, i.e. :

$$\partial\Phi(f) = \operatorname{conv}\left(\underset{P \in \mathcal{P}(W \times S)}{\arg\max} \{\langle P, f\rangle - H(P)\}\right)$$

By definition (3.5) of the subdifferential, it holds that $\forall\ g \in \mathcal{F}(W \times S)$:

$$\Phi(g) \geq \Phi(f) + \langle P, g - f\rangle$$

Then, we can define:

Definition 3.9. *We define the generalized Bregman divergence as*

$$\mathcal{B}_\Phi(g\|f) = \Phi(g) - \Phi(f) + \sup_{P \in \partial\Phi(f)} \langle P, f - g\rangle,$$

Note that this is a convex function of g, being a sum of convex functions and a supremum of affine functions.

Now, we are ready to state the following important result.

Theorem 3.10. *Let h be α-strongly convex with respect to the norm $\|\cdot\|$ and let $\|\cdot\|_*$ be its dual norm. The expected generalization error of a learning algorithm $\mathcal{A}$ can be controlled as:*

$$\left|\mathbb{E}\left[\operatorname{gen}\left(W_n, S_n\right)\right]\right| \leq \sqrt{\frac{4H(P_{W_n,S_n})\mathbb{E}\left[\|\bar{\ell}(\cdot,Z)\|_*^2\right]}{\alpha n}}$$

Prior to starting the proof, we introduce the following notation, analogous to the way we denoted the expectation of a function $g \in \mathcal{F}$ under $Q \in \mathcal{P}(W)$ (3.1), to denote the expectation of functions $f \in \mathcal{F}(W \times S)$ under distributions $P \in \mathcal{P}(W \times S)$:

$$\langle P, f \rangle = \mathbb{E}_{(W,S)\sim P}[f(W,S)] = \mathbb{E}_S[\langle P_{|S}, f(\cdot, S)\rangle] \tag{3.3}$$

This allows us to rewrite the generalization error as:

$$\mathbb{E}\left[\operatorname{gen}\left(W_n, S_n\right)\right] = \mathbb{E}_{(W,S)\sim P_{W_n,S_n}}\left[\bar{L}_n(W,S)\right] = \left\langle P_n, \bar{L}_n \right\rangle. \tag{3.4}$$

Before effectively starting the proof, we have to introduce a construction for what we will call "feasible convex set" $\Delta_n \subset \mathcal{P}(W \times S)$, that will be of key importance in one of the steps of the proof and justifies the particular choice of H defined through conditional distributions. In particular, we will consider the overfitting potential defined at (3.2) to be restricted to the set Δ_n, i.e. :

$$\Phi(f) = H^*(f) = \sup_{P \in \Delta_n} \{\langle P, f \rangle - H(P)\} \tag{3.5}$$

Furthermore, we define:

- the "ghost data set" $S'_n = \{Z'_i\}_{i=1}^n$ that consists of i.i.d. samples from the distribution μ and the independent from the training set $S_n = \{Z_i\}_{i=1}^n$
- $\forall\ i \in \{1, \ldots, n\}$ the "mixed bag" data set:

$$S_n^i = \{Z_1, Z_2, \ldots, Z_{i-1}, Z_i, Z'_i, Z'_{i+1}, \ldots, Z'_n\}$$

- $\forall i \in \{1, \ldots, n\}$ the output of the learning algorithm using as input the mixed bag data set: $W^i = \mathcal{A}(S_n^i)$
- the set of joint distributions $\{P_i\}_{i=1}^n$ where, $\forall i \in \{1, \ldots, n\}$ P_i is the joint distribution of (W^i, S_n)

Observe that, following this notation, we have $S_n^0 = S'_n$ and $S_n^n = S_n$ and $P_n = P_{W_n,S_n}$, $P_0 = P_{W_n} \otimes \mu^n$. Now, we can finally define the feasible set that we need and we introduced in (3.5):

$\forall i \in \{1, \ldots, n\}$ Δ_i is the convex hull of all distributions $\{P_j\}_{j=0}^i$

$$\Delta_i = \{P \in \mathcal{P}(W \times S) : \sum_{j=0}^{i} \alpha_j P_j, \alpha_j \geq \forall j, \sum_{j=0}^{i} \alpha_j = 1\} \tag{3.6}$$

Proof. (of Theorem 3.10) We take this problem from the perspective of convex analysis and we observe that the expected generalization error is a linear function of the joint input-output distribution P_n. We start the proof by evaluating the overfitting potential Φ (see 3.5) at $f = \eta\bar{L}_n$.

$$\Phi(\eta\bar{L}) = \sup_{P\in\Delta_n}\left\{\eta\left\langle P, \bar{L}_n\right\rangle - H(P)\right\} \geq \eta\left\langle P_n, \bar{L}_n\right\rangle - H\left(P_n\right)$$

From this observation it follows that we can bound the generalization error in the following way, for any convex function H:

$$\eta\mathbb{E}[\text{gen}(W_n, S_n)] = \eta\langle P_n, \bar{L}_n\rangle \leq H(P_n) + \Phi(\eta\bar{L}_n) \tag{3.7}$$

We, now, have to proceed and bound the two terms on the RHS. Let's start with the following result.

Theorem 3.11. $\forall\eta \in \mathbb{R}$, *the overfitting potential satisfies*

$$\Phi\left(\eta\bar{L}_n\right) \leq \textstyle\sum_{i=1}^n \mathcal{B}_\Phi\left(\eta L_i \| \eta L_{i-1}\right)$$

Proof. (of Theorem 3.11) First, we write $\Phi(\eta\bar{L}_n)$ as

$$\begin{aligned}
\Phi\left(\eta\bar{L}_n\right) &= \sum_{i=1}^n \left(\Phi\left(\eta\bar{L}_i\right) - \Phi\left(\eta\bar{L}_{i-1}\right)\right) + \Phi(0) \\
&= \sum_{i=1}^n \left(\mathcal{B}_\Phi\left(\eta\bar{L}_i \| \eta\bar{L}_{i-1}\right) - \eta \sup_{P\in\partial\Phi\left(\eta\bar{L}_{i-1}\right)} \left\langle P, \bar{L}_{i-1} - \bar{L}_i\right\rangle\right) \\
&= \sum_{i=1}^n \left(\mathcal{B}_\Phi\left(\eta\bar{L}_i \| \eta\bar{L}_{i-1}\right) - \frac{\eta}{n} \sup_{P\in\partial\Phi\left(\eta\bar{L}_{i-1}\right)} \left\langle P, \frac{1}{n}\sum_{j=1}^{i-1}\bar{\ell}_j - \frac{1}{n}\sum_{j=1}^{i}\bar{\ell}_j\right\rangle\right) \\
&= \sum_{i=1}^n \left(\mathcal{B}_\Phi\left(\eta\bar{L}_i \| \eta\bar{L}_{i-1}\right) - \frac{\eta}{n} \sup_{P\in\partial\Phi\left(\eta\bar{L}_{i-1}\right)} \left\langle P, -\frac{\bar{\ell}_i}{n}\right\rangle\right) \\
&= \sum_{i=1}^n \left(\mathcal{B}_\Phi\left(\eta\bar{L}_i \| \eta\bar{L}_{i-1}\right) + \frac{\eta}{n} \inf_{P\in\partial\Phi\left(\eta\bar{L}_{i-1}\right)} \left\langle P, \bar{\ell}_i\right\rangle\right)
\end{aligned}$$

where the second line uses the definition of the generalized Bregman divergence $\mathcal{B}_\Phi$ and also that $\Phi(0) = \sup_{P\in\mathcal{P}(\Delta_n)}\{0 - H(P)\} = 0$ because H is minimized with value 0 at $P_0 = P_{W_n} \otimes \mu^n \in \Delta_n$, due to the fact that, at the beginning of the discussion, we assumed that the conditional dependence measure was to satisfy $h(P_{W_n}) = 0$ and, by construction: $H(P_0) = \mathbb{E}_S[h(P_{0|S})] = \mathbb{E}_S[h(P_{W_n})] = \mathbb{E}_S[0] = 0$.

So, now to finish the proof of Theorem (3.11), we have to show that the last term in the above sum in non positive. We proceed in two main steps:

(i) We first show that $\forall\, i$, the subdifferential of Φ includes at least one element of Δ_{i-1}. Precisely, we show the following claim: $\forall P \in \Delta_n\ \exists P^+ \in \Delta_{i-1}$ such that

$$\eta\left\langle P^+, \bar{L}_{i-1}\right\rangle - H\left(P^+\right) \geq \eta\left\langle P, \bar{L}_{i-1}\right\rangle - H(P) \tag{3.8}$$

(i.e. P^+ belongs to the subdifferential of P, evaluated in $\bar{L}_{i-1}$). This implies $\exists\ P_i^* \in \Delta_{i-1} \cap \underset{P\in\Delta_n}{\arg\max}\{\langle P, f\rangle - H(P)\}$).

To show this, consider a fixed $P \in \Delta_n$. From the definition of the convex hull of distributions Δ_n we can write it as $P = \sum_{k=0}^n \alpha_k P_k$. Then, if we choose $P^{+} = \sum_{k=0}^{n} \alpha_k P_{\min\{k,i-1\}}$ we are able to prove the claim. First, we show that $\left\langle P^{+}, \bar{L}_{i-1} \right\rangle = \left\langle P, \bar{L}_{i-1} \right\rangle$. $\forall k \geq i$, we have

$$\begin{aligned}
\left\langle P_k, \bar{L}_{i-1} \right\rangle &= \frac{1}{n} \sum_{t=1}^{i-1} \mathbb{E}_{W^{(k)}, S_n} \left[\bar{\ell} \left(W^{(k)}, Z_t \right) \right] \\
&= \frac{1}{n} \sum_{t=1}^{i-1} \mathbb{E}_{Z_{1:i-1}} \left[\mathbb{E}_{W^{(k)}, S_n} \left[\bar{\ell} \left(W^{(k)}, Z_t \right) \mid Z_{1:i-1} \right] \right] \\
&= \frac{1}{n} \sum_{t=1}^{i-1} \mathbb{E}_{Z_{1:i-1}} \left[\mathbb{E}_{W^{(i-1)}, S_n} \left[\bar{\ell} \left(W^{(i-1)}, Z_t \right) \mid Z_{1:i-1} \right] \right] \\
&= \frac{1}{n} \sum_{t=1}^{i-1} \mathbb{E}_{W^{(i-1)}, S_n} \left[\bar{\ell} \left(W^{(i-1)}, Z_t \right) \right] \\
&= \left\langle P_{i-1}, \bar{L}_{i-1} \right\rangle,
\end{aligned}$$

where, after applying the definition of the bi-linear map $\langle \cdot, \cdot \rangle$ introduced at (3.4), we used the law of total expectation in the second line and the fact that the conditional distribution of $W^{(k)} | Z_{1:i-1}$ is the same as that of $W^{(i-1)} | Z_{1:i-1}$ in the third line, following from the definition of $W^{(i)}$ given $k \geq i$. This implies, recalling that $P = \sum_{k=0}^{n} \alpha_k P_k$, that,

$$\begin{gathered}
\left\langle P, \bar{L}_{i-1} \right\rangle = \tfrac{1}{n} \sum_{k=0}^{n} \alpha_k \sum_{j=1}^{i-1} \mathbb{E} \left[\bar{\ell} \left(W^{(k)}, Z_j \right) \right] = \\
\tfrac{1}{n} \sum_{k=0}^{n} \alpha_k \sum_{j=1}^{i-1} \mathbb{E} \left[\bar{\ell} \left(W^{(\min\{k,i-1\})}, Z_j \right) \right] = \left\langle P^{+}, \bar{L}_{i-1} \right\rangle
\end{gathered}$$

In this way, we can simplify the two terms $\eta \left\langle P^{+}, \bar{L}_{i-1} \right\rangle$ and $\eta \left\langle P, \bar{L}_{i-1} \right\rangle$ in (3.8) and it remains to prove that: $-H\left(P^{+}\right) \geq -H(P)$. To prove this, recalling the definition the probability kernel $P_{W|S} = k(\cdot, s) = \mathbb{P}(\mathcal{A} \in \cdot)$ that characterizes the randomized output of $\mathcal{A}(s)$ for any dataset $s \in Z^n$, we define the conditional distribution as:

$$\begin{gathered}
P_{|S} = \sum_{k=0}^{n} \alpha_k \mathbb{E} \left[\kappa \left(\cdot, S^{(k)} \right) \mid S \right] \quad \text{and} \\
P_{|S}^{+} = \sum_{k=0}^{n} \alpha_k \mathbb{E} \left[\kappa \left(\cdot, S^{(\min\{k,i-1\})} \right) \mid S \right]
\end{gathered}$$

Thus, now we can write the dependence measure $H(P)$ as:

$$H(P) = \mathbb{E}_S \left[h \left(P_{|S} \right) \right] = \mathbb{E}_S \left[h \left(\sum_{k=0}^{n} \alpha_k \mathbb{E} \left[\kappa \left(\cdot, S^{(k)} \right) \mid S \right] \right) \right]$$

Now, we apply Jensen's inequality conditionally:

$$\begin{gathered}
\mathbb{E}_S \left[h \left(\sum_{k=0}^{n} \alpha_k \mathbb{E} \left[\kappa \left(\cdot, S^{(k)} \right) \mid S \right] \right) \right] \geq \\
\mathbb{E}_{S_{1:i-1}} \left[h \left(\sum_{k=0}^{n} \alpha_k \mathbb{E}_{i:n} \left[\kappa \left(\cdot, S^{(k)} \right) \mid S_{1:i-1} \right] \right) \right] = \\
\mathbb{E}_S \left[h \left(\sum_{k=0}^{n} \alpha_k \mathbb{E} \left[\kappa \left(\cdot, S^{(\min\{k,i-1\})} \right) \mid S \right] \right) \right] = H\left(P^{+}\right)
\end{gathered}$$

This concludes the proof of claim (3.8), assuring that $\Delta_{i-1} \cap \partial \Phi(\eta \bar{L}_{i-1}) \neq \emptyset$.

(ii) We conclude the proof of Theorem (3.11) by establishing that $\inf_{P\in\partial\Phi(\eta L_{i-1})}\left\langle P,\eta\bar{\ell}_i\right\rangle \leq 0$: we take $P^*\in\Delta_{i-1}\cap\partial\Phi(\eta\bar{L}_{i-1})$. Since $P^*\in\Delta_{i-1}$ we can write is as $P^*=\sum_{k=0}^{i-1}\alpha_k^*P_k$. Since $P^*\in\partial\Phi(\eta\bar{L}_{i-1})$, then:

$$\begin{aligned}\inf_{P\in\partial\Phi(\eta L_{i-1})}\left\langle P,\eta\bar{\ell}_i\right\rangle \leq \left\langle P^*,\eta\bar{\ell}_i\right\rangle &= \sum_{k=0}^{i-1}\alpha_k^*\left\langle P_k,\eta\bar{\ell}_i\right\rangle = \sum_{k=0}^{i-1}\alpha_k^*\mathbb{E}[\eta\bar{\ell}_i]\\ &= \sum_{k=0}^{i-1}\alpha_k^*\mathbb{E}\left[\ell\left(W^{(k)},Z_i\right)-\ell\left(W^{(k)},Z_i'\right)\right]=0\end{aligned}$$

where we applied again the definition of $\langle\cdot,\cdot\rangle$ introduced at (3.4) and, in the second line, we observed that $W^{(k)}$ is independent of Z_i by definition for $k<i$.

QED

Let's go back to the proof of Theorem (3.10). At the moment, with Theorem (3.11) we have been able to rewrite the bound on the generalization error obtained at (3.7) in the following way:

$$\eta\mathbb{E}[\text{gen}(W_n,S_n)]\leq H(P_n)+\textstyle\sum_{i=1}^n\mathcal{B}_\Phi\left(\eta L_i\|\eta L_{i-1}\right)$$

To conclude the proof, it remains to handle to Bregman divergence that appears on the RHS. We are going to do it by proving the following lemma.

Lemma 3.12. *Let h be α-strongly convex with respect to the norm $\|\cdot\|$ whose dual norm is denotes as $\|\cdot\|_*$. Then, $\forall i$ and $\forall$ η, it holds that:*

$$\mathcal{B}_\Phi\left(\eta L_i\|\eta L_{i-1}\right)\leq\frac{\eta^2\mathbb{E}_Z\left[\|\bar{\ell}(\cdot,Z)\|_*^2\right]}{\alpha n^2}$$

Proof. (of Lemma 3.12) For this proof we have to define some objects.

Definition 3.13. *We call "lifted" norm the following:*

$$\|P-P^{'}\|_\mu=(\mathbb{E}[\|P_{|S}-P^{'}_{|S}\|^2])^{\frac{1}{2}}$$

We observe that the following property holds for this norm:

Lemma 3.14. *If h is α-strongly convex with respect to $\|\cdot\|$, then H is also strongly convex on Δ_n with respect to the lifted norm.*

Proof. (of Lemma 3.14) We start by considering: $H(\lambda P+(1-\lambda)P')$, since we want to prove that H is strongly convex. Using the definition of H and then the result of Proposition (3.1) on the affinity of the conditional distribution in the joint distribution, we obtain:

$$H(\lambda P+(1-\lambda)P')=\mathbb{E}_S\left[h\left((\lambda P+(1-\lambda)P')_{|S}\right)\right]=\mathbb{E}_S\left[h\left(\lambda P_{|S}+(1-\lambda)P'_{|S}\right)\right]$$

Then, since h is α-strongly convex, we have:
$\mathbb{E}_S\left[h\left(\lambda P_{|S}+(1-\lambda)P'_{|S}\right)\right] \leq \mathbb{E}_S\left[\lambda h\left(P_{|S}\right)+(1-\lambda)h\left(P'_{|S}\right)-\frac{\alpha}{2}\left\|P_{|S}-P'_{|S}\right\|^2\right]$

Applying linearity of expectation and the definition of H one more time we obtain the claimed result:

$$\mathbb{E}_S\left[\lambda h\left(P_{|S}\right)+(1-\lambda)h\left(P'_{|S}\right)-\tfrac{\alpha}{2}\left\|P_{|S}-P'_{|S}\right\|^2\right]=$$

$$\lambda H(P)+(1-\lambda)H\left(P'\right)-\tfrac{\alpha}{2}\mathbb{E}_S\left[\left\|P_{|S}-P'_{|S}\right\|^2\right]$$

QED

The next step of the proof of Lemma (3.12) is prove that the α-strong convexity of H on Δ_n implies $\frac{1}{\alpha}$-strong smoothness of its Legendre-Fenchel conjugate $\Phi = H^*$ with respect to the *dual seminorm*. We first introduce the following new notion.

Definition 3.15. *The dual seminorm is defined, $\forall f \in \mathcal{F}(W \times S)$ as:*

$$\|f\|_{\mu,*}=\left(\mathbb{E}_S[\|f(\cdot,S)\|_*]\right)^{\frac{1}{2}}$$

This seminorm satisfies all properties of a norm expect positive definiteness because it can be zero even when f is not identically zero but only on a set of measure zero with respect to μ.

The dual seminorm just defined satisfies the following property.

$$\forall P, P' \in \Delta_n \text{ and } \forall f \in \mathcal{F}(W \times S): \tag{3.9}$$

$$\langle P-P', f\rangle = \mathbb{E}_S\left[\left\langle P_{|S}-P'_{|S}, f(\cdot,S)\right\rangle\right] = \mathbb{E}_S[\|P_{|S}-P'_{|S}\|\langle\frac{P_{|S}-P'_{|S}}{\|P_{|S}-P'_{|S}\|}, f(\cdot,S)\rangle]$$

$$\leq \mathbb{E}_S\left[\left\|P_{|S}-P'_{|S}\right\|\|f(\cdot,S)\|_*\right] \leq \left(\mathbb{E}_S\left[\left\|P_{|S}-P'_{|S}\right\|^2\right]\right)^{1/2}\cdot\left(\mathbb{E}_S\left[\|f(\cdot,S)\|_*^2\right]\right)^{1/2}$$

$$=\|P-P'\|_\mu\|f\|_{\mu,*}$$

where we used the notation for $\langle\cdot,\cdot\rangle$ introduced in (3.4), the definition of the norm $\|\cdot\|$ and the definition of the dual norm $\|\cdot\|_*$ in the first and second line, and the Cauchy-Schwarz inequality at the end of the second line and, finally, the definition of "lifted" norm and dual seminorm.

Now we are ready to prove the already mentioned above result about the duality between strong-convexity and smoothness.

Lemma 3.16. *Let f and f' be two integrable functions under all distributions in Δ_n and let $P \in \partial\Phi(f)$ and $P' \in \partial\Phi\left(f'\right)$. Suppose that h is α-strongly convex with respect to $\|\cdot\|$; it follows by Lemma (3.14) that H is α-strongly convex on Δ_n with respect to $\|\cdot\|_\mu$. Then, for the Legendre-Fenchel conjugate $\Phi = H^*$ it holds:*

$$\mathcal{B}_\Phi(f\|f') \leq \frac{1}{\alpha}\left\|f-f'\right\|_{\mu,*}^2.$$

Proof. (of Lemma 3.16) Let $g \in \partial H(P)$ and $g' \in \partial H\left(P'\right)$. Then, by first-order optimality of P and P', we have

$$\left\langle s_P - f, P - P' \right\rangle \leq 0 \quad ; \left\langle s_{P'} - f', P' - P \right\rangle \leq 0.$$

If we sum the two inequalities, we get

$$\begin{aligned} \langle g - f, P \rangle - \langle g - f, P \rangle + \left\langle g' - f', P' \right\rangle - \left\langle g' - f', P \right\rangle \leq 0 \\ \left\langle g - g', P \right\rangle + \left\langle f' - f, P \right\rangle + \left\langle g' - g, P' \right\rangle + f - \left\langle f', P' \right\rangle \leq 0 \\ \left\langle g - g', P - P' \right\rangle + \left\langle f' - f, P - P' \right\rangle \leq 0 \\ \Rightarrow \left\langle g - g', P - P' \right\rangle \leq \left\langle P' - P, f' - f \right\rangle . \end{aligned}$$

Now, using the definition of strong convexity of H, we get

$$\begin{array}{l} H(P) \geq H\left(P'\right) + \left\langle g', P - P' \right\rangle + \frac{\alpha}{2} \left\| P - P' \right\|_\mu^2 \\ H\left(P'\right) \geq H(P) + \left\langle g, P' - P \right\rangle + \frac{\alpha}{2} \left\| P - P' \right\|_\mu^2 \end{array}$$

Summing these two inequalities then gives

$$\alpha \left\| P - P' \right\|_\mu^2 \leq \left\langle g - g', P - P' \right\rangle .$$

Combining both the inequalities above and the property (3.9) of the dual seminorm, we obtain

$$\alpha \left\| P' - P \right\|_\mu^2 \leq \left\langle P - P', f - f' \right\rangle \leq \left\| P - P' \right\|_\mu \left\| f - f' \right\|_{\mu,*} ,$$

It follows the property:

$$\left\| P - P' \right\|_\mu \leq \frac{1}{\alpha} \left\| f - f' \right\|_{\mu,*} . \tag{3.10}$$

By the mean value theorem, there exists an $f_\lambda = \lambda f + (1-\lambda) f'$ with $\lambda \in [0,1]$ such that $P_\lambda \in \partial \Phi\left(f_\lambda\right)$ and

$$\begin{aligned} \Phi(f) &= \Phi\left(f'\right) + \left\langle P_\lambda, f - f' \right\rangle \\ &= \Phi\left(f'\right) + \left\langle P', f - f' \right\rangle + \left\langle P_\lambda - P', f - f' \right\rangle \\ &\leq \Phi\left(f'\right) + \left\langle P', f - f' \right\rangle + \left\| P_\lambda - P' \right\|_\mu \left\| f - f' \right\|_{\mu,*} \\ &\leq \Phi\left(f'\right) + \left\langle P', f - f' \right\rangle + \frac{1}{\alpha} \left\| f_\lambda - f' \right\|_{\mu,*} \left\| f - f' \right\|_{\mu,*} \\ &= \Phi\left(f'\right) + \left\langle P', f - f' \right\rangle + \frac{\lambda}{\alpha} \left\| f - f' \right\|_{\mu,*}^2 \\ &\leq \Phi\left(f'\right) + \left\langle P', f - f' \right\rangle + \frac{1}{\alpha} \left\| f - f' \right\|_{\mu,*}^2 \end{aligned}$$

where in the third line we applied again property (3.9) for the dual seminorm and in next line we applied property (3.10).

We recall that the Bregman divergence was defined as:

$$\mathcal{B}_\Phi(f\|f') = \Phi(f) - \Phi(f') + \sup_{P\in\partial\Phi(f')} \langle P, f' - f\rangle$$

Since $P' \in \partial\Phi\left(f'\right)$, bounding $\langle P', f - f'\rangle$ by $\sup_{P\in\partial\Phi(f')}\langle P, f' - f\rangle$ and then moving the first two terms of the RHS to the left we get the desired result:

$$\mathcal{B}_\Phi(f\|f') \leq \frac{1}{\alpha}\left\|f - f'\right\|^2_{\mu,*}.$$

QED

Now, we are able to conclude the proof of Lemma (3.12), by applying the latter result:

$$\mathcal{B}_\Phi\left(\eta L_i\|\eta L_{i-1}\right) \leq \frac{\left\|\eta\left(\bar{L}_i - \eta\bar{L}_{i-1}\right)\right\|^2_{\mu,*}}{\alpha} = \frac{\eta^2\mathbb{E}_S\left[\left\|\bar{\ell}_i(\cdot,S)\right\|^2_*\right]}{\alpha n^2} = \frac{\eta^2\mathbb{E}_S\left[\left\|\bar{\ell}\left(\cdot,Z_i\right)\right\|^2_*\right]}{\alpha n^2}$$

where for the equalities we applied the definitions of dual seminorm and of i-th sample loss. QED

After this long discussion, we are finally able to conclude the proof of the main Theorem (3.10). Starting from:

$$\eta\mathbb{E}[\text{gen}(W_n,S_n)] = \eta\langle P_n, \bar{L}_n\rangle \leq H(P_n) + \Phi(\eta\bar{L}_n) \leq H(P_n) + \sum_{i=1}^{n}\mathcal{B}_\Phi\left(\eta L_i\|\eta L_{i-1}\right)$$

We apply Lemma (3.12) and we obtain:

$$\eta\left\langle P_n, \bar{L}_n\right\rangle \leq H\left(P_n\right) + \frac{\eta^2\mathbb{E}_Z\left[\|\bar{\ell}(\cdot,Z)\|^2_*\right]}{\alpha n}$$

To obtain an upper bound we optimize for $\eta > 0$:

$$\left\langle P_n, \bar{L}_n\right\rangle \leq \frac{H\left(P_n\right)}{\eta} + \frac{\eta\mathbb{E}_Z\left[\|\bar{\ell}(\cdot,Z)\|^2_*\right]}{\alpha n} \leq \sqrt{\frac{4H\left(P_n\right)\mathbb{E}_Z\left[\|\bar{\ell}(\cdot,Z)\|^2_*\right]}{\alpha n}}$$

A lower bound can be obtained by an analogous derivation for $\eta < 0$.

This concludes the proof of Theorem (3.10). QED

3.2 Applications of the result

There are numerous dependence measures satisfying the hypothesis we imposed at the beginning of Section 3, i.e.:

- h is convex on $\mathcal{P}(W)$
- h$(P_{w_n}) = 0$

- $H(P) = E_S[h(P_{|S})]$.

In particular, since we discussed the results of Russo and Zou [1], Xu and Raginsky [2] in Section 2, it is interesting to see how the mutual information behaves in this *convex analysis* framework. In the notation of this section, its definition is $h(Q) = \mathcal{D}_{KL}(Q\|Q_0)$, where $Q_0 = Q_{W_n} \otimes \mu^n$ in this case denote the marginal distribution of the hyphotesis. It is well known that this function is 1-strongly convex with respect to the total variation distance, i.e. $\|Q - Q'\|_{TV} = \sup_{f:\|f\|_\infty \leq 1}\langle f, Q - Q'\rangle$, whose dual norm is the supremum norm $\|f\|_\infty = \sup_{w\in W}|f(w)|$. One can compute the associated dependence measure H and obtain, $\forall P \in \Delta_n$: $H(P) = \mathbb{E}_S[\mathcal{D}_{KL}(P_{|S}\|Q_0)] = \mathcal{D}_{KL}(P\|P_0)$, where in this case $P_0 = P_{W_n} \otimes \mu^n$ denote the product of the marginal distributions, with the notation introduced above (3.6). Applying Theorem (3.10) we get the following generalization bound:

Corollary 3.17. *The generalization error of any learning algorithm satisfies*

$$|\mathbb{E}\left[\operatorname{gen}\left(W_n, S_n\right)\right]| \leq \sqrt{\frac{4\mathcal{D}_{\mathrm{KL}}\left(P_{W_n,S_n}\|P_{W_n}\otimes\mu^n\right)\mathbb{E}_Z\left[\|\bar{\ell}(\cdot,Z)\|_\infty^2\right]}{n}}$$

We observe that this bound depends on the second moment of the centered losses $\|\ell(\cdot, Z) - \mathbb{E}\left[\ell\left(\cdot, Z'\right)\right]\|_\infty$ in terms of the random data point Z. The cool thing about this result is that this quantity can be finite even for heavy-tailed loss distributions whose higher moments may not exist, relaxing the sugbaussianity assumption on the loss function imposed by Xu and Raginsky [2], which explicitly doesn't allow heavy-tailed losses.

Instead, we can directly recover the guarantees of Xu and Raginsky by increasing the domain from Δ_n to $\mathcal{P}(W \times S)$, in the formula of the overfitting potential, i.e. going back from the "restricted" version we gave at (3.5) to the one that followed Definition (3.8) at (3.2), and by recalling the Donsker-Varadhan variational representation of the Kullback-Leibler divergence introduced with Proposition (2.8). In fact, in this case we have that the measurable functions ψ are taken over the space $\mathcal{C} = \mathcal{F}(W \times S)$, leading to:

$$H(P) = \mathcal{D}_{KL}(P\|P_0) = \sup_{\psi\in\mathcal{F}(W\times S)}\left(\int_{W\times S}\psi\, dP - \log\int_{W\times S} e^{\psi}\, dP_0\right)$$

that, re-written in the same notation of this Section:

$$H(P) = \mathcal{D}_{KL}(P\|P_0) = \sup_{\psi\in\mathcal{F}(W\times S)}\left(\langle P,\psi\rangle - \log\langle e^{\psi}, P_0\rangle\right)$$

Now, plugging this in the definition of overfitting potential:

$$\begin{aligned}\Phi\left(\eta\bar{L}_n\right) &\leq \sup\nolimits_{P\in\mathcal{P}(\mathcal{W}\times\mathcal{S})}\left\{\eta\left\langle P,\bar{L}_n\right\rangle - H(P)\right\} \\ &= \sup_{P\in\mathcal{P}(\mathcal{W}\times\mathcal{S})}\left\{\eta\left\langle P,\bar{L}_n\right\rangle - \sup_{\psi\in\mathcal{F}(W\times S)}\left(\langle P,\psi\rangle - \log\langle e^{\psi},P_0\rangle\right)\right\}\end{aligned}$$

The supremum outside is achieved if inside we take as $\psi = \eta \bar{L}_n$. Then, we get, whenever the losses are σ-subgaussian:

$$\Phi\left(\eta \bar{L}_n\right) \leq \log \mathbb{E}\left[e^{\eta \bar{L}_n}\right] \leq \frac{\eta^2 \sigma^2}{2n}$$

This makes possible to write the proof of Xu and Raginsky [2] in the notation of Section 3 , in fact, recalling that for the proof of Theorem (3.10) we started from:

$$\eta \mathbb{E}[\text{gen}(W_n, S_n)] = \eta \langle P_n, \bar{L}_n \rangle \leq H(P_n) + \Phi(\eta \bar{L}_n)$$

Now, using the fact the we are considering the Kullback Leibler divergence as a dependence measure and the bound just derived for the overfitting potential whenever the losses are subgaussian:

$$\eta \mathbb{E}[\text{gen}(W_n, S_n)] = \eta \langle P_n, \bar{L}_n \rangle \leq \mathcal{D}_{KL}(P_n \| P_0) + \frac{\eta^2 \sigma^2}{2n}$$

Optimizing with respect to η one gets the same bound as in Theorem (2.6).

4 Single-Letter guarantees

Bu et al. [6] provided information-theoretic upper bounds on the generalization error of supervised learning algorithms constructed in terms of the mutual information between each individual training sample and the output of the learning algorithm. This work propose amendments to the previous results of Section 2 in a different way from the one just analysed in Section 3. One of the main problems that can arise is that the mutual information may be extremely large when the algorithm leaks too much information of the data into the output. For instance, if we consider empirical risk minimization (ERM), then if w^* is the unique minimizer of $L_S(w)$ in W; it means that w^* is a deterministic function of S and it follows that the mutual information is $I(w^*; S) = \infty$. Therefore, in cases such this, it is possible to tighten the information-theoretic generalization bound based on the individual sample mutual information $I(W; z_i)$. This idea is motivated by the definition of $point-wise\ stable$ algorithms.

Definition 4.1. *An algorithm is said to be point-wise stable if the expectation of the loss function $\ell(W, z_i)$ does not change much with the substitution of any individual training sample z_i.*

Then, the idea is that if an algorithm is point-wise stable, it generalizes well. This bound is derived under more general conditions for the loss function than sub-gaussianity and is applicable to a broader range of problems but, under the same conditions of Theorem (2.6) we have:

Proposition 4.2. *Suppose that $\ell(w, z)$ is σ-subgaussian under μ $\forall w \in W$, then:*

$$|\mathbb{E}\left[\text{gen}(W_n, S_n)\right]| \leq \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \sqrt{2\sigma^2 I(z_i, W_n)}$$

There are several examples where the previous bound provides a more accurate characterization of the generalization error than the bound in Theorem (2.6) but we will not go into many details here. It is also worth observing that the individual sample information $I(W; z_i)$ is between two vectors whose dimension do not scale with the sample size n and, therefore, in some cases the bound of Proposition (4.2) is much easier to be evaluated empirically in practice.

4.1 Single-Letter Guarantees via Convex Analysis

We can find analogous results to Proposition (4.2) also in the framework presented in Section 3. In particular, by making some slightly different choices of the dependence measure $H(P)$ and some adjustments in the proof we can get the following result single-letter version of Theorem (3.10).

Theorem 4.3. *Let h be α-strongly convex respect to the norm $\|\cdot\|$ and let $\|\cdot\|_*$ be its dual norm. The expected generalization error of a learning algorithm $\mathcal{A}$ can be controlled as:*

$$|\mathbb{E}\left[\text{gen}\left(W_n, S_n\right)\right]| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{h\left(P_{W_n,S_n|Z_i}\right)\mathbb{E}\left[\|\bar{\ell}(\cdot,Z)\|_*^2\right]}{\alpha}}$$

Proof. (of Theorem 4.3) This result can be proved by choosing:

- $H(P) = \frac{1}{n}\sum_{i=1}^{n} h(P_{|Z_i})$, instead of the previous choice $H(P) = \mathbb{E}_S[h(P_{|S})]$ (recall Definition 3.3)

Then, one can verify that all the steps of Theorem (3.11) continue to be valid. The bound of Lemma (3.12), instead, continues to hold if one properly adjust the definition of the lifted dual norm $\|\cdot\|_{\mu,*}$. In this case the choice for the "lifted" norm and the dual lifted norm are the followings, respectively:

- $\|P - P'\|_\mu = \frac{1}{n}\sum_{i=1}^{n}\|P_{|Z_i} - P'_{|Z_i}\|^2$
- $\|f\|_{\mu,*} = (\frac{1}{n}\sum_{i=1}^{n}\|f(\cdot, Z_i)\|_*)^{\frac{1}{2}}\ \forall f \in \mathcal{P}(W \times S)$

. With these few adjustments, the whole proof of Theorem (3.10) can be reproduced and obtain the result shown in Theorem (4.3). QED

5 High-probability bounds

In the previous sections we provided upper bounds on the expected generalization error. Since we are often interested in analyzing the tail behavior of the absolute generalization error $|L_{S_n}(W_n) - E[l(W_n, Z)|W_n]|$, in this section we want to establish some high probability bounds for the generalization error using concentration inequalities. There are the so called PAC (i.e. probably approximately correct) bounds.

In this section, we will mainly navigate in a framework more similar to the one of Section 2. Let's start, first, with a trivial result, that will be useful to introduce some typical tools and will serve as a first example.

Proposition 5.1. *Suppose that S and W are independent. For any fixed $w \in W$, if $\ell(w, Z)$ is σ^2-sub-Gaussian, the Chernoff-Hoeffding bound gives* $\mathbf{P}\left[|L_\mu(w) - L_Z(w)| > \alpha\right] \leq 2e^{-\alpha^2 n/2\sigma^2}$. *Thus a sample size of*

$$n \geq \frac{2\sigma^2}{\alpha^2} \log\left(\frac{2}{\beta}\right)$$

suffices to guarantee

$$\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] \leq \beta$$

Proof. (of Proposition 5.1) This simply follows applying Chernoff Bound to the subgaussian random variable.

Remark 5.2. *The generic Chernoff Bound for a random variable X is obtained applying Markov's Inequality to $\varphi(X) = e^{\lambda X}$:*

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{t\lambda}}$$

We can also write it, taking $\inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{t\lambda}} = \psi_X(\lambda) = \log \mathbb{E} e^{\lambda X}$*, as:*

$$\inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{t\lambda}} = \mathbb{E}\left[\exp\left(-\sup_{\lambda \geq 0}\left(\lambda t - \psi_X(\lambda)\right)\right)\right]$$

We can obtain the Cramer-Chernoff inequality from the previous Chernoff bound by defining the Cramer Transform of the random variable X.

Definition 5.3. *We define the Cramer Transform of the random variable X as:*

$$\psi_X^*(t) = \sup_{\lambda \geq 0}(\lambda t - \psi_X(\lambda))$$

Then, we can rewrite the Chernoff bound as:

$$\mathbb{P}(X \geq t) \leq \exp(\psi_X^*(t))$$

obtaining the Cramer-Chernoff inequality.

In this just introduced framework, we identify a centered random variable X to be subgaussian with parameter σ^2 if $\psi_X(\lambda) = \log \mathbb{E} e^{\lambda X} \leq \frac{\lambda^2 \sigma^2}{2}$. We observe that when this holds, then it is true that $\psi_X^*(t) \geq \frac{t^2}{2\sigma^2}$. In fact $\psi_X^*(t) = \sup_{\lambda \geq 0}(\lambda t - \psi_X(\lambda)) \geq \sup_{\lambda \geq 0}(\lambda t - \frac{\lambda^2\sigma^2}{2}) = \frac{t^2}{2\sigma^2}$, where we have applied the definition of subgaussianity and since the supremum is achieved for $\lambda = \frac{t}{\sigma^2}$.

Thus, it's easy to verify applying the Cramer-Chernoff inequality that the probability bound of Proposition (5.1) holds. In fact, if $\bar{\ell}$ is σ^2 subgaussian, then, from a simple computation it follows that $\bar{L}_n$ is $\frac{\sigma^2}{n}$ subgaussian (see the proof of Theorem 2.6). Since $\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] = \mathbb{P}\left[\left|\bar{L}_n(W_n, S_n)\right| > \alpha\right]$ by the definitions we introduced in Section 1, then:

$$\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] \leq 2e^{-\alpha^2 n/2\sigma^2}.$$

QED

Now, we will show that, even when W is dependent on S, it still suffices a sample complexity polynomial in $\frac{1}{\alpha}$ and logarithmic in $\frac{1}{\beta}$ to guarantee that $\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] \leq \beta$, under the hyphotesis that $I(S; W)$ is sufficiently small. It should be noted that, indeed, we will use a stronger assumption than $I(S; W)$ being small, since we will ask $I(\Lambda_W(S); W)$ to be small, where:

$$\Lambda_W(S) := (L_S(w))_{w \in W} \tag{5.1}$$

is the collection of training errors of the hypothesis in W. This is stronger because, if we see $\Lambda_W(S)$ - S - W as a Markov chain, which is correct since $\Lambda_W(S)$ and W are independent conditionally on S and, $\forall w \in W$, $L_S(w)$ is a function of S, then it holds that:

$$I(\Lambda_W(S); W) \leq I(S; W)$$

Now we can state the result analogous to Proposition (5.1) without the independence assumption, but requiring that $I(\Lambda_W(S); W)$ to be small.

Theorem 5.4. *Let $\ell(w, Z)$ be σ-subgaussian under μ $\forall w \in W$. If a learning algorithm $\mathcal{A}$ satisfies $I(\Lambda_W(S); W) \leq \varepsilon$, then for any $\alpha > 0$ and $0 < \beta \leq 1$, we can guarantee $\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] \leq \beta$ by a sample complexity of*

$$n \geq \frac{8\sigma^2}{\alpha^2}\left(\frac{\varepsilon}{\beta} + \log\left(\frac{2}{\beta}\right)\right)$$

Proof. (of Theorem 5.4) To prove the theorem, we first need two results. The proof will then be based on these two results and on Lemma (2.7). We introduce the following notation:

- $S_{n,i}$ for $i = 1, \ldots, m$ to denote m independent datasets of n observations drawn from the same probability distribution μ
- $W_{n,i} = \mathcal{A}(S_{n,i})$ to denote the output of one execution of an independent copy the algorithm $\mathcal{A}$,that takes as input one of the independent datasets $S_{n,i}$

We start by stating the first one of the two results.

Lemma 5.5. *Consider the parallel execution of m independent copies of $\mathcal{A}(S_n)$ on independent datasets $S_{n,1}, \ldots, S_{n,m}$. Let $S^m := (S_{n,1}, \ldots, S_{n,m})$ be the overall dataset. If under μ, $\mathcal{A}(S_n)$ satisfies $I(\Lambda_W(S); W) \leq \varepsilon$, then the overall algorithm $W^m = \mathcal{A}(S^m)$ satisfies $I\left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right)\right) \leq m \cdot \varepsilon$.*

Proof. (of Lemma 5.5) The proof of this lemma simply follows from the fact, for the way it was constructed, that the couples $(S_{n,i}, W_{n,i})\,, i = 1, \ldots, m$ are independent, and from the chain rule of mutual information, which we recall below.

Proposition 5.6. *(Chain rule of Mutual Information) Let X, Y and Z be random variables. Then:*

$$I(X, Y; Z) = I(X; Z) + I(Y; Z \mid X)$$

By induction, this chain rule also implies that for a set of random variables $X_1, X_2, \cdots X_n, Z$:

$$I\left(X_1, X_2, \cdots X_n; Z\right) = \sum_i I\left(X_i; Z \mid X_1, \cdots X_{i-1}\right)$$

In our cause we consider the following set of random variables:

$$\Lambda_W(\{S_{n,i}\}_{i=1}^m, W^m)$$

Plugging this random variables in the chain rule and using independence gives us the desired result stated in Lemma (5.5). QED

Now, we proceed by stating the second of the two results that we need. With the following Lemma, we introduce the so-called "monitor technique". We imagine a new algorithm, i.e. the monitor, which takes as input the overall dataset S^m. The monitor outputs the name of the query with the worst generalization error. The challenge that we will need to face in order to prove Theorem (5.4) lies, then, in showing that the monitor can find which copy overfit and finding what this implies for the sample size, and then deriving a contradiction.

Lemma 5.7. *Let $S^m := (S_{n,1}, \ldots, S_{n,m})$, where $S_{n,i} \sim \mu^{\otimes n}$. If an algorithm $\mathcal{A} : S^{m \times n} \to W \times [m] \times \{\pm 1\}$ satisfies $I\left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right); W, T, R\right) \leq \varepsilon$, and if $\ell(w, Z)$ is σ-subgaussian for all $w \in W$, then*

$$\mathbb{E}[R \cdot \bar{L}_{n,T}] \leq \sqrt{\frac{2\sigma^2\varepsilon}{n}}.$$

Proof. (of Lemma 5.7) This result is an application of Lemma (2.7) (the decoupling estimate we saw in Section 2). We can see it by choosing: $X = \left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right)\right), Y = (W, T, R)$, and

$$f\left(\left(\Lambda_W\left(s_{n,1}\right), \ldots, \Lambda_W\left(s_{n,m}\right)\right), (w, t, r)\right) = r L_{s_{n,t}}(w).$$

If $\ell(w, Z)$ is σ-subgaussian under $S \sim \mu$ for all $w \in W$, then $\frac{r}{n} \sum_{i=1}^n \ell\left(w, z_{t,i}\right)$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian for all $w \in \mathrm{W}, t \in [m]$ and $r \in \{\pm 1\}$, where $z_{t,i}$ denote the i-th sample of the t-th copy of the dataset S_n, and hence $f(\bar{X}, \bar{Y})$ is $\frac{\sigma}{\sqrt{n}}$-subgaussian. Then, from Lemma (2.7) it follows that

$$\mathbb{E}\left[R \cdot \bar{L}_{S_{n,T}}(W)\right] \leq \sqrt{\frac{2\sigma^2 I\left(\Lambda_{\mathrm{W}}\left(S_1\right), \ldots, \Lambda_{\mathrm{W}}\left(S_m\right); W, T, R\right)}{n}}$$

and this proves Lemma (5.7) . QED

With these two lemmas, we can now effectively prove Theorem (5.4).

Let $W^m = \mathcal{A}(S^m)$ be the output of the parallel exectution of m independent copies of $W = \mathcal{A}(S)$ on independent datasets $S_{i=1}^m$. Now, given S^m and W^m we consider the sample (W^*, T^*, R^*), drawn from t $W \times [m] \times \{\pm 1\}$ imposing:

$$(T^*, R^*) = \arg\max_{t \in [m], r \in \{\pm 1\}} r(\bar{L}_{S_{n,t}}(W_t)) \text{ and } W^* = W_{n,T^*} \tag{5.2}$$

This implies:

$$R^*(\bar{L}_{S_{n,T^*}}(W^*)) = \max_{t \in [m]} |\bar{L}_{S_{n,t}}(W_{n,t})|$$

. If we take the expectation on both sides we get:

$$\mathbb{E}\left[R^*(\bar{L}_{S_{n,T^*}}(W^*))\right] = \mathbb{E}\left[\max_{t \in [m]} |\bar{L}_{S_{n,t}}(W_{n,t})\right] \tag{5.3}$$

We observe that, given the output of the parallel execution W^m, then (W^*, T^*, R^*), conditional on that, can take only up to $2m$ values, which means that:

$$I\left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right); W^*, T^*, R^* | W^m\right) \leq \log(2m)$$

We notice that, since by assumption the algorithm $\mathcal{A}$ satisfies $I(\Lambda_W(S); W) \leq \varepsilon$, then by Lemma (5.5):

$$I\left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right); W^m\right) \leq m\varepsilon$$

Next, we recall a property of the mutual information that follows from the chain rule stated with Proposition (5.6), the so-called "data processing inequality".

Proposition 5.8. *(Data processing inequality) Let X, Y and Z be random variables that form a Markov chain in the order $X \longrightarrow Y \longrightarrow Z$, then the mutual information between X and Y is greater than or equal to the mutual information between X and Z. In mathematical terms:*

$$I(X; Y) \geq I(X; Z)$$

Intuitively, this means that you cannot get more information out of a set of data then was there to begin with, or in other words, no clever transformation of the received code Y can increase the information that Y contains about X. In our case, consider: $X = \left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right)\right), Z = (W^*, T^*, R^*), Y = W^m$. Therefore, by the data processing inequality (Proposition 5.8) and the chain rule of mutual information (Proposition 5.6), we obtain:

$$\begin{aligned} I\left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right); W^*, T^*, R^*\right) &\leq I\left(\Lambda_W\left(S_{n,1}\right), \ldots, \Lambda_W\left(S_{n,m}\right); W^*, T^*, R^*, W^m\right) \\ &\leq m\varepsilon + \log(2m) \end{aligned}$$

Then, we can use Lemma (5.7) and the above train of inequalities to get:

$$\mathbb{E}[R \cdot \bar{L}_{n,T^*}(W^*] \leq \sqrt{\frac{2\sigma^2(m\varepsilon + \log(2m))}{n}}. \tag{5.4}$$

Recalling the equality stated with (5.3), the above inequality gives:

$$\mathbb{E}\left[\max_{t\in[m]} |\bar{L}_{S_{n,t}}(W_{n,t})\right] \leq \sqrt{\frac{2\sigma^2(m\varepsilon + \log(2m))}{n}}. \tag{5.5}$$

Now, we choose $m = \lfloor \frac{1}{\beta} \rfloor$ and we proceed by contradiction.

Suppose that the algorithm $\mathcal{A}$ does not satisfy the generalization bound that we want to prove, meaning:

$$\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] \geq \beta$$

Since the couples $(S_{n,t}, W_{n,t})\,, t = 1, \ldots, m$ are independent by construction, we get:

$$\mathbb{P}\left[\max_{t\in[m]} \left|\bar{L}_{S_{n,t}}(W_{n,t})\right| > \alpha\right] > 1 - (1-\beta)^{\lfloor 1/\beta \rfloor} > \frac{1}{2}$$

This leads to:

$$\mathbb{E}\left[\max_{t\in[m]} \left|\bar{L}_{S_{n,t}}(W_{n,t})\right|\right] > \frac{\alpha}{2}$$

The above inequality in expectation, together with equation (5.5), gives the following inequality:

$$\frac{\alpha}{2} < \sqrt{\frac{2\sigma^2}{n}\left(\frac{\varepsilon}{\beta} + \log\frac{2}{\beta}\right)}$$

which implies that

$$n < \frac{8\sigma^2}{\alpha^2}\left(\frac{\varepsilon}{\beta} + \log\frac{2}{\beta}\right),$$

which contradicts the condition on the sample complexity given in Theorem (5.4). As a consequence, under the sample complexity condition stated in Theorem (5.4), it can never hold that:

$$\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] \geq \beta$$

. This leads us to the fact that it has to hold, necessarily:

$$\mathbb{P}\left[|\mathrm{gen(W_n, S_n)}| > \alpha\right] \leq \beta$$

, which completes the proof. QED

5.1 Challenges with High-Probability bounds

Bassily et al. [7] proved a result equivalent to Theorem (5.4) in a more restrictive setting, for the task of classification, where the loss is not subgaussian but bounded, being binary.

Theorem 5.9. *Let $\mathcal{A}$ be a learning algorithm that maps a set $S_n = Z_1, ...Z_n$ of n i.i.d. data points drawn from a distribution μ to the label set $W_n = 0, 1$ such that the its mutual information is bounded, i.e. $I(S; \mathcal{A}(S)) \leq C$, where C is a positive constant. Then:*

$$\mathbb{P}[|\operatorname{gen}(W_n, S_n)| > \alpha] < \frac{C+1}{2n\alpha^2 - 1}$$

The theorem above states that achieving confidence β requires $n \geq \frac{C+1+\beta}{2\beta\alpha^2}$ training examples. It's worth noting that this sample complexity bound is sharp. In fact, Bassily et al. [7] proved the existence of a learning problem and of a learning algorithm which has its mutual information bounded by a constant that has a generalization error of at least $\frac{1}{2}$ with probability at least $\frac{1}{n}$.

The issue that this paper raises is that it's not always sufficient to restrict the information used by the algorithm to make the learning possible. Even in settings that are simple or where generalization is easy to prove, in fact, mutual information can be infinite. For instance, let us consider $\mathcal{X} \subset \mathbb{R}$ and let $\mathcal{T} \subseteq \{0, 1\}^{\mathcal{X}}$ be the set of all thresholds, i.e. $\mathcal{T} = \{f_k\}_{k \in \mathcal{X}}$ where $f_k(x) = \mathrm{I}_{x \geq k}$. In the paper, it is shown that any consistent and proper learning algorithm $\mathcal{A}$ for threshold functions must reveal $I(\mathcal{A}(S); S) \geq \frac{\log \log |\mathcal{X}|}{n^2}$. So, we have a lower bound on the number of bits of information that an algorithm must reveal that depends on the size of the domain. Notice that the dependence on the size of the domain $\mathcal{X}$ is not very strong but, if it's of infinite size, then the mutual information is unbounded, making impossible implying generalization with the previous results. Since mutual information is a particular case of the framework of Section 3, it follows that the strong convexity condition is also insufficient for achieving such strong results.

6 Conclusion

The results showed in this article are only a small part of the work present in the literature. There are several other results that have been proposed as improvements of the results showcased in Section 2 that use chaining mutual information, conditional mutual information and other dependence measures such as α-Réyni divergence, f-divergence, Wasserstein distances, maximal leakage, etc. Some of those are included in the general framework presented in Section 3, some others not and constitute an area of further improvement. There are also a lot of open questions and room for improvements for high-probability bounds. In particular, it is interesting to understand if and how the techniques showed in Section 3 can be extended to obtain PAC-Bayes bounds, a question of not easy answer in the face of the challenges presented in Section 5.1.

References

[1] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 51:1232–1240, 2016. URL `https://proceedings.mlr.press/v51/russo16.html`.

[2] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017. URL `https://proceedings.neurips.cc/paper/2017/file/ad71c82b22f4f65b9398f76d8be4c615-Paper.pdf`.

[3] Gergely Neu and Gábor Lugosi. Generalization bounds via convex analysis. *arXiv preprint arXiv: 2202.04985*, 2022. doi: 10.48550/ARXIV.2202.04985. URL `https://arxiv.org/abs/2202.04985`.

[4] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv : 1912.13213*, 2019. URL `http://arxiv.org/abs/1912.13213`.

[5] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. Fundamentals of convex analysis. *Springer Berlin, Heidelberg*, 2001. doi: https://doi.org/10.1007/978-3-642-56468-0. URL `https://link.springer.com/book/10.1007/978-3-642-56468-0`.

[6] Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information based bounds on generalization error. *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591, 2019. doi: 10.1109/ISIT.2019.8849590.

[7] Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. *Proceedings of Machine Learning Research*, 83:25–55, 2018. URL `https://proceedings.mlr.press/v83/bassily18a.html`.